Volume 42 Number 4 December 2018 ISSN 0350-5596

Informatica

An International Journal of Computing and Informatics

Editorial Boards

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

https://doi.org/10.31449/inf.v42i4.1571 Informatica **42** (2018) 485–505 **485**

Clinical Decision Support Systems: A Visual Survey

Kamran Farooq
Computing Science and Mathematics Division, University of Stirling, Scotland, UK
E-mail: kfa@cs.stir.ac.uk

Bisma S Khan and Muaz A Niazi
Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan
E-mail: bis.sarfraz@gmail.com, muaz.niazi@ieee.org

Stephen J Leslie
Raigmore Hospital, Inverness, Scotland, UK
E-mail: stephen.leslie@nhs.net

Amir Hussain
Computing Science and Mathematics Division, University of Stirling, Scotland, UK
E-mail: ahu@cs.stir.ac.uk

Overview Paper

Keywords: cardiovascular decision support systems, CiteSpace, clinical decision support system, scientometrics, visualization

Received: March 28, 2017

Clinical Decision Support Systems (CDSS) form an important area of research. In spite of its importance, it is difficult for researchers to evaluate the domain primarily because of a considerable spread of relevant literature in interdisciplinary domains. Previous surveys of CDSS have examined the domain from the perspective of individual disciplines. However, to the best of our knowledge, no visual scientometric survey of CDSS has previously been conducted which provides a broader spectrum of the domain from the perspective of multiple disciplines. While traditional systematic literature surveys focus on analysing literature using arbitrary results, visual surveys allow for the analysis of domains by using complex network-based analytical models. In this paper, we present a detailed visual survey of CDSS literature using important papers selected from highly cited sources on the Clarivate Analytics' Web of Science. Our key results include the discovery of the articles which have served as key turning points in literature. Additionally, we have identified highly cited authors and the key countries of origin of top publications. We also present the universities with the strongest citation bursts. Finally, our network analysis also identifies the key journals and subject categories both in terms of centrality and frequency. It is our belief that this paper will thus serve as an important guide for researchers as well as clinical practitioners interested in identifying key literature and resources in the domain of clinical decision support systems.

Povzetek: Predsatavljen je pregled sistemov za podporo odločanju v zdravstvu.

1 Introduction

Clinical decision support system (CDSS) is a significant field of health information technology. It is designed to assist clinicians and other healthcare professionals in the diagnosis and decision-making. CDSS uses healthcare data and the patient's medical history to make recommendations. By using a predefined set of rules, CDSS intelligently filters knowledge from complex data and presents at an appropriate time [1]. By adopting CDSS, healthcare can become more accessible to large populations. However, it also implies that at times, CDSS may be used by people having literal medical knowledge [2].

Several researchers have contributed in the form of systematic literature reviews (SLR) and surveys to provide readers with an insightful information about CDSS, as demonstrated below in Table 1.

Despite the considerable variety of literature available, a key problem, researchers facing is the inability to understand the dynamics of CDSS-related literature. This is compounded due to the fact that this literature is spread across several related disciplines. Consequently, it is challenging to locate available information from a corpus of peer-reviewed articles. It is also difficult for researchers as well as clinical practitioners to comprehend the evolution of the research area.

SLR may be outdated, may not meet our requirements, may not exist for new and emerging fields, and may be written for specific areas of interest. Whereas, the visual survey gives a scientometric overview of the

Author	Ref.	Study Period	Survey Type	Study Area	Papers Reviewed
Ali et al. (2016)	[3]	2000-2014	Systematic Review	Randomised control trials of CDSS	38
Vaghela et al. (2015)	[4]	1987-2014	Survey	Classification techniques of CDSS	18
Son et al. (2015)	[5]	1979-2014	Visualisation	E-Health	3023
Njie et al. (2015)	[6]	1975-2012	Systematic Review	CDSS and prevention of cardiovascular diseases	45
Madara (2015)	[7]	1950-2014	Systematic Review	CDSSs to improve medication safety in long-term care homes	38
Martínez-Pérez et al. (2014)	[8]	2007-2013	Literature and Commercial Review	Mobile CDSS and applications	92
Loya et al. (2014)	[9]	2004-2013	Systematic Review	Service-oriented architecture for CDSS	44
Fatima et al. (2014)	[10]	2003-2013	Systematic Review	CDSSs in the care asthma and COPD patients	19
Diaby et al. (2013)	[11]	1960-2011	Bibliometric	MCDA in healthcare	2156
Kawamoto et al. (2005)	[12]	1966-2003	Systematic Review	Features of CDSS important for improving clinical practices	70
Chuang et al. (2000)	[13]	1975-1998	Methodological Review	Clustering in CDSS	24

Table 1: The existing literature review in the domain of clinical decision support systems.

scientific literature, which provides a broader spectrum by embracing publications across multiple disciplines of the domain. The visual survey allows us to explore various trends and patterns in the bibliographic literature more efficiently and keeps our knowledge up to date.

In this paper, we present a visual survey of key literature from Web of Science (WoS) to provide a meaningful and valuable reference for further study in the field. We have used CiteSpace a key visually analytical tool for information visualisation [14]. Although, CiteSpace has been used in a variety of disciplines, such as visual analysis of aggregation operator [15], agent-based computing [16], digital divide [17], anticancer research [18], tech mining [19], and digital medicine [20], etc. To the best of our knowledge, until now, there is no current review of recent literature on CDSS, which uses a scientometric analysis of networks formed from highly cited and important journal papers from the Web of Science (WoS).

The key contribution of this paper is the visual analysis of citations to give a scientometric overview of the diversity of the domain across its multiple sub-domains and the identification of core concepts. The ideas of visual analysis and survey stem from Cognitive Agent-based Computing framework [29] – a framework which allows for modelling and analysis of natural and artificial Complex Adaptive Systems.

In summary, the current paper identifies various important factors including the identification of emerging trends and patterns through exploring central nodes, pivot points, turning points, bursting nodes, and landmark nodes, the most important cluster in the cited references, visual analysis of the key authors, highly cited authors, key journals, core subject categories, countries of the origin of manuscripts, and the institutions from the bibliographic literature of the domain. We hope that this work will assist researchers, academicians, and practitioners to learn about the key literature and developments in the CDSS domain.

The rest of the CDSS survey is structured as: In Section II, we give a brief background of the visualisation techniques. Next, in Section III, we present the methodology section including data collection and an overview of CiteSpace. This is followed by Section IV, containing results and discussion. In Section VII, correlation from actual literature is provided. Finally, Section VII concludes the paper.

2 Background

This section presents some of the commonly used techniques for the analysis of bibliographic networks.

A bibliographic network is a network composed of authors, journals, categories, terms, articles, or cited references and interaction among them. Nodes in the bibliographic networks may be authors, institutions, countries, articles, terms, cited references, or categories and edges may be interactions among them, such as co-citation, collaboration, coupling, or co-occurrence.

From bibliographic dataset, a variety of networks can be generated. Types of bibliographic networks include co-authorship networks of authors/organizations, co-citation networks of articles/authors/journals, coupling networks of authors/journals/articles/organizations, and co-occurrence networks of categories/terms/keywords.

2.1 Bibliometrics and scientometrics

Bibliometrics and Scientometrics are closely related fields which focus mainly on the analysis of bibliographic

literature. They explore the bibliographic literature to measure the evolution of the scientific domain.

The bibliometric is defined as "the application of mathematics and statistical methods to books and other media of communication" [21, 22]. Bibliometrics concentrates specifically on books and publications.

The scientometric is defined as "the quantitative methods of the research on the development of science as an informational process" [21, 23]. Scientometric concentrates specifically on the study of all aspects of the dynamics of scientific literature and technology [24].

Bibliometric research measures and evaluates the impact of scientific literature in qualitative and quantitative manners [25]. In addition to this, certain features of scientific publications are analysed to obtain various scientific communication-related findings from the bibliometric study.

Currently, various scientific techniques and methods are introduced for bibliometric studies. SNA is also one of the frequently used technique in bibliometric studies

2.2 Social Network Analysis (SNA)

SNA is an approach used to study personal relationships or social interactions among individuals or organizations. The main goal of SNA is to investigate patterns of interaction, structure, and organization of the social networks [26]. A social network is a network of personal relationships or social interactions among individuals or organizations. The nodes and edges in the social networks are referred to as actors and ties.

The most important concepts used in SNA are centrality, network density, and community structure. Here, we demonstrate different performance measures to understand network fundamentals.

- *Centrality* of a node depicts its topological importance in the network [27]. Some of the commonly used centralities are described below:
  - *Degree Centrality* of a node is the number of links incident to it [27].
  - *Closeness Centrality* is based on the average distance. It focuses on how close a node is to the rest of the nodes in the network [27].
  - *Betweenness Centrality* of a vertex measures the extent to which that vertex lies in the geodesics of pairs of all other vertices in the network [27]. A *Geodesic* is the shortest path between a node pair [27].
  - *Eigenvector Centrality* measures the influence of a vertex based on degree centralities of its neighbours. The eigenvector centrality of a vertex is high if it is linked to the important nodes with higher degree centralities [28].
  - *Eccentricity centrality* of a vertex is the maximum geodesic distance between that vertex and all other vertices [29].
- *Density* of a network is the actual connections in the network divided by the possible connections in the network [30].
- *Component* is a maximally connected subnetwork. There is at least one path between every node pair of the component [27].
- *Giant component* is the largest connected component in the network [27].
- *K-core* is a maximally connected subnetwork in which every node has degree at least *k* [27].
- *Clique* is a maximally complete subnetwork of three or more nodes, in which each node is connected directly to every other node [27].
- *Bridge* is a crucial link whose deletion increases the number of disconnected components in the network [27].
- A *cut-vertex* (or *cutpoint*) is such a vertex whose deletion increases the number of disconnected components in the network [27].
- *Communities* in a network are the dense groups of the nodes which are highly connected to each other inside the group and sparsely connected to the nodes outside the group [30].
- *Affiliation* networks are two-mode networks, which represent the involvement of a set of actors in a set of events [27].

After presenting the background, the next section presents the methodology used in this research.

3 Methodology

In Figure 1, we illustrate the proposed methodology for the visual analysis of bibliographic literature in the domain of CDSS to uncover emerging patterns and trends.

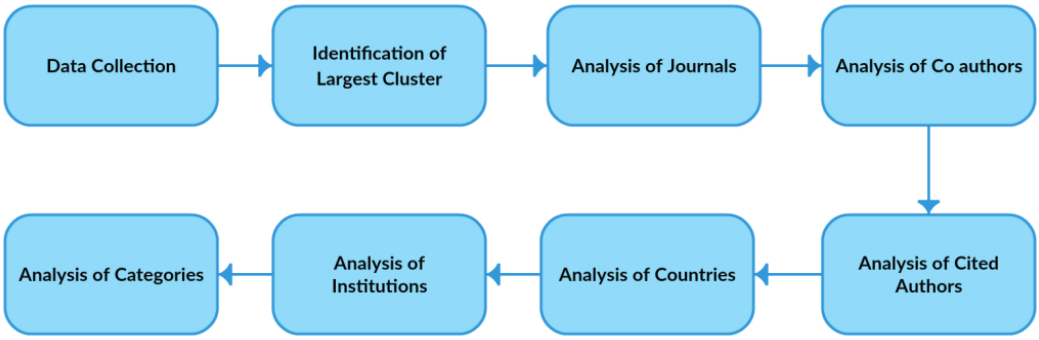


Figure 1: The proposed methodology (adapted from [2, 3]) for the visual analysis of clinical decision support systems for the discovery of emerging patterns and trends in the bibliographic data of the domain.

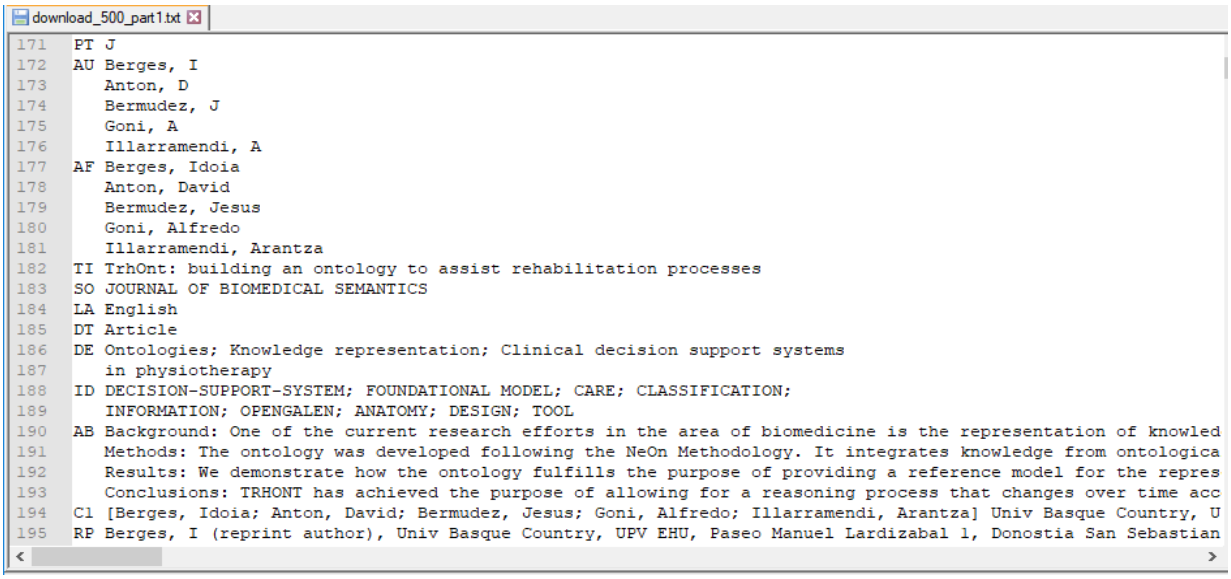


Figure 2: An example of input data from our dataset collected from Clarivate Analytics' Web of Science between the period of 2005-2016. The two-character field tags, such as AU and FN represent fields in the records.

3.1 Data collection

The input dataset was collected from the Clarivate Analytics' Web of Science [31] between the timespan of *2005* to *2016*. Data were retrieved on *11 Nov 2016*, by an extended topic search for CDSSs including the Web of Science. The databases searched include SCI-Expanded, SSCI, and A&HCI. The search was confined to document types including articles, reviews, letters, and editorial material published in the English language. Each data record includes information as titles, authors, abstracts, and references. The input dataset contains a total of *1,945* records.

Figure 2 shows an example of our input data. The two-character field tags in the input data, identify fields in the records. The detailed description of the field tags [32] can be found in Table 2.

It is pertinent to note here that there is a problem in data collected from Web of Science. The WoS data identified two cited-authors named as "Anonymous" and "Institute of Medicine." In terms of frequency,

Field Tags	Fields in Record	Field Tags	Fields in Record
FN	File Name	Z9	Total Times Cited Count
VR	Version Number	U1	Usage Count (Last 180 Days)
PT	Publication Type	U2	Usage Count (Since 2013)
AU	Authors	PU	Publisher
AF	Author Full Name	PI	Publisher City
TI	Document Title	PA	Publisher Address
SO	Publication Name	SN	ISSN
LA	Language	J9	29-Character Source Abbreviation
DT	Document Type	JI	ISO Source Abbreviation
DE	Author Keywords	PD	Publication Date
ID	Keywords Plus®	PY	Year Published
AB	Abstract	VL	Volume
C1	Author Address	IS	Issue
RP	Reprint Address	DI	DOI
EM	E-mail Address	PG	Page Count
RI	Researcher ID Number	WC	Web of Science Categories
OI	ORCID Identifier	SC	Research Areas
FU	Funding Agency and Grant Number	GA	Document Delivery Number
FX	Funding Text	UT	Accession Number
CR	Cited References	PM	PubMed ID
NR	Cited Reference Count	ER	End of Record
TC	Web of Science Core Collection Times Cited Count	EF	End of File

Table 2: The field tags representing record fields in the input data [32]

Anonymous is the landmark node. However, "Anonymous" itself is not an author, however, as a whole it is indicating all articles with missing author names. The larger diameter of the node "anonymous" indicates that several publications have missing author names. Whereas on an extensive search on the internet, we found multiple papers having "Institute of Medicine" as an author.

3.2 Why CiteSpace?

A variety of tools are available for network analysis and visualization. In this section, we will give a brief overview of some of the most commonly used and freely available tools. Gephi and Pajek are the most common tools used for the general analysis of the networks. However, they require other software tools for extracting scientometric data from WoS. For Pajek, WoS2Pajek can be used and for Gephi, Sci2 is used for this purpose. Pajek is focused less on network visualization and more on network analysis, whereas Gephi is focused more on network visualization and less on network analysis.

VOSview, CiteNetExplorer, and CiteSpace are specifically developed for visualization of bibliometric networks. However, VOSview has computational limitations and memory constraints. It also ignores the time dimension. Whereas, CiteNetExplorer is designed only for the visualization of citation networks of publications [33]. Unlike other tools, CiteSpace provides visualization of dynamic networks. CiteSpace is extensively used for network visualization.

In this research, we have used CiteSpace a key visually analytical tool for information visualisation [14].

3.3 CiteSpace: an overview

CiteSpace is custom designed for visual analysis of citations. It uses colour coding to capture some details, which otherwise cannot be captured easily by using any other tool. In CiteSpace users can specify the years' range and the length of the time slice interval to build various networks. CiteSpace is based on network analysis and visualisation. It enables interactive visual analysis of a knowledge domain in different ways. By selecting display of visual attributes and different parameters, a network can be viewed in a variety of ways. CiteSpace has been used to analyse diverse domain areas such as agent-based computing [16], cloud computing [34], cross-language information retrieval [35], and clinical evidence [36].

One of the key features of CiteSpace is the calculation of betweenness centrality [14]. The betweenness centrality score can be a useful indicator of showing how different clusters are connected [37]. In CiteSpace, the range of betweenness centrality scores is [*0, 1*]. Nodes which have high betweenness centrality are emphasised with purple trims. The thickness of the purple trims represents the strength of the betweenness centrality. The thicker the purple trim, the higher the betweenness centrality. A pink ring around the node indicates centrality >= *0.1*.

Burst identifies emergent interest in a domain exhibited by the surge of citations [16]. Citation bursts indicate the most active area of the research [37]. Burst nodes appear as a red circle around the node.

3.3.1 Colours used

CiteSpace is designed for visualisation; it extensively relies on colours, therefore the description in this paper is based on colours.

The colours of the co-citation links personify the time slice of the study period of the first appearance of the co-citation link. Table 3 demonstrates CiteSpace's use of colour to visualise time slices. The blue colour is for the earliest years, the green colour is for the middle years, and orange and red colours are for the most recent years. A darker shade of the same colour corresponds to earlier time-slice, whereas lighter shades correspond to the later time slice.

Link Colours	Corresponding Time Slice
Blue	Earliest years
Green	Middle years
Orange and Redish	Most recent years
Darker shade of the same colour	Earliest time-slice
Lighter shade of the same colour	Later time-slice

Table 3: CiteSpace's use of colours to visualise links, and time slices.

3.3.2 Node types

The importance of a node can be identified easily by analysing the topological layout of the network. Three most common nodes, which are helpful in the identification of potentially important manuscripts are *i)* hub node, *ii)* landmark node and *iii)* pivot node [14].

- Landmark nodes are the largest and most highly cited nodes. In CiteSpace, they are represented by concentric circles with largest radii. The concentric citation tree rings identify the citation history of an author. The colour of the citation ring represents citations in a single time slice. The thickness of a ring represents the number of citations in a particular time slice.
- Hub nodes are the nodes with a large degree of co-citations.
- Pivot nodes are links between different clusters in the networks from different time intervals. They are either gateway nodes or shared by two networks. Whereas turning points refer to the articles which domain experts have already identified as revolutionary articles in the domain. It is a node which connects different clusters by same coloured links.

4 Results and discussion

This section briefly demonstrates the results of our analysis.

4.1 Identification of the largest clusters in document co-citation network

To identify the most important areas of research, here we used cluster analysis. CiteSpace is used to form the

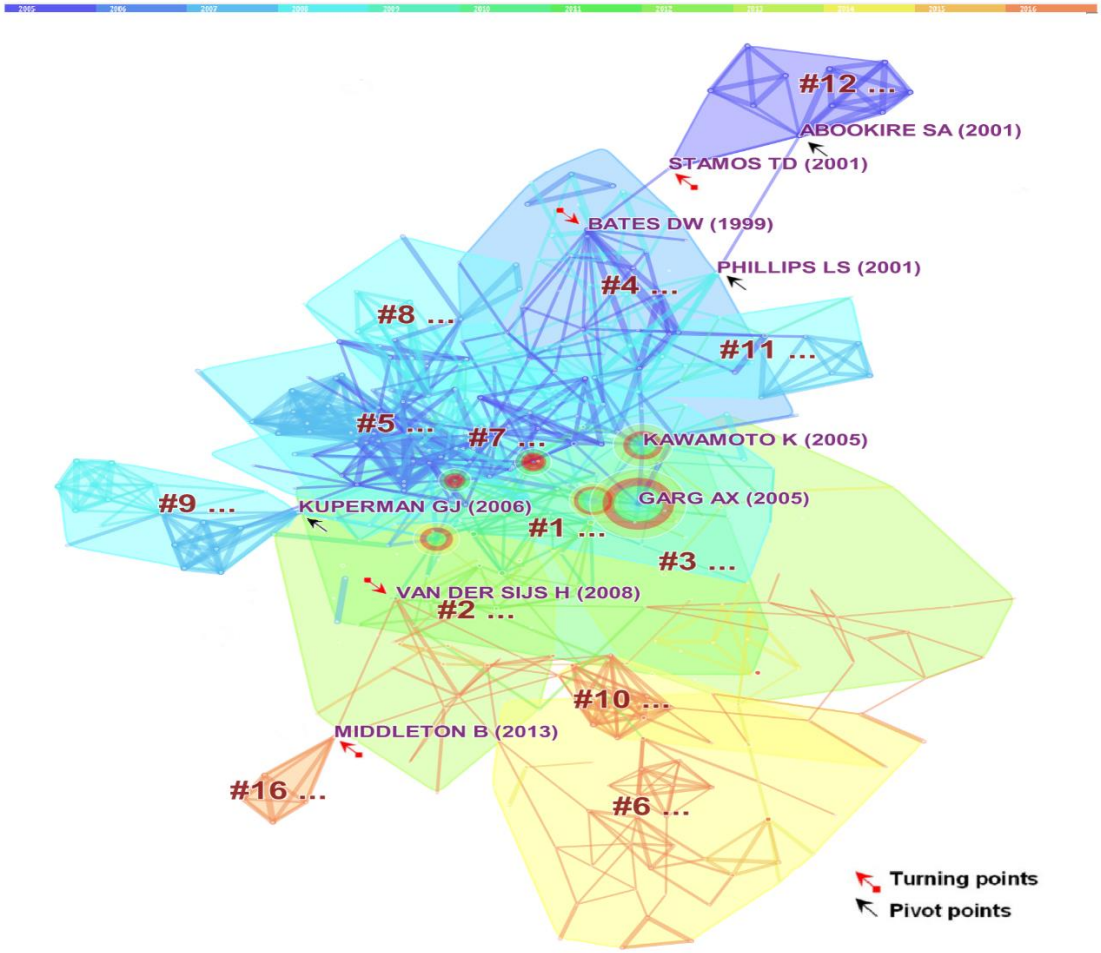


Figure 3: A merged network of cited references with 611 nodes and 1958 links on our CDSS dataset (2005-2016) based on 1-year time slices. The largest component of connected clusters is divided into 13 smaller clusters. The largest cluster is "computerised decision support" and the smallest is "computerised prescriber order entry." The diameter of the circle corresponds to the frequency of the node. Whereas red circle indicates high citation burst of the article. The article by Garg AX has the highest frequency and highest citation burst among other articles of the domain.

clusters. It uses time slice to analyse the clusters. The merged network of cited references is partitioned into some major clusters of articles. In Figure 3, years from *2005* to *2016* show up as yearly slices represented by unique colours. We have selected top *50* cited references per one-year time slice. The links between the nodes also represent the particular time slices. In [14] authors noted clusters with the same colours are indicative of co-citations in any given time slice. The cluster labels start from *0*; the largest cluster is labelled as (*#0*), the second largest is labelled as *(#1)*, and so on. The largest cluster is the indicator of the major area of research.

It can also be noticed in Figure 3 that the articles of David W. Bates (1999) and Thomas D. Stamos (2001) are the intellectual turning points, which join two linked clusters: (cluster #4) "combination" and (cluster #12) "family practice" respectively. Similarly, articles of Heleen Van Der Sijs (2008) and Blackford Middleton (2013) are the intellectual turning points, which join two linked clusters: (cluster #2) "decision support" and (cluster #16) "computerised prescriber order entry" respectively. After a gap of five years, Middleton B has cited a paper of Van Der Sijs H, which drew the interest of many researchers in the field of "decision support".

It is interesting to note that a half-life of the article of Bates DW is 7 years and a half-life of the article by Thomas D. Stamos is 4 years. Whereas a half-life of Van Der Sijis H's article is 5 years and a half-life of Middleton B's paper is 3 years.

In Table 4, details of top five cited references are given in terms of high citation frequency. By observing this table, we observed that the top five articles have low centrality but are still significant by having more frequency. The article by Amit X. Garg (2005) has the highest frequency of citations among all the cited references. Following it are the articles of Kensaku Kawamoto and Gilad J. Kuperman published in 2005 and 2007 respectively. The articles of Van Der Sijs H and Basit Chaudhry are also included in the top five articles in this domain.

In Table 4, it is also interesting to note that the article by Amit X. Garg (2005) is the landmark node with the largest radii. Amit X. Garg's article also has the highest citation burst of 20.71, which indicates that it has attracted

huge attention from the research community. It has 223 citations and 6-year half-life. It has 2357 citations on Google Scholar. Following it is the article of Kensaku Kawamoto (2005) with 15.46 citation burst, 151 citations, and a half-life of 6 years. It has 1684 citations on Google Scholar. Next is the article by Kuperman GJ (2007) with 3.48 citation burst, 135 citation frequency, and a half-life of 5 years. It has 547 citations on Google Scholar. It is closely followed by the Van Der Sijs H (2007) with a citation burst of 15.09, citation frequency 116, and a half-life of 5 years. It has 690 citations on Google Scholar. The article by Basit Chaudhry (2006) has the lowest citation burst of 2.99 among top five articles in the domain. It has a citation frequency of 112 and a half-life of 6 years. It has 2491 citations on Google Scholar.

F	CB	AU	PY	J	V	PP	HL	CL	GSC
223	20.71	Garg AX	2005	JAMA-J AM MED ASSOC	V293	P1223	6	3	2357
151	15.46	Kawamoto K	2005	BRIT MED J	V330	P765	6	3	1684
135	3.48	Kuperman GJ	2007	J AM MED INFORM ASSN	V14	P29	5	2	547
116	15.09	Van der Sijs H	2006	J AM MED INFORM ASSN	V13	P138	6	2	690
112	2.99	Chaudhry B	2006	ANN INTERN MED	V144	P742	5	1	2491

Table 4: The summary table of cited references sorted in terms of frequency includes frequency (F), citation burst (CB), author (AU), publication year (PY), journal (J), volume (V), page no. (PP), a half-life (HL), cluster ID (CL), and Google Scholar Citations (GSC) of the top 5 cited references.

Table 5 contains cited documents in terms of betweenness centrality. The article by Basit Chaudhry (*2006*) is the most influential document with the highest centrality score of *0.43*. The half-life of this article is *5* years and it has *2491* citations on Google Scholar. Following it is the article by Ross Koppel (*2005*) with *0.24* centrality, and a half-life of *5* years. It has *1995* citations on Google Scholar. Next is the article by Amit X. Garg (*2005*) with *0.18* betweenness centrality and a half-life of *6* years. It has *2357* citations on Google Scholar. It is closely followed by Jerome A. Osheroff (*2007*) with betweenness centrality of *0.16* and a half-life of *5* years. It has *357* citations on Google Scholar. Finally, we have an article by Gilad J. Kuperman (*2007*) with lowest betweenness centrality of *0.14* among top five articles in the domain. It has a half-life of *5* years. It has *547* citations on Google Scholar.

The merged network in Figure 3 contains a total of *611* cited references and *1,958* co-citation links. The largest cluster, i.e. (*#0*) of the network is disconnected from the largest component of the network. In this analysis, we will consider only the largest component.

BC	AU	PY	J	V	PP	HL	CL	GSC
0.43	Chaudhry B	2006	ANN INTERN MED	V144	P742	5	4	2491
0.24	Koppel R	2005	JAMA-J AM MED ASSOC	V293	P1197	5	0	1995
0.18	Garg AX	2005	JAMA-J AM MED ASSOC	V293	P1223	6	4	2357
0.16	Osheroff JA	2007	J AM MED INFORM ASSN	V14	P141	5	4	357
0.14	Kuperman GJ	2007	J AM MED INFORM ASSN	V14	P29	5	1	547

Table 5: The summary table of cited documents sorted in terms of Centrality includes betweenness centrality (BC), author (AU), publication year (PY), journal (J), Volume (V), page no. (PP), a half-life (HL), cluster ID (CL), and Google Scholar Citations (GSC) of the top 5 cited references.

The largest component of connected clusters contains *442* nodes, which is *72%* of the network. The largest component is further divided into *13* smaller clusters of different sizes. Table 6 illustrates the details of these clusters.

Cluster *#1* (largest cluster) contains *65* nodes, which are *10.628%* of all nodes in the network. The average publication year of the literature in this cluster is *2007*. The mean silhouette score of *0.737* indicates relatively high homogeneity in the cluster.

Cluster *#2* contains *57* nodes, which are *9.328%* of all nodes in the network. The average publication year of the literature in this cluster is *2009*. The mean silhouette score of *0.7* indicates relatively high homogeneity in the cluster.

Cluster *#3* contains *56* nodes, which are *9.165%* of all nodes in the network. The average publication year of the literature in this cluster is *2008*. The mean silhouette score of *0.722* indicates relatively high homogeneity in the cluster. It is interesting to note that cluster *#3* ("AIDS") contains several articles with strongest citation burst, which indicates it is an active or an emerging area of research.

Cluster *#4* contains *52* nodes, which are *8.51%* of all nodes in the network. The average publication year of the literature in this cluster is *2001*. The mean silhouette score of 0.791 indicates average homogeneity in the cluster. It is interesting to note that most of the highly influential articles are the members of cluster *#4*.

Cluster *#5* contains *49* nodes, which are *8.01%* of whole nodes in the network. The average publication year of the literature in this cluster is *2003*. The mean silhouette score of 0.772 indicates relatively high homogeneity in the cluster.

Cluster *#6* contains *45* nodes, which are *7.364%* of whole nodes in the network. The average publication year of the literature in this cluster is *2012*. The mean silhouette

score of *0.955* indicates very high homogeneity in the cluster.

Cluster *#7* contains *40* nodes, which are *6.546%* of all nodes in the network. The average publication year of the literature in this cluster is *2002*. The mean silhouette score of *0.73* indicates relatively high homogeneity in the cluster.

Cluster *#8* contains 19 nodes, which are *3.10%* of all nodes in the network. The average publication year of the literature in this cluster is *2003*. The mean silhouette score of *0.854* indicates high homogeneity in the cluster.

Cluster ID	Size	Silhouette	Mean (Year)	Label (Log-Likelihood Ratio)	Terms (Mutual Information)
1	65 (10.638%)	0.737	2007	Impact; Adverse Drug Event; Physician Order Entry	Computerized Decision Support
2	57 (9.328%)	0.7	2009	Alert; Ambulatory Care; Safety Alert	Drug Administration
3	56 (9.165%)	0.722	2008	Patient Outcome; Management; Guideline	Aid
4	52 (8.51%)	0.791	2001	Decision Support System; Primary Care; Expert System	Combination
5	49 (8.01%)	0.772	2003	Adverse Drug Event; Medication Error; Prevention	Chronic Illness
6	45 (7.364%)	0.955	2012	Personalized Medicine; Pharmacogenomics; Computed Tomography	ACR Appropriateness Criteria
7	40 (6.546%)	0.73	2002	Prevention; Intervention; Adverse Drug Event	Acute Kidney Failure
8	19 (3.10%)	0.854	2003	Emergency Medicine; ASHP; Systems Analysis	Intra Cluster Correlation Coefficient
9	18 (2.945%)	0.976	2004	Personal Digital Assistant; Resource; PDA	Consultation
10	13 (2.127%)	0.976	2011	Medication Alert System; Interview; Observational Study	Surgery
11	12 (1.963%)	0.944	2002	Guideline Implementation; Adverse Event; Clinical Practice Guideline	Factor-V-Leiden
12	11 (1.800%)	0.979	1999	Statin; Cholesterol Reduction; Treatment Panel III	Family Practice
16	5 (0.818%)	0.995	2010	Smoking Cessation; Control Intervention; Usability	Computerized Prescriber Order Entry

Table 6: The summary table of largest clusters of the cited authors. It contains the ID of the cluster, the size of the cluster, the average publication year of the literature in the cluster, and title terms of the clusters. The merged network contains *611* nodes and *1958* connections.

Cluster *#8* contains *19* nodes, which are *3.10%* of all nodes in the network. The average publication year of the literature in this cluster is *2003*. The mean silhouette score of *0.854* indicates high homogeneity in the cluster.

Cluster *#9* contains *18* nodes, which are *2.945%* of all nodes in the network. The average publication year of the literature in this cluster is *2004*. The mean silhouette score of *0.976* indicates very high homogeneity in the cluster.

Cluster *#10* contains *13* nodes, which are *2.127%* of all nodes in the network. The average publication year of the literature in this cluster is *2011*. The mean silhouette score of *0.976* indicates very high homogeneity in the cluster.

Cluster *#11* contains *12* nodes, which are *1.963%* of all nodes in the network. The average publication year of the literature in this cluster is *2002*. The mean silhouette score of *0.944* indicates very high homogeneity in the cluster.

Cluster *#12* contains *11* nodes, which are *1.800%* of all nodes in the network. The average publication year of the literature in this cluster is *1999*. The mean silhouette score of *0.979* indicates very high homogeneity in the cluster.

Cluster *#16* (smallest cluster) contains *5* nodes, which are *0.818%* of all nodes in the network. The average publication year of the literature in this cluster is *2010*. The mean silhouette score of *0.955* indicates very high homogeneity in the cluster.

After an overview of the identification of clusters in the cited reference network, next, we move to the analysis of the journals.

4.2 Analysis of journals

In this section, we visualise cited journals. Out of *1,945* records in the dataset, the *60* most cited journals were selected per one-year slice to build the network.

The pink rings around the nodes depicted in Figure 4 indicate that there are five nodes in the network with centrality > 0.1. "Journal of the American Medical Informatics Association" has the largest number of highly cited publications. The second largest number of publications is associated with "The Journal of the American Medical Association." "Proceedings of the AMIA Symposium" (2005) has the strongest citation burst among authors from the period of 2005.

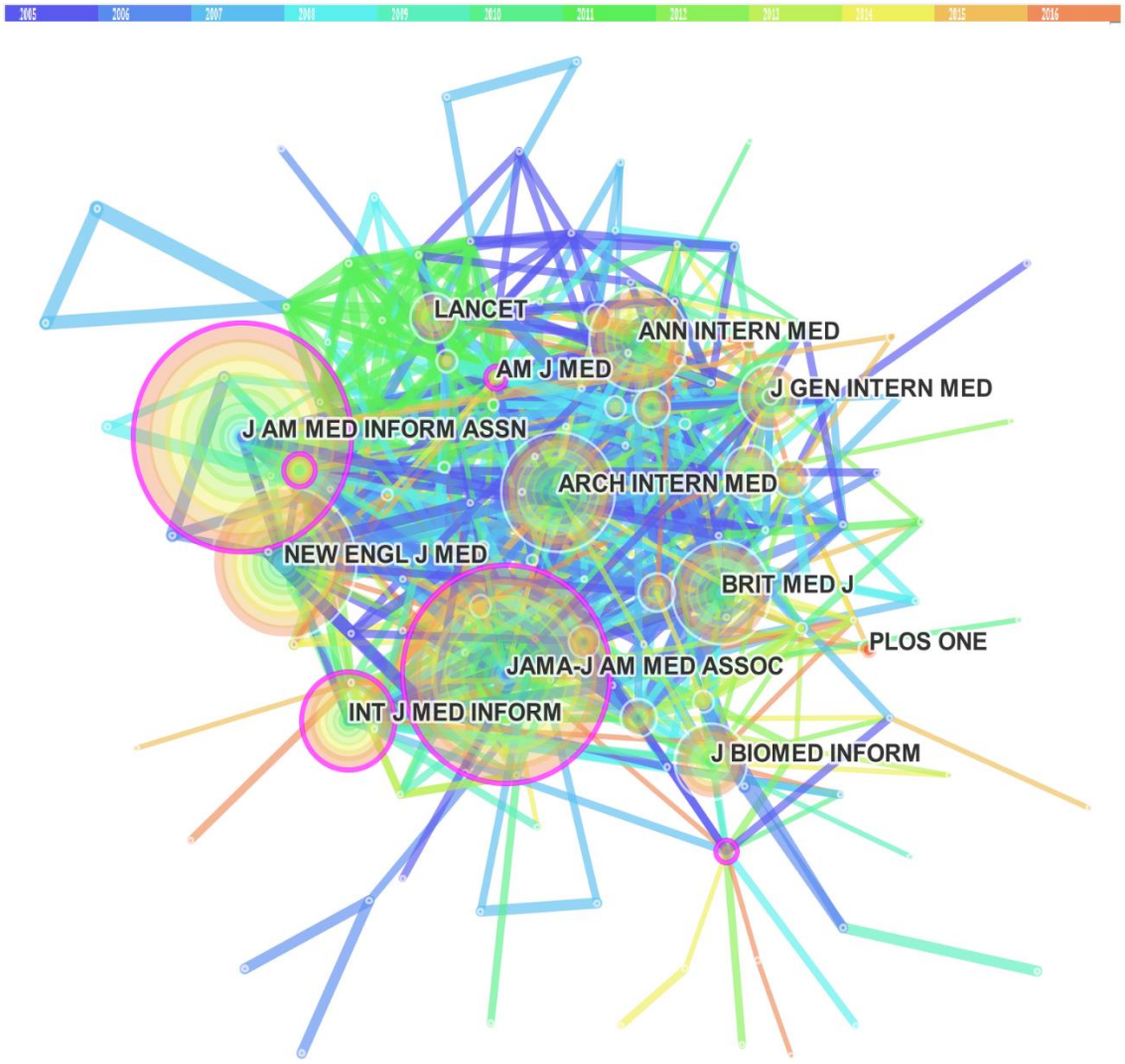


Figure 4: Journals' network in terms of centrality. Concentric citation tree rings indicate the citation history of the publications of a journal. The colours of the circles in the tree rings represent citations in a corresponding year. The red rings indicate the citation burst of the publication. The colours of the links correspond to the time slice. The pink rings around the nodes indicate the centrality >= 0.1. The "J AM MED INFORM ASSN" is the highly cited journal, whereas the "Jama-j AM MED ASSOC" is the most central journal of the domain.

Centrality	Title	Abbreviated Title	Impact Factor
0.14	The Journal of the American Medical Association	Jama-j AM MED ASSOC	37.684
0.13	Journal of the American Medical Informatics Association	J AM MED INFORM ASSN	3.428
0.13	International Journal of Medical Informatics	Int J MED INFORM	2.363
0.13	The American Journal of Medicine	Am J MED	5.610
0.13	Artificial Intelligence in Medicine (AIIM)	Artif INTELL MED	2.142

Table 7: In terms of centrality, the five most productive journals in the bibliographic literature of the CDSS domain. Jama-j AM MED ASSOC is the most central journal with a centrality score of *0.14*, whereas Artif Intell Med is the least central journal with a centrality score of *0.13*.

Publication frequency	Journal full title	Abbreviated title	Impact Factor (2016)
1169	Journal of the American Medical Informatics Association (JAMIA)	J AM MED INFORM ASSN	3.428
1096	The Journal of the American Medical Association	Jama-j AM MED ASSOC	37.684
819	The New England Journal of Medicine (NEJM)	New ENGL J MED	59.558
687	Archives of Internal Medicine	Arch INTERN MED	17.333
655	Annals of Internal Medicine Journal	Ann INTERN MED	16.593

Table 8: The five most productive journals in the bibliographic literature of the CDSS domain based on frequency. J AM MED INFORM ASSN is the most cited journal with frequency 1169, whereas Ann INTERN MED is the least cited journal with frequency 655.

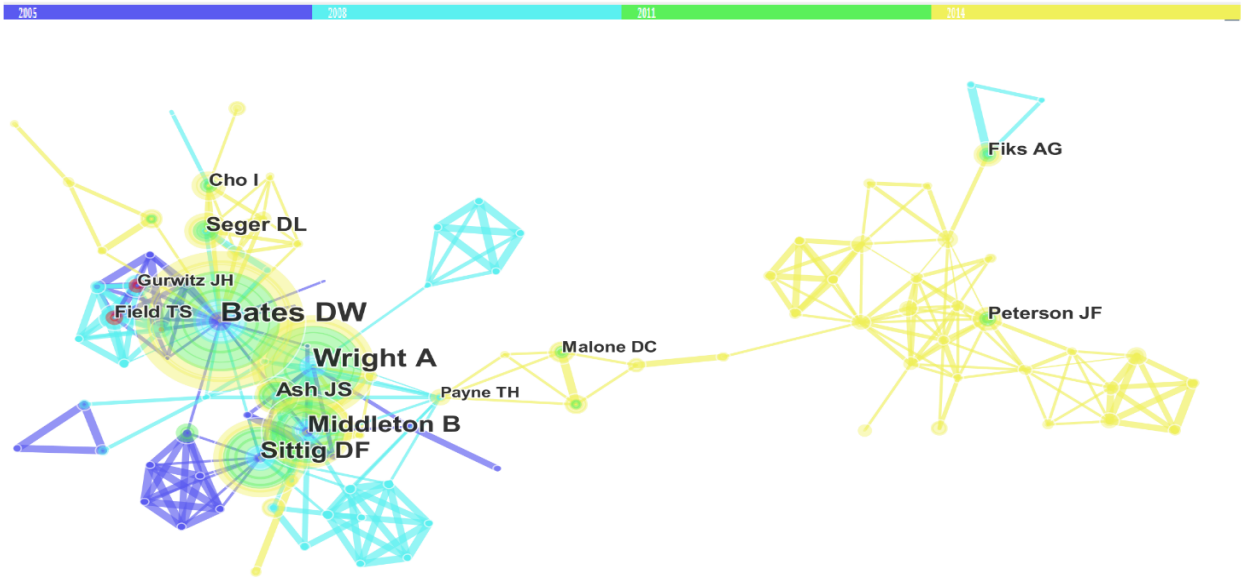


Figure 5: Co-authors network visualisation. The merged network contains *346* nodes and *719* links. Top *20%* nodes are selected per slice (of length *3*). Burst nodes appear as a red circle around the node. Concentric tree rings indicate the history of the publications of an author. David BW is the highly productive author with the frequency of *59*, whereas Payne TH is the most central node with a centrality score of *0.08*. Gurwitz JH and Field TS have longest publication burst periods.

Table 7 gives details of the top *5* key journals based on centrality. "The Journal of the American Medical Association" has the highest centrality score of *0.14* among all the other journals. It has *37.684* impact factor. In addition, it could be seen that in terms of centrality, the "Journal of the American Medical Informatics Association," the "International Journal of Medical Informatics," "The American Journal of Medicine" and the "Artificial Intelligence in Medicine" are also some of the productive journals of this domain with a centrality score of *0.13* and impact factor of *3.428*, *2.363*, *5.610*, and *2.142* respectively.

Table 8 gives details of the top *5* key journals based on their frequency of publications. It is interesting to note that the table organised in terms of frequency of publications gives a somewhat different set of key journals. The "Journal of the American Medical Informatics Association" is at the top with the frequency of *1169* publications and *3.428* impact factor. This is followed by "The Journal of the American Medical Association", "The New England Journal of Medicine," "The Archives of Internal Medicine", and the "Annals of Internal Medicine Journal" with frequencies *1961*, *819*, *687*, and *655* and impact factor *37.684*, *59.558*, *17.333*, and *16.593* respectively.

After a visual analysis of the journals, in the next section, we will analyse the authors' network.

4.3 Analysis of co-authors

This section analyses the author collaborative network. Figure 5 displays the visualisation of the core authors of the domain. The merged network contains *346* authors and *719* co-authorship links. As shown in Figure 5, burst nodes appear as a red circle around the node. The citation burst in authors network specifies the authors who have rapidly grown the number of publications. As shown in Figure 5, in terms of frequency, David BW is the landmark node with largest radii of the node. Payne TH is the most central author of this domain.

Visualisation in Figure 6 illustrates the authors who have the strongest publication bursts and years in which it took place. It can be seen that Ali S. Raja (*2014)* from "Harvard Medical School, USA" has the strongest burst among the top *5* authors since *2005*. Ivan K. Ip (*2005)* from "Harvard Medical School, USA" has the second strongest burst, which took place in the period of *2013* to *2016*. Following him are Terry S. Field (*2005)* from Meyers Primary Care Institute, Ramin Khorasani (*2014)* from "Brigham and Women's Hospital", and Jerry H. Gurwitz (*2005)* from "Meyers Primary Care Institute, USA."

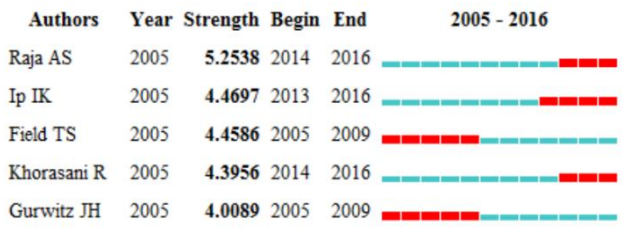


Figure 6: The top *5* Co-authors associated with strongest publication bursts. The history of the burstness of authors includes names of the authors, publication year, burst strength, starting, and ending year of the citation burst. "Raja AS" has strongest publication burst among all other authors. "Field TS" and "Gurwitz JH" have the longest burst period.

Even though this visualisation gives a general picture of the several authors, Table 9 also illustrates a comprehensive analysis of the authors' network.

Frequency	Author	Abbreviations
395	David Bates	BATES DW
296	Amit X. Garg	GARG AX
255	Kensaku Kawamoto	KAWAMOTO K
180	Rainu Kaushal	KAUSHAL R
173	Gilad J. Kuperman	KUPERMAN GJ

Table 9: The top 5 Authors in terms of the frequency of joint publications. David Bates is the most productive author with *395* publications.

Here we can notice that the most productive author in the network is David Bates with *59* joint publications. David Bates is a Prof. of Medicine at "Harvard Medical School, USA." His areas of interest are medication safety, patient safety, quality, medical informatics, and clinical decision support. Next is Adam Wright, an Assoc. Prof. of Medicine, "Harvard Medical School, USA" and "Brigham and Women's Hospital, USA." His areas of interest are health information technology, medical informatics, biomedical informatics, clinical information systems, and CDS. Dean F. Sittig is the Cristopher Sarofim Family Prof. of Bioengineering, "Biomedical Informatics, and UTHealth, USA." CDS, electronic health records, medical informatics, and biomedical informatics are his areas of interest. Next is Blackford Middleton, an Instructor, "Harvard TH Chan School of Public Health, USA". His areas of interest include personal health record, clinical informatics, CDS, knowledge management, and electronic medical record. Finally, we have Ramin Khorasani, MD, PhD, "Brigham and Women's Hospital, USA."

Centrality	Author	Abbreviations
0.08	Thomas Payne	Payne TH
0.07	David Bates	Bates DW
0.07	Richard D Boyce	Boyce RD
0.07	Robert R Freimuth	Freimuth RR
0.07	Matthias Samwald	Samwald M

Table 10: The top *5* Co-Authors in terms of centrality. Payne TH is the most central author with a centrality score of *0.08*, whereas the rest of the authors have the same centrality score of *0.07*.

For additional relative analysis, we have observed the collaborative authors based on centrality, as depicted in Table 10. Thomas Payne a Prof. of Medicine, "University of Washington, USA." His areas of interest are clinical informatics and clinical computing. Richard D Boyce, Asst. Prof. of "Biomedical Informatics, University of Pittsburgh, USA." His areas of interest are Pharmacoepidemiology, medication safety, knowledge representation, comparative effectiveness research, and semantic web. Next is Robert R Freimuth, "Mayo Clinic, USA." His areas of interest include genomics CDS, personalised medicine, genetic variation, data integration, Pharmacogenomics, data integration and interoperable infrastructure. Matthias Samwald, "Medical University of Vienna, Austria." His interest is in biomedical informatics.

After analysing authors' network, in the next section, we have visualised the cited authors' network.

4.4 Analysis of cited-authors

This section analyses the authors' co-citation network. Figure 7 displays the visualisation of the cited authors of this domain. The merged network contains *211* cited authors and *656* co-citation links. Burst nodes appear as a red circle around the node; the citation burst in cited-authors network specifies the authors who have rapidly grown the number of citations. In terms of frequency, David BW is the landmark node with largest radii of the citation ring. The pink ring around David BW indicates that it is also the most central author of this domain.

Even though this visualisation gives a general picture of the several authors, Tabel 11 also illustrates a comprehensive analysis of authors' network.

Here we can notice that a highly cited author in the network is David Bates with *460* citations. Next is Amit X. Garg, a Prof. of Medicine (Nephrology), Biostatics & Epidemiology, "Western University, Canada". His areas of interest are kidney diseases, kidney donation, and clinical research. Following him is Kensaku Kawamoto, an Asst. Prof. of Biomedical Informatics and Assoc. CMIO in the "University of Utah, USA". Knowledge management, CDS, and standards and interoperability are his areas of interest. Next is Rainu Kaushal, "Departments of Medicine, Quality Improvement, Risk Management, and Children's

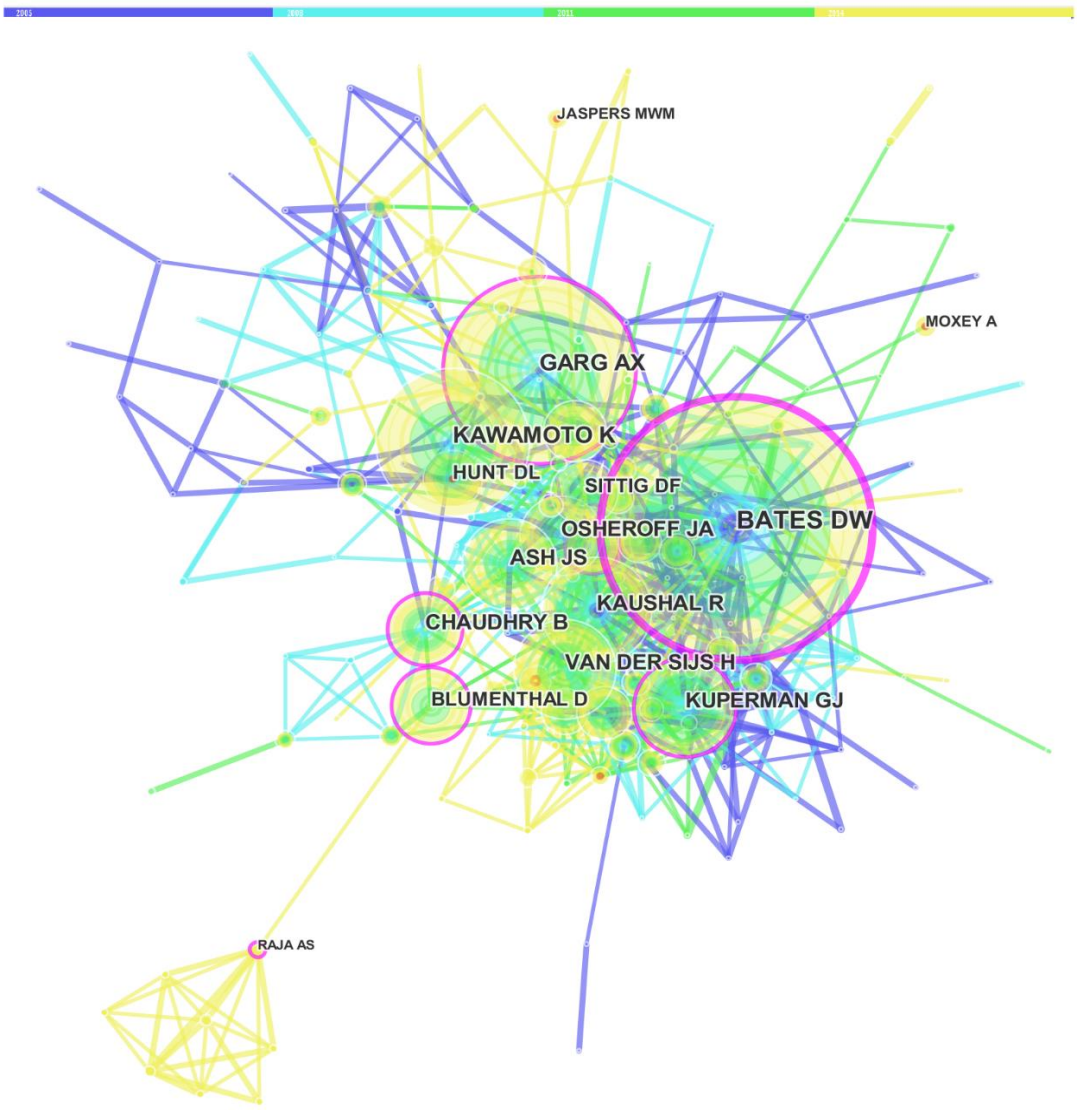


Figure 7: Cited-authors network visualisation. The merged network contains *211* nodes and *656* links. Burst nodes appear as a red circle around the node. Concentric citation tree rings indicate the citation history of the publications of an author. The pink rings around the node indicate the centrality score >= *0.1*. Bates DW is the landmark with largest radii and is also the hub node with the highest degree.

Hospital, Boston, Massachusetts, USA." Finally, we have Gilad J. Kuperman, an Adjunct Assoc. Prof. of Biomedical Informatics, "Columbia University Clinical Informatics, USA".

For additional comparative analysis, we have observed the top-cited authors in terms of centrality. Fresh names which enter in Table 11 are David Blumenthal from the "Harvard Medical School, USA" and Basit Chaudhry from the "University of California, USA."

After analysing authors' network, in the next section, we will visualise the countries of the origin of the key publications of the domain.

Frequency	Author	Abbreviations
460	David Bates	Bates DW
338	Amit X. Garg	Garg AX
280	Kensaku Kawamoto	Kawamoto K
207	Rainu Kaushal	Kaushal R
198	Gilad J. Kuperman	Kuperman GJ

Table 12: The top *5* cited-authors in terms of the frequency of citations. David Bates is the most cited author with *460* citations, whereas Kuperman GJ is the least cited author with *198* citations.

Centrality	Author	Abbreviations	Year
0.29	David Bates	Bates DW	2005
0.13	Gilad J. Kuperman	Kuperman GJ	2005
0.13	Amit X. Garg	Garg AX	2005
0.13	David Blumenthal	Blumenthal D	2009
0.12	Basit Chaudhry	Chaudhry B	2007

Table 11: The top *5* cited-authors in terms of centrality. Bates DW is the most central author with a centrality score of *0.29*, whereas Chaudhry B is the least central author with a centrality score of *0.12*.

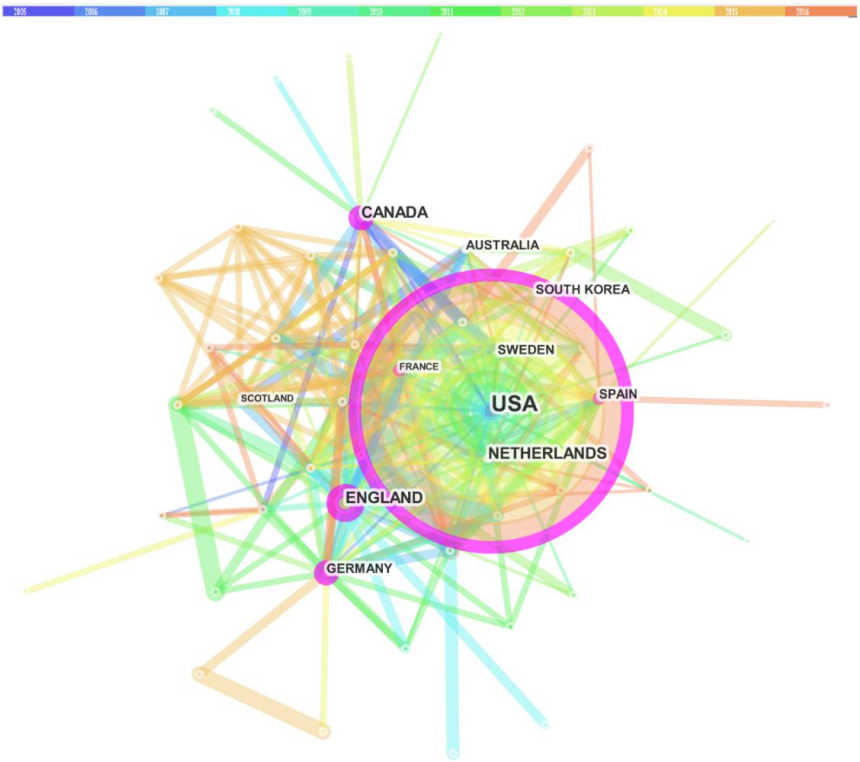


Figure 8: Countries network of *55* nodes and *263* links. The burst nodes appear as a red circle around the node. Concentric tree rings indicate the history of the publications of a country. The pink circle around the node represents the centrality >= *0.1*. The USA is the highly cited node, whereas Canada is the most central node and Scotland has strongest publication burst.

4.5 Analysis of countries

In this section, we demonstrate a visual analysis of the spread of research in the domain from different countries. For this visualisation, top *30* countries are chosen from the entire time span of *16* years (i.e. *2005-2016*) for each one-year time slice. In Figure 8, the concentric rings of different colours represent papers published in different time slices. The diameter of the ring thus indicates the publication frequency of the country. From the display, it can be seen that the "United States" has the highest publication frequency, which indicates that the origin of key publications in the domain is the "United States". This is followed by articles originating from England, Canada, Netherlands, and Australia. The pink circle around the node represents the centrality >= *0.1*. As depicted in Figure 8, Canada has the highest centrality value. This is followed by the US, England, Germany, and Spain. Red circles represent the publication burst. Scotland has the strongest publication burst, which provides the evidence that the articles originating in the domain from Scotland have attracted a degree of attention from its research community.

After a visual analysis of countries, we will present a visual analysis of institutions.

4.6 Analysis of institutions

In this section, the visualisation of institutions is performed. Figure 10 contains a merged network of institutions of *319* institutions and *844* collaboration links. We have selected top *50* nodes per one-year length time slice from *1,945* records. The "Harvard" is the most central, as well as the most productive node among all other institutions. Following it is the "Brigham and Women's Hospital, USA." Whereas, the "University of Massachusetts, USA" has the strongest publication burst.

A visual analysis of the history of the burstness of institutions identifies universities that are specifically active in the research in this domain.

Institutions	Year	Strength	Begin	End	2005 - 2016
Univ Massachusetts	2005	**6.2883**	2006	2009	
Indiana Univ Sch Med	2005	**6.1261**	2013	2016	
Cleveland Clin Fdn	2005	**4.1361**	2010	2011	
Johns Hopkins Univ	2005	**4.0639**	2012	2013	
Weill Cornell Med Coll	2005	**3.9329**	2011	2012	

Figure 9: History of the burstness of institutions includes names of institutions, years of publication, the strength of burstness, beginning and ending year of the citation burst.

As shown in Figure 9, the "University of Massachusetts, USA" has the strongest and longest publication burst among all other institutes in the timespan of *2006* to *2009*. The "Indiana University School of Medicine, USA" also has the longest period of the burst

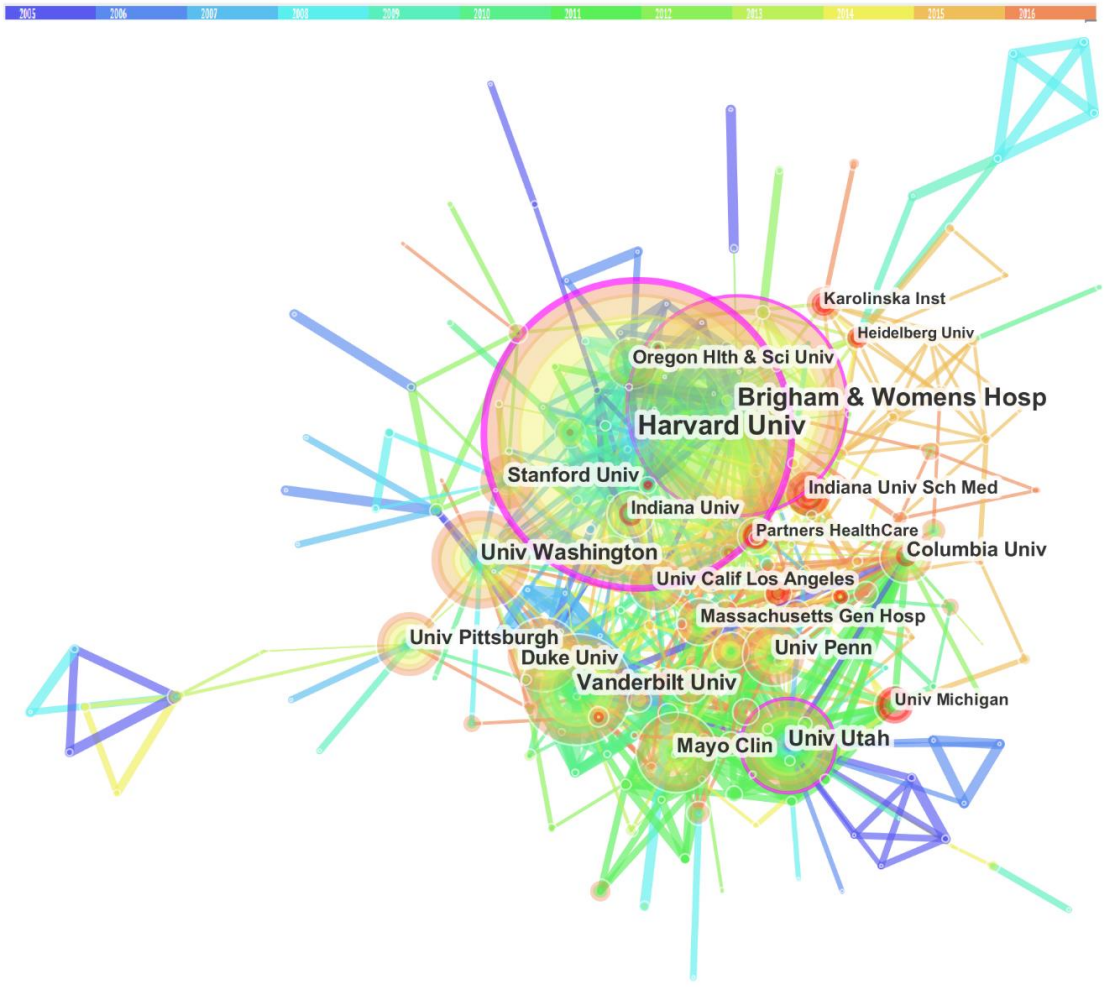


Figure 10: The network of Institutions, containing *319* nodes and *844* edges. Concentric citation tree rings demonstrate the citation history of the publications of an institution. The purple circle represents betweenness centrality. The thicker the purple ring, the higher the centrality score. The "University of Massachusetts" has the strongest burst. The Harvard is the highly cited and most central institution of the domain.

from *2013* till *2016*. Whereas, the "Weill Cornell Graduate School of Medical Sciences, USA" has shortest publication burst.

Next, we performed an analysis in terms of the frequency of publications associated with the institutions.

Table 13 represents the top five institutions based on the frequency of publications. The "Havard, USA" has the highest ranking with the frequency of *165* publications. The "Brigham & Women's Hospital, USA" followed it closely with the frequency of *122* publications. Next is the "Vanderbilt University, USA" with the frequency of *62* publications. With *56* publications, next, we have the "University of Utah, USA". Following it, we have the "University of Washington, USA" with the frequency of *55* publications.

Here, we performed another analysis in terms of the centrality of the publications. Table 14 contains the list of the top five universities based on the centrality. It is interesting to note that the top two universities the "Harvard" and "Brigham & Women's Hospital, USA" with centrality scores *0.3* and *0.17* respectively are also the highly cited institutions. Following them is the "University of Utah, USA" with a centrality score of *0.14*.

Frequency	Institution	Countries
165	Harvard University	USA
122	Brigham and Women's Hospital	USA
62	Vanderbilt University	USA
56	University of Utah	USA
55	University of Washington	USA

Table 13: The top institutions in terms of frequency of publications. "Harvard" has the highest publication frequency of *165*, whereas the "University of Washington" has the lowest frequency of *55*.

Centrality	Institutions	Countries
0.3	Harvard University	USA
0.17	Brigham and Women's Hospital	USA
0.14	University of Utah	USA
0.09	University of Washington	USA
0.07	Heidelberg University	Germany

Table 14: The top 5 institutions in terms of the betweenness centrality. Topmost University has a centrality score of *0.3*, whereas the "Heidelberg University" has the lowest centrality score of *0.07*.

Next is the "University of Washington, USA" with a centrality score of *0.09*. With centrality value *0.07*, it seems however that the "Heidelberg University, USA" has the lowest centrality score among all other institutions.

After visualisation of institutions, in the next section, we will present an analysis of subject categories of the domain.

4.7 Analysis of categories

In this section, our next analysis is to discover publications associated with various categories. Figure 11 depicts the temporal visualisation of categories in the domain. This merged network contains *95* categories and *355* links (co-occurrences). We have selected top *50* nodes per one-year time slice. The detailed analysis based on the centrality and frequency is given below.

Table 15 lists the top *5* categories based on centrality. The category "Health Care Sciences & Services" leads over other categories with centrality value *0.29*. It is closely followed by "Engineering" with centrality *0.28*. Next is "Computer Science" with a centrality score of *0.25*. Following it is the "Surgery" with centrality *0.18*. Subsequently, we have "Nursing" with a centrality score of *0.24*.

For relative analysis, we have also analysed these categories in terms of frequency of occurrence in manuscripts. The outcomes of this analysis are illustrated underneath in Table 16.

Table 16 lists the top *5* categories based on the frequency of occurrence. With the frequency of *658*, "Medical Informatics" leads the rest of the categories. Following it is the "Computer Science" with a frequency of *545*. Next is "Health Care Sciences & Services" with a frequency of *495*, which is followed by "Computer Science, Information Systems" and "Computer Science, Interdisciplinary Applications" with frequencies *320* and *318* respectively.

After visually analysing co-authors, journals, co-cited authors, countries, institutions, and subject categories, in the end, we are presenting the summary of the results.

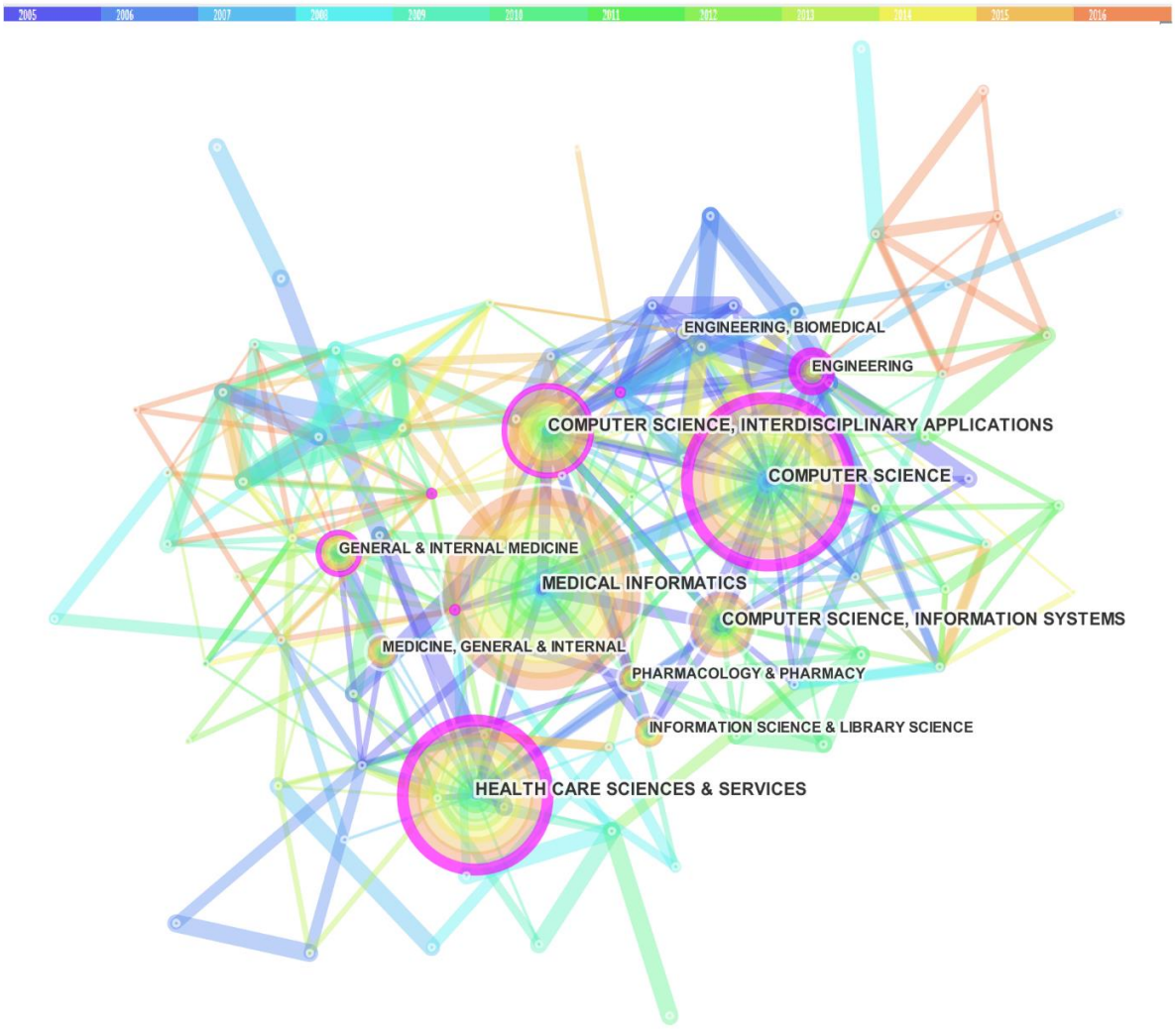


Figure 11: The category network containing *95* categories and *355* co-occurence. Concentric citation tree rings demonstrate the citation history of the co-occurrence of categories. The purple circle represents betweenness centrality. The thicker the purple ring, the higher the centrality score. Medical Informatics is the category with the highest co-occurence, whereas Health Care Sciences and Services is the most central category.

Centrality	Category
0.29	Health Care Sciences and Services
0.28	Engineering
0.25	Computer Science
0.18	Surgery
0.16	Nursing

Table 15: The top 5 categories based on centrality. The subject category "Health Care Sciences & Services" leads over other categories with a centrality score of *0.29.*

Frequency	Category
658	Medical Informatics
545	Computer Science
495	Health Care Sciences & Services
320	Computer Science, Information Systems
318	Computer Science, Interdisciplinary Applications

Table 16: The top *5* categories based on the frequency of occurrence. The subject category "Medical Informatics" leads over other categories with a frequency of *658*.

5 Summary of results

In this paper, we have utilised CiteSpace for the analysis of various types of visualisation to identify emerging trends and abrupt changes in scientific literature in the domain over time. In this section, we give an overview of the key results of the visual analysis performed in this study.

Firstly, using clustering of cited references we observed Cluster *#1*, the "computerised decision support" is the largest cluster, which contains *65* nodes that are *10.638%* of whole nodes in the network. The articles by Bates DW (*1999*), Stamos TD (*2001*), Van Der Sijs H (*2008*), and Middleton B (*2013*) are the key turning point. The half-life of these articles is *7, 4, 5, and 3* years respectively.

Subsequent analyses verified that there is conducted diversity in authors, journals, countries, institutions, and subject categories.

In the analysis of journals, we observed that the "Journal of the American Medical Informatics Association" has the largest number of highly cited publications in the domain and "Journal of the American Medical Association" is the most central journal among all other journals.

In terms of the analysis of the author's network, we observed that since *2005* Ali S. Raja (*2014*) has the strongest burst among the top authors of the domain. We also observed that most collaborative author in the network is David Bates, a Prof. of Medicine at the "Harvard School", has *59* publications is also the most central author with a centrality score of *0.33*. His areas of interest are medication safety, patient safety, quality, medical informatics, and clinical decision support. It is interesting to note that David Bates is also the highly cited and most central cited author of this domain.

In the analysis of countries, top *30* countries were chosen from the entire time span of *2005-2016* for each one-year time slice. We observed that the United States has the highest frequency of publications, which indicates the origin of key publications in the domain. Whereas, Canada has the highest centrality score. Scotland has the strongest publication burst, which provides the evidence that the articles originating in the domain from Scotland have attracted a degree of attention from its research community.

On the visual analysis of institutions, we found that "The University of Massachusetts" has the strongest and longest publication burst in the timespan of *2006* to *2009*. The "Indiana University School of Medicine" also has the longest period of the burst among all other institutes from *2013* till *2016*. Harvard has a top ranking with a frequency of *165* publications. It is interesting to note that the Harvard is also the most central institution with the centrality score of *0.3*.

In the analysis of categories, we observed that the category "Health Care Sciences & Services" leads over other categories with centrality value *0.29*. Whereas with a co-occurrence frequency of *658*, the category "Medical Informatics" leads the rest of the categories.

6 Correlation from actual literature

This section presents the necessary background of the Decision Support System (DSS) and CDSS.

6.1 Decision support system

The idea of DSS is very broad and different authors have defined it differently based on their research and roles DSS plays in the decision-making process [38-47]. DSS applications are adopted in several areas, such as business management [48], finance management [49], forest management [50], medical diagnosis [51], waste management [52, 53], oral anticoagulation management [54], ship routing [55], ecosystem management [56], value-based management [57], World Wide Web [58], diagnosis and grading of brain tumour [59], agent-based medical diagnosis [60, 61], and so on.

We intend to provide insight to CDSS researchers and practitioners about historical trends, current developments, and future directions of the CDSS domain.

6.2 Clinical decision support system

Since the beginning of computers, physicians and other healthcare professionals have expected the time when machines would aid them in clinical decision-making and other restorative procedures. "CDSS provides clinicians, patients, or individuals with knowledge and person-specific or population information, intelligently filtered or presented at appropriate times, to foster better health processes, better individual patient care, and better population health" [62].

There exist two main types of CDSS. The first one is derived from expert systems and uses knowledge base. The knowledge base depends on the inference engine to implement the rules, such as if-then-else on the patient

data and presents the findings to end-users [2]. The second type of CDSS is based on non-knowledge based systems, which depends on machine learning techniques for the analysis of clinical based data [63].

CDSSs are considered as an important part in the modern units of healthcare organisations. They facilitate the patients, clinicians, and healthcare stakeholders by providing patient-centric information and expert clinical knowledge [64]. To improve the efficiency and quality of healthcare, the clinical decision-making uses knowledge obtained from these smart clinical systems. The automated DSSs of Cardiovascular are available in primary health care units and hospital in order to fulfil the ever-increasing clinical requirements of prognosis in the domain of coronary and cardiovascular diseases. The computer-based decision support strategies have already been implemented in various fields of cardiovascular care [65]. In the US and the UK, these applications are considered as the fundamental components of the clinical informatics infrastructures.

Many reviews have identified the benefits of the CDSSs, in particular, Computerized Physician Order Entry systems [66-68]. The CDSS as part of the Computerized Physician Order Entry has been found to alleviate adverse drug events and medication errors [69-71]. The key benefits of CDSS reported in the studies conducted in [72-76] are higher standards of patient safety, improving the quality of direct patient care, standardisation and conformance of care using clinical practice guidelines, and the collaborative decision-making.

CDSSs also have demonstrated to improve clinician performance, by way of promoting the electronic prescription of drugs, adherence to guidelines and to an extent the efficient use of time [69, 71]. CDSSs perform a key role in providing primary care and preventative measures at outpatient clinics, e.g. by alerting caregivers of the need for routine blood pressure checking, to recommend cervical screening, and to offer influenza vaccination [67, 77].

The adoption of CDSSs in diagnosis and management of chronic diseases, such as diabetes [78], cancer [79], dementia [80], heart disease [81], and hypertension [82] have played significant clinical roles in the main healthcare organisations in the improvement of clinical outcomes of the organisations worldwide at primary and secondary care. These CDSS also provide a foundation to system developer and knowledge expert to collate and build domain expert knowledge for screening by clinicians and clinical risk assessment [72, 83].

Ontology-driven DSSs are also used widely in the clinical risk assessment of chronic diseases. The ontology-driven clinical decision support (CDS) framework for handling comorbidities in [84] presented remarkable results in the disease management and risk assessment of breast cancer patients, which was deployed as a CDSS handling comorbidities in the healthcare setting for primary care clinicians in Canada.

The ontology-driven recommendation and clinical risk assessment system could be used as a triage system in the cardiovascular preventative care which could help clinicians prioritize patient appointments after reviewing the snapshot of a patient's medical history containing patient demographics information, cardiac risk scores, cardiac chest pain and heart disease risk scores, recommended lab tests and medication details.

7 Conclusions and future work

In this paper, we have demonstrated a comprehensive visual and scientometric survey of the CDSS domain. This research covers all Journal articles in Clarivate Analytics from the period *2005-2016*. Our survey is based on real data from the Web of Science databases. This allowed us to comprehend all publications in the domain of CDSSs.

Our analysis has produced many interesting results. The CDSS has gained the interest of the research community from the era of *2005*. David Bates is the highly cited author in the literature of CDSS, whereas Ali S. Raja is the author who has rapidly grown the number of publications during the period of study. The "Journal of the American Informatics Medical Association" is the top-ranking source journal. It contributes *1169* publications during the period of study. The United States has contributed the highest number of publications, whereas the United Kingdom is the second highest productive country. Most of the contributions came from Harvard, whereas the "University of Massachusetts" remained specifically active in the research in this domain. The "Health Care Sciences & Services" leads the rest of the categories in CDSS.

A significant dimension of future work is to conduct scientometric analysis for identifying disease patterns, specifically in the cardiovascular, breast cancer, and diabetes domains.

8 Acknowledgement

This research project is funded by the EPSRC (Grant Ref. No. EP/H501584/1) and Sitekit Solutions Ltd. We would like to thank Professor Warner Slack from Harvard Medical School for providing useful insights and for his support and encouragement.

9 References

[1] J. A. Osheroff, J. M. Teich, B. Middleton, E. B. Steen, A. Wright, and D. E. Detmer, "A roadmap for national action on clinical decision support," *Journal of the American medical informatics association*, vol. 14, pp. 141-145, 2007. https://doi.org/10.1197/jamia.M2334

[2] J. T. Ahn, G. H. Park, J. Son, C. S. Lim, J. Kang, J. Cha, *et al.*, "Development of test toolkit of hard review to evaluate a random clinical decision support system for the management of chronic adult diseases," *Wireless Personal Communications*, vol. 79, pp. 2469-2484, 2014. https://doi.org/10.1007/s11277-014-1835-7

[3] S. M. Ali, R. Giordano, S. Lakhani, and D. M. Walker, "A review of randomized controlled trials of medical record powered clinical decision support system to improve quality of diabetes care,"

International journal of medical informatics, vol. 87, pp. 91-100, 2016. https://doi.org/10.1016/j.ijmedinf.2015.12.017

[4] C. Vaghela, N. Bhatt, and D. Mistry, "A Survey on Various Classification Techniques for Clinical Decision Support System," *International Journal of Computer Applications*, vol. 116, 2015. https://doi.org/10.5120/20498-2369

[5] Y.-J. Son, S. Jeong, B.-G. Kang, S.-H. Kim, and S.-K. Lee, "Visualization of e-Health research topics and current trends using social network analysis," *Telemedicine and e-Health*, vol. 21, pp. 436-442, 2015. https://doi.org/10.1089/tmj.2014.0172

[6] G. J. Njie, K. K. Proia, A. B. Thota, R. K. Finnie, D. P. Hopkins, S. M. Banks, *et al.*, "Clinical decision support systems and prevention: a community guide cardiovascular disease systematic review," *American journal of preventive medicine*, vol. 49, pp. 784-795, 2015. https://doi.org/10.1016/j.amepre.2015.04.006

[7] K. M. Marasinghe, "Computerised clinical decision support systems to improve medication safety in long-term care homes: a systematic review," *BMJ open*, vol. 5, p. e006539, 2015. https://doi.org/10.1136/bmjopen-2014-006539

[8] B. Martínez-Pérez, I. de la Torre-Díez, M. López-Coronado, B. Sainz-De-Abajo, M. Robles, and J. M. García-Gómez, "Mobile clinical decision support systems and applications: a literature and commercial review," *Journal of medical systems*, vol. 38, pp. 1-10, 2014. https://doi.org/10.1007/s10916-013-0004-y

[9] S. R. Loya, K. Kawamoto, C. Chatwin, and V. Huser, "Service oriented architecture for clinical decision support: A systematic review and future directions," *Journal of medical systems*, vol. 38, pp. 1-22, 2014. https://doi.org/10.1007/s10916-014-0140-z

[10] M. Fathima, D. Peiris, P. Naik-Panvelkar, B. Saini, and C. L. Armour, "Effectiveness of computerized clinical decision support systems for asthma and chronic obstructive pulmonary disease in primary care: a systematic review," *BMC pulmonary medicine*, vol. 14, p. 1, 2014. https://doi.org/10.1186/1471-2466-14-189

[11] V. Diaby, K. Campbell, and R. Goeree, "Multi-criteria decision analysis (MCDA) in health care: a bibliometric analysis," *Operations Research for Health Care*, vol. 2, pp. 20-24, 2013. https://doi.org/10.1016/j.orhc.2013.03.001

[12] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success," *Bmj*, vol. 330, p. 765, 2005. https://doi.org/10.1136/bmj.38398.500764.8F

[13] J.-H. Chuang, G. Hripcsak, and R. A. Jenders, "Considering clustering: a methodological review of clinical decision support system studies," in *Proceedings of the AMIA Symposium*, 2000, p. 146.

[14] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *Journal of the American Society for information Science and Technology*, vol. 57, pp. 359-377, 2006. https://doi.org/10.1002/asi.20317

[15] D. Yu, "A scientometrics review on aggregation operator research," *Scientometrics*, vol. 105, pp. 115-133, 2015. https://doi.org/10.1007/s11192-015-1695-2

[16] M. Niazi and A. Hussain, "Agent-based computing from multi-agent systems to agent-based models: a visual survey," *Scientometrics*, vol. 89, pp. 479-499, 2011. https://doi.org/10.1007/s11192-011-0468-9

[17] S. Zhu, H. H. Yang, and L. Feng, "Visualizing and Understanding the Digital Divide," in *International Conference on Hybrid Learning and Continuing Education*, 2015, pp. 394-403. https://doi.org/10.1007/978-3-319-20621-9_33

[18] P. Xie, "Study of international anticancer research trends via co-word and document co-citation visualization analysis," *Scientometrics*, vol. 105, pp. 611-622, 2015. https://doi.org/10.1007/s11192-015-1689-0

[19] F. Madani, "Technology Mining'bibliometrics analysis: applying network analysis and cluster analysis," *Scientometrics*, vol. 105, pp. 323-335, 2015. https://doi.org/10.1007/s11192-015-1685-4

[20] Y. Fang, "Visualizing the structure and the evolving of digital medicine: a scientometrics review," *Scientometrics*, vol. 105, pp. 5-21, 2015. https://doi.org/10.1007/s11192-015-1696-1

[21] J. Mingers and L. Leydesdorff, "A review of theory and practice in scientometrics," *European Journal of Operational Research*, vol. 246, pp. 1-19, 2015. https://doi.org/10.1016/j.ejor.2015.04.002

[22] J. Pritchard, "Statistical bibliography or bibliometrics?," *Journal of documentation*, vol. 25, pp. 348-349, 1969. https://doi.org/10.1108/eb026482

[23] V. V. e. Nalimov and Z. M. Mul′chenko, "Measurement of science. Study of the development of science as an information process," Foreign technology div wright-patterson AFB Ohio, 1971.

[24] W. Hood and C. Wilson, "The literature of bibliometrics, scientometrics, and informetrics," *Scientometrics*, vol. 52, pp. 291-314, 2001. https://doi.org/10.1023/A:1017919924342

[25] Z. Taşkın and A. U. Aydinoglu, "Collaborative interdisciplinary astrobiology research: a bibliometric study of the NASA Astrobiology Institute," *Scientometrics*, vol. 103, pp. 1003-1022, 2015. https://doi.org/10.1007/s11192-015-1576-8

[26] M. Rahman and R. Karim, "Comparative study of different methods of social network analysis and visualization," in *Networking Systems and Security (NSysS), 2016 International Conference on*, 2016, pp. 1-7. https://doi.org/10.1109/NSysS.2016.7400702

[27] W. De Nooy, A. Mrvar, and V. Batagelj, *Exploratory social network analysis with Pajek* vol. 27: Cambridge University Press, 2011. https://doi.org/10.1017/CBO9780511996368

[28] P. Hage and F. Harary, "Eccentricity and centrality in networks," Social networks, vol. 17, pp. 57-63, 1995. https://doi.org/10.1016/0378-8733(94)00248-9

[29] L. C. Freeman, "Centrality in social networks conceptual clarification," Social networks, vol. 1, pp. 215-239, 1978. https://doi.org/10.1016/0378-8733(78)90021-7

[30] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1-44, 2016. https://doi.org/10.1016/j.physrep.2016.09.002

[31] T. Reuters, "Web of science. Online factsheet Thomson Reuters, Philadelphia, Pennsylvania," ed, 2008.

[32] *Web of Science Core Collection Field Tags*. Available: https://images.webofknowledge.com/images/help/WOS/hs_wos_fieldtags.html

[33] N. J. Van Eck and L. Waltman, "CitNetExplorer: A new software tool for analyzing and visualizing citation networks," *Journal of Informetrics*, vol. 8, pp. 802-823, 2014. https://doi.org/10.1016/j.joi.2014.07.006

[34] J. Wu and Q. Chen, "Mapping the emerging field of cloud computing: Insights from a visualization analysis," in *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, 2012, pp. 1794-1799. https://doi.org/10.1109/ICSMC.2012.6377998

[35] Z. Rongying and C. Rui, "Visual analysis on the research of cross-language information retrieval," in *Uncertainty Reasoning and Knowledge Engineering (URKE), 2011 International Conference on*, 2011, pp. 32-35. https://doi.org/10.1109/URKE.2011.6007900

[36] C. Chen and Y. Chen, "Searching for clinical evidence in CiteSpace," in *AMIA Annual Symposium Proceedings*, 2005, p. 121.

[37] C. Chen, CiteSpace: A Practical Guide for Mapping Scientific Literature. New York, USA: Nova Science Publishers, 2016.

[38] C. W. Holsapple, "DSS architecture and types," in *Handbook on Decision Support Systems 1*, ed: Springer, 2008, pp. 163-189. https://doi.org/10.1007/978-3-540-48713-5_9

[39] M. J. Druzdzel and R. R. Flynn, "Decision support systems. Encyclopedia of library and information science. A. Kent," *Marcel Dekker, Inc.*, vol. 10, p. 2010, 1999.

[40] D. Power, "What is a DSS," *The On-Line Executive Journal for Data-Intensive Decision Support*, vol. 1, pp. 223-232, 1997. http://dssresources.com/papers/whatisadss/index.html

[41] D. J. Power, "Web-based and model-driven decision support systems: concepts and issues," *AMCIS 2000 Proceedings*, p. 387, 2000. http://aisel.aisnet.org/amcis2000/387

[42] J. Ralph, H. Sprague, and H. Watson, "Decision Support Systems: Putting Theory into Practice," ed: Prentice Hall, 1986.

[43] D. Power, "Decision Support Systems: Concepts and Resources for Managers," *Studies in Informatics and Control*, vol. 11, pp. 349-350, 2002.

[44] G. Marakas, "Decision Support Systems in the Twenty-First Century, 1999," ed: Prentice Hall, Inc. Upper Saddle River, NJ, ISBN: 0-13-744186-X.

[45] M. S. Silver, Systems that support decision makers: description and analysis: John Wiley & Sons, Inc., 1991.

[46] V. Sauter, Decision support systems: an applied managerial approach: John Wiley & Sons, Inc., 1997.

[47] A. M. Schroff, An approach to user oriented decision support systems, 1998.

[48] R. Bose and V. Sugumaran, "Application of intelligent agent technology for managerial data analysis and mining," *ACM SIGMIS Database*, vol. 30, pp. 77-94, 1999. https://doi.org/10.1145/342251.342270

[49] C. Zopounidis, M. Doumpos, and N. F. Matsatsinis, "On the use of knowledge-based decision support systems in financial management: a survey," *Decision Support Systems*, vol. 20, pp. 259-277, 1997. https://doi.org/10.1016/S0167-9236(97)00002-X

[50] G. A. Mendoza, W. Sprouse, W. G. Luppold, P. Araman, and R. J. Meimban, "An integrated management support and production control system for hardwood forest products," *Computers in Industry*, vol. 16, pp. 343-351, 1991. https://doi.org/10.1016/0166-3615(91)90074-J

[51] E. Alickovic and A. Subasi, "Medical decision support system for diagnosis of heart arrhythmia using dwt and random forests classifier," *Journal of medical systems*, vol. 40, pp. 1-12, 2016. https://doi.org/10.1007/s10916-016-0467-8

[52] G. Bertanza, P. Baroni, and M. Canato, "Ranking sewage sludge management strategies by means of Decision Support Systems: A case study," *Resources, Conservation and Recycling*, vol. 110, pp. 1-15, 2016. https://doi.org/10.1016/j.resconrec.2016.03.011

[53] V. Inglezakis, M. Ambăruş, N. Ardeleanu, K. Moustakas, and M. Loizidou, "Waste management in romania: current data and application of a decision support tool," *Environmental Engineering & Management Journal (EEMJ)*, vol. 15, 2016.

[54] D. Fitzmaurice, F. Hobbs, B. Delaney, S. Wilson, and R. McManus, "Review of computerized decision support systems for oral anticoagulation management," *British journal of haematology*, vol. 102, pp. 907-909, 1998. https://doi.org/10.1046/j.1365-2141.1998.00858.x

[55] Y. Dong, D. M. Frangopol, and S. Sabatino, "A decision support system for mission-based ship routing considering multiple performance criteria," *Reliability Engineering & System Safety*, vol. 150, pp. 190-201, 2016. https://doi.org/10.1016/j.ress.2016.02.002

[56] H. M. Rauscher, "Ecosystem management decision support for federal forests in the United States: a

review," *Forest ecology and management*, vol. 114, pp. 173-197, 1999. https://doi.org/10.1016/S0378-1127(98)00350-8

[57] G. J. Hahn and H. Kuhn, "Designing decision support systems for value-based management: A survey and an architecture," *Decision Support Systems*, vol. 53, pp. 591-598, 2012. https://doi.org/10.1016/j.dss.2012.02.016

[58] W.-C. Chen, T.-P. Hong, and R. Jeng, "A framework of decision support systems for use on the World Wide Web," *Journal of Network and Computer Applications*, vol. 22, pp. 1-17, 1999. https://doi.org/10.1006/jnca.1999.0078

[59] A. R. Tate, J. Underwood, D. M. Acosta, M. Julià-Sapé, C. Majós, À. Moreno-Torres, *et al.*, "Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra," *NMR in Biomedicine*, vol. 19, pp. 411-434, 2006. https://doi.org/10.1002/nbm.1016

[60] B. L. Iantovics, "Agent-based medical diagnosis systems," *Computing and Informatics*, vol. 27, pp. 593-625, 2012.

[61] B. Iantovics, "An Agent-Based Hybrid Medical Complex System," *International Information Institute (Tokyo). Information*, vol. 16, p. 3709, 2013.

[62] J. A. Osheroff and A. M. I. Association, "A roadmap for national action on clinical decision support," Journal of the Amercian medical informatics association, vol. 14, pp. 141-145, 2007. https://doi.org/10.1197/jamia.M2452

[63] M. Alther and C. K. Reddy, "Clinical decision support systems," ed: Healthcare Data Analytics, Chapman and Hall/CRC Press, 2015.

[64] D. C. Classen, S. Phansalkar, and D. W. Bates, "Critical drug-drug interactions for use in electronic health records systems with computerized physician order entry: review of leading approaches," *Journal of patient safety*, vol. 7, pp. 61-65, 2011. https://doi.org/10.1097/PTS.0b013e31821d6f6e

[65] G. J. Kuperman, A. Bobb, T. H. Payne, A. J. Avery, T. K. Gandhi, G. Burns, *et al.*, "Medication-related clinical decision support in computerized provider order entry systems: a review," *Journal of the American Medical Informatics Association*, vol. 14, pp. 29-40, 2007. https://doi.org/10.1197/jamia.M2170

[66] S. Eslami, N. F. de Keizer, and A. Abu-Hanna, "The impact of computerized physician medication order entry in hospitalized patients—a systematic review," *International journal of medical informatics*, vol. 77, pp. 365-376, 2008. https://doi.org/10.1016/j.ijmedinf.2007.10.001

[67] D. L. Hunt, R. B. Haynes, S. E. Hanna, and K. Smith, "Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review," *Jama*, vol. 280, pp. 1339-1346, 1998. https://doi.org/10.1001/jama.280.15.1339

[68] G. Zuccotti, F. Maloney, J. Feblowitz, L. Samal, L. Sato, and A. Wright, "Reducing risk with clinical decision support," *Appl Clin Inform*, vol. 5, pp. 746-756, 2014. https://doi.org/10.4338/ACI-2014-02-RA-0018

[69] T. J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R. R. Coeytaux, *et al.*, "Effect of clinical decision-support systems: a systematic review," *Annals of internal medicine*, vol. 157, pp. 29-43, 2012. https://doi.org/10.7326/0003-4819-157-1-201207030-00450

[70] M. A. Steinman, S. M. Handler, J. H. Gurwitz, G. D. Schiff, and K. E. Covinsky, "Beyond the prescription: medication monitoring and adverse drug events in older adults," *Journal of the American Geriatrics Society*, vol. 59, pp. 1513-1520, 2011. https://doi.org/10.1111/j.1532-5415.2011.03500.x

[71] M. W. Jaspers, M. Smeulers, H. Vermeulen, and L. W. Peute, "Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings," *Journal of the American Medical Informatics Association*, vol. 18, pp. 327-334, 2011. https://doi.org/10.1136/amiajnl-2011-000094

[72] A. Wright, D. F. Sittig, J. S. Ash, D. W. Bates, J. Feblowitz, G. Fraser, *et al.*, "Governance for clinical decision support: case studies and recommended practices from leading institutions," *Journal of the American Medical Informatics Association*, vol. 18, pp. 187-194, 2011. https://doi.org/10.1136/jamia.2009.002030

[73] R. B. Haynes and N. L. Wilczynski, "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: Methods of a decision-maker-researcher partnership systematic review," *Implement Sci*, vol. 5, p. 12, 2010. https://doi.org/10.1186/1748-5908-5-12

[74] K. Kawamoto, G. Del Fiol, C. Orton, and D. F. Lobach, "System-agnostic clinical decision support services: benefits and challenges for scalable decision support," *The open medical informatics journal*, vol. 4, p. 245, 2010. https://doi.org/10.2174/1874431101004010245

[75] G. Ivbijaro, L. Kolkiewicz, L. McGee, and M. Gikunoo, "Addressing long-term physical healthcare needs in a forensic mental health inpatient population using the UK primary care Quality and Outcomes Framework (QOF): an audit," *Mental health in family medicine*, vol. 5, p. 51, 2008.

[76] M. A. Musen, B. Middleton, and R. A. Greenes, "Clinical decision-support systems," in *Biomedical informatics*, ed: Springer, 2014, pp. 643-674. https://doi.org/10.1007/978-1-4471-4474-8_22

[77] L. Ahmadian, M. van Engen-Verheul, F. Bakhshi-Raiez, N. Peek, R. Cornet, and N. F. de Keizer, "The role of standardized data and terminological systems in computerized clinical decision support systems: literature review and survey," *International journal of medical informatics*, vol. 80, pp. 81-93, 2011. https://doi.org/10.1016/j.ijmedinf.2010.11.006

[78] P. J. O'Connor, J. M. Sperl-Hillen, W. A. Rush, P. E. Johnson, G. H. Amundson, S. E. Asche, *et al.*, "Impact of electronic health record clinical decision support on diabetes care: a randomized trial," *The Annals of Family Medicine,* vol. 9, pp. 12-21, 2011. https://doi.org/10.1370/afm.1196

[79] S. B. Clauser, E. H. Wagner, E. J. A. Bowles, L. Tuzzio, and S. M. Greene, "Improving modern cancer care through information technology," *American journal of preventive medicine,* vol. 40, pp. S198-S207, 2011. https://doi.org/10.1016/j.amepre.2011.01.014

[80] H. Lindgren, "Integrating clinical decision support system development into a development process of clinical practice–experiences from dementia care," in *Artificial Intelligence in Medicine*, ed: Springer, 2011, pp. 129-138. https://doi.org/10.1007/978-3-642-22218-4_17

[81] R. F. DeBusk, N. Houston-Miller, and L. Raby, "Clinical validation of a decision support system for acute coronary syndromes," *Journal of the American College of Cardiology,* vol. 10, p. A132. E1240, 2010.

[82] S. H. Luitjes, M. G. Wouters, A. Franx, H. C. Scheepers, V. M. Coupé, H. Wollersheim, *et al.*, "Study protocol Open Access," 2010.

[83] P. Khong and R. Ren, "Healthcare information system: building a cyber database for educated decision making," *International Journal of Modelling, Identification and Control,* vol. 12, pp. 133-140, 2011. https://doi.org/10.1504/IJMIC.2011.037842

[84] S. Abidi, J. Cox, S. Abidi, and M. Shepherd, "Using OWL ontologies for clinical guidelines based comorbid decision support," in *System Science (HICSS), 2012 45th Hawaii International Conference on*, 2012, pp. 3030-3038. https://doi.org/10.1109/HICSS.2012.629

https://doi.org/10.31449/inf.v42i4.2037 Informatica **42** (2018) 507–514 **507**

Feature Extraction Trends for Intelligent Facial Expression Recognition: A Survey

Sajid Ali Khan, Hafiz Syed Ahmed Qasim and Irfan Azam
Department of Computer Science
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan
E-mail: sajidalibn@gmail.com, syed.ahmed@szabist-isb.edu.pk, irfan.azam@szabist-isb.edu.pk

Overview Paper

Keywords: facial expressions, feature extraction, classification methods

Received: November 18, 2017

Human facial expression is important means of non-verbal communication and conveys a lot more information visually than vocally. In human-machine interaction facial expression recognition plays a vital role. Still facial expression recognition through machines like computer is a difficult task. Face detection, feature extraction and expression classification are the three main stages in the process of Facial Expression Recognition (FER). This survey mainly covers the recent work on FER techniques. It especially focuses on the performance including efficiency and accuracy in face detection, feature extraction and classification methods.

Povzetek: V prispevku je predstavljena primerjalna študija tehnik prepoznavanja izrazov obraza.

1 Introduction

Social psychology says facial expressions are means of coordinating conversations and communication. With the advancements in artificial intelligence and pattern recognition, people started considering Facial Expression Recognition (FER) as the most important technology of intelligent human interactive interface [1]. Beside differences, the expressions of different people are still recognizable. Facial information from human face, mostly provide clues for the better depiction of user mind. This increases greatly the human-computer interaction. Scientists have been working on facial expression classification and recognition for the past few decades. Problem-solving abilities and vast applications of a particular discipline act as an inspiration for further exploration and research. The urge to make the visual data useful, is the motivation for all image processing and computer vision algorithms. The FER has same motivation in the domain of computer vision. Its applications in the HCI (Human-Computer Interaction), visual look of human, touch sensations (moods), sight and voice utilization at the same time increases its requirement and value today. Moreover, it has applications like disables emotion detection system, assistance systems for autistic system [2] for detection of pain and stress in psychological studies [3], for instructor feedback an intelligent tutoring system [4], social and emotionally intelligent robot [17] etc. These applications reveal that facial expression detection systems work in effectively unless they are bound to do so in real-time. Face detection and tracking, feature extraction and tracking, feature classification and reduction etc. are the phases involved in FER. Each individual phase uses distinct algorithms, researchers tried to classify basic six expressions using these specific algorithms. This research mainly discusses algorithms for the phases of facial expression. Algorithms like adaptive skin color are used for detection and tracking [5, 6], mean shift algorithms [6], Stereo Active Appearance Model (STAAM) [7] etc. For feature extraction and tracking some algorithms are used like Local Binary Pattern (LBP) [8], Guided Particle Swarm Optimization (GPSO) [9], Gabor feature [8] etc., and there are few algorithms used for feature reduction like Principal Component Analysis (PCA), AdaBoost [10] etc. along classifiers like support vector machine [11-13], Hidden Markov Model (HMM) [14] etc. Accuracy and efficiency are two important aspects in real time environment for FER. Efficiency includes time complexity, space complexity and computational complexity. Due to very high computational complexity (Gabor feature and mean shift algorithm) and space complexity (LBP), it becomes difficult to work in real time environment for most of the mentioned algorithms. At the level of feature extraction, tracking or reduction efficiency can be improved. There are some appropriate algorithms for real-time environment like Optical flow calculation [6], Pixel Pattern-Based Texture Feature (PPBTF) [11], Adaboost [7, 10, 11], Pyramid LBP [12], Haar classifier [8, 10, 13, 15], PCA [10]. This survey mainly covers performance aspects of FER domain. We have mentioned the current approaches for FER and our views related to limitations of these approaches with respect to its execution in real time.

2 Literature review

Owusu et al. presented discussion and study about improving execution time and recognition accuracy [18].

In this work, Viola and Jones algorithm was used for face detection and Basel transformation is utilized for feature extraction. Thousands of facial features are extracted using Gabor feature extraction technique and also those features represent different facial detection patterns. To improve classification speed Adaboost Hypothesis is applied which will select few hundred features from thousands of extracted features. A three layers neural network classifier is then used to further process the selected features. JAFFE and YALE facial expression database are used to train the system and also for its testing. Combination of Basel and AdaBoost is used for the reduction of expression dataset. You will be amazed to know that Basel downsampling is never been used before for FER. So, it is an innovation for improving speed and accuracy. Proposed technique gave an accuracy of 96.83 and the average recognition rate of 92.22% for mentioned databases JAFFE and YALE and execution time required for 100x100 pixel sizes is 14.5ms. The results show that neutral expression has the weakest accuracy 92.23% in JAFFE and 86.16% in YALE.

In [19], Xijian and Tjahjadi extracted spatial pyramid histogram of gradients to three-dimensional facial features. They captured both spatial and motion information of facial expression by integrated the extracted features with dense optical flow. Support vector machine was used in this study with one-to-one strategy for training and testing. Investigation on CK+ and MMI datasets proved that integrated framework gives better performance than using individual descriptors. Contribution of this paper includes an integrated framework that captures dynamic information from deformation of facial regions and also facial landmarks movements, PHOG-TOP facial feature, dense optical flow having fused weighted PHOG-TOP and proposed framework analysis using contribution of different Fabian sub-regions. Canny edge detector is used for the detection of edges. Then to enhance spatial information an image is segmented into number of 3D subregions. PHOG-TOP is employed to whole face in a video sequence and also on 4 different sub-regions (forehead, eye, mouth and nose). Optical flow is implemented in order to extract dynamic info in video sequence. Dense facial points are equally distributed on mid of the face. Grid size is responsible for the efficiency of computing the optical flow. Average accuracy rate is 83.7% on CK+ and 73.1% on MMI dataset. It is also observed that happiness and surprise are easy to detect than remaining expressions. Proposed framework has limitations of generalizing to other datasets and it is also difficult to detect expressions other than happiness and surprise. This is due to the reason of smash and expressed datasets. Another limitation is that it is unable to detect faces wearing glasses or changed hairstyle.

Proposed technique by Zhang [20] includes following databases consisting of videos and images from movies and websites (AFEW, SFEW, HAPPEI, CENKI-4Kand QUT FER). Database used for FER falls in two different categories. One category consists of data collected in laboratory environment and the second category which collects data from broadcast TV and World Wide Web includes different databases. They used two lab-based facial FER databases for comparisons. They first applied video selection and segmentation process. Videos are captured from real-world environment and then segmentation is done using video splitter software. Annotator's pre-training consists of different students. Results from these students are also tested and the cycle is repeated again and again until 96% accuracy is achieved. Clips and vectors are classified by annotators and tested by experienced members. Detection applied using Viola Jones and ASM, failure rate of both algorithms is 4.7%. In Bench Mark FER approach face detection and tracking is applied on selected images using Viola Jones and ASM algorithms then texture features and geometric features are separated using SIFT, FAP and ASM algorithms. Features from texture features are selected using mRMP. Selected features and geometric features are then transferred to feature level fusion. Next, SVM classifier with Radial Basis Function (RBF) is trained for classification of facial expressions. Six basic universal emotions and three categorized emotions (positive, negative and neutral) are extracted using this method. For detection of six basic emotions SIFT+FAP gives 70% accurate results while, detection of three categories (positive, negative, neutral) gives 65% accurate result on realistic QUT images. Realistic QUT video clips have accuracy of 52.9 % for (SIFT+FAP), SIFT 48 %, FAP 50.6% on detection of six universal emotions. Accuracy for three categorized emotions (positive, negative, neutral) is given as SIFT+FAP 62.9 %, SIFT 61% and FAP 56.3 %. On the other hand, performance on lab-base data, FEEDTUM and NIVE database are used which gives 61.0% and 82% accuracy. Classification of three categorized emotions is slightly difficult from six universal expressions. Fear and sadness is highly affected due to nature of data as compared to other expressions.

Fang, Hui, et al., [21]is about automatic facial expression which extracts prominent features from videos without any preprocessing of subjective and without requiring any additional data for frames selection. Proposed technique uses machine learning methods in parallel with human reasoning to achieve dynamic changes in expressions in a better way. It is mandatory to detect facial regions first and for this purpose Viola-Jones detector is used. Face detection here is used only for the initiation of group registration. After face detection our next step is to detect features and for this purpose key feature is landmarked (eye, lips and nose). To align faces in static or dynamic data these landmark features are used to eliminate rotation and scaling effect. Through these landmarks deformations in video could be captured for further feature analysis. Traditional base algorithm is then used here to select neutral face image or first frame as template and wrap all images on it. Then global sharp model, appearance, local texture can be used as a knowledge to short the searching. Machine learning method can be used to locate optional landmarks (i.e. linear regression, graphical models). For facial sequences, group-wise

registration is applied. Successfully displacement between landmarks and other relative measurements like lip curvature, eye size is taken for expression recognition after landmarks tracking, and geometric features can be extracted from result. In following four regions (cheek region, eyebrow region, outer eye corner wrinkle, and forehead region) a Gabor filter is applied to extract an energy value that helps to obtain texture feature for learning expressions. These textures and geometric features are used for each video sequence. For classification of gathered data two techniques can be used 50% stratified split, one half for testing and other one is for training. The other technique contains stratified 10x10 fold cross validation which produces models using given data. Six classifiers are used J48, FRNN, VQNN, random forest, SMO-SVM and logistic and database used for this purpose is MMI. Accuracy by proposed method is 71.56%. It is noticed that happiness and surprise are easily identified by automatic classifiers and also human participants but there are difficulties in identification in the remaining expressions.

Zang Wang et al., in [22] proposed that face image is usually presented in high dimensional space as a data point. Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) methods are used for dimension reduction. Both can reveal global Euclidean structure but cannot manifold structure. Various manifold learning-based methods has been developed to get discriminant features for image detection and classification. One of the dimensional reduction methods is Local Fisher Discriminant Analysis (LFDA) but this method was not sparse like others and has no discriminant information so it was removed. A new system with Sparse LFDA (SLFDA) was introduced. From LFDA the minimum L1 normalization solution was obtained as a sparse solution. L1 minimization problem overcomes by Bregman. Therefore, Bregman method is applied to obtain sparse projection vector. Original features weight can be controlled by SLFDA. Moreover, in multimode problem SLFDA works well and the contrasting power of LFDA enhanced by it. In dimension reduction methods competitive to others SLFDA can achieve performance shown by the experiments performed on the databases (JAFFE and Yale B). The best recognition rate of SLFDA is 77.92%. As the proposed method strength is in dimension reduction and also gives reasonable interpretation of extracted features but their only focus was on supervised learning, so there is possibility to extend this approach to semi-supervised learning framework as well as to find the fast-numerical algorithm to solve L1-normalization problem.

By using both appearance and geometric features performance improved but most of existing algorithms are based on geometric features such algorithms track the facial components like eyes, lips, corner etc. as well as shapes and size of the face. Happy et al. [23] discussed that major issue is landmarks selection for which a relative geometric distance-based approach described to detect landmarks. Deformable model also become famous for landmark detection but high computation cost is a hurdle for them in real-time applications. It is a learning free approach to detect eyes, nose, lips etc. in face image and mark the required region. Some salient patches are extracted in training stage and within-pair of expressions; features having maximum variants are selected. Then multi-class classifier divides these selected features into basic six classes of expressions (those are fear, sadness, happiness, disgust, surprise and anger). In near frontal image with less computational complexity the results of facial landmark system are similar to the state of art method. Accurate emotion recognition in low-resolution image is ultimate goal of effective computing so this facial landmark detection technique along with salient patches-based FER framework performance is good in different image resolutions. Accuracy rate on JAFFE database is 91.8% and 94.1% on Cohn-Kanade (CK+) database which is satisfactory result but they are just considering few facial patches not whole face and also analyzing facial features without considering facial hairs. There is possibility of improvement in performance with partially occluded images and by using different appearance features. Moreover, an un-optimized MATLAB code is used for execution time. However, to improve the computational cost and real-time expression recognition with good accuracy rate an optimal implementation is required.

Ying tong et al. [24] discussed facial expressions of human and feature extraction method. Although Gabber wavelet, LBP are also used for feature extraction methods but they are time-consuming and dimensions increased significantly. In [24], the authors made binary coding for the separate block images which results a LGC statistics histogram. For identification feature they linked together the resultant histograms. In order to obtain the LGC-HD operator (LGC based on Horizontal and Diagonal gradient prior principle) more optimization is provided on the LGC operator. This reduces the computational complexity without losing the main expression information from face texture and also reduces the characteristic dimension. By using JAFFE database, the recognition average time in seconds are 90 for 8x8 block size with LGC-HD operator. Even after comparison of LBP, LBP uniform pattern, Gabor filter and LGC-HD the recognition average rate of LGC-HD is 90 %, which is higher than others. Experimental results show the weakness that either the block is larger or smaller will impact the recognition rate. The very small block number has smaller sub-block which cannot be accurately extracted. The redundancy of LGC-HD factor will affect the classification which results in inaccurate expression characteristics of large area.

The common universal facial expression recognizes cross-cultural facial expression including Japanese, Chinese, European and American. Ali et. al. in [25] claims that instead of using whole face only consider facial components (eyebrows, eye, etc.) as a lot of work has been done on those and through such technique satisfactory result is gained. In the race of solving problems of facial expressions consider multiple classifier decisions instead of using single classifier decision. Moreover, neural network-based ensemble classifier is made to enhance the accuracy of classifier.

Multi-objective genetic algorithm is also used. Acquisition and representation of multicultural dataset are major problems of multicultural facial expression classification. Three databases JAFFE, TFED and RadBoud used to overcome these problems. To check the presence of expression they used KNN, NB (Nave Bayes classifier) and SVM classifier. Furthermore, they used established dataset verification strategy for system performance evaluation. Four types of classifier considered for the classification of expressions those are BNN, KNN, SVM and Naïve Bayes classifier so that finally this plan worked out accurately. Experimental result 93.75% got with the combination of NNE collection, with NB predictor and using HOG descriptor. This is best in order to get satisfactory result of multicultural FER. But on few confuse facial expression still future work is needed due to visual representation, facial structure and difference in number of samples.

For recognition of facial expression, theoretical description of face operations often highlights the feature shapes as a primary visual signal. Although, facial surface characteristics can also be affected by changes in facial expression. Mladen Sormaz et al., in [26] examined that in the recognition of facial expression this surface knowledge can also be used. Firstly, facial expressions from images with distinct shapes and surface characteristics are identified by the participants. Mladen Sormaz et al., said various expressions depend on properties of shape and surface. They Further elaborated that facial expression categorization is feasible in any type of image. Moreover, in order to categorize the facial expressions, they evaluated the corresponding contributions of surface and shape information. This involves a correlative method in which shape properties and surface properties both are taken from different expressions. The experimental results show that in hybrid images categorization of facial expressions basically depends on the properties of surface and shape of image. Collectively, all the data directly demonstrate that recognition of facial expression is done through significant contribution of both the surface and the shape properties.

Andres Hernandez-Matamoros et al., mainly focus facial expression algorithm that significantly encounter facial image, present in color picture and segment divided into two Regions of Interest (ROI) i.e. the forehead and the mouth [27]. Both these regions are further segmented into non-overlying NXM blocks. Then dimension reduction is carried out after inserting the matrix in the Principal Component Analysis (PCA) module. Lastly, the resultant matrix generates the feature vectors. These vectors are then incorporated into the low complexity classifier, which uses the congregation and fizzy logical techniques. Though this classifier gives similar recognition rate as the high-performance classifiers but it gives least computational complexity. Results show that when feature vector from only one ROI is used in the proposed system then the recognition rate was increased to 97%. But the usage of feature vector of both ROI increases recognition rate up to 99%. It means overall 97% recognition rate in proposed system can only be achieved by clogging only one ROI.

In another work, Khan et al., in [28] highlighted the importance of local descriptors like Weber Local Descriptor (WLD) and LBP for recognition of facial expressions which is robust to illumination and pose changes. They argue that the local descriptors cannot be used to store the data of face image. In order to handle this problem, they proposed a framework in which the first they extracted features from face image using WLD and LBP and then fuse both type of features.

Similarly, a novel framework known as Weber Local Binary Image Cosine Transform (WLBI-CT) is proposed in [29] to recognize facial expression from multi-scale images. The results of this framework are robust to image resolution and image orientation.

Recently, in [30] Munir et al., utilized Merged Binary Pattern Code (MBPC) descriptor for face feature extraction. MBPC descriptor is capable to capture the prominent face feature. In this study, the results are compared with the variants of LBP.

Many other works [31]-[32] is presented in literature to describe the importance of face recognition and facial expression recognition.

3 Analysis table

References	Techniques Used	Pros	Cons
	• Viola Jones • Basel transformation • Gabor feature Extraction technique • Combination of Basel and Adaboost • 3 layers neural network classifier	• Improves execution time and recognition accuracy • Basel downsampling is used here first time. • Proposed technique gives an accuracy of 96.83 and 92.22%	• Neutral expression has the weakest accuracy
Xijian et al. [19]	• Multi-class support vector machine Based classifier. • PHOG-TOP facial Feature • PHOG descriptor • Canny edge detector	• Integrated framework gives better performance than using individual descriptors • Average accuracy rate on CK+ 83.7% and 73.1% on MMI dataset	• Limitations of generalizing to other datasets • Difficult to detect expressions other than happiness and surprise • Unable to detect faces wearing glasses or changed hairstyle.
Zhang et al. [20]	• Video splitter software • Annotator's • Viola Jones and ASM • SIFT, FAP • mRMP • SVM classifier with • Radial Based Function (RBF)	• Facial Expression is applied to real-world dynamic data which is quite difficult to handle.	• Fear and sadness is highly affected due to nature of data as compared to other expressions • Classification of 3 categorized emotions are slightly difficult from 6 universal expressions
Fang, Hui, et al. [21]	• Viola jones detector • Traditional base Algorithm used • Global sharp model, appearance, local texture • Machine Learning Method • Group-wise registration • Gabor filter • 6 classifiers are used J48, FRNN, VQNN, random forest, SMO-SVM and logistic.	• Extracts prominent features from videos without any preprocessing of subjective and without requiring any additional data for frames selection. • Uses machine learning method in parallel with human reasoning to achieve dynamic changes in expressions in a better way.	• Happiness and surprise are easily identified by automatic classifiers and also human participants but remaining expressions faces difficulties in identification.
Zang Wang et al. [22]	• PCA • LDA • Manifold learning based method • LFDA • SLFDA	• Dimension reduction • Interpretation of extracted features • In multimode problem SLFDA work well • The best recognition rate of SLFDA is 77.92%.	• Only focused was on supervised learning so possibility to extent semi-supervised learning • Find the fast-numerical algorithm to solve L1-normalization problem
Happy et al. [23]	• Deformable model • Landmark detection Technique	• Landmark detection • Well performance of FER in low resolution image	• Not considering the whole face • Analyzing facial feature without considering the facial

		• Accuracy rate on JAFFE database is 91.8% and 94.1% on Cohn-Kanade (CK+) database	hairs • Un-optimized MATLAB code is used
Ying tong et al. [24]	• LGC-HD operator	• Efficient without losing the main expression information from face texture • Reduce the characteristics dimension. • Recognition average rate is 90%	• Block is larger or smaller will impact the recognition rate • The LGC-HD features of each sub-block will show redundancy
Ghulam Ali et al. [25]	• SVM • Neural Network(NN) • K Nearest Neighbor (KNN) • Rule-based classifier • Nave Bayes classifier (NB) • Binary Neural Network (BNN) • NB predictor • HOG descriptor	• Result 93.75% got with the combination of NNE collection, with NB predictor and using HOG descriptor • Solved the representation of multicultural dataset problems of diverse face expression classification	• Future work is needed due to some confuse facial expression
Mladen Sormaz et al. [26]	• Complementary or converging methods • Hybrid image	• Both shape and surface play important role in the recognition of facial expressions	
Andres Hernandez-Matamoros et al. [27]	• Voila jones • PCA • Gaussian functions • Gabor filters • Empirical mode decomposition (EMD) • SVM (Support Vector Machine) classifier	• Facial recognition rate higher than 97% • Dimension reduction, • Similar recognition performance obtained in both using YUV color space and RGB color space • ROI extraction is accurate even if the brightness changes	• It only requires Full face without occlusion or partially occluded face for expression recognition

4 Conclusion and future work

During communication transmission facial expressions are produced so that the images can be obtained in unmanageable condition i.e. occlusion (effect of makeup, glasses, facial hair, hijab which can also affect the rate of recognition), illumination of light, posed expressions and variations in expression etc. This paper presented a survey on current work done in the domain of FER. Some techniques of feature extraction were explained. In addition, comparisons were also done which can help other researchers to advance and polish the present methods for getting accurate and better results in future. In future we are intended to investigate the local descriptor in frequency domain for real-world FER.

5 References

[1] Shuai-Shi Liu, Y. Tian and Dong Li, "New research advances of facial expression recognition," *2009 International Conference on Machine Learning and Cybernetics*, Hebei, 2009, pp. 1150-1155. https://doi.org/10.1109/ICMLC.2009.5212409

[2] R. Alazrai and C. S. G. Lee, "Real-time emotion identification for socially intelligent robots," *2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, 2012*, pp. 4106-4111. https://doi.org/10.1109/ICRA.2012.6224587

[3] Bee Theng Lau. "Portable real time emotion detection system for the disabled." *Expert Systems with Applications*. 37(9): 6561-6566, 2010.

https://doi.org/10.1016/j.eswa.2010.02.130

[4] X. Lu, "Image Analysis for Face recognition", [online] Available: http:/www.cse.msu.edu.

[5] P. Zhao-yi, Z. Yan-hui and Z. Yu, "Real-time Facial Expression Recognition Based on Adaptive Canny Operator Edge Detection," *2010 Second International Conference on Multimedia and Information Technology*, Kaifeng, 2010, pp. 154-157.
https://doi.org/10.1109/MMIT.2010.100

[6] C. Mayer et al., "A real time system for model-based interpretation of the dynamics of facial expressions," *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, 2008*, pp. 1-2.
https://doi.org/10.1109/AFGR.2008.4813440

[7] Z. Peng, Z. Wen and Y. Zhou, "Application of Mean Shift Algorithm in Real-Time Facial Expression Recognition," *2009 International Symposium on Computer Network and Multimedia Technology*, Wuhan, 2009, pp. 1-4.
https://doi.org/10.1109/CNMT.2009.5374770

[8] S. L. Happy, A. George and A. Routray, "A real time facial expression classification system using Local Binary Patterns," *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, Kharagpur, 2012, pp. 1-5.
https://doi.org/10.1109/IHCI.2012.6481802

[9] J. N. Bailenson, E. D. Pontikakis, I. B. Mauss, J. J. Gross, M. E. Jabon, C. A.C. Hutcherson, C. Nass, O. John, "Real-time classification of evoked emotions using facial feature tracking and physiological responses." *International journal of human-computer studies*. 66 (5): 303-317, 2008.
https://doi.org/10.1016/j.ijhcs.2007.10.011

[10] D. Ghimire and J. Lee "Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines." *Sensors*, 13(6): 7714-7734, 2013.
https://doi.org/10.3390/s130607714

[11] Jaewon Sung, Sangjae Lee and Daijin Kim, "A Real-Time Facial Expression Recognition using the STAAM," *18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, 2006, pp. 275-278.
https://doi.org/10.1109/ICPR.2006.158

[12] Lu, H-C., Y-J. Huang, and Y-W. Chen. "Real-time facial expression recognition based on pixel-pattern-based texture feature." *Electronics letters*. 43(17): 916-918, 2007.
https://doi.org/10.1049/el:20070362

[13] R. A. Khan, A. Meyer, H. Konik, S. Bouakaz, "Framework for reliable, real-time facial expression recognition for low resolution images." *Pattern Recognition Letters*, 34(10): 1159-1168, 2013.
https://doi.org/10.1016/j.patrec.2013.03.022

[14] B. M. Ghandi, R. Nagarajan and H. Desa, "Particle Swarm Optimization algorithm for facial emotion detection," *2009 IEEE Symposium on Industrial Electronics & Applications*, Kuala Lumpur, 2009, pp. 595-599.
https://doi.org/10.1109/ISIEA.2009.5356389

[15] A. Punitha, M. K. Geetha, HMM Based Real Time Facial Expression Recognition, *International Journal of Emerging Technology and Advanced Engineering*, 3(1), 180-185, 2013.

[16] A. M. Adeshina, S. Lau and C. Loo, "Real-time facial expression recognitions: A review," *2009 Innovative Technologies in Intelligent Systems and Industrial Applications, Monash*, 2009, pp. 375-378.
https://doi.org/10.1109/CITISIA.2009.5224179

[17] A. Geetha, V. Ramalingam, S. Palanivel, B. Palaniappan. "Facial expression recognition–A real time approach." *Expert Systems with Applications*. 36(1): 303-308, 2009.
https://doi.org/10.1016/j.eswa.2007.09.002

[18] Ebenezer Owusu, Yongzhao Zhan, Qi Rong Mao "A neural-AdaBoost based facial expression recognition system." *Expert Systems with Applications*, 41(7): 3383-3390, 2014.
https://doi.org/10.1016/j.eswa.2013.11.041

[19] F. Xijian, and T. Tjahjadi "A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences." *Pattern Recognition*, 48(11): 3407-3416, 2015.
https://doi.org/10.1016/j.patcog.2015.04.025

[20] Z. Ligang, D. Tjondronegoro, and V. Chandran "Facial expression recognition experiments with data from television broadcasts and the World Wide Web." *Image and Vision Computing*, 32(2): 107-119, 2014.
https://doi.org/10.1016/j.imavis.2013.12.008

[21] Hui Fang, Neil Mac Parthaláin, Andrew J. Aubrey, Gary K.L. Tam, Rita Borgo, Paul L. Rosin, Philip W. Grant, David Marshall, Min Chen, "Facial expression recognition in dynamic sequences: An integrated approach." *Pattern Recognition* ,47(3): 1271-1281, 2014.
https://doi.org/10.1016/j.patcog.2013.09.023

[22] Wang, Zhan, Qiuqi Ruan, and Gaoyun An "Facial expression recognition using sparse local Fisher discriminant analysis." *Neurocomputing*, 174: 756-766, 2016.
https://doi.org/10.1016/j.neucom.2015.09.083

[23] S.L. Happy, A. Routray. "Automatic facial expression recognition using features of salient facial patches." *IEEE transactions on Affective Computing*, 6(1): 1-12, 2015.
https://doi.org/10.1109/TAFFC.2014.2386334

[24] T. Ying, R. Chen, and Y. Cheng "Facial expression recognition algorithm using LGC based on horizontal and diagonal prior principle." *Optik-International Journal for Light and Electron Optics*, 125(16): 4186-4189, 2014.
https://doi.org/10.1016/j.ijleo.2014.04.062

[25] G. Ali, M. A. Iqbal, and Tae-Sun Choi. "Boosted NNE collections for multicultural facial expression recognition." *Pattern Recognition*, 55: 14-27, 2016.
https://doi.org/10.1016/j.patcog.2016.01.032

[26] Timothy J. Andrews, Heidi Baseler, Rob Jenkins, A. Mike Burton, Andrew W. Young. "Contributions of feature shapes and surface cues to the recognition and neural representation of facial identity." *Cortex*, 83: 280-291, 2016. https://doi.org/10.1016/j.cortex.2016.08.008

[27] A.H Matamoros, A. Bonarini, E.E.Hernandez, M.Nakano-Miyatake, H. Perez-Meana. "A facial expression recognition with automatic segmentation of face regions." *International Conference on Intelligent Software Methodologies, Tools, and Techniques*. 529-540, 2015. https://doi.org/10.1007/978-3-319-22689-7_41

[28] S.A. Khan, A. Hussain, M. Usman, "Facial expression recognition on real world face images using intelligent techniques: A survey." *Optik-International Journal for Light and Electron Optics*, 127(15): 6195-6203, 2016. https://doi.org/10.1016/j.ijleo.2016.04.015

[29] S.A. Khan, A. Hussain, M. Usman, "Reliable facial expression recognition for multi-scale images using weber local binary image-based cosine transform features." *Multimedia Tools and Applications*, 77(1): 1133–1165, 2018. https://doi.org/10.1007/s11042-016-4324-z

[30] A. Munir, A. Hussain, S.A. Khan, M. Nadeem, S. Arshid, "Illumination invariant facial expression recognition using selected merged binary patterns for real world images" *Optik-International Journal for Light and Electron Optics*, 158: 1016-1025, 2018. https://doi.org/10.1016/j.ijleo.2018.01.003

[31] S.A. Khan, M. Ishtiaq, M. Nazir, M. Shaheen, "Face recognition under varying expressions and illumination using particle swarm optimization," *Journal of Computational Science*, 28: 94-100, 2018. https://doi.org/10.1016/j.jocs.2018.08.005

[32] S.A. Khan, S. Hussain, S. Xiaoming, S. Yang "An Effective Framework for Driver Fatigue Recognition Based on Intelligent Facial Expressions Analysis", *IEEE ACCESS*. https://doi.org/10.1109/ACCESS.2018.2878601

https://doi.org/10.31449/inf.v42i3.2173 Informatica **42** (2018) 515–525 **515**

Automated Self-Learning Chatbot Initially Built as a FAQs Database Information Retrieval System: Multi-level and Intelligent Universal Virtual Front-Office Implementing Neural Network

Alessandro Massaro, Vincenzo Maritati and Angelo Galiano
Dyrecta Lab, Via V. Simplicio 45, 70014 Conversano (BA), Italy
http://www.dyrecta.com
E-mail: alessandro.massaro@dyrecta.com, vincenzo.maritati@dyrecta.com, angelo.galiano@dyrecta.com

Keywords: chatbot, artificial intelligence, neural networks, multi-level self-learning information system.

Received: February 2, 2018

The method proposed in this paper is based on dynamical information system capable toimplement a universal multi-level virtual front-office made by FAQs and chatbot self-learning systems.We describe statistics and tests necessary to validate the solution, and report acomparison between neural network and AIML results.

Povzetek: V prispevku je opisana nova metoda za izdelavo virtualnega asistenta iz arhiva vprašanj in odgovorov.

1 Introduction

Robot and artificial intelligence appeared a lot of time ago in the design of the "*Leonardo's mechanical knight*" [1] and earlier in the 12th century by Al-Jazari [2]. Artificial Intelligence (AI) begins eight centuries later with the conceptualizations of Alan Turing (Turing's test) [3]. The merging of robot and artificial intelligence is much more recent and opens perspectives with very strong potentials and related concerns. Virtual assistants as "*Chatbot*" (also known as a talkbot, chatterbot, Bot, IM bot, interactive agent, or Artificial Conversational Entity) can dialogue as real personal assistants; their main current use is on smartphones [4] to retrieve information useful for everyday life, to set the alarm or an appointment on the agenda, to send mail or text messages, find places or browse or launch apps. As robots enables mechanical automatisms, chatbot implementing AI allows the information automatism.

A chatbot is a computer program designed to simulate conversation with human users, especially over Internet; it acts like a human computer interface created to facilitate communication between human and computer, understanding natural language questions and answering with actual answers. Although chatbot is a current hot topic, it has been object for the past fifty years. The chatbot systems idea originated in M.I.T back in 1966, where professor Joseph Weizenbaum implemented the ELIZA chatbot to emulate a psychotherapist [5]. After ELIZA have been developed a lot of chatbots, for example to simulate the interaction with different personalities [6], matching with web-based search engines (AskJevees) [7], and with open-source initiatives like ALICE [8] [9] implementing artificial intelligent applications called AIML (Artificial Intelligence Markup Language) [8]. AIML is a widely adopted standard for creating chatbots and mobile virtual assistants like ALICE [8], Mitsuku [10], English Tutor [11], The Professor [12] and many more. Over the years, chatbots have become a sophisticated tool, able to perform natural conversations and make users happy for the quick support they can provide. Although a chatbot cannot handle all customer queries, it can be used to deal with many of the routine queries activating service requests.

The knowledge bases of existing chatbots are mostly built manually [13], thus requiring a long time, and are difficult to adapt to new domains. There are several researches on the extraction of knowledge from different types of data sets [14] [15] [16] but these approaches use the characteristics of their domains and are therefore only suitable for their specific tasks, limiting the possibility to transform directly their methods in a general knowledge extraction approach.

Nowadays there are tens of thousands chatbots available online, over 30K already on the Messenger platform alone [17], and their number is growing rapidly thanks to the ease of implementation and distribution provided by services such as Facebook Messenger, Telegram, etc. Users around the world are logging into messaging apps to not only chat with friends but also to connect with brands, to browse merchandise, and to watch content (in the mid-2014 the number of messages exchanged on the main four platforms of instant messaging has exceeded the number of messages exchanged on the 4 main social networks [18]).

The approach described in this paper involves the implementation of a smart virtual front-office model that exploits the synergy between a list of FAQs and a chatbot [19], [20]. The main features of this model are:

- be able to self-learn their knowledge base from an archive of questions and answers (FAQs) organized in a static tree structure;
- the use of machine learning algorithms to dynamically generate its own knowledge base;
- the speeding up of the management of most user requests;
- the possibility to acquire different feedbacks thus improving the efficiency of the system, so as to be able to easily reshape the available data according to greater effectiveness in supporting users;

- allowing full control over the management activities of the system itself.

Microsoft today offers a service loading knowledge base (KB) for a chatbot by copying and pasting phrases of FAQs [21]. Recently some researchers implemented long short-term memory (LSTM) neural networks [22], which automatically generate responses for users requests on social media. Other authors highlighted that a chatbot does not understand colloquial usage [23], and cannot yet simulate the full range of intelligent human conversation

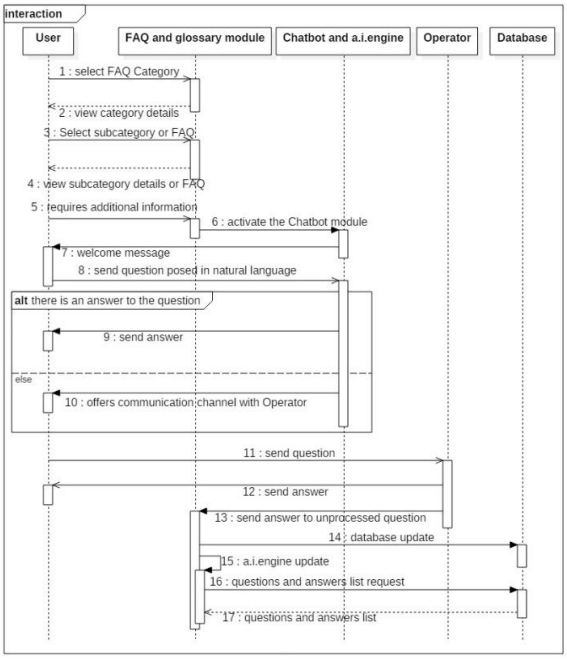


Figure 2 - System's interactions

[24].

In the proposed research are presented the following claims with respect to the state of the art:

-automatic loading and construction of the knowledge base of the chatbot system (automatic generation of the AIML files constructing the KB by an automatic data entry);

-automatic updating of the training dataset, progressively reducing the probabilistic error of request recognition (see self-learning multi-level model enabling automatic storing of the training dataset described in the next section);

-gradually overcoming the barrier due to the training of model including colloquial usage forms.

2 Case study

The proposed case study consists of improving customer support through a website by exploiting the FAQs archive available to its users, and useful for the automated creation of a chatbot. The management of requests for assistance from a chatbot will reduce the workload of human operators dedicated to customer support and will allow the assistance service to be operational even in the absence of available operators.

If the system does not find a response to a user question, it can redirect the user to a human operator or store the user questions and data in order to send these information to the first available operator.

Figure 1 shows a functional scheme of the whole system.

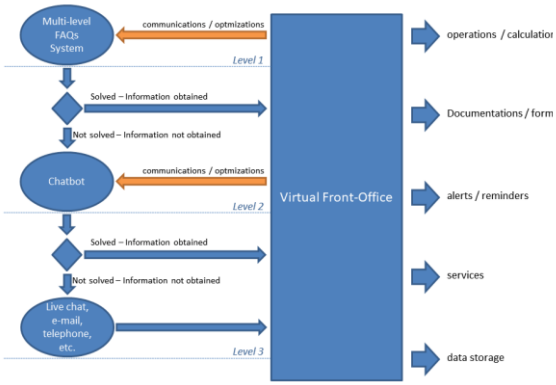


Figure 1 - Functional scheme

The multi-level access to the different modules of the system allows the optimization of available resources and, at the same time, a better user-experience:

- according to the sections consulted by users, a series of relevant FAQs, organized in a tree structure, are proposed (***level 1***);
- when the user requests a contact with an operator to ask further information, he is put in touch with the chatbot that, according to the user's requests and related information, proposes the most relevant answers (***level 2***);
- if the chatbot does not find a response for the user, it puts him in contact with a specific human operator, selected by the system on an available operators list, on one of the available channels (e.g. e-mail, chat, telephone – ***level 3***), thus optimizing the response time; the response provided by the operator is memorized by the system and is exploited by a self-learning algorithm to increase the knowledge base of the system.

Operators are able to improve the knowledge base and the overall performance of the system taking advantage of:

- the archive of questions to which the chatbot does not respond (the training dataset is gradually enriched by adding a new FAQ formulated through of the three level responses if Fig. 1);
- a series of tools, tests and indications provided by the system in a fully automatic way in order to make the operators capable of optimizing the knowledge base (operators is guided by a platform allowing an automatic update of the training dataset).

The system’s interactions are sketched in the diagram of Figure 2.

The diagram of Figure 3 represents the interactions between the chatbot module and the artificial intelligence engine.

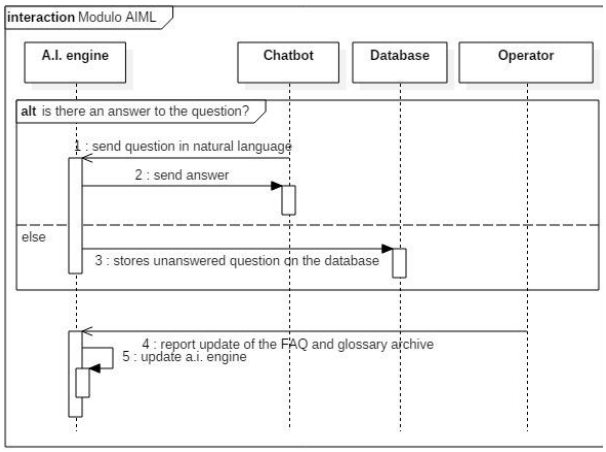


Figure 3 - The interactions between the Chatbot and the artificial intelligence engine

3 System design

The system has been designed in a modular way to allow fast integration into any web-based platform and isn't bound to a particular environment. It can be logically divided into three macro-modules:

- FAQ: management by authenticated users (operators and administrators) and display by unauthenticated users of FAQs and Glossary items;
- Chatbot: management and use of the "Chatbot" application that will allow interaction between users and the system by simulating communications in natural language;
- Back-office: management of data and the consequent information in the system available to users; system supervision; statistics; system optimization.

These three macro-modules are detailed in the use case diagram of Figure 4, where there are explained the main actions performed; this diagram shows the actors of the system:

- the end user, which accesses to the FAQ and chatbot system to find answers to some of his questions;
- the operator of the system that has access to the back-office interfaces for managing the information that the system makes available to the users;
- the system administrator who has access to back-office interfaces to supervise the system's activities.

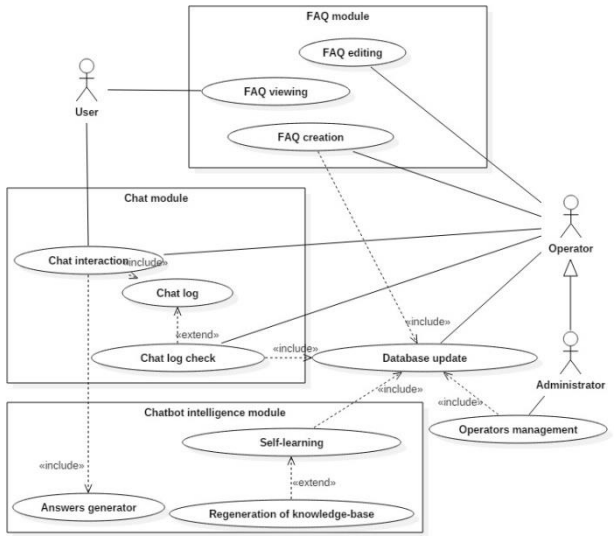


Figure 4 - Use case diagram

The database is used as aggregator of the data coming from the system modules and provides the necessary information for data processing. In this way it is possible to decouple the logic operations from the data administration, as a function of a smart management of all the underlying functionalities.

During a user's visit, the system stores its preferences and the sections visited so that these information can be used to better evaluate which answers to generate to the user's questions.

Each component of the system cooperates synergistically with the others in order to improve the user-experience; at the same time this integration allows to update the system's knowledge base by simply operators tools, exploiting an archive of questions to which the system has not been answered, and providing useful indications on how to improve the knowledge base in function of a better performance.

The diagram of Figure 5 shows the interaction between the system components.

The **autoresponder with self-learning module** consists of two modules: an artificial intelligence (A.I.) module, and a chatbot module; the A.I. module generates the answers of the user questions, and allows to train the knowledge model. Both modules are connected to the database but only the chatbot module modifies it, by inserting new data useful for statistical purposes and for the optimization of the knowledge base. The use of machine learning algorithms based on neural networks has been compared with the use of AIML during the design of the artificial intelligence part of this module, in order to validate the proposed final solution. The chatbot module also manages user questions and system-generated responses to make them available to administrators for control purposes, and to generate statistics on system usage. An algorithm has been modeled to read the knowledge base available in the database and to transform these data into a format that is exploitable by neural networks able to generate AIML patterns.

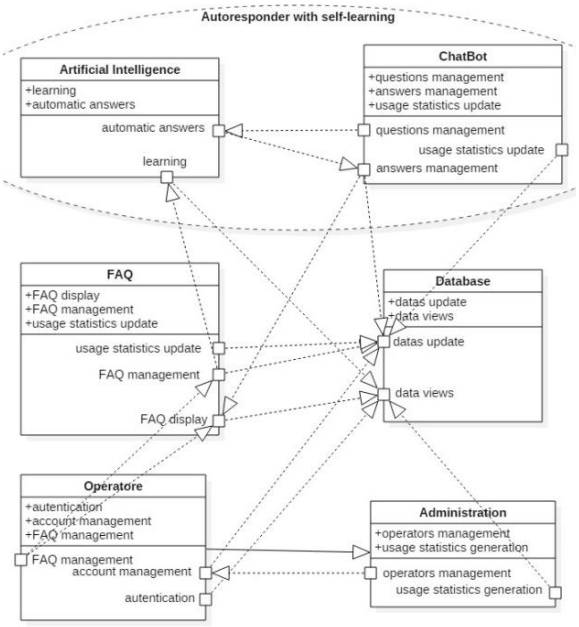


Figure 5 - Composite structure diagram

The database module deals with exchanging data with the database server by recycling and caching queries in order to optimize the performance; this module provides a single interface for all modules of the system connected to the database.

The FAQ module allows the complete management of the FAQs and glossary entries; this module updates the system usage statistics regarding the operators activities.

The operator module manages all the actions concerning the operators of the system, by allowing the authentication of each operator or administrator, the modification of his personal data and, in the case of users authenticated as administrators, the management of other accounts through the administrator module; the operator and the administration write on the database all the actions performed by a user for statistical purposes.

3.1 Data model

The entity-relationships model of Figure 6 highlights the data necessary for each system entities and how these are associated with each other.

This simplified diagram shows a clearer description of the main characteristics of the database architecture.

From Figure can be deduced the following data structure:

- FAQs are grouped into categories that can be grouped into other parent categories themselves, creating a tree structure that simplifies the selection and use of FAQ data;
- every user can have different roles in the system (*tables: User, UserRole* and *Role*)
- there may be more different answers to a question and

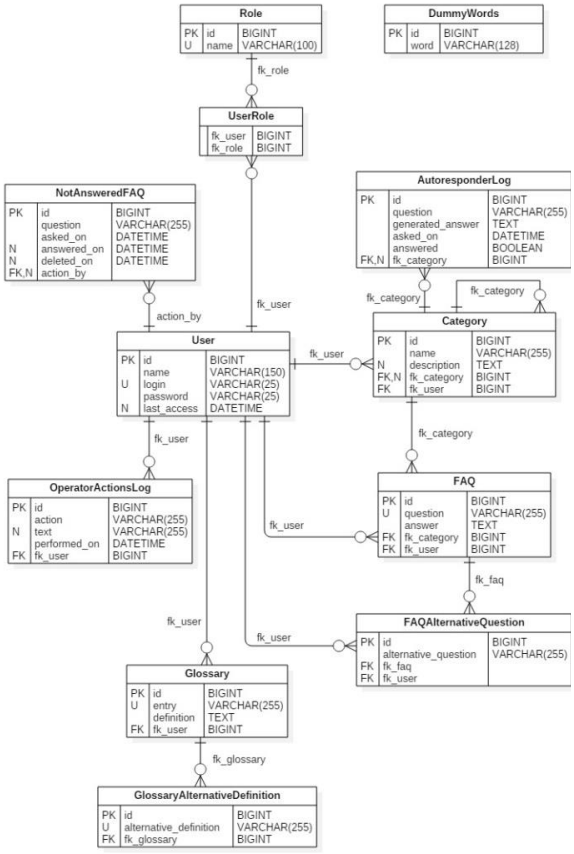


Figure 6 - Entity-relationship diagram

there may be different definitions for a glossary entry (*tables: FAQAlternativeQuestion* and *GlossaryAlternativeDefinition*);

- words without semantic content are stored in a specific table (*DummyWords*);
- all the users' questions to which the chatbot has not answered are traced (*table: NotAnsweredFAQ*); the operators can use these questions to create new FAQs or to add new answers to existent FAQs;
- all the chatbot answers and the corresponding user questions are traced (*table: AutoresponderLog*);
- all the actions performed by the operators on the system are traced (*table: OperatorActionsLog*) for statistical and control purpose.

The dynamic training dataset of the chatbot system (initial knowledge base) is built from the previous described data model: the dataset is constructed by a starting database containing glossary, by the FAQs database, and by the *DummyWords* list (words to ignore).

3.2 Self-learning algorithm

Upon receipt of a natural language question asked by a user, the first operation to be performed is the elimination of all words without semantic meaning according with the context in which the chatbot operates; this operation is performed through a comparison with a list of words, characters and symbols to be ignored.

Once the words without semantic meaning are removed from the original question, a list of remaining keywords remains is passed to the A.I. engine thus activating the answer searching process.

If the A.I. engine finds a response to the keyword sequence, this is sent to the user; otherwise the question is stored to be subsequently verified by a human operator. According to the last user choices and the context in which the chat has been opened, the question is passed to a specific operator.

When an operator answers to an user question, the response provided by the human operator is stored, and it is used to regenerate the knowledge-base.

This algorithm depicted in the next diagram of Figure 7.

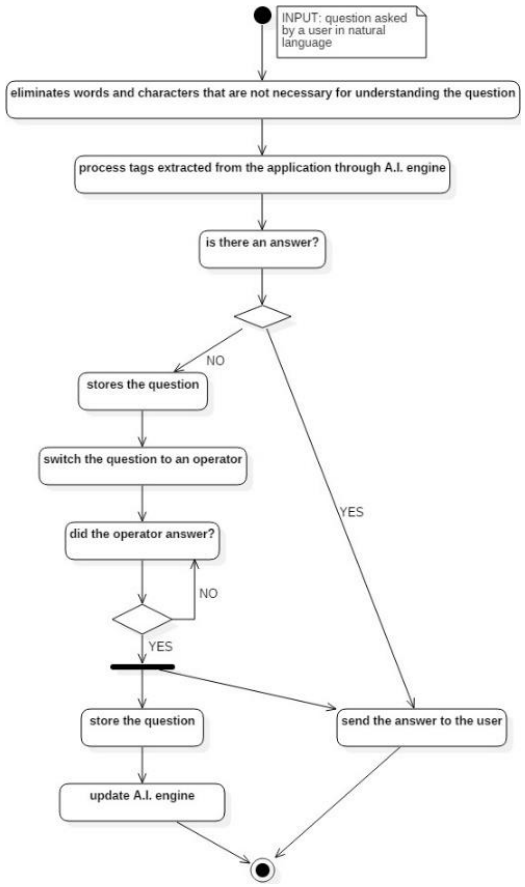


Figure 8 - Self-learning algorithm

Second action of the previous flow-chart "process tags extracted from the application through A.I. engine" underlies two A.I. pre-processing algorithms: one for the automatic production of standard AIML patterns, and one for setting the right weights of the neural network. Pre-processing algorithms take into account user question generating outputs for feeding to the respective A.I. engine. The reply of these engines has the same data type (a natural language phrase and a Boolean value indicating response validity).

The AIML A.I. engine is based on an open-source library compatible with AIML 1.0.1 standard. The artificial neural network (ANN) is a fully connected neural network that is used as a universal functions approximator performing the training, and behaving as a classifier of questions.

3.2.1 Automated AIML patterns production

The automatic production of AIML patterns based on system knowledge-base (FAQs, glossary entries, dummy words), is managed by an iterative algorithm that extracts keywords from the input text and using them to construct AIML patterns (see Figure 8). All the generated AIML patterns are stored in a file ready to be processed by the AIML engine.

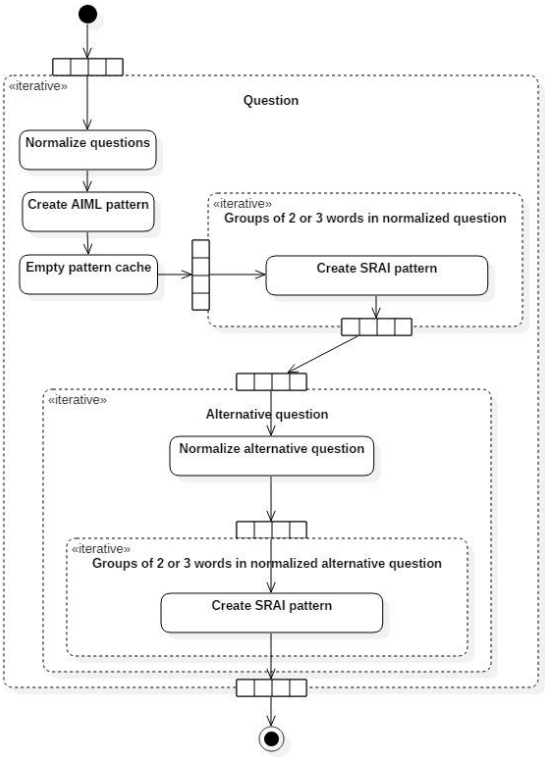


Figure 7 – Automatic AIML patterns creation

For the automatic reconstruction of AIML files, all the questions (between FAQ, Glossary and alternative questions of both) are analyzed in a cycle that, for each question:

1. normalize the question in the archive (remove words with no semantic meaning and unwanted characters such as punctuation, dashes, etc.);
2. create an AIML pattern on the normalized question;
3. clean the AIML pattern cache;
4. for each group of 2 and 3 words taken from words in the normalized question of point (1):
   - create a Symbolic Reduction in Artificial Intelligence (SRAI) pattern on the AIML pattern;
   - verify that the SRAI pattern has not already been created based on the AIML pattern cache;
5. for each alternative question to the current question:
   - normalize the alternative question, as for the original question;

- for each group of 2 and 3 words taken from words in the normalized alternative question:
  - create a SRAI pattern on the AIML pattern of the original question;
  - verify that the newly created SRAI pattern has not already been created based on the AIML pattern cache.

This process is started automatically when an operator changes the knowledge base of the system.

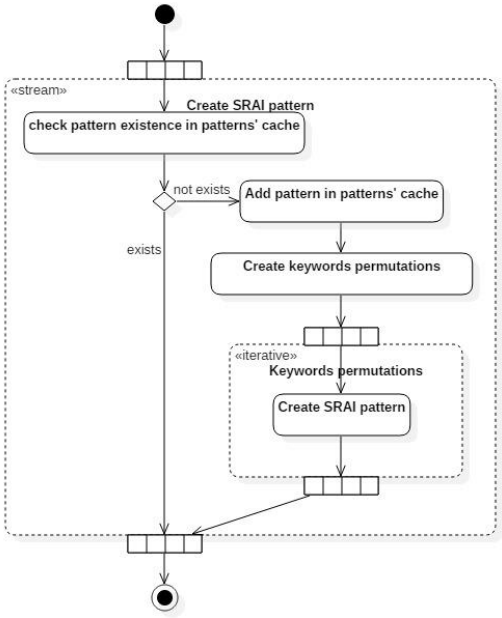


Figure 9 - SRAI pattern creation

It is possible to disable the automatic creation of AIML patterns in order to allow massive changes to the knowledge base and to re-enable this functionality at the end of all changes: in this way the operators avoid making unnecessary reconstruction of the AIML patterns until all the necessary changes have been made to the knowledge base.

The diagram in Figure 9 exhibits the steps for the creation of each SRAI pattern related to each original question and to all the alternative questions generated by the original question.

By compared Figure 9 with Figure 8, the only passage to be explained in this diagram is the "Create keywords permutations" action, having the task to create the SRAI pattern in different combinations by the 2 or 3 selected keywords using the wildcard character [*] provided by the AIML standard.

3.2.2 AIML optimizer

In order to find the best combination of patterns for the construction of the AIML file that will manage the chatbot, an automatic algorithm has been implemented. The chatbot:

- is able to find the best combination of keywords / patterns among all those that can be created;
- can find any redundant questions, which lead to different answers while being syntactically similar (these questions are reported to the operators so that they can verify and modify them according with an appropriate knowledge base).

This algorithm is called for each editing process in the knowledge-base. The AIML optimizer algorithm is described in the next flow-chart diagram of Figure 10, where it is highlighted the optimization procedure that aims to remove some keywords between those obtained in the analyzed question thus activating the system response.

3.2.3 AIML vs neural network architecture

An AIML module is based on the recognition of precise textual patterns. By using a defined logic in the patterns definition, the AIML allows the construction of recursive patterns and the use of wildcards that subtend one or more words, making AIML a standard ready to simulate a conversation in natural language.

Although the process of creating AIML patterns can be automated, the final result is a series of static patterns, which will respond to the inputs according to precise rules and structures.

Better results can be achieved using an artificial neural network instead of an AIML-based system, because:

- there will be a better correspondence between the questions proposed to the system and the answers provided by it (depending on the implemented artificial neural network architecture, it is possible to construct complex maps between the input and the output text generated by the system, exceeding the limits and the static nature of AIML patterns and handling specific or difficult cases);
- the execution speed is greater, especially in conjunction with large archives of texts to be managed (while the number of texts to be managed increases, the number of AIML patterns necessary to manage all the possible cases increases; an artificial neural network does not need the addition of new neurons or new layers in order to manage a new texts; moreover the recognition of AIML patterns occurs through the direct comparison of text strings which is a very slow operation, especially when compared to the additions and multiplications used in neural networks, operations in which the electronic computers excel thus requiring parallel computing). In Figure 10 is illustrated the flowchart of the AIML optimizer.

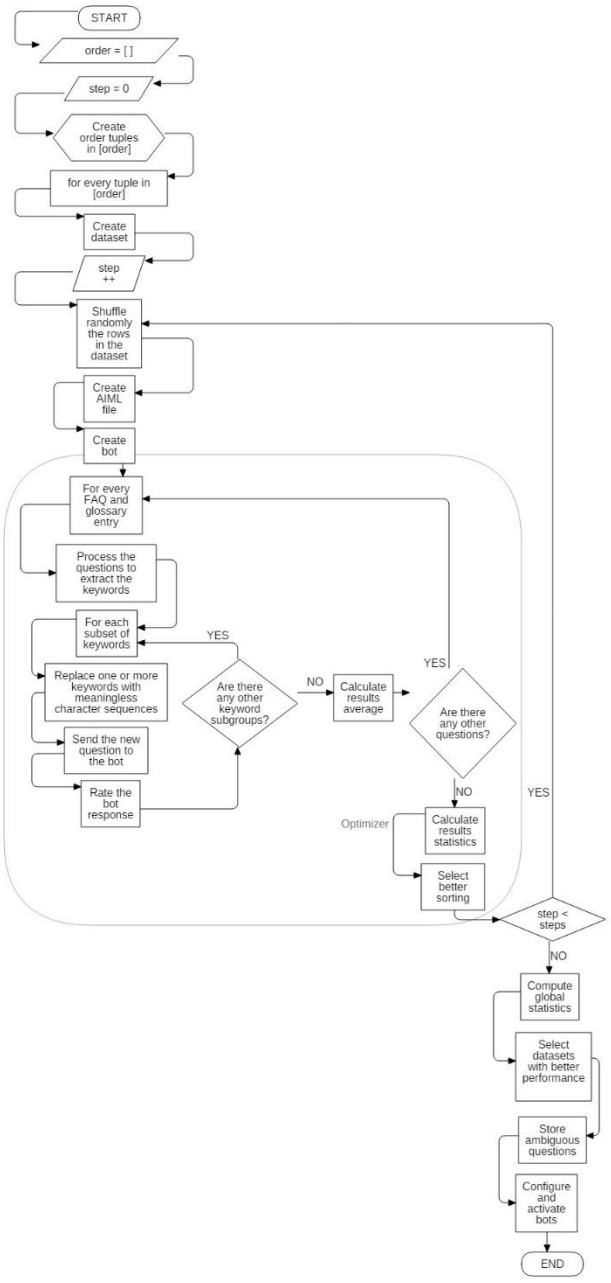


Figure 10 - AIML optimizer

3.2.4 Automated neural network

The automatic creation of the neural network has been implemented by exploiting some algorithms able to process the knowledge base of the system without the intervention of human actors. The main parameters of the neural network, like the number of neurons used in the hidden layers, are automatically set on the basis of the available dataset, so as to allow an adaptation of the solution to any new dataset suitable for the system implementation.

The technique called "one hot encoding" [25] (discussed in the following paragraph) has been adopted to make the text of the questions usable by the underlying neural network. Mathematically, one hot encoding generates a balanced matrix, which is easy to understand during complex computations inside algorithms. One hot encoding technique is then used to encode categorical features: one-hot is a group of bits among which the allowed combinations of values are only those with a single high bit (1), and all the others are low (0) [25].

3.2.4.1 One-hot encoding

The words found in the questions archive, allow to create a dictionary. In this dictionary, each word will correspond to an index. In order to transform any new sentence into binary vectors, a binary vector is created having a length equal to the number of words of the dictionary and having all the elements equal to zero. Below is shown an example useful to understand the one-shot encoding approach.

For example, taking the following three sentences:

```
[ today is a beautiful day ]
[ she is a beautiful girl ]
[ that girl has a red car ]
```

is created the following dictionary of 11 words:

```
[today,is,a,beautiful,day,she,girl,that
,has,car,red]
```

A new sentence like:

```
[ your new red car is simply beautiful ]
```

would be converted to the following binary vectors:

```
[ 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1 ]
```

which turned into text would be for the four weighted keywords:

```
[ new beautiful car red ]
```

The semantic meaning of the sentence has been lost during the transformation, but a fully connected ANN processing this input can still find a combination of its weight values to match the best answer to each question ignoring the order in which the words of the sentence are placed: the combination of words is processed by taking into account only the maximum weight of their occurrence into a sentence.

Another way is to create a binary vector for each word of the input question and thus create a binary matrix able to bring back the order in which the single words succeed each other in the original question.

The `[ your new red car is simply beautiful ]` sentence would be converted to the following binary array:

```
[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ]
[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ]
[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0 ]
[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1 ]
[ 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0 ]
[ 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 ]
[ 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0 ]
```

where the units represent the keyword presence structured in a predetermined sized matrix, allowing also to extract information of the word position into the sentence (important aspect for colloquial usage forms). In this matrix each line corresponds to a word of the sentence, and the lines are ordered following the order of the analyzed sentence. In each row, the position of a "1" indicates which word to consider in the sentence sequence

This array make possible to quickly process the relative textual information. If the array is processed by the recurrent neural network, will be also possible to analyze the order of the input keywords.

To obviate the problem of the management of sentences composed by different number of words, it is used the technique of zero-padding: by this approach are added null vectors having a number equal to the number of missing words up to a value N, where N is the number of words managed; if a sentence has a number of words that exceeds the limit of the words managed, only the first N words will be considered for the data processing.

3.2.4.2 Neural network based chatbot

In our model a normalization function on the texts to be processed is considered. This function clean the words with no semantic meaning and not useful for the project aims. All input texts are cleaned by articles, adverbs, prepositions, special characters, accented letters, etc. in order to analyze the input text exclusively by the key words useful in tracing the most appropriate answer for the input question.

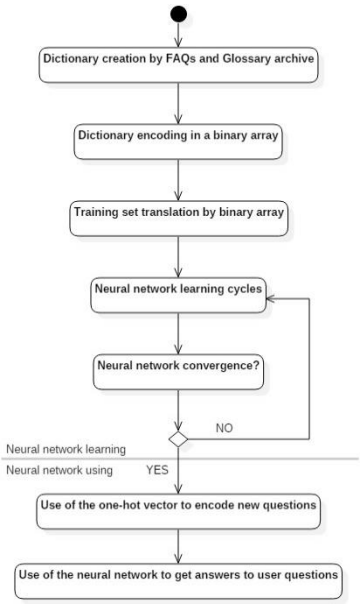


Figure 11 - Neural network learning and using model

In Figure 11 is illustrated the diagram of the used model.

The learning phase (upper part of Figure 11) is started at each database variation made by an operator updating the questions/answers knowledge base. The learning phase duration depends on the number of questions to be analyzed, but is faster in the execution if compared with the equivalent algorithm of the AIML patterns creation.

Another advantage in the execution time is also checked when the neural network is used to answer new questions.

3.2.4.3 Neural network model

The ANN model used in the project is reported in the following Python script based on the Keras library:

```
  model = Sequential()
  model.add(Dense(input_features,
activation='relu',
input_dim=input_features ))

model.add(Dense(int(input_features/1.5),
activation='relu'))

model.add(Dense(int(input_features/1.5),
activation='relu'))
  model.add(Dense(int(input_features/2),
activation='relu'))
  model.add(Dense(int(input_features/2),
activation='relu'))
  model.add(Dense(int(input_features/3),
activation='relu'))
  model.add(Dense(int(input_features/3),
activation='relu'))
  model.add(Dense(output_class,
activation='softmax'))
  model.compile(loss='categorical
crossentropy', optimizer='nadam',
metrics=['accuracy'])
```

The main characteristics of the proposed model are:

- parameterization of the number of neurons for each layer based on the dictionary size;
- use of the ReLu activation function to speed up the learning process [26], [27];
- use of Nesterov Adam optimizer [28] to obtain a faster convergence of the neural network;
- reduction of the number of neurons per layer to induce the ANN to behave like an encoder able to classify the questions according to the available answers.

4 Tools and development

The system is based on Python 3.x, using the Django web framework. The database is built in MySQL Community Edition 5.7.

The Python libraries used are:

- Keras - deep learning library;
- Tensorflow - an open source software library for Machine Intelligence;
- MySQLdb - thread-compatible interface to MySQL database server;
- NumPy - a package for scientific computing;
- PyAIML - an interpreter package for AIML.

Some features of the system are performed by exploiting Jupiter Notebook on an Anaconda distribution; this development method allows to quickly test thus validating some modules of the system before inserting them as components of the final prototype system (preliminary test phase).

Eclipse Oxygen was used as IDE, with PyDev plugin installed to support Django's modules development.

Some useful functions are adopted to test the system and to optimize the underlying dataset. By using these functions, it is possible to evaluate the efficiency of the AIML Chatbot system vs Neural Network Chatbot system without an huge dataset of new questions to be submitted to the system.

The most relevant function for the optimization of the dataset used for the learning stage of the A.I. module, is a function that adds random noise in the questions used for training and that uses the new produced questions to verify the effectiveness of A.I. algorithm. This function is based on the optimizer of the algorithm described in 3.2.2 where the only difference is that the ANN is used instead of the AIML engine for the evaluation of the chatbot performance.

Comparing the answers generated by the system on all the questions in the archive, and modifying these questions in different ways, it is possible to construct a metric able to indicate the redundancy degree of the questions sored into the archive. Staring to the results provided by this function, it is possible to create visual interfaces for operators able to highlight any questions to be modified. These interfaces are useful to build a more appropriate knowledge base.

The following table shows the average results obtained by applying the described algorithm applied on 10 datasets of questions and answers concerning 5 different application domains; each dataset includes 180÷300 questions and answers whereas a set of questions can be referred to a single answer:

Keeped Keywords	Accuracy	Mistaked Answers	No reply
All	98,925%	0%	1,075%
None	0%	0%	100%
50%	37,634%	3,226%	59,14%
66,6%	83,871%	0%	16,129%
75%	89,247%	1,075%	9,677%
80%	98,925%	0%	1,075%
83,3%	98,925%	0%	1,075%
85,71%	98,925%	0%	1,075%

where “all keeped keywords” refers to the results of the original dataset, the “none keeped keywords” refers to the obtained results skipping every keywords (the system has not knowledge base), the “50% keeped keywords” refers to the obtained results skipping a keyword every two keywords, the “66,6% keeped keywords” refers to the obtained results skipping a keyword every three keywords and so on.

Analyzing the questions to which the system responded by proposing an incorrect answer, it is possible to make a correction on the same questions, exploiting the keywords highlighted as equivocal to replace these questions with similar ones but based on different keywords. The analysis of the unanswered questions generated by the system provide an indication to the operators on which topics to generate new questions and / or answers in order to enrich the system knowledge base in an optimized way. The accuracy index provides a useful value to compare the system's performance before and after any substantial changes to the knowledge base so as to report to the operators any final performance decays.

5 Results

The approach described in this document allows to obtain a highly automated management system of front-office services applicable in any context characterized by a knowledge base consisting only of an archive of questions and answers like a FAQs archive. The main advantage in adopting this approach is given by the high automation of the different processes underlying the system, which allows:

- an easier management of front-office procedures by implementing an efficient multi-level architecture able to automatize the self-learning training process;
- an easier data entry procedure interfaced with a portal framework (the user adds requests online into a web page thus enriching automatically the training dataset and eliminating other manual steps for the dataset construction);
- a lower workload for operators, enabled by the chatbot that replaces operators in most users' questions (when the training dataset will be efficient the operators will be totally theoretically replaced by the virtual assistant);
- the applicability in any context, not dependent on the technologies used (the self-learning process could be applied in different applications where users require responses);
- the usage of the chatbot system also for colloquial usage forms.

The comparison between AIML and ANN for the automatic construction of chatbot based on the same input datasets, allowed to verify the best performances of the second approach. Using the metrics described in the previous paragraph it has been possible to find a better accuracy, a smaller number of wrong answers and a smaller number of missed answers through the use of an ANN through to the use of AIML, as shown in the following table:

Artificial intelligence module	Accuracy	Mistaked Answers	No reply
AIML	85%	5%	10%
Fully connected neural network	99%	0%	1%

An improvement to the proposed solution may be provided by the use of a recurrent ANN, in order to exploit the memory of the neural network to better associate the responses available in the system to the questions of its users, thus recognizing the order in which the keywords are arranged in the sequences created from a user’s question.

6 Conclusion

Authors proposed in this work an innovative self-learning multi-level virtual front office, by improving a structured FAQ procedure together with an innovative chatbot system. The proposed model is suitable for industrial applications requiring the optimization of human resources activities. The goal of the self-learning model is

to eliminate totally the humans responses represented by level 3 of the model of Fig. 1, by constructing dynamically and automatically the training dataset. The model could be applied to Big Data and Machine to Machine (M2M) systems [29]. The experimental dataset is reported in [30].

7 Acknowledgement

The work has been developed in the frameworks of the Italian projects: "*Piattaforma universale multilivello di front office virtuale intelligente*" (Universal multi-level platform of intelligent virtual front office) -Multi-Level Virtual Front Office-. Authors gratefully thanks V. Antonacci, L. D'Alessandro, G. Lonigro, G. Ronchi, G. Sicolo and E. Valenzano. Special thanks are addressed to N. Calamita and P. Re David for their support in the definition of the application's scenario.

8 References

[1] M. E. Rosheim (2006) "Leonardo's Lost Robots" Springer-Verlag Berlin Heidelberg 2006.

[2] W. Aguilar, G. Sanatamaria-Bonfil, T. Froese and Carlos (2014) "The Past, Present, and Future of Artificial Life" Frontiers in Robotics and AI, Vol. 1.

[3] Turing, A. (1950). "Computing Machinery and Intelligence" Mind vol. LIX N° 236. https://doi.org/10.1093/mind/LIX.236.433

[4] S. V. Doshi, S. B. Pawar, A. G. Shelar and S. S. Kulkarni (2017) "Artificial Intelligence Chatbot in Android System using Open Source Program-O" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6 (4), 816-821. https://doi.org/10.17148/IJARCCE.2017.64151

[5] Weizenbaum, J. (1966) "ELIZA - A Computer Program for the Study of Natural Language Communication between Man and Machine" Communications of the Association for Computing Machinery 9: 36-45. https://doi.org/10.1145/365153.365168

[6] Colby K.M. (1999) Human-Computer Conversation in A Cognitive Therapy Program. In: Wilks Y. (eds) Machine Conversations. The Springer International Series in Engineering and Computer Science, vol 511. Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-5687-6_3 AskJevees. (2004). [Online]: https://uk.ask.com/

[7] Wallace, R. (2004). "The elements of AIML style." ALICE AI Foundation.

[8] R. S. Wallace (2003). [Online]: http://alice.pandorabots.com by ALICE AI Foundation

[9] Steve Worswick (2003). [Online]: http://www.mitsuku.com by Square Bear.

[10] English tutor (2011). [Online] https://www.eslfast.com/robot/english_tutor.htm by ESL Robot.

[11] The Professor (2012). [Online]: https://www.pandorabots.com/pandora/talk?botid=935a0a567e34523c by ALICE AI Foundation.

[12] J. Huang, M. Zhou and D. Yang (2007) "Extracting Chatbot Knowledge from Online Discussion Forums" IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence, 423-428.

[13] L. Shrestha and K. McKeown (2004). "Detection of question-answer pairs in email conversations." In Proceedings of Coling, pp.889-895. https://doi.org/10.3115/1220355.1220483

[14] R. Nishimura, Y. Watanabe and Y.Okada (2005). "A Question Answer System Based on Confirmed Knowledge Developed by Using Mails Posted to a Mailing List." In Proceedings of the IJCNLP 2005, pp.31-36.

[15] L. Zhou and E. Hovy (2005). "Digesting Virtual Geek Culture: The Summarization of Technical Internet Relay Chats" In Proceedings of ACL 2005, pp.298-305. https://doi.org/10.3115/1219840.1219877

[16] Edelman Digital's (2016). "2017 Trend Report" [online] https://edelmandigital.com/wp-content/uploads/2016/12/2017-Edelman-Digital-Trends-Report.pdf

[17] Business Insider (2015). "Messaging apps are now bigger than social networks" [online] http://uk.businessinsider.com/the-messaging-app-report-2015-11

[18] A. Shawar, B.A. Atwell, A. Roberts (2005). "FAQchat as in Information Retrieval System" Human Language Technologies as a Challenge. Proceedings of the 2nd Language and Technology Conference, Wydawnictwo Poznanskie.

[19] Z. Wang, A. Ittycheriah (2015) "FAQ-based Question Answering via Word Alignment," quesarXiv: 1507.02628v1 [cs.CL] 9 Jul 2015. Microsoft 2018.[Online]: https://www.qnamaker.ai/

[20] A. Xu, Z. Liu, Y. Guo, V. Sinha and R. Akkiraju (2017) "A New Chatbot for Customer Service on Social Media" CHI '17 Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 3506-3510. https://doi.org/10.1145/3025453.3025496

[21] A. Deshpande, A. Shahane, D. Gadre, M. Deshpande and P. M. Joshi (2017) "A Survey of Various Chatbot Implementation Techniques" International Journal of Computer Engineering and Applications, Vol. 11.

[22] J. Hill, W. R. Ford and I. G. Farreras (2015) "Real Conversations with Artificial Intelligence: A comparison between Human–Human Online Conversations and Human–Chatbot Conversations" Computers in Human Behavior, Vol. 49, 245–250. https://doi.org/10.1016/j.chb.2015.02.026

[23] D. Harris and S. Harris (2012) "Digital Design and Computer Architecture" Elsevier book 2nd edition, 129.

[24] Andrew L. Maas and Awni Y. Hannun and Andrew Y. Ng (2013) "Rectifier Nonlinearities Improve Neural Network Acoustic Models".

[25] Vinod Nair and Geoffrey Hinton (2010) "Rectified linear units improve restricted Boltzmann machines".

[26] T. Dozat (2016) “Incorporating Nesterov Momentum into Adam” Stanford ICLR 2016 workshop submission.

[27] A. Galiano, A. Massaro, D. Barbuzzi, L. Pellicani, G. Birardi, B. Boussahel, F. De Carlo, V. Calati, G. Lofano, L. Maffei, M. Solazzo, V. Custodero, G. Frulli, E. Frulli, F. Mancini, L. D’Alessandro, F. Crudele (2016) “Machine to Machine (M2M) Open Data System for Business Intelligence in Products Massive Distribution oriented on Big Data,” (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (3), 1332-1336. Github dataset .[Online]: https://github.com/DyrectaLab/faq-data .

https://doi.org/10.31449/inf.v42i4.1132 Informatica **42** (2018) 527–533 **527**

Automatic Estimation of News Values Reflecting Importance and Closeness of News Events

Evgenia Belyaeva*, Aljaž Košmerlj, Dunja Mladenić* and Gregor Leban
Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
Email addresses: firstname.lastname@ijs.si

*Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

Keywords: news values, newsworthiness, text mining, Apple

Received: January 27, 2016

This paper addresses a problem of automatic estimation of three journalistic news values, more specifically frequency, threshold and proximity, by applying various text mining methods. Although theoretical frameworks already exist in social sciences that identify if an event is newsworthy, these manual techniques require enormous amount of time and domain knowledge. Thus, we illustrate how text mining can assist journalistic work by finding news values of different international publishers across the world. Our experiments both on a collection of news articles from different publishers about Apple's launch of new iPhone 6 and Apple Watch and on a wider collection of documents confirm that some journalists still follow some of the well-known journalistic values. Furthermore, we acknowledge that news values are often orthodox and outdated, and no longer apply to all publishers. We also outline possible future implications of our approach to work on interaction between text mining and journalistic domains.

Povzetek: Članek obravnava problem avtomatske ocene novičarskih vrednosti, natančneje: pogostosti, prag pomembnosti ter bližine (oz. relevance), z uporabo različnih metod avtomatske analize teksta. Čeprav v družboslovju obstaja več teoretičnih ogrodij, ki določajo ali je nek dogodek vreden poročanja, te temeljijo na »ročnem« delu in zahtevajo veliko časa ter globoko poznavanje domene. V članku nakažemo, kako lahko avtomatska analiza teksta pomaga pri novinarskem delu z uvidom v novičarske vrednosti na globalnem nivoju in med velikimi, mednarodnimi mediji. Rezultati naših poskusov na zbirki člankov iz različnih virov (spletnih časopisov) o splovitvi izdelkov iPhone 6 in Apple Watch podjetja Apple potrjujejo, da novinarji še sledijo nekaterim uveljavljenim novičarskim vrednostim. Ob tem ugotavljamo, da so nekatere novičarske vrednosti preveč ortodoksne in zastarele ter ne veljajo več za vse novičarske vire. Na koncu orišemo načrtovano nadaljnje delo in možne implikacije uporabe avtomatskih metod analize teksta na novinarsko delo.

1 Introduction

Every news outlet has a different agenda for selecting which news stories to cover. Mass media have traditionally relied on the so-called news values to evaluate newsworthiness of a story i.e. what to publish and what to leave out, introduced firstly by two Norwegian scholars Galtung and Ruge (Galtung & Ruge, 1965). News values are certain guidelines to follow in producing a news story, so-called ideological factors in understanding decisions of journalists (Cotter, 2010). The more news values (12 in total) are present in a piece of information, the more likely that you will see the story featured in different mass media across the globe.

In the last years there has been a growing interest to work on the intersection of social and computer sciences (Greening, 2000). Text mining is emerging as a vital tool for social sciences and the trend will most likely increase (Amolfo & Collister, 2015). Due to the abundance of news information and with the advances in text mining, it is now possible to help journalists to process information in every day job and at the same time to prove old media theories and to discover old, often biased patterns in the news across the world.

Some research has been already done on detecting news bias (Ali & Flaounas, 2010), (Flaounas & Turchi, 2010), less attention has been paid to automatic detection of news values (De Nies & D'heer, 2012), (Al-Rawi, 2017). We argue that in order to understand and automatically detect news bias, it is first important to understand the logic of news selection processes (news values) and try to detect news values on a large-scale.

We make a first attempt to automate the detection of initially three news values by applying several text mining techniques from selected publishers and when reporting about Apple Corporation. Apple has a great impact on our lives and as any technology it has become newsworthy almost by default. Our goal is to distinguish if the theory of newsworthiness by Galtung and Ruge (Galtung & Ruge, 1965) is still a valid approach to predict news selection values. If yes, what are the relevant news criteria we find in our experiments and do we see some interesting recurring patterns in the news?

2 Data description

News articles analyzed in this paper were first aggregated by the Newsfeed and then analyzed by Event Registry[1] – a global media monitoring service that collects and processes articles from more than 100.000 news sources globally in more than 10 languages (Leban & Fortuna, 2014).

We extracted news about the Apple Corporation (iPhone 6 and Apple Watch launch) from 16 selected online outlets during the period of 01.09.2014 – 31.10.2014. We considered the occurrences of the company name Apple and the terms "iPhone" and "Apple Watch". The time range corresponds to the announcements of the launch of the two above-mentioned products and the start of sales. We also considered to include two months of coverage and to check the news about Apple before and after the big events in order to control for variation in media interest in the company. The time range of complete, uninterrupted two months is important because we are also interested in the extent to which editors and journalists write about Apple and how the coverage changes in case if a greater event. The sources under our analysis correspond to the most influential daily news websites, easily accessible, widely read in the following three languages: English (EN), German (DE) and Spanish (ES).

Apple represents an important context for the purpose of this study. The more new and novel ideas and products come from Apple, the more space will be allotted to it by the international media.

The Table 1 summarizes the total number of events and the total number of articles reporting on Apple collected and analyzed per publisher during the above-mentioned period including the information on the headquarters of each publisher. All journals listed in the Table 1 publish daily.

Important to note is that the websites were also selected with the purpose to cover different geographical places (Europe and the USA) in order to identify one of the three news values, i.e. proximity – geographical or cultural proximity of the event to the source. The core available piece of information for each article for our experiment included the date, the location of the event and the location of the publishers' headquarters, as well as the size of the events (i.e. the number of articles about them).

Publisher	Total No. Events	Total No. Articles	Headquarters
The Next Web	1064	1670	Amsterdam
Gizmodo	2007	3911	New York
The Guardian	14299	19997	London
BBC	15582	23852	London
USA Today	7692	13629	Tysons Corner
Wall Street J.	7197	18837	New York
Heise	4194	2190	Hannover
Chip online	907	1212	Munich
Stern	4194	10092	Hamburg
Die Zeit	3722	5600	Hamburg
Die Welt	14683	30359	Berlin
Der Spiegel	2261	2759	Hamburg
El Mundo	6707	8705	Madrid
ABC.es	7431	10388	Madrid
El Pais	686	979	Madrid
El Dia	6700	12752	Barcelona

Table 1: Publishers and Totals of Events/Articles on Apple.

3 Socio-cognitive perspective on news values

One of the most critical questions in the media research field is why certain aspects of reality are selected by journalists and eventually registered to become news. The news selection process is a very complex process influenced by economic, political, organizational and social factors. Over the years the most used explanation of the phenomena prevails to be the so-called theory of newsworthiness, which focuses on explaining the logic of the journalists and media organizations and on predicting what will most likely strike attention of the audience and be selected as news. News selection, i.e. selecting novel pieces of reality, is not purely a journalistic problem; it has its roots in psychology of perception and cognitive science. Looking for new information in order to reach an optimal level of stimulation is fundamental to human behavior (Martindale, 1981). Humans constantly search for arousal or stimulation, often driven by pleasure centers of the brain (Martindale, 1981), (Donohew, Sypher, & Higgins, 1988).

According to Van Dijk, the notion of news values is part of the social cognition since news values are shared by journalists and even by the public of the mass media in an indirect way. They "provide cognitive basis for decisions about selection, attention, understanding, representation, recall and the uses of news information." (Van Dijk, 2009). News audience is not only a vital part of the cognitive processes to label their psychological activation, but the arousal itself is the end product of cognitive process that often occurs automatically. Journalists only need to select an important piece of information and code it in a nice fashionable way so that it "facilitates recognition and heightens the impact of these cognitive processes" (Martindale, 1981). For most journalists, news criteria are something very physical, something that is always present in the back of mind and is an integral part of themselves (Schulz, 2007). These criteria are often unconscious in journalistic practice.

News values are considered to be ground rules or "distillation" of what an identified audience is interested in reading or listening (Richardson, 2007). The most influential contribution came from Galtung and Ruge who underlined a list of 12 news factors, which they divided into eight "culture-free" factors and four "culture-dependent" factors. The following Table 2 outlines Galtung and Ruge's theory of news selection

[1] http://eventregistry.org

News Values	**Short Description**
Frequency	***Time span of an event***
Threshold	***The size of an event***
Meaningfulness/Proximity	***Geographical closeness***
Unambiguity	Clarity of the meaning
Consonance	Conventional expectations of the audience
Continuity	Continuous over time
Unexpectedness	Unplanned/unexpected
Composition	Include other pieces of information
Reference to elite nations	Relate to famous nations
Reference to elite people	Relate to famous people
Negativity	Bad news, conflict-oriented
Personalization	Action of individuals

Table 2: News Values by Galtung and Ruge.

and its news values (Galtung & Ruge, 1965). Most of the news values have a common-sense perception and are simple to understand. The more an event satisfies the bellow-mentioned criteria, the more likely it will be reported in the news.

Later on there have been several attempts to revise the theory of news worthiness (MacShane, 1979), (Harrison, 2006), (Harcup & O'Neill, 2017) despite the existence of several new lists of news values, the theory by Galtung and Ruge has not really been challenged and is still taught at many Journalism schools across the globe. Various scholars from the social sciences field have discussed the theory of news values extensively. However, technologies have now added new "shades" of how we want to test the theory on a larger scale and if and of how the news values have changed over time.

4 Mining news values

Galtung and Ruge originally came up with a taxonomy of 12 factors, but due to the space limitations, the goal of this work is to identify automatically the first three news values: *frequency*, *threshold* and *proximity*. Frequency and threshold are both impact criteria, calculated through the number of articles per publisher (frequency) and a number of articles per events (threshold), whereas, the proximity criterion is considered more about the audience identification and geographical distance. The above-mentioned news values are convenient for the analysis by text mining methods and represent a starting point of a complete framework of automatic detection of 12 news values.

4.1 Frequency

Frequency as news value refers to the time-span of an event (Galtung & Ruge, 1965). For example, a single event on a certain day is more likely to be reported rather than a long process (Fowler, 1991). Since Apple has become a new religion of the 21st century, it is newsworthy by default and news about Apple exists in most outlets around the world on daily basis. In this paper, we understand frequency value as the frequency of all articles from the selected publishers mentioning iPhone 6 and Apple Watch (also known as Watch) respectively. We are interested in finding trends or particular patterns among publishers during the selected period of time. The Figure 1 summarizes the time distribution (i.e. frequency value) of the mentions related to iPhone 6.

The frequency measurement experiment indicates that there are two sudden busts in frequency among certain publishers; one peak corresponds to the announcement of the Apple Watch launch on the 9th of September 2014 and the second peak being the announcement of the iPhone 6 release on the 19th of September 2014. Both announcements received a much bigger coverage (especially, among the following publishers: Wall Street Journal, Stern Magazine and die Welt) in respect to the actual start of sales of the products at the end of October. Applying the logic of the media in our context – announcement of the launches and of the start of sales – announcement is news since the audience did not know when Apple would launch the products and announce the official sales date. The surprise is even larger if the organization is Apple, which is perceived to be the trend maker in technology in general.

The frequency distribution of new Apple Watch has a similar to iPhone 6 trend, having, however, less coverage (number of articles) per publisher, per day. The following Figure 2 outlines the frequency of Watch coverage among the selected publishers during 01.09.2014 – 31.10.2014. The two bursts are also visible in the coverage of the Watch due to the fact that journalists are more likely to complement news pieces with additional, background information (Bell & Garrett, 1998), in our case if a journalist is writing about the Apple Watch he is likely to mention Apple and another Apple product.

We also measured frequency between specific tech publishers in comparison to other international publishers. Our first assumption that technology-oriented outlets do publish more news with higher frequency (number of articles) on Apple was not confirmed as also seen from the Table 1 and Figure 2. This partially could be explained by the small size of editorial and journalistic teams working for tech publishers in comparison to big media corporations with journalists all over the world, for example, BBC, Wall Street Journal or USA Today.

4.2 Threshold

The threshold criterion often refers to the impact of an event and its effect on the readers, i.e. a size needed for an event to become news, for example, thousands of people buying a new iPhone 6 and not just one person buying it in a small local store will get more attention of

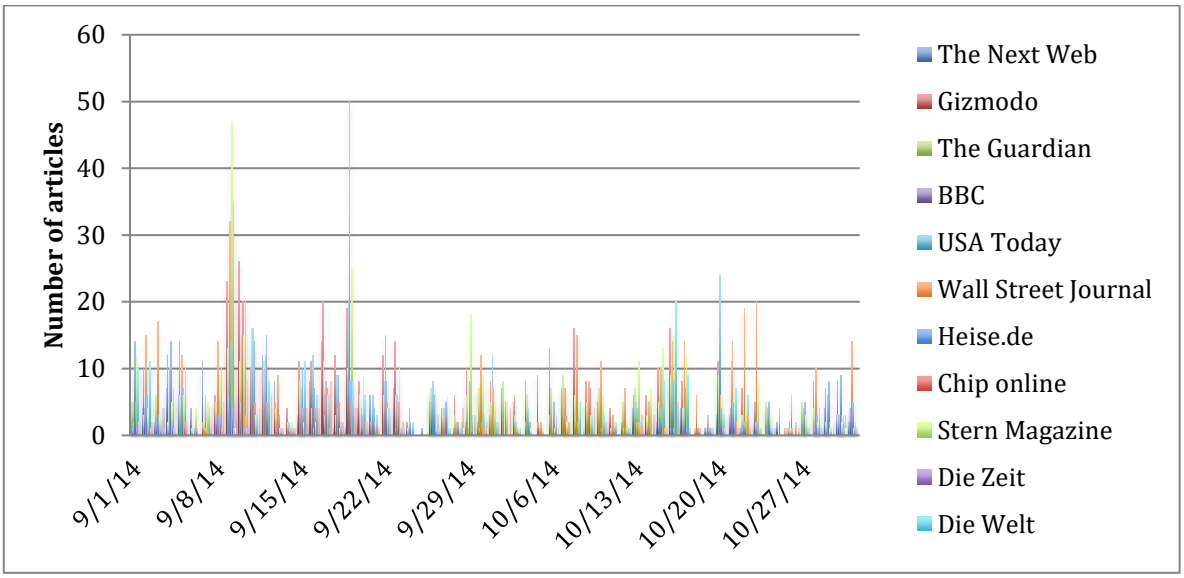


Figure 2: iPhone 6 Frequency distribution.

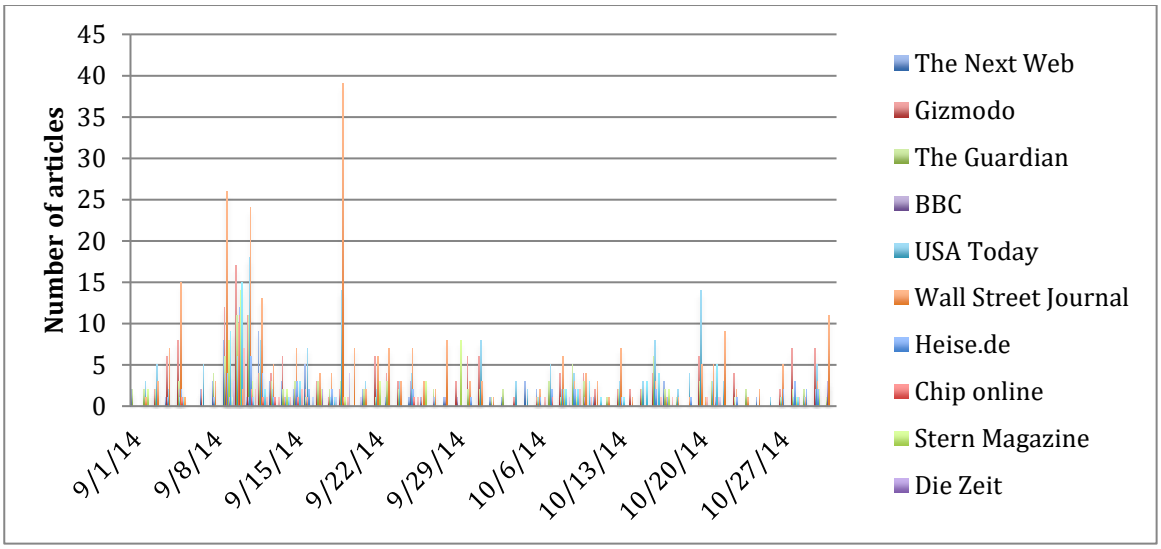


Figure 1: Apple Watch Frequency distribution.

the media. In other words, the bigger, the better, and the cooler a product is, the higher the "amplitude" and the more fuss it will create in the media.

It is indeed difficult to measure something that should have a larger effect on the readers. We understand that events can meet this threshold value either by being large in absolute terms i.e. having a higher frequency or an increase in reporting of a topic. In this experiment, we decided to look at the size of events among the selected publishers without limiting our search to reports about Apple in order to take in more data. The main reason we limit ourselves to Apple related stories in frequency analysis is that we can manually show that remarkable and frequent events like a new product launch draw more media attention.

An event is understood as a group of articles that are clustered to report on the same issues in the world (Leban, Kosmerlj, & Belyaeva, 2014). Our assumption is that a single article might not be very informative, but a group of articles on a certain issue, which is picked up by more publishers can form a part of a bigger story with more impact on the readers and thus match the threshold value. Note that frequency and threshold values are both impact criteria, we see threshold as the size of an event, whereas frequency should also be understood as events unfolding within production cycle of a news media and will be reported on repeatedly.

Therefore, for the threshold analysis we aim at capturing the size of clusters (number of articles of all publishers in event clusters) and assume to witness a greater number of articles that form an event. To note that news articles are first aggregated by the Newsfeed service[2] - a real-time stream of articles from more than 100.000 RSS-enabled websites in several major world languages, then we process the articles by a linguistic and a semantic analysis pipeline that provides semantic

[2] http://newsfeed.ijs.si

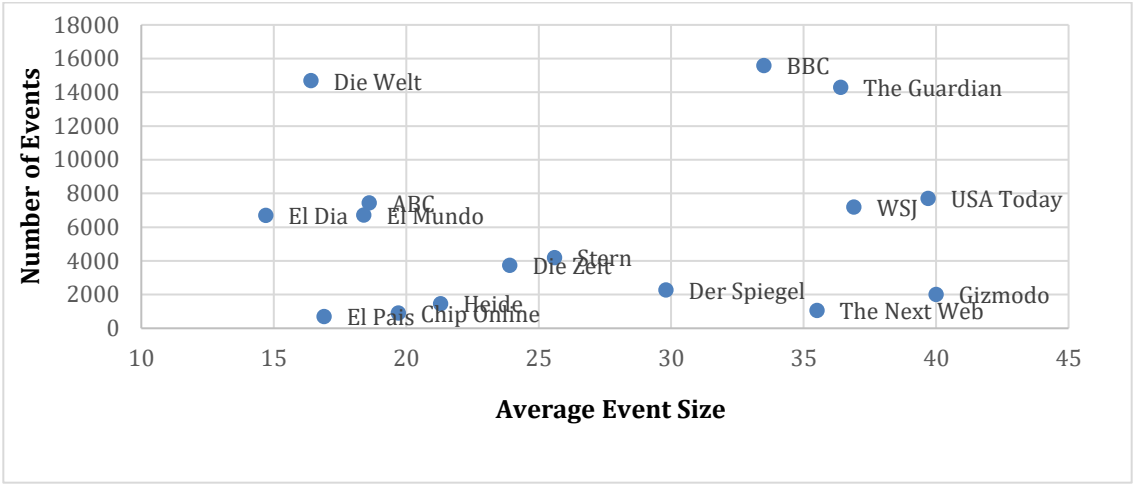


Figure 3: Threshold analysis per publisher

annotations. The semantic annotation tool developed within XLike project comprises three main elements: *named entity recognition* based on corresponding Wikipedia pages, *Wikipedia Miner Wikifier* – detecting similar phrases in any document of the same language as Wikipedia articles and *cross-lingual semantic analysis* that links articles by topics (Carreras & Padro, 2014). The analysis and clustering are then done by the Event Registry. The data in the following scatterplot shows the average event size per publisher during the same period of time and confirms our hypothesis: the higher the threshold (number of articles per event), the greater the impact of a publisher (i.e. The Guardian, BBC), more intense and more frequent the coverage about an event is.

If an article is written by an influential publisher other bigger and smaller publishers will most likely pick the event up and eventually it will form an event (cluster). Interestingly, Spanish and German publishers have a smaller average event size, which could be explained as those publishers are more interested in local events or events within their countries of origin.

4.3 Proximity

The Proximity value (also often referred to as Meaningfulness factor) corresponds to physical i.e. geographical or cultural (in terms of religion or language) closeness of a news story to a listener and thus the media publisher. Proximity helps readers to relate to a story on a more personal, familiar level. It can change over time and is open to subjective interpretations. However, proximity might also mean an emotional (fear, happiness, pride tec.) trajectory in the audience's eyes, regardless of where it takes place (Schults, 2005). An event may happen in a distant place but still be of interest in terms of a certain relevant meaning to the reader: Apple is an American company with headquarters in California, but as the largest technology company its products are praised and sold in every part of the world.

Our assumption when measuring this journalistic value is that the closer the geographical location of the story to the news publisher is, the more frequent and more intense (also higher threshold) the coverage is.

Event location detection is done automatically in the Event Registry in the following way: we first try to identify a dateline in the article (a piece of text at the beginning of every article) that names a location: we assume that this is the location of the event. When the dateline does not appear in the article, we check the Event Registry to use the event's location. A classification algorithm that considers all articles belonging to the same event determines it. In some cases, the Event Registry does not determine the location; we omit such cases in our analysis.

The headquarters of each publisher was manually searched for on the official websites of the selected publishers and Wikipedia[3]. We did not limit our search to the stories reporting about Apple since we assume to see some recurring proximity patterns of the selected publishers in spite of the story topic.

Since our system was not able to automatically identify location of all events, we use only a sub-selection of our data for each publisher for which we compute the distance in kilometers and calculate how many of them report from the same country and same city where the publisher is. The following Table 3 outlines the proximity experiment results: Total number of sub-Selection of Events where Country/City were detected and a Total number of events where publishers reported either on the same country or the same city where a publisher has headquarters.

It has been found that the coverage of most publishers is not local, they do not report on the events close to their headquarters: it can be explained by the fact that the selected publishers are not local publishers and are no longer "national" publishers, but some have become over time "international" outlets whose news is read across the world. Interesting to note that proximity value was not confirmed for, for example, The Next Web – technology oriented website with headquarters in Amsterdam, Netherlands, reported only once on the

[3] https://www.wikipedia.org

Publisher	Country sub-Selection	Same country	City sub-selection	Same city
The Next Web	178	1	174	1
Gizmodo	***371***	***204***	370	15
The Guardian	6563	2261	6510	462
BBC	7105	2909	7039	438
USA Today	***4299***	***2842***	4291	0
WSJ	3091	1194	1074	122
Heise	***586***	***262***	585	5
Chip online	211	71	211	1
Stern	***2704***	***1197***	2701	90
Die Zeit	***2505***	***1103***	2504	95
Die Welt	***9185***	***5340***	9182	1248
Der Spiegel	1592	630	1590	55
El Mundo	***4077***	***2269***	4076	861
ABC	***4493***	***2372***	4491	879
El Pais	337	46	337	7
El Dia	***3789***	***2399***	3785	154

Figure 4: Geographical proximity analysis per publisher.

events from their headquarter city. Whereas, some selected publishers, mainly Spanish and German (in italics in Table 3) dedicate more or less half of their attention to the news from the same country, which confirms relatively strong the proximity value. Not surprisingly, the Guardian, the BBC and the Wall Street Journal do not support journalistic proximity value since their geographical scope is scattered around the world. Publishers and some figures in italic letters in the Table 3 represent newspapers that confirm the proximity value that is they report on the events that happen in the same country of their headquarters in respect to the remaining publishers under selection.

5 Discussions and future work

We made an initial attempt to automate detection of journalistic news values, in particular, *frequency* in the context of Apple news, *threshold* and *proximity* in the context of selected publishers. We believe that using text-mining methods is an essential step of interaction between social and computer sciences approaches. This hybrid approach will not only help journalists in their everyday work, but it will also potentially help to identify various ideological patterns or news bias of various global publishers.

The study tested some news values from the theory of newsworthiness formulated by Galtung and Ruge. The assumption that a product launch is likely to get more coverage and thus meet the frequency value was confirmed for most outlets. No assumptions were confirmed on technology-oriented publishers writing more frequently and more intensely on Apple. No influences were found for the proximity news factor in case of international publishers; however, the proximity factor was confirmed for most Spanish and German publishers. These findings suggest continuing exploring the topic in depth.

Future work will include developing our framework further, which will automate the process of assessing newsworthiness of all 12 news values applied to different languages, as well as to different domains like conflicts, natural disasters, political crises etc. By detecting news values through text mining we also aim at confirming still existing ideological patterns i.e. news slant or bias of different publishers. Research designed more specifically and comprising automation of all news values could provide more answers to the problems of outdated and orthodox new values that keep on contributing to the news bias. To our knowledge, there are no automated systems to compare our approach with, thus, in the future we also plan on conducting several evaluations including manual to verify our results.

6 Acknowledgments

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under XLike (ICT-STREP-288342) and XLime (FP&-ICT-611346).

7 References

[1] Ali, O., & Flaounas, I. (2010). Automating News Content Analysis: An Application to Gender bias and Readibility. *JMLR: Workshop and conference Proceedings.*

[2] Al-Rawi, A. (2017). News values on social media: News organizations' Facebook use. *Journalism , 18* (7), 871-889. https://doi.org/10.1177/1464884916636142

[3] Amolfo, L., & Collister, S. (2015). Text mining and Social Media: when Quantative Meets Qualitative, and software meets human. In P. Halfpenny, & R. Procter, *Innovations in Digital Research Methods.* London: Sage.

[4] Bell, A., & Garrett, P. (1998). *Approaches to Media discourse.* Oxford: Blackwell Publishers.

[5] Carreras, X., & Padro, L. (2014). XLike project language analysis services. *Proceedings of EACL'14*, (pp. 9-12). https://doi.org/10.3115/v1/E14-2003

[6] Cotter, C. (2010). *News Talk. Investigatingthe Language of journalism.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511811975

[7] De Nies, T., & D'heer, E. (2012). Bringing Newsworthiness into the 21st Century. *Web of Linked entities Workshop, ISWC*, (pp. 106-117). Boston.

[8] Donohew, L., Sypher, H., & Higgins, T. (1988). *Communication, Social Cognition, and Affect.* London: Psycology Press.

[9] Flaounas, I., & Turchi, M. (2010). The Structure of the EU Mediasphere. *PLoS ONE , 5* (12). https://doi.org/10.1371/journal.pone.0014243

[10] Fowler, R. (1991). *Language in the News. discourse and Ideology in the Press.* London: Routledge.
[11] Galtung, J., & Ruge, M. (1965). Structuring and Selecting News. *Journal of International Piece Studies , 2* (1), 64-91.
[12] Greening, T. (2000). *Computer Science Education in the 21st Century.* New York: Springer. https://doi.org/10.1007/978-1-4612-1298-0
[13] Harcup, T., & O'Neill, D. (2017). What is news? News values revisited (again). *Journalism studies , 18* (12), 1470-1488. https://doi.org/10.1080/1461670X.2016.1150193
[14] Harrison, J. (2006). *News.* New York: Routledge.
[15] Leban, G., & Fortuna, B. (2014). Event Registry - Learning About World Events from News. *WWW*, (pp. 107-111).
[16] Leban, G., Kosmerlj, A., & Belyaeva, E. (2014). News reporting bias detection prototype. XLike Deliverable D5.3.1.
[17] MacShane, D. (1979). *Using the Media: how to deal with the press, television and radio.* London: Pluto.
[18] Martindale, C. (1981). *Cognition and consciousness.* Homewood, IL: Dorsey.
[19] Richardson, J. (2007). *Analysing Newspapers. An Approach from Critical discourse Analysis.* New York: Palgrave MacMillan. https://doi.org/10.1007/978-0-230-20968-8
[20] Schults, B. (2005). *Broadcast News Producing.* London: Sage Pubblications.
[21] Schulz, I. (2007). The Journalistic Gut Feeling. In I. Schulz, *Journalism Practice* (pp. 190-207). London: Routledge.
[22] Van Dijk, T. (2009). *News as Discourse.* New York: Routledge.

https://doi.org/10.31449/inf.v42i4.2066 Informatica **42** (2018) 535–544 **535**

KAIRÓS: Intelligent System for Scenarios Recommendation at the Beginning of Software Process Improvement

Ana Marys Garcia Rodríguez, Yadian Guillermo Pérez Betancourt , Juan Pedro Febles Rodríguez,
Yaimí Trujillo Casañola and Alejandro Perdomo Vergara
Universidad de las Ciencias Informáticas, La Habana, Cuba
E-mail1: agarcia@uci.cu, ygbetancourt@uci.cu, febles@uci.cu, yaimi@uci.cu, apvergara@uci.cu

Keywords: artificial intelligence, Critical Success Factors, Good Practices, intelligent system, software process improvement

Received: December 1, 2017

Software Process Improvement provides benefits to organizations. However, improvement efforts are not guided by the combined use of Good Practices and Critical Factors that influence success. Resources are dedicated without a prior analysis that guides the actions intentionally. The objective of this research is to support decision-making in Software Process Improvement. To achieve this, an intelligent system is conceived, which based on association rules, identifies dependencies between Good Practices and Critical Success Factors. In addition, this system implements a Genetic Algorithm to optimize improvement scenarios and an evolutionary Artificial Neural Network to predict success in Software Process Improvement. The methods used to validate the results corroborated the contribution and usefulness of the proposal.

Povzetek: Predstavljen je inteligentni sistem KAIROS, ki na osnovi metod umetne inteligence zasnuje scenarij izdelave sistema pred začetkom softverskega procesa.

1 Introduction

The analysis around the Critical Success Factors that influence the Software Process Improvement (SPI), allows to infer that its use in function of the organizational contexts, contributes to the success of the improvement project [1; 2]. In spite of the advances in the treatment of Critical Success Factors [3; 4; 5], insufficiencies associated with the reuse of knowledge persist. This hinders to obtaining evaluations that are close to reality and makes it difficult to provide scenarios that guide organizations in the improvement. Also, the influence performed by the combination between Critical Success Factors and Good Practices in the SPI is not analyzed [6]. The weights assigned to the Critical Success Factors are not adjustable and their relevance changes according to the context.

An analysis to guides organizations at the beginning of the SPI is appropriate. It is cumbersome to process information when a large number of elements affects the decision-making of an organization. An effective alternative is the application of artificial intelligence techniques that transform SPI experiences into useful knowledge to guide insertion in an improvement project.

The research problem is: how to recommend improvement scenarios from the use of Critical Success Factors and Good Practices, to support decision-making at the beginning of the SPI? The objective of this paper is to develop an intelligent system for scenarios recommendation, which combines the use of Critical Success Factors and Good Practices to decision-making support at the beginning of the SPI. To the development of this research were used some scientific methods:

- Historical-logical and dialectical to the critical analysis of researches associated with the use of Critical Success Factors and Good Practices in the SPI.
- Induction-deduction to the identification of the problem, as well as its solution variants.
- Hypothetical-deductive to the proposal of this research line.
- Analytical-synthetic to the decomposition of the problem in elements that allow its analysis.
- Bibliographic analysis for literature review.
- Survey to know the degree of customer satisfaction with the system developed.
- Experimental to evaluate the utility of the obtained results.
- Consult experts to the research validation.
- Focal group to the conceptualization of Good Practices and recommendations.
- Iadov technique to evaluate the solution satisfaction.
- Statistical methods to the analysis of applied surveys.

Scientific contributions of the research:

- An informatics system (KAIRÓS) which combines artificial intelligence techniques, to support decision-making in SPI. In order to achieve this, the system optimizes improvement scenarios and predicts their success in the SPI.
- A genetic algorithm (GA) to optimize improvement scenarios from the redefinition of selection and

crossover operators and from the definition of a new mutation operator.
- An evolutionary ANN that uses genetic algorithms to topology design, and integrates the backpropagation algorithm and genetic algorithms for the net learning. Also, it uses the Principals Components Analysis technique to handle the changeful number of neurons in the input layer.

2 Theoretical bases

To solve the introduced problem, it was realized a research study to clarify the approach of the Good Practices use in SPI. In addition, it was analyzed the artificial intelligence techniques that facilitate the reutilization of Good Practices and Critical Success Factors experiences.

2.1 A survey of Good Practices and Critical Success Factors association in SPI

The literature analysis reveals the need to apply Good Practices for a successful SPI projects execution. However, only four papers consider the influence of Good Practices on the behavior of Critical Success Factors [5; 7; 8; 9]. These papers contribute important elements, associated with the Good Practices incorporation to have a positive influence in Critical Success Factors behavior. However, there are some insufficiencies that affect the use of this relationship:

- The dependencies between Good Practices and Critical Success Factors are considered without detailing which are the relationships specifically.
- The experiences reuse is assumed, but based on the Critical Success Factors and without considering the influence of Good Practices or their combined use.
- Trujillo [5] defines the Critical Success Factors and their measurements, establishes the weighting coefficient of Critical Success Factors, but does not assess its dynamic treatment.
- Improvement scenarios are not offered to support decision-making of organizations in the SPI.

In this sense, it is necessary to extend the treatment of Good Practices and Critical Success Factors, and to consider the influence performed by the combination of Good Practices on Critical Success Factors. Also, the dynamism of Critical Success Factors relevance, must be taken into account.

2.2 Artificial intelligence applied to SPI

For the forecast and recommendation of scenarios before the investment in SPI, it was considered the application of artificial intelligence techniques. With the aim of supporting decision-making in SPI from two perspectives: guide the efforts of organizations towards better scenarios in the SPI and forecast the result prior to investing in the SPI, the experiences reuse, associated with Critical Success Factors and Good Practices is adopted. In this sense, three needs were identified:

- Recommendation of scenarios to improve an organization initial state, prior to invest in SPI. It is considered as an optimization problem and is solved with the implementation of a GA [10; 11].
- Identification of relationships between Good Practices and Critical Success Factors measurements. It is considered as an association problem, taking into consideration the dependencies identification between dependent variables (Critical Success Factors measurements) and independent variables (Good Practices), whether metric or non-metric. It is solved with the use of association rules [12; 13].
- Forecast of success or failure of scenarios in SPI. It is considered as a classification problem, where it is necessary to identify the tendency to success or failure of an organization in the SPI. It is solved by the implementation of an evolutionary Artificial Neural Network (ANN) [14; 15; 16].

2.2.1 Considerations for KAIRÓS optimization

The GA of KAIRÓS for scenarios optimization takes the functioning principles from the operators descripted by the literature and provides new operators, capable to solve the particularities of the problematic. It was considered to not provide solutions unattainable since the organization capabilities to the implementation of selection operator. To the crossover and mutation operators, it was considered to not change the values of the Critical Success Factors measurements, which are not affected by the good practices applicable by the organization.

In this research, the chromosomes represent the initial state and the improvement scenarios. Genes are measurements of the Critical Success Factors, where $M = \{m \in R,\ 0 \leq m \leq 1:$ *m is a measurement of Critical Success Factors*$\}$.

Selection operator redesign:

Several operators were analyzed where some was discarded and others partially satisfy the solution:
- *Selection by roulette* is ruled out, due to the randomness factor that it uses.
- *Selection by tournament* is ruled out, due to its high computational cost.
- *Hierarchical selection* and *selection by rank* partially satisfy the needs of the problem, because they do not necessarily obtain individuals close to the initial state.
- *Selection by rank* partially satisfies, because it does not consider fitness.

For those reasons, selection operator was redesigned from the hierarchical ordering of the population chromosomes, taking as criterion of order the fitness of each individual. Then a range of individuals is selected, which will be closest to the initial state.

Crossover operator redesign:

A set of operators were analyzed where were discarded:

- *Crossover by a point* and *Crossover by N points*, because the measurements influenced by Good Practices do not have a specific order or position within the chromosome.
- *Arithmetic crossover*, because it does not allow to establish which measurements will be modified, these are determined by the randomness.
- *Uniform crossover*, because it affects all the genes in the chromosome.

Finally, the authors determined that *uniform crossover* with binary mask is the operator that satisfies the most of the problem needs. However, the randomness when generating the binary mask does not make it feasible to the solution. Therefore, it was redesigned in such a way that the binary mask is intentionally generated from the dependencies between Good Practices to be applied and measurements of Critical Success Factors. The dependencies are obtained by applying association rules between these variables.

Mutation operator design:

A set of operators were analyzed where were discarded:

- *Binary mutation*, because it does not correspond to the coding of the chromosomes in this research.
- *Mutation to the edge* and *Uniform mutation*, exchange values of the attributes, which can alter genes not influenced by the Good Practices.

Therefore, a new operator was designed that uses the binary mask of the redesigned crossover operator and randomly identifies a position of the mask. If the position value is 1, proceed to mutate by 1% in this position.

2.2.2 Considerations for KAIRÓS dependences identification

The dynamic identification from the accumulated experience of the association relationships between Good Practices and Critical Success Factors measurements, is relevant. The objective is to know what measurements to enhance in the optimization process, based on the Good Practices that the organization can apply. To determine the dependencies association rules are applied, due to their potential to identify relationships between variables in combination, as well as the treatment of both metric and non-metric variables.

To generate the rules combinations the algorithm *Apriori* is used [17; 18; 19], with the aim of reducing the number of candidates through the technique of reduction by pruning. In this sense, all variants of rules whose elements are not frequent are discarded, because their combinations will not be.

2.2.3 Consideration for KAIRÓS classification

To predict the success of initial state and improvement scenarios, it must be taken into account that:

- The weights relevance associated to the Critical Success Factors, must have a dynamic treatment.
- The Critical Success Factors and their measurements can change over time.

Based on the above, an ANN is implemented because it favors learning by readjusting the weights associated with network connections. Considering that a classification problem is addressed, is appropriate to use supervised learning, specifically the multilayer perceptron. This architecture is usually trained using the backpropagation algorithm. However, an architecture that provides a solution to one problem can't be used to solve another [16]. Under the conditions of the problem, the construction of a self-adapting intelligent system based on ANN is required. This research considers the use of an evolutionary ANN, which allows adapting to the input patterns. Genetic algorithms are applied to the design and learning of the evolutionary network.

3 Intelligent system KAIRÓS

In this article, the Critical Success Factors and its measurements defined by Trujillo [5] are assumed for the processing of KAIRÓS. In addition, with the aim of defining Good Practices to improve the behavior of the Critical Success Factors, the bibliographic review, Delphi and focus group methods were applied.

For the identification of Good Practices, a bibliographic review of 77 articles and documented experiences was made, of which 15 allude to the use of Good Practices to diminish the influence of the Critical Success Factors in the SPI [4; 8; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 31; 32]. Then it was refined with the help of experts, managers and members of software development organizations, using the Delphi method in a first round. The results were submitted to the exploratory focal group, where the proposal was enriched with the recommendations for the execution of the Good Practices. Finally, a second round of Delphi method was applied with the refined information. As a result, 49 Good Practices and 127 recommendations that guide its application were defined [33].

KAIRÓS [34; 35] has the purpose of processing and automating the information associated with Critical Success Factors and Good Practices to support decision-making in SPI. The following describes its components.

3.1 GA for scenarios optimization in SPI

For the proposal of improvement scenarios, a GA is conceived [34]. The restriction of the optimization problem is associated to achieve a distance between the initial state and the improvement scenario in an affordable range, attending to the Good Practices that the organization can apply.

GA description:

Step 1. Generate initial population: the size of the initial population sample is calculated with the probabilistic method of calculating the population sample size, knowing the population size [36]. The individuals of this population are taken randomly from the knowledge base.

Step 2. Select scenarios: the scenarios of the initial population are assessed with the evaluation function (equation 1). The scenarios are ordered hierarchically, according to their fitness (value of evaluation function).

Then, a population sample corresponding to a new calculation of sample size is selected.

$$f_{eval} = \sum_{i=1}^{n} f\big(Sn(i), Sm(i)\big)\, Dg\big(Ov(i), Sm(i)\big)/n \quad (1)$$

Where:
f_{eval} is the evaluation function.
n is the amount of measurements (genes) of the scenario (chromosome).
i is the position of the measurement (gene) subject to analysis.
Sn(i) is the value of the measurement (gene) in the position *i* of initial state (chromosome).
Sm(i) is the value of the measurement (gene) in the position *i* of the scenario (chromosome) to be analyzed.
f(Sn(i), Sm(i)) is the aptitude function for the measurement (gene) *i* and is determined by the equations 2 and 3.
Ov(i) is the optimum value achievable for *Sm(i)* (obtained from the analysis of association between Good Practices and Critical Success Factors measurements).
Dg(Ov(i), Sm(i)) is the degree of improvement for the measurement (gene) *i*.

The aptitude function *f(Sn(i), Sm(i))* for each measurement, it is calculated depending on whether the measurement under analysis *i*, belongs or not to the set of measurements that are affected by the association between Good Practices and Critical Success Factors measurements.
Being:
M= {m∈ R, 0≤m≤1: m is a measurement of Critical Success Factors}
MGP= {mgp∈ R, 0≤mgp≤1: mgb is a measurement affected by Good Practices}
MGP⊂M
i is the position of the measurement (gene) subject to analysis.
m(i) is the measurement in position *i*, represented to evaluate whether the measurement in a given position is affected by the association with Good Practices, *m(i)∈M*.
If m(i)∈MGP

$$f\big(Sn(i), Sm(i)\big) = \begin{cases} 1 & si \quad Sn(i) < Sm(i) \\ 0{,}5 & si \quad Sn(i) = Sm(i) \\ 0 & si \quad Sn(i) > Sm(i) \end{cases} \quad (2)$$

Where:
Sn(i) is the value of the measurement (gene) in the position *i* of initial state (chromosome).
Sm(i) is the value of the measurement (gene) in the position *i* of the scenario (chromosome) to be analyzed.
If m(i)∉MGP

$$f\big(Sn(i), Sm(i)\big) = \begin{cases} 1 & si \quad \delta\big(Sn(i), Sm(i)\big) = 0 \\ 0 & si \quad \delta\big(Sn(i), Sm(i)\big) = 1 \end{cases} \quad (3)$$

Where:
Sn(i) is the value of the measurement (gene) in the position *i* of initial state (chromosome).
Sm(i) is the value of the measurement (gene) in the position *i* of the scenario (chromosome) to be analyzed.
δ(Sn(i),Sm(i)) is the distance between the value of *Sn(i)* y *Sm(i)* and it is calculated by:

$$\delta\big(Sn(i), Sm(i)\big) = \begin{cases} 1 & si \quad Sn(i) - Sm(i) \neq 0 \\ 0 & si \quad Sn(i) - Sm(i) = 0 \end{cases} \quad (4)$$

The improvement degree is calculated by subtracting one from the normalization of the difference between the optimum value achievable by the measurement and the value of said average.

$$Dg(Ov(i), Sm(i)) = 1 - |Ov(i) - Sm(i)| \quad (5)$$

Where:
Sm(i) is the value of the measurement (gene) in the position *i* of the scenario (chromosome) to be analyzed.
Ov(i) is the optimum value achievable for *Sm(i)* (obtained from the analysis of association between Good Practices and measurements of the Critical Success Factors).

Step 3. Check if the solution is among the selected scenarios: if the last scenario of the population sample fitness exceeds the 0,75 threshold, returns the first and last chromosome from the population, else step 4 is executed.

Step 4. Cross scenarios: the binary mask is generated by assigning 1 to the positions of measurements favored by the Good Practices and 0 to the rest of the positions. For each gene of the scenario if its position corresponds with value 1 in the binary mask, the gene of the scenario being analyzed is added to the new scenario. If its position corresponds with value 0 in the mask, the gene of the initial state is added to the new scenario. Finally, the new scenario is added to the set of crossed scenarios.

Step 5. Mutate scenarios: the same binary mask used in the crossover is applied for mutation. A random number greater than 0 and less than the number of Critical Success Factors measurements that compose the initial state, is generated. If the random number coincides with the position of some measurement affected by the association, the value of this measurement will increase by 1%, otherwise the scenario in the mutation process will be ignored. Finally, the new scenario is added to the set of mutated scenarios.

Step 6. Increase population: the crossed and mutated scenarios are added to the population.

Step 7. Execute step 2.

3.1.1 Association rules to identify dependencies between Good Practices and Critical Success Factors measurements

Several practices can influence more than one measurement of the Critical Success Factors. It is considered relevant to identify dynamically the association relationships between Good Practices and Critical Success Factors measurements from the accumulated experience. For this, association rules are applied in the present research.
GP= {gp: gp is an SPI action that decreases the negative influence of Critical Success Factors}
M= {m∈ R, 0≤m≤1: m is a measurement of Critical Success Factors}

The association rules are represented as: X → Y, Where:

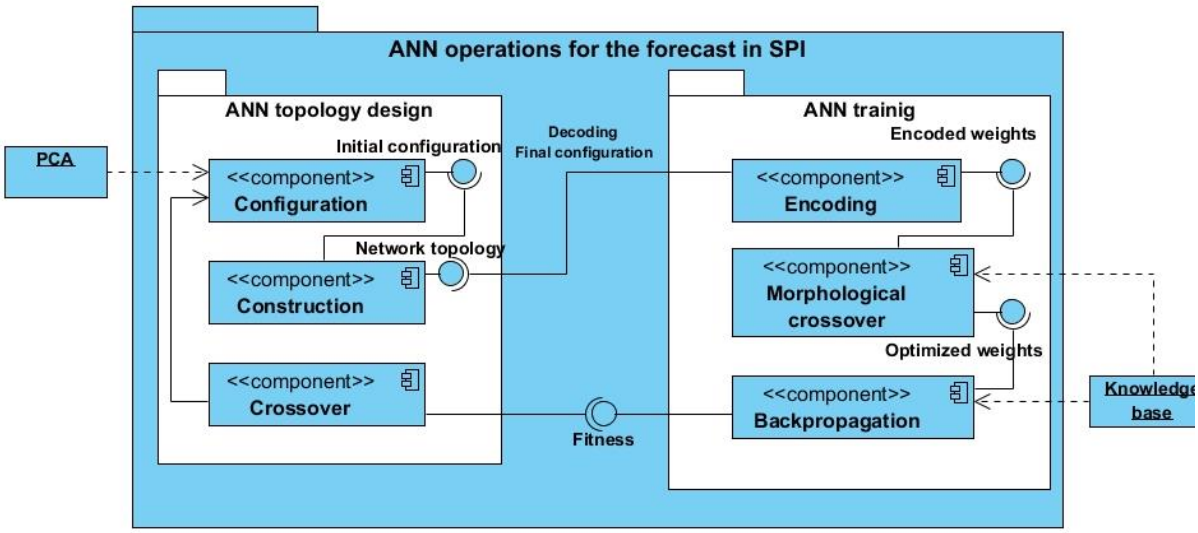


Figure 1: ANN operations for the forecast in SPI

being *X* and *Y* sets of elements, where: $X \subset GP$, $Y \subset M$ y $X \cap Y = \emptyset$.

Step 1. Transformation of knowledge in transactions: a search is made in the knowledge base about the measures of Critical Success Factors, which evolved positively from the initial state to the improvement scenario reached, as well as the Good Practices applied by the organization for the change. The recovered information is stored as transactions in a temporary list for further processing. T is a set of transactions where $T=GP \cup M$.

$T= \{GP; M: gp_1, gp_2, \ldots, gp_n, m_1, m_2, \ldots, m_m\}$, example: $\{GP_1, GP_2, M_1, M_2, M_3\}$.

Step 2. Calculation of support indexes: being the rule $X \rightarrow Y$, where $X \subset GP$ and $Y \subset M$, the support of the rule is calculated as:

$$Sup\ (X \rightarrow Y) = \frac{Nt(XY)}{T_t} \quad (6)$$

Where:
Sup (X→Y) is the support of the rule $X \rightarrow Y$.
$N_t(XY)$ represents the number of transactions that contain the elements of *X* and *Y*.
T_t represents the total of transactions of *T*.

Step 3. Identification of frequent elements sets: elements sets with equal or greater support than the established threshold (0,75) are identified.

Step 4. Generation of candidate rules: combinations of candidate rules are generated. The *Apriori* algorithm is applied [17; 18; 19; 37; 38] to reduce the number of candidates, through reducing by pruning. All the rules whose elements are not frequent are discarded, because their combinations will not be.

Step 5. Calculation of confidence indexes: being the rule $X \rightarrow Y$, where $X \subset GP$ and $Y \subset M$, the confidence index is calculated from the equation 7.

$$Conf(X \rightarrow Y) = \frac{N_t(XY)}{N_t(X)} \quad (7)$$

Conf (X→Y) is the confidence of the rule $X \rightarrow Y$.
$N_t(XY)$ represents the number of transactions that contain the elements of *X* and *Y*.
$N_t(X)$ represents the number of transactions that contain the elements of *X*.

Step 6. Obtaining association rules: rules with a confidence index lower than the defined threshold (0,75) are discarded and then, the association rules are generated.

Step 7. Application of association rules: the information of the association rules generated, is provided to the GA. This information is about which Critical Success Factors measurements are favored by which Good Practices.

3.2 ANN for the forecast in SPI

Considering the characteristics of the classification problem, the implementation of an evolutionary ANN based on the execution of GA for its design and learning is required. The ANN operations are represented in figure 1 [35]. The Critical Success Factors measurements are input patterns and they can be dynamic, the output layer responds to the success or failure in SPI.

The design of the network topology is done in the *Configuration* component, where the initial configuration is created to build the network topology in the *Construction* component. Later the decoding of the network and the final configuration are performed. Subsequently, the network training begins.

The *Codification* component encodes the weights of the ANN, which are used in the *Morphological Crossover* component to perform the evolution of the weights for the network topology obtained. These values are used as initial weights in the *Backpropagation* component. Once the data of the knowledge base is obtained, the training of the ANN is realized. The fitness (mean square error of the network) is calculated as a value that allows to determine how effective is the network. Then, the *Crossover* component is executed to obtain a new ANN topology.

The entire process is repeated until an ANN architecture of lower fitness is obtained.
The design and training of the ANN are described in the following steps:

Step 1. Principal Component Analysis (PCA) is applied to reduce the number of input measurements to the ANN. PCA is the problem of fitting a low-dimensional affine subspace to a set of data points in a high-dimensional space. It has become one of the most useful tools for data modeling, compression, and visualization [39]. A binary matrix *M* of dimension Dim_x x Dim_y is created, which is initialized by 0. The size of the matrix depends on the number of input and output neurons. Dim_x (rows) is the number of input neurons *n* plus the number of output neurons *m* (in this case 1), and Dim_y (columns) corresponds to the maximum number of hidden neurons to consider.

In the matrix *M*, which represents the topology of the ANN with one hidden layer, the meaning for the position *(i, j)* is defined as follows. Being *n* the number of input neurons, if $i \leq n$ then *(i, j)* represents a connection between the input neuron *i* and the *j-th* hidden neuron; if $i > n$, *(i, j)* represents a connection between the *j-th* hidden neuron and the *(i - n)-th* output neuron. The individuals or chromosomes (seeds of growth and pruning) of the population are generated randomly and in random positions.

Step 2. The growth seeds are located in the matrix according to the values of their genes. The initial configuration of the network is performed, replicating the growth seeds sequentially over its quadratic neighborhood. During replication if a new seed has to be placed in a position previously occupied by another seed, the first one will be replaced.

Step 3. Each chromosome is decoded and converted into network locations, where each seed is represented by two genes (X, Y), corresponding to the coordinates in the matrix.

Step 4. The algorithms of growth and pruning of seeds are applied.

Being $a_{i,}$ the value in the position *(i, j)* of the matrix and *S* the set of growth seeds, $S = \{s_k: s_1, s_2, \dots, s_n\}$, a seed s_k is copied, which grows when a position is inactive ($a_{i,j}$=0) and there are at least three seeds that grow identical in their quadratic neighborhood.

The pruning configuration is performed. The pruning seeds are placed in the positions where $a_{i,}$ =0. The pruning rule is designed to eliminate the seeds that grow in the network.

Being $a_{i,j}$ the value in the position *(i, j)* of the matrix and *D* the set of pruning seeds, $D = \{d_r: d_1, d_2, \dots, d_n\}$, a s_k growth seed is extracted, when two contiguous neighboring positions contain identical growth seeds and another neighboring position contains a pruning seed d_r. If two pruning seeds are present in the vicinity, the rule is not activated.

Step 5. The matrix is decoded and the compliance of the necessary restrictions is verified, to obtain the model of the ANN architecture.

- Every seed of growth takes value 1 and pruning seeds acquire value 0. Every 1 in the matrix is interpreted as a connection and 0 as absence of connection.
- Columns with values 0 in the matrix are eliminated. If the elements of the column of order *k* are 0, there are no connections from the inputs to the hidden neurons *k-th* and there are no connections from the *k-th* hidden neuron to the outputs.
- The columns where the value $a_{i,}$ =0, if $i>n$ (where *n* is the number of input neurons), are eliminated. If a neuron in the hidden layer has no connection with the output layer, it is eliminated, as it will have no influence on the outputs.
- The rows with values 0 in the matrix are eliminated. When there is a neuron of the input layer without any connection with the hidden layer, it is eliminated, because it will not influence the outputs.

Step 6. The ANN weights are initialized for the defined architecture, in order to obtain a fast convergence in a multilayer perceptron. The weights are encoded by real coding, which allows to explore the domain of the evaluation function (medium square error) with great precision.

Step 7. The ANN is trained through the evolution of connection weights (morphological crossover) to find the best weights configuration.

The selection is made by tournament. They are chosen at random, as many individuals (weights) of the population as has been prefixed in the size of the tournament (given by the number of input neurons). The best individual of the tournament group is selected and the process is repeated until the desired number of individuals to be selected is obtained. Individuals with the best initial weights are considered, to be used by the backpropagation. Subsequently, the morphological crossover is performed, which reinterprets the morphological gradient operation, to obtain a measure of the genetic diversity.

The morphological crossover operates with populations of λ individuals constituted by chains of real numbers with length *l*. Starting from an odd *n* number of progenitor chains ($n \leq \lambda$), obtained without repetition of the current population, a set of intervals called crossing intervals (C_i), is obtained. The descendant chains of the operator are generated from the crossing intervals. The following actions are carried out for the morphological crossover:

- Calculation of the measure of genetic diversity, gene to gene from the *n* individuals taken as parents. Being *G* the progenitor matrix with dimension *(n x l)*, for the *l* columns of *G*, the one-dimensional vector f_i is defined. f_i contains the *n* values of the *n* progenitors for the gene *i*.

$$G = \begin{bmatrix} a_{10} & a_{11} & \cdots & a_{1l-1} \\ a_{20} & a_{21} & \cdots & a_{2l-1} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n0} & a_{n1} & \cdots & a_{nl-1} \end{bmatrix}$$

$$f_i = (a_{1,i}, a_{2,i}, \ldots, a_{n,1}) \quad with \quad i = 0, \ldots, l-1$$

It is defined as a measure of genetic diversity of gene *i* in the population, the value $g_i \in [0,1]$ calculated as:

$$g_i = g(E(n/2)+1) = (f_i \oplus b)(E(n/2)+1) - (f_i \ominus b)(E(n/2)+1) \quad (8)$$

Where:
g_i is the measure of genetic diversity.
E(n/2) +1 is the component located in the middle position of the vector f_i.
f_i is the one-dimensional vector.
n is the number of progenitors for the gene *i*.
$fi \oplus b$ is the dilation of vector f_i on the point *E(n/2) +1*, with the structuring element *b*. The result is the maximum value of the components of the vector, because the structuring element *b* iterates through f_i from the component E (n/2) + 1−E (n/2) = 1, to E (n/2) + 1 + E (n/2) = 2 E (n/2) + 1 = n (where *n* is odd).
$fi \ominus b$ is the erosion of vector f_i on the point *E(n/2) +1*, with the structuring element *b*. It is obtained in the same way as dilation, but calculating the minimum value of the components of vector f_i.

- The crossing intervals are calculated, determining the lower and upper bounds of the crossing interval C_i denoted by $C = \{C_0, \ldots, C_{l-1}\}$. The maximum gene is calculated from the equation 9 and the minimum gene by the equation 10. The crossing intervals $C_i = [g_i min, g_i max]$.

$$g_i max = max(f_i) - \phi(g_i) \quad (9)$$

Where:
$g_i max$ is the maximum gene of the crossing interval C_i.
$max(f_i)$ is the dilation of the vector f_i at the midpoint of C_i.
$\phi(g_i)$ is the value of the exploration / exploitation function at the point g_i.

$$g_i min = min(f_i) + \phi(g_i) \quad (10)$$

Where:
$g_i min$ is the minimum gene of the crossing interval C_i.
$min(f_i)$ is the erosion of the vector f_i at the midpoint of C_i.
$\phi(g_i)$ is the value of the exploration / exploitation function at the point g_i.

- Obtaining descendants. It is the final result of the morphological crossover operator. The descendants are determined by:

$o = (o_0, \ldots, o_{l-1})$ and $o' = (o'_0, \ldots, o'_{l-1})$.
o_i is a random value of the crossing interval C_i
o'_i is obtained by the equation 11:

$$o'_i = \big(min(f_i) + max(f_i)\big) - o_i \quad (11)$$

Where:
$i = 0, 1, \ldots, l-1$.
$min(f_i)$ is the erosion of the vector f_i at the midpoint of C_i.
$max(f_i)$ is the dilation of the vector f_i at the midpoint of C_i.

- Then, the worst individuals of the starting population are replaced with the new descendants, taking as an evaluation function the mean square error.
- Subsequently, the selection by tournament is made again, obtaining a new progenitor matrix. This procedure is carried out in several iterations until obtaining the values of the connection weights, which minimize the mean square error for the configuration of the network in question.

The application of GA for the evolution of weights is not very efficient in local searches, but it is effective in global search. Therefore, training can be improved with the incorporation of the local search method, backpropagation. It is very appropriate to perform a combination where the GA searches for a suitable region in the search space and then the backpropagation refines the solution found, obtaining a result closer to the optimum in said region.

Step 8. The training of the RNA is refined using the weights optimized for the architecture obtained, through backpropagation. The fitness defined by the mean square error is obtained to determine the network efficiency. This step is carried out in several iterations until obtaining the values of refined connection weights, which minimize the mean square error.

Step 9. The ANN resulting from the previous step is encoded.

Step 10. From the set of chromosomes used in the *Configuration* component, the genes that compose the chromosomes are crossed and new populations of topologies are obtained, which will be used in next iterations in *Configuration* component.

Step 11. The steps from 3 to 10, are executed through different iterations, until obtaining in the intermediate step between *Backpropagation* and *Crossover*, an ANN with fitness lower than the established threshold (0,05). This ANN will be used for the forecast of result in the SPI of an initial state or improvement scenario.

In this way, KAIRÓS automates the processing in combination of the Critical Success Factors and Good Practices, to support the decision-making in the SPI. It implements artificial intelligence techniques for the scenarios optimization, the proposal of recommendations, the forecast of the state of organizations to face an SPI project and the generation of association rules between Good Practices and Critical Success Factors measurements.

4 Solution validation

To assess the effect of implementation on decision-making support, a quasi-experiment of multiple chronological series was developed with two pre-tests, two post-tests and a control group in 12 software development centers of the University of Informatics Sciences, with degrees of manipulation (with and without stimulus).

- Pre-test 1: the initial diagnosis was applied and it was identified with KAIRÓS, that the forecast of the initial state was failure for four centers in both groups. The forecast of the minimum improvement scenario was successful in four centers for both groups. No significant differences were identified.
- Pre-test 2: five days after the application of the diagnosis, the improvement plans of the centers were analyzed. The objective of this test was to evaluate the ratio between the improvement actions associated with the Good Practices and the recommendations proposed by KAIRÓS for the minimum scenario. There were no significant differences (significance level of 0,065).
- Post-test 1: the processing results of KAIRÓS were presented to the experimental group. After 15 days, the recommendations proposed by the system in the Improvement Plan of the experimental group had been incorporated, which didn't happen in the control group. Significant differences were identified between the groups (significance level of 0,003).
- Post-test 2: after the application of the stimulus, it was observed that the ratio between improved Critical Success Factors measurements and measurements that should be intentional according to KAIRÓS, ranged between 0,14 y 0,31 in control group, and in the experimental group between 0,87 and 1,00. In the control group, four successful minimum scenarios were predicted and after two months, only the center with initial status predicted as success could maintain this condition. In the experimental group, four successful minimum scenarios were predicted and after two months, successful states were reached by these centers. Significant differences were identified between the groups (significance level of 0,004).

To assess the applicability and satisfaction, six quality consultants and seven managers of software development centers were surveyed. The variables evaluated were customer satisfaction, applicability and utility through the use of Iadov. A group satisfaction index of 0,92 was obtained. There was a concordance of 84,62% with "Excellent" qualification for the utility and a 92,31% with "Excellent" qualification for the applicability in real environments. About its contribution to the decision-making at the beginning of SPI, there was a concordance of 92,31% with "Excellent" qualification. The rest of qualifications was "Good".

5 Conclusions

Based on the results obtained, it is considered that experiences reuse for the scenarios recommendation and the forecast before the investment in SPI, favor the decision-making in SPI. For the analysis of the information associated with Good Practices and Critical Success Factors combined, it is necessary to lean on artificial intelligence techniques, which facilitate the information processing for decision-making support in SPI.

KAIRÓS intelligent system, automates the processing of Critical Success Factors and Good Practices combined, through the integration of artificial intelligence techniques. The implementation of a GA favors the optimization towards better scenarios in SPI. The association rules allow to identify dependencies between Good Practices and Critical Success Factors measurements. The use of an evolutionary ANN, helps to predict the results of organizations in SPI.

The validation results of the solution corroborate that its application contributes to support decision-making at the beginning of SPI, through the combination treatment of Critical Success Factors and Good Practices. A high satisfaction with the solution is evidenced, in the positive criteria about the contribution of the system and in the evaluation of its implementation effect.

3 References

[1] DOUNOS, P. and G. BOHORIS (2010). Factors for the design of CMMI-based software process improvement initiatives. Conference on Informatics (PCI), 2010 14th Panhellenic. Tripoli IEEE Xplorer Digital Library: 43-47. 1424478383. https://doi.org/10.1109/pci.2010.46

[2] MONTONI, M. A. and A. R. ROCHA (2010). Applying grounded theory to understand software process improvement implementation. Conference on the 2010 Seventh International Quality of Information and Communications Technology (QUATIC). IEEE Computer Society: 25-34. 1424485398. https://doi.org/10.1109/quatic.2010.20

[3] NIAZI, M.; M. A. BABAR and J. M. VERNER (2010). Software Process Improvement barriers: A cross-cultural comparison. Information and software technology, 52(11): 1204-1216. https://doi.org/10.1016/j.infsof.2010.06.005

[4] NIAZI, M.; D. WILSON and D. ZOWGHI (2006). Critical success factors for software process improvement implementation: an empirical study. Software Process: Improvement and Practice, 11(2): 193-211. https://doi.org/10.1002/spip.261

[5] TRUJILLO-CASAÑOLA, Y.; A. FEBLES-ESTRADA and G. LEÓN-RODRÍGUEZ (2014). Modelo para valorar las organizaciones al iniciar la mejora de procesos de software. Ingeniare. Revista chilena de ingeniería, 22(3): 412-420. https://doi.org/10.4067/s0718-33052014000300011

[6] FERNÁNDEZ DÍAZ, H.; N. MILÁN CRISTO; A. M. GARCIA RODRÍGUEZ and Y. TRUJILLO CASAÑOLA. (2016). Bases teóricas para un procedimiento que evalúe cuantitativamente la influencia de los Factores Críticos de Éxito en la Mejora de Procesos. Informática 2016. VII Taller Internacional de Calidad en las Tecnologías de la Información y las Comunicaciones. La Habana, XVI Convención y Feria Internacional INFORMÁTICA 2016.

[7] CLARKE, P. and R. O'CONNOR (2010). Harnessing ISO/IEC 12207 to Examine the Extent of SPI Activity in an Organisation. European

Conference on Software Process Improvement. Springer: 25-36. https://doi.org/10.1007/978-3-642-15666-3_3

[8] NIAZI, M.; D. WILSON and D. ZOWGHI (2005). A maturity model for the implementation of software process improvement: an empirical study. Journal of systems and software, 74(2): 155-172. https://doi.org/10.1016/j.jss.2003.10.017

[9] NIAZI, M.; D. WILSON and D. ZOWGHI (2005). A framework for assisting the design of effective software process improvement implementation strategies. Journal of systems and software, 78(2): 204-222. https://doi.org/10.1016/j.jss.2004.09.001

[10] GOLDBERG, D. E. (1989). Genetic Alogorithms in Search. Optimization & Machine Learning. Choice Reviews Online, 27(2): 27–0936 – 0927–0936. https://doi.org/10.5860/choice.27-0936

[11] PAVEZ-LAZO, B.; J. SOTO-CARTES; C. URRUTIA and M. CURILEM (2009). Selección determinística y cruce anular en algoritmos genéticos: aplicación a la planificación de unidades térmicas de generación. Ingeniare. Revista chilena de ingeniería, 17(2): 175-181. https://doi.org/10.4067/s0718-33052009000200006

[12] MARTÍN, D.; A. ROSETE; J. ALCALÁ-FDEZ and F. HERRERA (2014). QAR-CIP-NSGA-II: A new multi-objective evolutionary algorithm to mine quantitative association rules. Information Sciences, 258: 1-28. https://doi.org/10.1016/j.ins.2013.09.009

[13] OVIEDO CARRASCAL, E. A.; A. I. OVIEDO CARRASCAL and G. L. VÉLEZ SALDARRIAGA (2015). Minería de datos: aportes y tendencias en el servicio de salud de ciudades inteligentes. Revista Politécnica, 11(20): 111-120.

[14] TALLÓN-BALLESTEROS, A. J. (2014). New training approaches for classification based on evolutionary neural networks. Application to product and sigmoidal units. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, 17(54). https://doi.org/10.4114/intartif.vol17iss54pp30-34

[15] TALLÓN-BALLESTEROS, A. J.; C. HERVÁS-MARTÍNEZ; J. C. RIQUELME and R. RUIZ (2013). Feature selection to enhance a two-stage evolutionary algorithm in product unit neural networks for complex classification problems. Neurocomputing, 114: 107-117. https://doi.org/10.1016/j.neucom.2012.08.041

[16] TOSTADO SÁNCHEZ, S. E.; M. ORNELAS RODRÍGUEZ; A. ESPINAL JIMÉNEZ and H. J. PUGA SOBERANES (2016). Implementación de Algoritmos de Inteligencia Artificial para el Entrenamiento de Redes Neuronales de Segunda Generación. JÓVENES EN LA CIENCIA, 2(1): 6-10.

[17] YANG, G.; H. ZHAO; L. WANG and Y. LIU (2009). An implementation of improved apriori algorithm. Conference on 2009 International Machine Learning and Cybernetics. IEEE: 1565-1569. 1424437024. https://doi.org/10.1109/icmlc.2009.5212246

[18] SINGH, J.; H. RAM and D. J. SODHI (2013). Improving efficiency of apriori algorithm using transaction reduction. International Journal of Scientific and Research Publications, 3(1): 1-4.

[19] YABING, J. (2013). Research of an improved apriori algorithm in data mining association rules. International Journal of Computer and Communication Engineering, 2(1): 25-27. https://doi.org/10.7763/ijcce.2013.v2.128

[20] BADDOO, N. and T. HALL (2002). Motivators of Software Process Improvement: an analysis of practitioners' views. Journal of systems and software, 62(2): 85-96. https://doi.org/10.1016/s0164-1212(01)00125-x

[21] BLANCO, K. R.; A. S. BATISTA; D. P. MONTALVÁN; D. N. AGÜERO; A. F. ESTRADA; R. D. MARTÍNEZ and M. M. ROJA (2011). Experiencias del programa de mejora de procesos en la Universidad de las Ciencias Informáticas. Revista Cubana de Ciencias Informáticas, 5(2).

[22] CAPOTE, J.; C. J. LLANTÉN; C. PARDO; A. J. GONZÁLEZ and C. A. COLLAZOS (2008). Gestión del conocimiento como apoyo para la mejora de procesos software en las micro, pequeñas y medianas empresas. Ingenieria e investigacion, 28(1): 137-145.

[23] CONRADI, R.; T. DYBÅ; D. I. SJØBERG and T. ULSUND (2003). Lessons learned and recommendations from two large norwegian SPI programmes. European Workshop on Software Process Technology. Springer: 32-45. https://doi.org/10.1007/978-3-540-45189-1_4

[24] DEL VILLAR, B. L. D. and M. A. M. MATA (2016). Selección de estrategias para la implementación de Mejoras de Procesos Software. ReCIBE, 2(3).

[25] DYBA, T. (2000). An instrument for measuring the key factors of success in software process improvement. Empirical software engineering, 5(4): 357-390.

[26] GONZALO, C.; M. JEZREEL; M. MIRNA and S. F. TOMÁS (2010). Experiencia en la mejora de procesos de gestión de proyectos utilizando un entorno de referencia multimodelo. RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação(6): 87-100.

[27] JALOTE, P. (2002). Lessons learned in framework-based software process improvement. Software Engineering Conference, 2002. Ninth Asia-Pacific. IEEE: 261-265. 0769518508. https://doi.org/10.1109/apsec.2002.1182995

[28] MAS, A. and E. AMENGUAL (2005). La mejora de los procesos de software en las pequeñas y medianas empresas (pyme). Un nuevo modelo y su aplicación a un caso real. Revista Española de Innovación, Calidad e Ingeniería del Software, 1(2): 7-29.

[29] PANTOJA, W. L.; C. A. COLLAZOS and V. M. R. PENICHET (2013). Entorno Colaborativo de Apoyo a la Mejora de Procesos de software en Pequeñas Organizaciones de Software. Dyna, 80(177): 40-48.

[30] PINO, F.; F. GARCÍA and M. PIATTINI (2006). Revisión sistemática de mejora de procesos software en micro, pequeñas y medianas empresas. Revista Española de Innovación, Calidad e Ingeniería del Software, 2(1): 6-23.

[31] SANTOS, G.; M. MONTONI; J. VASCONCELLOS; S. FIGUEIREDO; R. CABRAL; C. CERDEIRAL; A. E. KATSURAYAMA; P. LUPO; D. ZANETTI and A. R. ROCHA (2007). Implementing software process improvement initiatives in small and medium-size enterprises in Brazil. Conference on the 6th International Quality of Information and Communications Technology, 2007. QUATIC 2007. IEEE: 187-198. 0769529488. https://doi.org/10.1109/quatic.2007.22

[32] YÉPEZ VARGAS, W.; C. PRIMERA LEAL and M. TORRES SAMUEL (2013). Mejoras al proceso de planificación de proyectos de software usando el Modelo de Madurez de Capacidad Integrado (CMMI). Compendium, 16(30).

[33] GARCIA RODRÍGUEZ, A. M.; Y. MILANÉS ZAMORA; Y. TRUJILLO CASAÑOLA; J. P. FEBLES RODRÍGUEZ and I. J. SÁNCHEZ GONZÁLEZ (2018). Asociación entre Buenas Prácticas y Factores Críticos para el éxito en la MPS. Revista Cubana de Ciencias Informáticas, 12(2): 89-103.

[34] GARCÍA RODRÍGUEZ, A. M.; Y. TRUJILLO CASAÑOLA and A. PERDOMO VERGARA (2016). Optimización de estados en la mejora de procesos de software. Enl@ ce: Revista Venezolana de Información, Tecnología y Conocimiento, 13(2): 9-27.

[35] GARCIA RODRÍGUEZ, A. M.; Y. TRUJILLO CASAÑOLA and L. ARZA PÉREZ (2016). Pronóstico de éxito en la Mejora de Procesos de Software. Revista Cubana de Ciencias Informáticas, 10: 15-30.

[36] TORRES, M.; K. PAZ and F. SALAZAR. (2006). Tamaño de una muestra para una investigación de mercado, Universidad Rafael Landívar.

[37] PRADHAN, T.; S. R. MISHRA and V. K. JAIN (2014). An effective way to achieve excellence in research based learning using association rules. Conference on 2014 International Data Mining and Intelligent Computing (ICDMIC). IEEE: 1-4. 1479946745. https://doi.org/10.1109/icdmic.2014.6954226

[38] LIN, X. (2014). Mr-apriori: Association rules algorithm based on mapreduce. Conference on 2014 5th IEEE International Software Engineering and Service Science (ICSESS). IEEE: 141-144. 1479932795. https://doi.org/10.1109/icsess.2014.6933531

[39] VIDAL, R.; Y. MA and S. S. SASTRY. (2016). Principal component analysis. en: Generalized Principal Component Analysis. Springer: 25-62.

The Heteroskedasticity Tests Implementation for Linear Regression Model Using MATLAB

Lyudmyla Malyarets, Katerina Kovaleva, Irina Lebedeva, Ievgeniia Misiura and Oleksandr Dorokhov
Simon Kuznets Kharkiv National University of Economics, Nauky Avenue, 9-A, Kharkiv, Ukraine, 61166
E-mail: aleks.dorokhov@meta.ua
http://www.hneu.edu.ua/

Keywords: regression model, homoskedasticity, testing for heteroskedasticity, software environment MATLAB

Received: September 23, 2017

The article discusses the problem of heteroskedasticity, which can arise in the process of calculating econometric models of large dimension and ways to overcome it. Heteroskedasticity distorts the value of the true standard deviation of the prediction errors. This can be accompanied by both an increase and a decrease in the confidence interval. We gave the principles of implementing the most common tests that are used to detect heteroskedasticity in constructing linear regression models, and compared their sensitivity. One of the achievements of this paper is that real empirical data are used to test for heteroskedasticity. The aim of the article is to propose a MATLAB implementation of many tests used for checking the heteroskedasticity in multifactor regression models. To this purpose we modified few open algorithms of the implementation of known tests on heteroskedasticity. Experimental studies for validation the proposed programs were carried out for various linear regression models. The models used for comparison are models of the Department of Higher Mathematics and Mathematical Methods in Economy of Simon Kuznets Kharkiv National University of Economics and econometric models which were published recently by leading journals.

Pozvetek: Avrorji prispevka se ukvarjajo s problemi ekonometričnih modelov z veliko dimenzijami, kjer je izračun problematičen. Razvijejo metodo v MATLABu za multifaktorske regresijske modele.

1 Introduction

In econometrics, a linear regression model is often used to describe different processes and phenomena. Using matrix notation, the linear model regression can be given as:

$$\mathbf{Y} = \mathbf{XB} + \varepsilon \tag{1}$$

where $\mathbf{Y}$ and ε are $n \times 1$ matrices, $\mathbf{X}$ is $n \times (m+1)$, and $\mathbf{B}$ is $(m+1) \times 1$; n is the number of measurements (sample size); m is the number of independent variables in the regression model.

For the i^{th} row of $\mathbf{X}$ (the i^{th} observation) the linear regression model can be written as follows:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_m x_{im} + \varepsilon_i \tag{2}$$

where y_i are the values of the dependent variable, $y_i \in \mathbf{Y}$; i is the experiment identification number, $i = \overline{1, n}$; x_{ij} are the values of the independent variable $x_j \in \mathbf{X}$ ($j = \overline{1, m}$) in the i^{th} experiment; b_0 is the constant term of the equation; b_j are the regression coefficients, $b_0, b_j \in \mathbf{B}$; ε_i are the residuals (model errors).

An error term is introduced in a regression model because the model does not fully represent the actual relationship between the variables of the model. As a result of this incomplete relationship, there are differences between the observed responses (values of the variable being predicted) in the given dataset and those predicted by a linear function of a set of explanatory variables. The error term is the amount at which the equation may differ from measurements. In other words that is the 'white noise'.

As a rule, the building a linear regression model is done by the method of ordinary least squares (OLS). This method for estimating the unknown parameters is based on the minimization of the sum of the squares of the model errors. The estimators of model parameters determined by OLS are known as best linear unbiased estimators (BLUE). The variances of the model parameters are determined by:

$$S^2_{b_j} = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-m-1} z_{jj} = \sigma_e^2 \cdot z_{jj} \tag{3}$$

where z_{jj} is the diagonal element of matrix $\mathbf{Z} = (\mathbf{X}'\mathbf{X})^{-1}$ which corresponds to the parameter b_j; σ_e is the standard error.

The OLS application requires the realization of a number of conditions [1 – 3]. Only if these conditions are met, the estimates calculated by such a model will be unbiased, efficient and well-off. These conditions are formulated in the form of the Gauss – Markov theorem.

According to this theorem there are four principal assumptions which admit the using of linear regression models for research and prediction. One of them is the homoskedasticity (constant variance) of the errors in relation to any independent variable.

Homoskedasticity makes the assumption that the errors have a constant variance: $\text{var}(\varepsilon) = const$ and independent of causal variables: $\text{cov}(x_j, \varepsilon) = 0$ for all j, $j = \overline{1, m}$. The error ε is a random variable distributed according to the normal law: $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ where the mathematical expectation of the error term is zero and the variance is constant. Failure to comply with this requirement leads to bias in the estimates obtained using such a regression model. Thus [4] indicate that estimation uncertainty may increase dramatically in the presence of conditional heteroskedasticity.

The requirement of homoskedasticity also exists in the construction of the econometric model using the maximum likelihood method [5 – 7].

When the scatter of the errors is different, varying depending on the value of one or more of the independent variables, the error terms are heteroskedastic. Namely the distribution law of errors remains normal with a mathematical expectation equal to zero, but the errors of the model are a function of the values of the independent variables: $\varepsilon \sim N(0, f(\mathbf{X}))$, where $f(\mathbf{X})$ is a function that describes the change in the variance of errors as a function of the values of the independent variables.

A similar problem arises during the building of semiparametric [8 – 10] and nonparametric [11, 12] models.

Heteroskedasticity makes difficult to gauge the true standard deviation of the forecast errors. The OLS estimates are no longer BLUE. Thus, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. In particular, heteroscedasticity does not allow us to use equation 3 for the computation of S_{b_j}, since it assumes a uniform dispersion of the errors. Under heteroskedasticity, the sample variance of OLS estimator is

$$\mathbf{Var}(\hat{b_j}) = \sigma_e^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\Omega\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \qquad (4)$$

where **Ω** is the covariance matrix, the elements of which are defined as the variance of the model parameters. Under homoskedasticity, **Ω**= I. Equation 4 is correct if there is no autocorrelation.

For these reasons, all the conclusions obtained on the basis of the corresponding t – statistics and F – statistics, as well as interval estimates, will be unreliable.

A unified approach to the estimation of heteroscedasticity is lacking. To solve this problem, a large number of different tests and criteria have been developed: the Spearman rank correlation test, the Park test, the Glaser test, the Goldfeld – Quandt test, the Breusch – Pagan test, the Leven's test, the White test, and so on.

The application of all the above tests is very difficult for the so-called ‘manual’ account, and for a large set of initial data it is completely impossible.

There are a lot of software with which you can identify heteroscedasticity. These are professional packages (SAS, BMDP), universal packages (STADIA, OLIMP, STATGRAPHICS, STATISTICA, SPSS) and specialized packages (DATASCOPE, BIOSTAT, MESOSAUR).

When using economic data researcher can face two main problems. Firstly, all the listed software are quite expensive and price of the product may be an insurmountable barrier for the young researcher. Secondly, company-developer never provides the source code, considering that this is not necessary for an ordinary user. Therefore, we can not modify the built-in algorithms to detect and eliminate heterosquadity.

Another drawback of the above program products is the outdated conceptual approaches to econometric methods, which are constantly being improved.

For example, the program products SPP and MICROSTAT calculate the coefficient of multiple correlations as the square root of the coefficient of determination. STATGRAPHICS calculates it as the square root of the adjusted coefficient of determination [13]. While in theory the coefficients of multiple correlations is estimated using elements of the correlation matrix [2].

Another important aspect that should be taken into account is the existence of different algorithms to identify heteroskedasticity and the specific problem of division by zero [14].

Ideal option would be to create your own software product that would take into account the research tasks.

However, to write such a program, the economist should be an expert in algorithmic programming. But this happens rarely.

In this article, we carry out a comparative analysis of the tests most often used to detect heteroskedasticity [1, 2, 14] and give their source code. The use of program code allows you to modify the program in accordance with the objectives of the study.

2 Analysis of literary data and the formulation of the problem

Before starting the construction of the regression model, it is necessary to verify whether the conditions of the Gauss-Markov theorem are fulfilled.

One of the main methods of preliminary research on heteroskedasticity is a visual analysis of the graph of residues. On these graphs, the scattering of points can vary depending on the value of the independent variables [14, 15].

To estimate heteroskedasticity, are used such quantitative tests [15 – 17] as the White test, the Goldfeld – Quandt test, the Breusch – Pagan test, the Park test, the Glazer test and also the Spearman test. Unlike other tests the Spearman rank correlation test is a nonparametric statistical test for the heteroskedasticity of random errors in the econometric model. The test algorithm can be studied in detail in [18, 19]. However, it is still not implemented in software products which are used to build multiple models [20 – 27].

In this paper we examined the software packages most commonly used in economic activity, which contain tests for heteroskedasticity [15, 28]. Indeed, these software products do not contain the Spearman rank correlation test.

The most widely used for evaluating heteroscedasticity is the Park test [20, 21]. However, the Park test contains the assumption that the change in the remnants of the model is described by a functional dependence of a certain type. It was noted in [24, 25] that this can lead to unreasonable conclusions. Therefore, the authors propose to consider the Park test together with other tests.

The software implementation of the Park test for multiple models also does not exist [28]. As far as we know software implementation of the Park test for multifactor models also does not exist.

Another test that the authors of the article implemented in the MATLAB environment is the Goldfeld – Quandt test. This test to check for heteroskedasticity of random errors is used when there is reason to believe that the standard deviation of errors is proportional to some variable.

The test statistics has a Fisher distribution [18, 27]. The Goldfeld – Quandt test can also be used if there is an assumption of intergroup heteroskedasticity, when the variance of errors takes, for example, only two possible values. In this case, for the application of the test, there is a need for its software implementation, since applied commercial software has not taken this possibility into account [25, 28].

In scientific articles ~~on~~ for the problem of detecting heteroskedasticity, the Breusch – Pagan test is often considered [10, 29]. We also carried out research this problem. But it oversteps this article.

Analysis of literature sources shows that all tests of heteroskedasticity detection are difficult for 'manual' application and require the development of special software. In turn, the software of econometric research does not contain built-in functions for heteroskedasticity testing with open source code.

That is way the authors of this article attempted to implement the above tests for heteroskedasticity in the construction of multifactor econometric models in the MATLAB software environment.

It should be noted that MATLAB does not contain ready-made software implementation to verify compliance of homoskedasticity. We chose it as a programming environment. For this purpose, other programming environments can also be used, for example, ~~a~~ the free software environment R.

The authors have chosen MATLAB by the following reasons. First, MATLAB is used as a high-level programming language for writing scripts (Spearman.m, Parks.m and Gold_Quan.m). Secondly, MATLAB includes built-in functions for constructing regression models (Econometric toolbox), which gave the authors relief from the duty of programming the standard functions of regression analysis. Thirdly, the authors worked with data structures based on matrices.

3 Aims and objectives of the study

The purpose of the article is to present functions to check for heteroskedasticity in multifactor regression models. The implementation is made in MATLAB.

To achieve this purpose, it is necessary to solve a number of problems. Namely:

- writing the program code in the MATLAB programming environment;
- planning and execution of computer calculations;
- completion of programs;
- analysis and interpretation of results;
- comparison with the results of calculations using software products of leading companies.

4 Practical implementation of the criteria for the detection of heteroskedasticity in econometric models in the MATLAB

4.1 Spearman's rank correlation test for multiple regression models

The use of the Spearman's test assumes that the variance of model errors will increase (or decrease) with increasing values of the independent variable.

This means that the absolute values of errors ε_i $(i=\overline{1,n})$ and the values x_{ij} of the independent variable x_j $(j=\overline{1,m})$ will correlate with each other.

To check whether heteroskedasticity is statistically significant the Spearman's test provides for the following stages:

1) Estimation of the parameters of the econometric model by the OLS:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_m x_{im}, \quad (5)$$

where $\hat{y}_i$ is the predicted response in accordance with the model when the independent variables are $(x_{i1};\ x_{i2};\ ...x_{im})$;

2) Calculate model errors as the difference between the empirical and the ratchet value of the dependent variable:

$$\varepsilon_i = y_i - \hat{y}_i \tag{6}$$

where y_i is the value of the dependent variable in the i^{th} experiment;

3) The pairs (x_{ij}, ε_i) are ranked in order of increasing values of the independent variable x_j;

4) The coefficient of rank correlation between ε_i and x_{ij} is calculated as

$$r_{x\varepsilon} = 1 - 6 \cdot \frac{\sum_{i=1}^{n} d_i^2}{n \cdot \left(n^2 - 1\right)}, \tag{7}$$

where d_i is the difference between the two ranking;

5) The significance of $r_{x\varepsilon}$ is tested by using $t-$ statistic:

$$t = \frac{r_{x\varepsilon}\sqrt{n-2}}{\sqrt{1-r_{x\varepsilon}^2}} \tag{8}$$

6) In accordance with the predetermined confidence probability p (where $\alpha = 1 - p$) the tabulated value of $t_{cr.} = t_{0.5\alpha}(n-2)$ is found. Then the calculated value is compared with the critical one.

If the *t*-statistic value is greater than the critical value, we must say that heteroscedasticity is statistically significant. Here α is the significance level which is chosen to test the null hypothesis: $\rho_{x\varepsilon} = 0$. In the opposite case, the null hypothesis is non-contradictory.

As an example of the implementation of this test, we can suggest the following m-file named *Spearman*:

```
========================================
% initialization:
X1 = load('data1.scv');
X2 = load('data2.scv');
X3 = load('data3.scv');
X4 = load('data4.scv');
X5 = load('data5.scv');
Y = load('data.scv');
```

% Formation of the source data array:

```
X = [ones(n,1) X1' X2' X3' X4' X5'];
% Construction of a linear multifactor
% model by OLS - method:
[b,bint,r,rint,stats] = regress(Y,X,0.05);
y_p = b(1) + b(2).*X1 + b(3).*X2+
b(4).*X3+b(5).*X4+b(6).*X5
sprintf('Model:')
fprintf('y_p = %f + %f *X1+%f *X2+%f *X3+%f
*X4+%f *X5',b)
% Calculation of model remains:
e = Y - y_p';
% Preparing an array for further work:
X = [X1' X2' X3' X4' X5'];
[n,m] = size(X);% Determining the size of the
source data
%=======================================
%% Spearman rank correlation test
% Ranking of factors:
[Xs I] = sort(X)
Dx = zeros(n,m);
for j = 1:m
    for i = 1:n
        Dx(i,j) = i;
    end
end
TMP = zeros(n,m);
% Filling an array of factors with ranks
% taking into account their sequence numbers:
for j = 1:m
    for i = 1:n
        i1 = I(i,j);
        TMP(i1,j) = Dx(i,j);
    end
end
X = [X TMP]% Output array
% Ranking of remains:
[es I] = sort(e);
es = [es ones(size(e),1)];
e = [e ones(size(Y),1)];
sprintf(' critical values t:','\n')
t_r(:,j) = (r(:,j)*sqrt(n-1))/sqrt(1 -
r(:,j)^2);
end
t_r % output array t-Statistics by Spearman
% Comparative analysis and conclusions:
c = 0;
for i = 1:size(e)
    es(i,2) = i;
end
% Filling an array of remains with ranks
% taking into account their sequence numbers:
for i=1:size(e)
    e(I,2) = es(:,2);
end
e% an array of remains which contains ranks
r = zeros(1,m);
d = zeros(n,m);
% Calculating the difference of ranks
for j = 1:m
    for i = 1:n
        d(i,j) = TMP(i,j) - e(i,2);
    end
end
d % difference in rank
% The square of the difference of ranks:
for j = 1:m
    d(:,j) = d(:,j).^2;
end
d
Sd = zeros(1,m);
% The sum of the difference of ranks squares
% by the corresponding columns of ranks:
for j = 1:m
    Sd(:,j) = sum(d(:,j));
end
Sd % output array
% Calculating Spearman's Statistics:
for j = 1:m
    r(:,j) = 1 - (6*Sd(:,j))/(n*(n^2-1));
end
r % Output array
t_r = zeros(1,m);
%% Testing of the significance of the Spearman
coefficient:
t_t = tinv(0.975,n-2)% tabulated value t
for j = 1:m
  if abs(t_r(:,j)) < abs(t_t)
    sprintf(' Heteroskedasticity is absent ')
  else
    sprintf(' Heteroscedasticity is present ')
    c = c + 1;
  end
end
========================================
```

4.2 Park's test for multiple regression models

R. Park proposed a test to check for heteroskedasticity, which is based on some formal dependencies. Namely, it assumes that the heteroskedasticity may be proportional to some power of an independent variable x_j in the multiple models.

Since the variance of errors $\sigma_i^2 = \sigma^2(\varepsilon_i)$ is a function of the $i-$th value x_{ij} of the explanatory variable x_j, and for its description Park proposed the this dependence: $\sigma_i^2 = \sigma^2 x_{ij}^{\beta} \varepsilon^{v_i}$.

After computing its logarithms, we obtain the following relation: $\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln x_{ij} + v_i$. Since the variances σ_i^2 are usually unknown, they are replaced by their estimates ε_i^2.

The Park's test provides for such effectuation stages:

1) Estimation of the parameters of the econometric model by the OLS (Equation 5);

2) Calculation of the value $\ln \varepsilon_i^2 = \ln(y_i - \hat{y}_i)^2$ for each observation;

3) Building the regression model:

$$\ln \varepsilon_i^2 = \alpha + \beta \ln x_{ij} + \upsilon_i, \tag{9}$$

where $\alpha = \ln \sigma^2$. For the case of multiple regressions, this dependence is constructed for each explanatory variable;

4) Verification of statistical significance of the coefficient β on the basis of $t-$statistics:

$$t = \left| \beta / \sigma_{\beta} \right|. \tag{10}$$

5) In accordance with the predetermined confidence probability p (where $\alpha = 1 - p$) the tabulated value of $t_{cr.} = t_{\alpha}(n-m-1)$ is found. Then the calculated value is compared with the critical one.

If $t > t_{\alpha}(n-m-1$, then at the level of significance α the coefficient β is statistically significant and there is a link between $\ln \varepsilon_i^2$ and $\ln x_i$. It means that heteroskedasticity is present in statistical data.

The M-file named *Park's* which is implementation of the Park test has the form:

```
=======================================
% initialization:
X1 = load('data1.scv');
X2 = load('data2.scv');
X3 = load('data3.scv');
X4 = load('data4.scv');
X5 = load('data5.scv');
Y = load('data.scv');
% Formation of the source data array:
X = [ones(n,1) X1' X2' X3' X4' X5'];
[n, m] = size(X);
% ==========  Park Test Algorithm =======
%  1 stage of the Park test
% Construction of a linear multifactor
% model by OLS - method:
[b,bint,r,rint,stats] = regress(Y,X,0.05);
y_p   =   b(1)   +   b(2).*X1   +   b(3).*X2+
b(4).*X3+b(5).*X4+b(6).*X5
sprintf('Model:')
fprintf('y_p  =  %f  +  %f  *X1+%f  *X2+%f  *X3+%f
*X4+%f *X5')
%   2 stage of the Park test
ln_eps = log((Y' - y_p).^2)
%   3 stage of the Park test
for j=1:m
    for i = 1:n
        X(i,j) = log(X(i,j));
    end
end
%  4 stage of the Park test
for i = 2:m
 [bet, dev,stat] = glmfit(X(:,i),ln_eps);
    t_t = tinv(0.95, n-2);
    t_r = stat.t(2);
% Comparative analysis and conclusions:
        if abs(t_r) < abs(t_t)
            sprintf('  Heteroskedasticity  of  %i
factor is absent \n',i-1)
        else
            sprintf('  Heteroskedasticity  of  %i
factor is present\n',i-1)
        end
end
=======================================
```

The Park test's weakness is that it assumes the heteroskedasticity has a particular functional form.

4.3 Goldfeld – Quandt test for multiple regression models

When using the Goldfeld-Quandt test for heteroscedasticity, it is assumed that model errors σ_{ε} depend on one of the external variables x_j: $\sigma_{\varepsilon_i}^2 = \sigma^2 x_{ij}^2$

It is also assumed that errors ε_i are distributed according to the normal law, there is no autocorrelation.

The Goldfeld-Quandt test provides for such effectuation stages:

1) Estimation of the parameters of the econometric model by the OLS (Equation 5);

2) Ranking of all n observations in magnitude of the independent variable x_j;

3) Segregation this ordered sample into three approximately equal parts k, $n-2k$, k, respectively;

4) For each part of the sample that has a volume k, its regression equation is constructed and the sums of the squares of the deviations determine:

$$RSS_1 = \sum_{i=1}^{k} \varepsilon_i^2 \tag{11}$$

and

$$RSS_3 = \sum_{i=n-k+1}^{n} \varepsilon_i^2 . \tag{12}$$

Than empirical meaning of the $F-$statistic is calculated:

$$F = \frac{RSS_1/(k-m-1)}{RSS_3/(k-m-1)} \quad (13)$$

5) Evidence of heteroskedasticity is based on a comparison of the residual sum of squares (*RSS*) using the F – statistic. The calculated value is compared with the critical value $F_{cr.} = F_\alpha(k-m-1; k-m-1)$ in accordance with the predetermined confidence probability p (where $\alpha = 1-p$).

If $F < F_\alpha(k-m-1; k-m-1)$, this means that at the level of significance α the hypothesis that there is no heteroskedasticity does not have grounds to reject. In the opposite case, the hypothesis of the absence of heteroskedasticity is rejected.

For multiple regressions, we performed tests for all factors. The M-file named *Gold_Quan* which is the implementation of the Goldfeld – Quandt test has the form:

```
=======================================
% initialization:
X1 = load('data1.scv');
X2 = load('data2.scv');
X3 = load('data3.scv');
X4 = load('data4.scv');
X5 = load('data5.scv');
Y = load('data.scv');
% Formation of the source data array:
X = [ones(n,1) X1' X2' X3' X4' X5'];
[n, m] = size(X);
%=======================================
%% Goldfeld – Quandt test:
[Xsort Is] = sort(X);
for i=1:size(Y)
    Ysort(i,1) = Y(Is(i),1);
end
Dat = [Xsort Ysort];
c = fix(4*n/15);
k = fix((n - c)/2);
if floor(k) > 0.4
    k = k+1;
end
k
% Selective aggregate 1:
Dat1 = Dat(1:k,:);
[b1,dev1,stats1] = glmfit(Dat1(:,1),Dat1(:,2));
S1 = sum(stats1.resid.^2);
% Selective aggregate 2:
Dat2 = Dat(n-k+1:n,:);
[b2,dev2,stats2] = glmfit(Dat2(:,1),Dat2(:,2));
S2 = sum(stats2.resid.^2);
% Testing the hypothesis:
if S1 > S2
    Fp = S1/S2;
else
    Fp = S2/S1;
end
Ft = finv(0.95,k-m-1,k-m-1);
if Fp > Ft
     sprintf(Heteroscedasticity is present ')
else
    sprintf(Heteroscedasticity is absent ')
end
=======================================
```

A weakness of the Goldfeld – Quandt test is that the result is dependent on the criteria chosen for separating the sample measurements into their representative groups.

5 Results of numerical experiments

The problem of detecting heteroskedasticity in various multifactor econometric models was considered.

For carrying out numerical simulation experiments we used both the models of the Department of Higher Mathematics, Economic and Mathematical Methods of KhNEU [30 – 33], and econometric models which were published recently by leading journals [34 – 36].

To check for heteroscedasticity, we used real data. This is one of the advantages of this paper. However, it is possible to use the data obtained with the Monte Carlo simulation [6, 7, 37 – 39].

Numerical experiments were performed on the configuration AMD Athlon 64 3200+1.5Gb Ram, graphic accelerator – Nvidia GeForce GTX 560 2Gb with using technology NVIDIA CUDA 4.2.

Let's look at a concrete example of what happens to an eccentric model, if you do not take into account heteroskedasticity.

As a model problem, the linear regression model was calculated for the cost of electronic textbooks produced by the Department Higher Mathematics and Mathematical Methods in Economy. The initial data and designations used in the process of correlation-regression analysis are shown in Figure 1, where Y is the resulting factor Y (cost of the electronic textbook).

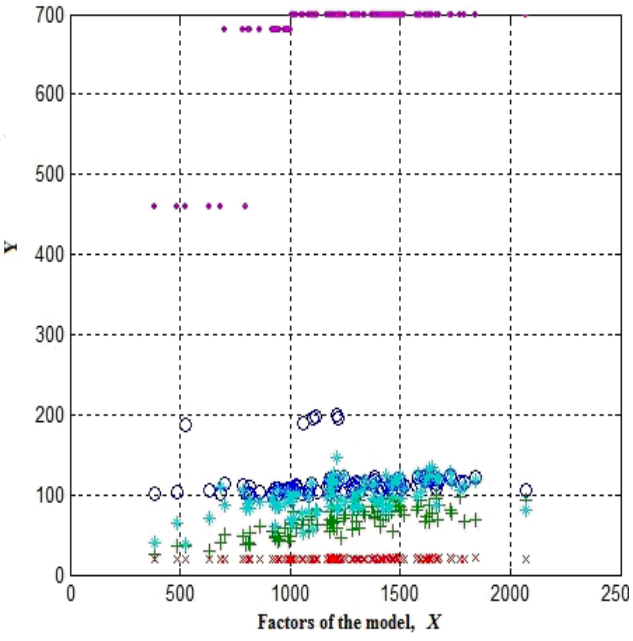


Figure 1: Initial data for model building

Figure 1 shows the dependence of the cost of the electronic textbook (Y) on such external factors:

○ - X1 (average cost of developers' wages);
+ - X2 (publication volume);
× - X3 (average CD recording price);

* - X4 (storage and distribution costs);
• - X5 (cost of the use of licensed software).

The regression model was constructed using the built-in function Matlab-regress (y, X, alpha) with the code:

```
=========================================
% The program for multiple regression model building, if heteroskedasticity is not taken into account :
[b,bint,r,rint,stats] = regress(Y,X,0.05);
y_p = b(1) + b(2).*X1 + b(3).*X2+ b(4).*X3+b(5).*X4+b(6).*X5;
sprintf(' Heteroskedasticity is not taken into account:')
fprintf('y_p = %f + %f *X1+%f *X2+%f *X3+%f *X4+%f *X5',b)
=========================================
```

The program for constructing multiple regressions, if you do not take into account heteroskedasticity, gives such a result:

$$\hat{y} = -1864.06 + 0.33 \cdot x_1 + 10.61 \cdot x_2 + \\ + 70.90 \cdot x_3 + 3.33 \cdot x_4 + 0.87 \cdot x_5. \quad (14)$$

The results of calculating the errors of the model represented by the Equation 10 are shown in Figure 2.

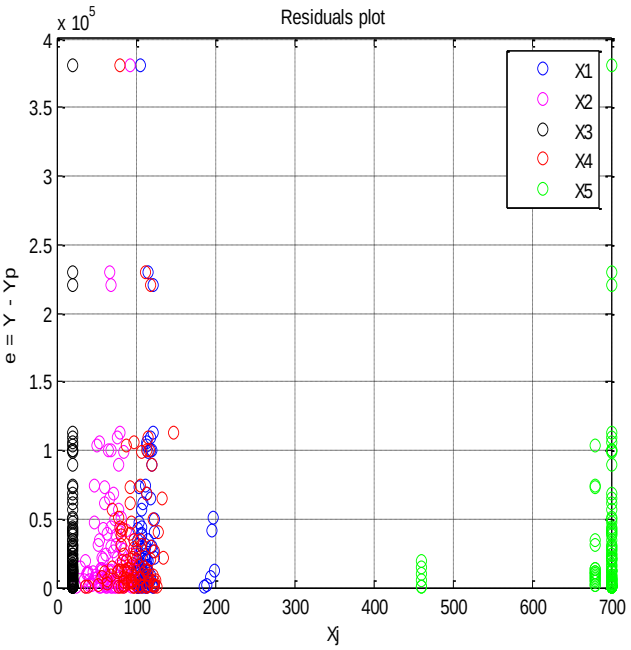


Figure 2: Graphic illustration of the remnants of the model

Analysis of the remnants of the model indicates that for this model the dispersion of remnants increases with an increasing of the value of external factors, that is, heteroskedasticity can not be ignored.

Using the program procedures developed by the authors to identify heteroskedasticity, the following results were obtained:

```
=========================================
   ans = Heteroskedasticity 1 is absent
   ans = Heteroskedasticity 2 is absent
   ans = Heteroskedasticity 3 is absent
   ans = Heteroskedasticity 4 is absent
    ans = Heteroskedasticity 5 is present
=========================================
```

The construction of the regression model, which takes into account the heteroskedasticity, was performed using the built-in function MATLAB: robustfit (X, y, wfun, tune,const).

It should be emphasized that the presence or absence of heteroskedasticity in the initial data is determined automatically by using the check box.

For this we used the code:

```
=========================================
%c is a parameter that takes the value 0 or 1
%(where 0 - Heteroscedasticity is absent, 1 -
% Heteroscedasticity is present),
%c depends on the result of the scripts' work
if c > 0
X = [X1' X2' X3' X4' X5'];
[b,stats3] = robustfit(X,Y,'fair',0.001,'on');
y_p = b(1) + b(2).*X1 + b(3).*X2+
b(4).*X3+b(5).*X4+b(6).*X5;
sprintf('Heteroskedasticity is taken into account:')
fprintf('y_p = %f + %f *X1+%f *X2+%f *X3+%f
*X4+%f *X5',b)
   end
=========================================
```

The program for multiple regression model building, if heteroskedasticity is taken into account yields this result:

$$\hat{y} = 27.85 + 0.94 \cdot x_1 + 10.33 \cdot x_2 - \\ - 29.16 \cdot x_3 + 4.18 \cdot x_4 + 0.80 \cdot x_5. \quad (15)$$

Thus, the above procedure allows eliminating heteroskedasticity. In this case, the resulting models will be able to adequately reflect the reality.

Table 1 shows the results of numerical experiments on testing of programs which are presented in this article on various multifactor models.

As can be seen from Table 1, software products developed by us using MATLAB can be proposed both for constructing multifactor econometric models, and for investigating the latter for the presence of heteroskedasticity.

In doing so, we used new numerical algorithms, developed on the basis of well-known tests of heteroskedasticity detection.

Open source code allows the researcher to use this software to solve their own problems.

Multiple Models	Theoretical results	The results of the work of the authors' programs		
		Spirmen.m	*Park.m*	*Goldfeld – Quandt.m*
Model [28]: Linear approximation	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent
Power approximation	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent
Hyperbolic approximation	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present
Model [30]	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present
Model [31]	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is present
Model [33]	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent
Model [34]	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent	Heteroskedasticity is absent
Model [35]	Heteroskedasticity is present	Heteroskedasticity is present	Heteroskedasticity is absent*	Heteroskedasticity is present

** The conclusion is not justified, since the test uses a monotonically increasing function*

Table 1: Results of testing programs on multiple models

6 Conclusion and future work

The article examined one of the key problems of regression analysis, which consists in verifying the fulfillment of the requirement of homoskedasticity of the remainders of the model. To this end we used various statistic tests.

Analysis of literature sources and our own studies confirm the complexity of using all existing tests for detecting heteroskedasticity in the 'manual account' mode. Therefore, we gave our own implementation in MATLAB for tests used for detecting heteroskedasticity.

This problem was successfully solved, as shown results of numerical experiments which are presented in the article. We represent all software products we have created with open source code, which enables each researcher to customize the program to their problems.

In conclusion, we want to note that the work presented in this article is an on going work having the final purpose to create a complete and effective software for detecting-heteroskedasticity in regression models.

Another further development consists in developing a complete econometric toolbox in MATLAB.

7 Reference

[1] Dougherty C (2016). Elements of econometrics http://www.londoninternational.ac.uk/sites/default/files/programme_resources/lse/lse_pdf/subject_guides/ec2020_ch1-4.pdf

[2] Dougherty C (2016). *Introduction to Econometrics* (5th edition) University Press: Oxford

[3] Hansen B (2018). Econometrics. University of Wisconsin http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf

[4] Brüggemann R, Jentsch C and Trenkler C (2016). Inference in VARs with conditional heteroskedasticity of unknown form *Journal of econometrics* 191 pp. 69-85. http://dx.doi.org/10.1016/j.jeconom.2015.10.004

[5] Cordeiro G (2008). Corrected maximum likelihood estimators in linear heteroskedastic regression models *Brazilian Review of Econometrics* 28 pp. 11–16.

[6] Hayakawa K and Pesaran H (2015). Robust standard errors in transformed likelihood estimation of dynamic panel data models with cross-sectional heteroskedasticity *Journal of econometrics* 188 pp. 111-134. http://dx.doi.org/10.1016/j.jeconom.2015.03.042

[7] Chen S, Khan S and Tang X (2016). Informational content of special regressors in heteroskedastic binary response models *Journal of econometrics* 193 pp. 162-182. http://dx.doi.org/10.1016/j.jeconom.2015.12.018

[8] Kai B, Li R and Zou H (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models *The annals of statistics* 39 pp. 305-332

[9] Pelenis J (2014) Bayesian regression with heteroscedastic error density and parametric mean function *Journal of econometrics* 178 pp. 624-638. http://dx.doi.org/10.1016/j.jeconom.2013.10.006

[10] Norets A (2015). Bayesian regression with nonparametric heteroskedasticity *Journal of econometrics* 185 pp. 409-419. http://dx.doi.org/10.1016/j.jeconom.2014.12.006

[11] Wei C and Wan L (2015). Efficient estimation in heteroscedastic varying coefficient models *Econometrics* 3 pp. 1-7.

[12] Shen S, Cui J and Wang C (2014). Testing heteroscedasticity in nonparametric regression based on trend analysis *Journal of Applied Mathematics* 2014 pp. 1-5. http://dx.doi.org/10.1155/2014/435925

[13] Pen R (2015). Planirovaniye experimenta v Statgraphics Centurion. Mezdunarodnye zurnal eksperimentalnoy obrazovanija pp. 160–161 (in Russian)

[14] Malyarets L (2014). *Economico-mayematechni metody i modeli*. KhNEU im. S Kuznetz: Kharkiv (in Ukrainian)

[15] Williams R (2015). Heteroskedasticity University of Notre Dame https://www3.nd.edu/~rwilliam/stats2/l25.pdf
[16] Kolchinskaya E (2015). Vlijanie transportnoj infrastrucrury na promyshlennoe razvitie regionov Rossii. *Aktualnye problemy ekonomiki* 34 pp. 77-82 (in Russian)
[17] Radkovskaya E (2015). Matematicheskie metody v sovremennyh ekonomicheskih issledovaniyah. *Vestnik Yugorskogo gosudarstvennogo universiteta* 37 pp. 37-40 (in Russian)
[18] Gmurman V (2001). *Teorija verojatnostej i matematicheskaya statistika* (7th edition). Vyshaya shkola: Moscow (in Russian)
[19] Heteroskedasticity http://gauss.stat.su.se/gu/e/slides/Lectures%208-13/Heteroscadasticity.pdf
[20] Redace R (2017). Use the Park test to check for heteroskedasticity http://www.dummies.com/education/economics/econometrics/use-the-park-test-to-check-for-heteroskedasticity/
[21] Mazorchuk M (2014). Osobennosti vybora metodov izmereniya nadezhnosti pedagogicheskih tekstov. *Radioelectrohhi i komp'jutorni sistemy* pp. 131-137 (in Russian)
[22] Kim D, El-Tawil S and Naaman A (2007). Correlation between single fiber pullout and tensile response of FRC composites with high strength steel fibers. Fifth International RILEM Workshop on High Performance Fiber Reinforced Cement Composites (HPFRCC5). RILEM, ed H W Reinhardt and A E Naaman: Paris pp. 67-76.
[23] Baranov N and Sorokin L (2015). Komp'yuternye prikladnye programmy v formatirovanii stilja myshleniya budushchego spetsialista. *Mezhdunarodnyj nauchno-issledovatekskij zhurnal* 42 pp. 60-62 (in Russian)
[24] Kazanskaya A and Kompanietz V (2009) Opyt issledovanija metodov klasternogo analiza iz paketa Statistica 6. 0 na primere vuborki gorodov. *Izvestiya YuFU, Tehnicheskie nayki* pp. S103-110 (in Russian)
[25] Kosonogov V (2014). The psychometric properties of the Russian version of the empathy *Quotient Psychology in Russia* pp. 196-104
[26] Yüce M (2017). An Asymptoic test for the detection of heteroskedasticity http://eidergisi.istanbul.edu.tr/sayi8/iueis8m2.pdf
[27] Redace R (2017). Test for heteroskedasticity with the Goldfeld – Quandt test http://www.dummies.com/education/economics/econometrics/test-for-heteroskedasticity-with-the-goldfeld-quandt-test/
[28] Krasilnikov D (2011). Programmnoe obespechenie ekonometricheskogo issledovanija Econometric Software. *Vestnik Nizhegorodskogo universiteta im. N I Lobachevskogo* pp. 231-238 (in Russian)
[29] Halunga A, Orme C and Yamagata T (2017). A heteroskedasticity robust Breusch–Pagan test for contemporaneous correlation in dynamic panel data models *Journal of econometrics* 198 pp. 209-230. https://doi.org/10.1016/j.jeconom.2016.12.005
[30] Ponomarenko V, Malyarets L and Dorokhov A (2011). Obespechenie kontrolya logisticheskoyj deyatelnosti s minimizatsiey logisticheskih zatrat. *Izvestija IGEA* pp. 137-142 (in Russian)
[31] Malyarets L (2011). *Matematychni metody v suchasnyh tkonomichnih doslidzhennyah* KhNEU im. S Kuznetz: Kharkiv (in Ukrainian)
[32] Kovaleva E (2015). Regressionnaya model sebestoimosti elektronnyh multimedijnih izdaniy *Vestnik NTU KhPI. Mehaniko-tehnologichni sistemy i kompleksy* pp. 55-60 (in Russian)
[33] Malyarets L (2016). *Matematychni metody i modeli v upravlinni ekonomichnymy protsesamy* KhNEU im. S Kuznetz: Kharkiv (in Ukrainian)
[34] Degtyareva T, Buresh O and Chepasov V (2003). Statisticheskiy analiz transportnogo kompleksa regiona na osnove regressionnyh modelej. *Voprosy statistiki* pp. 65-67 (in Russian)
[35] Jacob J and Lamari M (2012). Factors influencing research productivity in higher education: an empirical investigation *Foresight* 6 pp. 40-50
[36] Krasnobokaya I (2011). Analiz formirovaniya sebestoimosti produktsii proizvodstvennogo predprijatiya s ispolzovaniem mnogofaktornyh ekonometricheskih modeley *Ekonomicheskiy analiz: teoriya i praktika* pp. 38-47 (in Russian)
[37] Chao J, Hausman J, Newey W, Swanson N and Woutersen T (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity *Journal of econometrics* 178 pp. 15-21. https://doi.org/10.1016/j.jeconom.2013.08.003
[38] Bekker P and Crudu F (2015). Jackknife instrumental variable estimation with heteroskedasticity *Journal of econometrics* 185 pp. 332-342. https://doi.org/10.1016/j.jeconom.2014.08.012
[39] Cavalierea G, Nielsen M and Taylor A (2015). Bootstrap score tests for fractional integration in heteroskedastic ARFIMA models, with an application to price dynamics in commodity spot and futures markets *Journal of econometrics* 187 pp. 557-579. https://doi.org/10.1016/j.jeconom.2015.02.039

https://doi.org/10.31449/inf.v42i4.1510 Informatica **42** (2018) 555–561 **555**

Analysing RPC and Testing the Performance of Solutions

Sandor Kiraly
Eszterhazy Karoly University, Eger, Eszterhazy ter 1., Hungary
E-mail: kiraly.sandor@uni-eszterhazy.hu

Szilveszter Szekely
Imperial College, Kensington, London SW7 2AZ, UK
E-mail: szekelyszilv@gmail.com

Keywords: remote procedure call, marshalling, Google protocol buffers, JSON-RPC, XML-RPC, performance test

Received: January 29, 2017

In distributed computing, network sockets provide mechanism for a process to establish a remote connection to another process and send messages back and forth. This interface makes possible a proper mechanism that allows a program running as a process on computer A to call a procedure or a function on remote computer B and pass parameters to it. In the case of synchronous Remote Procedure Call (RPC), processes on computer A need to wait for the finishing of execution of procedures on computer B. When the called procedure finishes, produces its result and passes it to the process on computer A that can continue execution. The question is what happens between the time of the remote procedure call and arrival of the returned values and how much the caller must wait for result. Prompted by the release of Protocol Buffers and gRPC by Google, this paper answers that question, describing the structure of third generation RPCs and analysing them putting the focus on performance and the way of marshalling parameters. To facilitate the choice between them this paper represents the results of performance tests carried out by the authors.

Povzetek: Podana je analiza oddaljenih klicev (RPC) v distribuiranih sistemih predvsem v smislu performans.

1 Introduction

While developing computer applications, using procedures and functions is very common. In most cases the subroutines work independently so they could even be run on a remote computer. To reach the remote subroutine (procedure or function) network communication is necessary that is performed via RPC mechanisms. Since the caller and callee procedures run on different machines, they execute them in different address spaces, and different operating system which cause complications. Parameters and results also have to be passed, which can be complicated, especially if the software architectures are not identical or the data structures are complex. Still, most of these can be dealt with, and RPC is a very popular technique that underlies many distributed systems. [1]

To understand the working of RPC it is necessary to examine how local procedure calls are implemented. Before calling a procedure the processor stores the local variables and the state of the caller procedure on the stack while the running of the current procedure will be suspended. To perform the call, the caller pushes the parameters onto the stack in order, last one first. The processor transfers the control to the address determined by the call. In the callee procedure, the compiler is responsible for saving the necessary registers, allocating stack space for local variables, and then restoring the registers and stack prior to the return from the callee. After the procedure has finished running the processor puts the return value in a register, removes the return address, and transfers control back to the caller. The caller then removes the current parameters from the stack, returning it to the original state.

This method cannot be performed if the callee procedure is stored on a remote computer since there are two different running contexts. To solve the problem, another function is used that looks like the remote procedure and it contains code for sending and receiving messages over the network. Its name is stub function. Figure 1 represents the working of remote procedure call for a function pow that returns a long value

More text of the introduction. More text of the introduction. More text of the introduction. More text of the introduction.

The sequence of operations labeled in Figure 1 is as follows:

The client calls a local function (1) that seems to be the actual function but it is the client stub function that serializes the parameters into a message (raw byte stream) (2), and then sends the message to the server machine (3) using socket interfaces. The server stub deserializes the parameters from the raw message (4), and then calls the server function (5) passing it the arguments that it received from the client using the standard calling sequence. After completing the server function, it passes the return value to the server stub (6) that serializes it into a message (7) to

send to the client stub. The message is sent back across the network (8) and the network layer passes the message to the client stub (9) that reads and deserializes it then returns the result to the client function (10).

Figure 1 represents a remote procedure call applying passing parameters by value which is simple since it just copies the value into the network message. Passing by reference is more complex. To enable this technique it is necessary to send a copy of the arguments over, place them in memory on the remote computer, pass a pointer to them to the server function, and finally send the object back to the client, copying it over the reference. For complex structures, it is needed to copy the structure into a pointerless representation, transmit it, and reconstruct the data structure on the remote machine. [2][3][4]

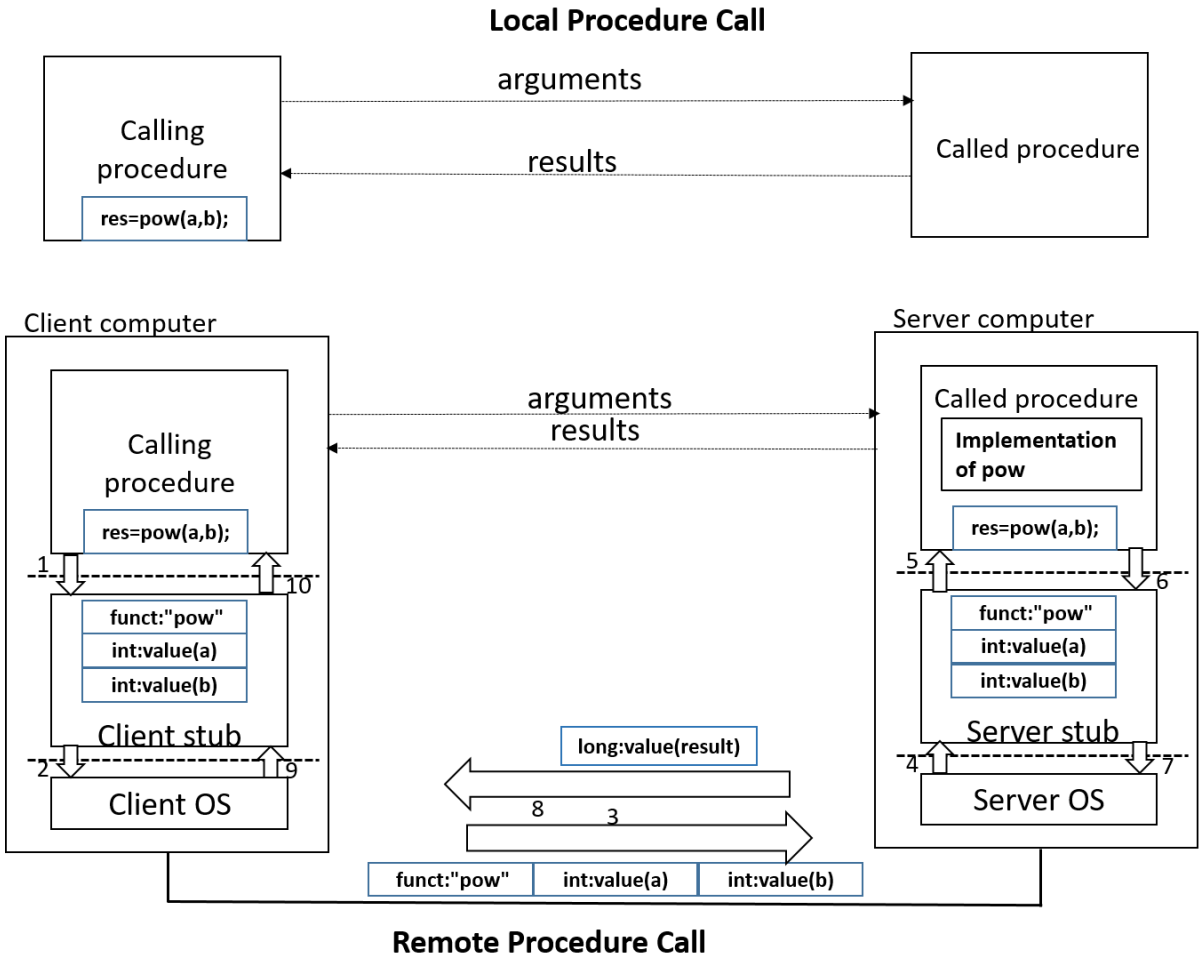


Figure 1: The RPC mechanism comparing with the Local Procedure Call.

Both the client program and the callee function see only ordinary, local procedure calls, using the normal calling conventions. Only the stubs know that the call is remote. It also means, the performance of RPC depends on the stub implementation apart from the network conditions.

Most languages were not designed to handle remote procedures natively with built in transparent stubs. That is the reason why they are not capable of generating the necessary stub functions. To enable them for performing remote procedure calls, the commonly adopted solution is to provide a separate compiler that can generate both the client and server stub functions. The input of this compiler comes from the remote procedure call interfaces written by a programmer. These are written in an interface definition language (IDL) for example proto3 in gRPC. After the RPC compiler is run, the server and client programs can be compiled and linked with the appropriate stub functions. Both the client and the server codes need to be changed to initialize the RPC mechanism.

2 RPC APIs

RPC implementations generally use supporting libraries to complete the stub operations. They must provide the following operations:

Name service operations: They must register themselves and support servers to advertise these bindings and clients to find them.

Binding operations: They establish client/server communications using the appropriate protocol.

Endpoint operations: They register endpoint information (protocol, port number, machine name) to the name server and listen for procedure call requests.

Security operations: They provide the authentication procedure and a secure communication channel between the two computers

Internationalization operations (possibly): They include functions to convert currency formats, time formats and language-specific strings through string tables.

Marshaling/data conversion operations: They pack data into package for transmitting onto a network and functions to reconstruct it. Sometimes, they have to serialize the messages as well.

Stub memory management and garbage collection: It may occur that stubs need to allocate memory for storing parameters, particularly in case of accomplishing pass-by-reference technique. RPC library needs to allocate and clean up such allocations. For RPC packages that support objects, the RPC system must provide the deletion of unnecessary references to objects.

Program ID operations: They allow applications to access identifiers of sets of RPC interfaces for communication.

Object and function ID operations: They support passing references to remote functions or remote objects to other processes. [5]

The more effective the implementation of these operations the faster the RPC solution will be.

3 Third generation RPCs and Web Services

Microsoft DCOM (Distributed Component Object Model) and CORBA (Common Object Request Broker Architecture) were the first RPC solutions that supported the object oriented programming techniques, and CORBA also includes IDL to specify the name of classes, their attributes, and their methods. It based on binary serialization. [5]

The increasing popularity of internet use led that web browsers became the dominant model for accessing information. Clients access the service via the HTTP protocol that allows services to be published, discovered, and used in a technology-neutral form.

Web server is configured to recognize the part of the URL pathname and pass the request to a specific plug-in module. This module can strip out the headers, parse the data (if needed), and call any other functions or modules as needed. [6][7]

XML-RPC

XML-RPC is one of the simplest web service approaches that was designed in 1998 as an RPC messaging protocol for serializing procedure requests and responses into human-readable XML. The XML format uses HTTP protocol to send data from a client computer to a server computer using traditional web ports for RPC.

XML-RPC does not define any standard methods for generating stub functions or handling remote procedures. It only focuses on messaging and therefore consists of only three small parts:

XML-RPC data model is a set of types used in passing parameters, return values, and faults (error messages).

XML-RPC request structures that contain method and parameter information for supporting HTTP requests.

XML-RPC response structures that contain return values or fault information for supporting HTTP responses.

For the performance test several libraries are available for example Apache XML-RPC that was selected to compare to other solutions.

3.1 SOAP and WSDL

The XML-RPC specification was used as a basis for creating SOAP (Simple Object Access Protocol) that is an open-standard, XML-based messaging protocol for exchanging information among computers. It is platform- and language-independent and enables client applications to easily connect to remote services and invoke remote methods. For creating a standardized messaging structure it is necessary to define a service definition document in WSDL (Web Services Description Language) so that to create and check the proper SOAP messages. Though, WSDL is an XML document, it is hard to create and read it by human, therefore tools such as *Java2WSDL* or *wsdl.exe* (in .NET) are used to generate template code for programmers. [5]

SOAP and WSDL are complex and highly-verbose formats, therefore their performances are naturally worse, than XML-RPC. Furthermore, if correctly implemented all XML-RPC libraries are compatible the same cannot be said about SOAP. The protocol has extensions which are not all implemented in all libraries. These properties make it somewhat unsuitable for our cross platform testing and was therefore omitted from the tests.

3.2 JSON-RPC

JSON (JavaScript Object Notation) is another marshaling format. JSON is based on JavaScript and does not need to be generated since it is human readable and writable, and it contains less redundancies. It was introduced as the "fat-free alternative to XML" as it has much less markup overhead compared to XML. This is just a messaging format and JSON do not offer RPC libraries and support for stub operations.

JSON-RPC is very similar to XML-RPC but encoded in JSON instead of XML. As XML-RPC was available before JSON-RPC this RPC has enjoyed less uptake. While JSON has less markup overhead the format is still textual and the savings are not large. This was also evident as for the example none of the available Ruby libraries had documentation. [5]

3.3 Google RPC and Google's Protocol Buffers

gRPC (Google RPC) is a cross-platform, language and platform independent, general-purpose infrastructure used by Google Inc. and they made it public in 2015. It can automatically generate idiomatic client and server stubs for service in a variety of languages and platforms. It uses Protocol Buffers that is a flexible, efficient, automated mechanism for binary serialization of structured data. [8]

Prompted by this newly released RPC solution, with this paper we aim to compare its use and performance to other popular solutions that predate it.

Users need to define how they want their data to be structured once in Protocol Buffers language (proto3) and the signature of the methods that will be called remotely. Then they can use a generated source code to easily write and read their structured data to and from a variety of data streams and using a variety of languages. Figure 2. shows the relevant sections of the proto file used for the performance test.

```
service Database {
  rpc Request(InfoRequest) returns (Info) {}
}
message Info {
  int32 id = 1;
  string first_name = 2;
  string last_name = 3;
  int32 age = 4;
  string email = 5;
  string phone = 6;
  bool newsletter = 7;
  float latitude = 8;
  float longitude = 9;
  bytes photo = 10;
}
message InfoList {
  repeated Info infos = 1;
}

message InfoRequest {
  int32 id = 1;
  bool photo = 2;
}
```

Figure 2: Services and messages defined in Protocol Buffers.

The defined data structure is stored in .proto files. Each protocol buffer message is a small logical record of information, containing a series of name-value pairs. Once the user defined their messages, they run the protocol buffer compiler for their application's language on their .proto file to generate data access classes. These provide simple accessors for each field as well as methods to serialize/parse the whole structure to/from raw bytes – so, for instance, if the chosen language is C++, running the compiler on the user's .proto file will generate a class. User can then use this class in his application to populate, serialize, and retrieve the class protocol buffer messages. The compiler also provides the stub implementations that can be inherited to code the remote function definition.

The protocol buffer message encoded in binary format is much smaller than its XML code but is not human-readable and human-editable. Protocol buffers result not only binary format but are 3 to 10 times smaller and 20 to 100 times faster than XML for serializing structured data that may one of the reasons for the higher performance of gRPC. [9]

4 The performance test of the implemented RPCs

Based on the structure of RPC the performance differences of the different RPC solutions must come from the differently implemented stub operations. The RPC solution that performs stub operations the fastest way and produces the shortest data for sending must have the best performance.

We have performed benchmarks to test the performance of each of these RPC methods and compare them against each other. (See the signature of the methods in Figure 2.)

With these benchmarks the aim was to measure the processing overhead of the RPC methods and their implementations.

For the request method we have written server and client implementations in C++, Java, and Ruby. The server part reads sample data that has multiple data formats, including strings, integers, floats, and 1MB of binary data. After the data has been read it starts listening for connections from the client. The client can only send one request to the server, which is requesting one of the data items with an option to specify whether to include the binary data part or not. The request method in the client program was invoked 100 times, the client program was run 10 times.

The data on the server component was serialized from memory where it was loaded previously, which was not part of the measurement. The client component did no processing on the data apart from printing receipt of request with the identifier from the current item to

standard output. This was to prevent potential elision of deserialization.

The RPC methods would usually be part of a system that further processes data in either a synchronous or asynchronous manner that would have different performance and latency implications. Asynchronous or non blocking systems are usually preferred for more optimal resource usage on both client and server side. With non blocking operations the components would send further requests needed to fulfill their answer, but they would not wait for the answer actively while holding up resources. Instead these systems store that a request is pending, suspend execution of the routine, and continue to do other outstanding operations that they have the necessary data for. When the answer arrives from the server, they load the previously stored request and execution state and continue from the point where execution was suspended.

Our implementation of the server and client do not follow this asynchronous model of operation, but instead blocks until the response arrives from the server. The reason for this is to have more reliable and stable measurements. As we focus on the RPC itself, the server and client components do minimal processing, there are no further requests to wait for. Using an asynchronous model would mean more outside effects on the measurements, as asynchronous signaling is less predictable than synchronous blocking operations.

For gRPC the gRPC and Protobuffers library were used, for XML RPC and JSON RPC the most popular library was selected for each language. These are: for XML-RPC in C++ xmlrpc-c[10], in Java Apache XMLRPC [11], in Ruby the standard library XMLRPC [12] for JSON-RPC in C++ jsonrpccpp[13], in Java JSON-RPC 2.0 by [d]zhuvinov [s]oftware [14], in Ruby jsonrpc2.0 with webrick [15]. The only restriction was that it needed to be able to start listening for connections without a large framework that it would be deployed part of. This means that for example Servlet based Java implementations were excluded.

Docker containers were created for each of these server and client implementations so that they had a runtime environment that is not dependant on the host system. This caused some overhead when starting the client programs, as a new Docker instance had to created for each run, but we found that this did not influence our overall conclusion.

We used a Linux rack to run the server instances and a commodity laptop to run the client instances to simulate somewhat real conditions and connected both of them to the subnet with 125 MBit/s wired connections to exclude the interference in WiFi or otherwise long distance internet connection.

With the RPC method we cross tested all of the languages with each other to get more measurements and lessen the influence of particular implementations on the overall results [16].

It also has to be noted while XML-RPC implementations were easy to find, JSON-RPC is not as widespread judging from the available libraries. The only server library available for Ruby had some issues and no documentation. Table 1 and Table 2 show the results.

		server		
	grpc small	**cxx**	**java**	**ruby**
client	**cxx**	1.446586288	1.538543454	1.843524988
	java	2.385020082	2.574738704	2.862809575
	ruby	2.335048487	2.357542191	2.329401348
	xmlrpc small	**cxx**	**java**	**ruby**
client	**cxx**	1.745901795	1.789834515	6.283093411
	java	1.872503282	1.932996376	6.336315439
	ruby	2.531053102	2.39889046	6.991524778
	jsonrpc small	**cxx**	**java**	**ruby**
client	**cxx**	1.746023073	5.997436199	6.166025146
	java	1.857895238	6.242739725	6.315714135
	ruby	2.245318859	6.360743144	6.627618996

Table 1: The measured average values in seconds after 100 invokes and 5 runs with small test data.

The overall results we have found in our test runs it that overall gRPC performed the best of all three, with XML RPC and JSON RPC having similar performance characteristics with the differences between mainly attributable to implementation details of the libraries. (Table 1)

With small test data, without the 1MB binary, we found that while the different methods had similar performance, in most cases the gRPC was slightly faster except, for example, in the java server java client case where the gRPC implementation did 2.5s while the XML RPC finished in under 2s. In XML-RPC, the Ruby server implementation almost tripled the amount of time required to run the tests regardless of client language. The same can be observed in JSON RPC with the Java and Ruby server implementation. With small test data, C++ implementations were faster than the Java or Ruby ones.

The languages, in which the stub operations are implemented also influences the performance. All RPC solutions performed better in C++ with small test data.

		server		
	grpc big	**cxx**	**java**	**ruby**
client	**cxx**	10.49837347	19.78831593	10.61724809
	java	11.50234759	11.65334163	20.65961758
	ruby	11.44378246	15.36368512	16.35586611
	xmlrpc big	**cxx**	**java**	**ruby**
client	**cxx**	16.87760948	23.02763573	17.56123346
	java	16.70705216	22.97051365	16.33389231
	ruby	31.4515749	36.40461197	26.93085479
	jsonrpc big	**cxx**	**java**	**ruby**
client	**cxx**	23.85746603	23.59290045	23.09974722
	java	18.68594349	21.19607064	18.29131204
	ruby	17.81739152	17.89413377	17.48205703

Table 2: The measured average values in seconds after 100 invokes and 5 runs with binary data.

With the inclusion of the binary data the differences were more pronounced (see Table 2). gRPC performed better except one case. How much faster it was depended on the language combination used. Only the Ruby server with the Java client did beat the time of the gRPC solution. The XML-RPC Ruby client was generally slower than other clients, taking almost twice the time to complete the test runs.

The increased performance of gRPC can be attributed to the transmission format. Both XML and JSON are textual formats. While binary versions exist, these are not as widely used and the RPC libraries do not use them. Because of their text nature to include binary data in them it needs to encoded to some representation that only uses printable ASCII characters, in most cases to Base64. This increases the data to be transmitted by 4/3 and the overhead of the markup structure is also not insignificant. gRPC uses Protobuffers as its wire format, which is a binary format. Binary data can be included as is, no conversion necessary. It also does not add much overhead to the structure, only field identifiers are added for backward compatibility.

5 Conclusions

In this paper, the structure of third generation RPCs was analysed to find answers for the differences in the performance of different RPC solutions: Google RPC, XML-RPC and JSON-RPC. The chosen libraries implemented the stub operations in different ways and used different formats for marshalling. gRPC with Protocol Buffers performed best in our tests because of the fast binary serialization method of structured data, that resulted in smaller sized encoded messages. Our tests proved, that the chosen computer language has an influence on the performance of RPC invocations. gRPC proved faster in C++ implementations than in Java or Ruby with small test data. In case of XML-RPC and JSON-RPC, Ruby server with Java client proved to be the fastest with large test data.

References

[1] Andrew D. Birrell and Bruce Jay Nelson (1984). Implementing Remote Procedure Calls. *ACM Transactions on Computer Systems*, Vol. 2, No. 1, February 1984, Pages 39-59. https://doi.org/10.1145/2080.357392

[2] Andrew S. Tanenbaum, Robbert van Renesse (1988). *A Critique of the Remote Procedure Call Paradigm.* Available at http://www.cs.vu.nl/~ast/Publications/Papers/euteco-1988.pdf

[3] Andrew S. Tanenbaum, Maarten van Steen (2016). *Distributed Systems: Principles and Paradigms*. Pearson Education Inc. ISBN:978-15-302817-5-6
Andrew D. Birrell (1985). Secure Communication Using Remote Procedure Calls. *ACM Transactions on Computer Systems*, Vol. 3, No. 1, February 1985, Pages 1-14. https://doi.org/10.1145/214451.214452
[4] Paul Krzyzanowski (2012). *Remote Procedure Calls* Available at https://www.cs.rutgers.edu/~pxk/417/notes/08-rpc.html
[5] Michael D. Schroeder and Michael Burrows1 (2006). *Performance of Firefly RPC*. http://web.mit.edu/6.826/www/notes/HO11.pdf
[6] Hakan Bagci and Ahmet Kara (2016). A Lightweight and High Performance Remote Procedure Call Framework for Cross Platform Communication. *ICSOFT-EA 2016 Abstracts*. Available at: http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=Rqt07DUDIy8=&t=. https://doi.org/10.5220/0005931201170124
[7] *What is gRPC?* Available at http://www.grpc.io/docs/guides/
[8] *Protocol Buffers.* Available at https://developers.google.com/protocol-buffers/docs/overview#whynotxml
[9] *XML-RPC for C and C++*. Available at http://xmlrpc-c.sourceforge.net/
[10] Apache XML-RPC. Avilable at https://ws.apache.org/xmlrpc/
[11] *XML-RPC for Ruby*. Available at https://github.com/ruby/xmlrpc
[12] *JSON-RPC 2.0. Essential Java libraries and tools for JSON-RPC 2.0development.* Available at http://software.dzhuvinov.com/json-rpc-2.0.html
[13] *JSON-RPC 2.0*. Available at http://software.dzhuvinov.com/json-rpc-2.0-client.html
[14] *JSON-RPC 2.0. for Ruby*. Available at https://github.com/chriskite/jimson
[15] *Downloadable programs and the environment for the benchmarks*: Available at https://github.com/ksanyi007/rpc

https://doi.org/10.31449/inf.v42i4.1282 Informatica **42** (2018) 563–576 **563**

A Category-theoretic Approach to Organization-based Modeling of Multi Agent Systems on the Basis of Collective Phenomena and Organizations in Human Societies

Abderrahim Siam
ICOSI Lab University, Abbes Laghrour khenchela BP 1252 El Houria 40004 Khenchela, Algeria
E-mail: siamabderrahim@gmail.com

Ramdane Maamri
Lab Lire University Mahri Abdelhamid Contantina, Algeria
E-mail: rmaamri @yahoo.fr

Keywords: multi-agent systems, organization, category theory

Received: June 30, 2016

This paper presents an idea of using category theory for developing organizational multi-agent systems by taking inspiration from collective phenomena and organizations in human societies. Category theory is used for studying and formalizing organizations and collective phenomena in human societies with the aim of capturing their logics into categorical models. Afterward, the captured models are mapped categorically to categorical MAS organizational models. This way of thinking allows studying properties of result MAS organizational models as well as properties of organizations in human societies such as stability and adaptation before taking them as landmarks for developing MAS organizational models.

Povzetek: Predstavljeno je modeliranje multiagentnih sistemov na osnovi kolektivnega delovanja človeških združb.

1 Introduction

Several characteristics have emerged as essential in nowadays computer applications seeing that the widespread of software use in the various fields and the pervasiveness of information processing tools in all equipments around us and which are embedded with more and more means of communication. Applications must be increasingly distributed, open, adaptable and robust. Moreover, a great complexity characterizes software and their development processes.

Multi-agent systems often abbreviated MASs and organizational MASs approaches, in which MASs are analyzed and designed as computational organizations using social concepts, present privileged solutions to develop applications outlined above. This is due to their interesting features such as the proposed abstractions for structuring the software as combinations of entities in interactions; the introducing of concepts with very high levels of abstraction as agents, plans, roles and organizations; and the flexible coupling that MASs offer through indirect interaction modes and the late binding between agents in such a way that the determination of the action to execute and the entity responsible for its execution can be postponed as late as possible.

The most interesting features of agent-based approaches include the possibility to combine agents and MASs with other development paradigms and technologies in order to strengthen agents and MASs with other interesting features. In several works, the agent based development paradigm is combined with the component based development [1]. Although agents and MASs are real technological advances if compared to software components, the component based development has reached a stage of maturity and preserves some assets such as good structuring of applications with variable granularities, reuse of components as well as possibilities of dynamic adaptation of component based applications by adding, removing, substituting components or reassembling and reconfiguring component assemblies. Agents and components are combined in various forms in several works. In [2],[3],[4], and [5] as well as in numerous other works, agents via their capacities of negotiation are used to assist component based development. For example, this assistance takes forms of classification, research and rapid selection of reusable components, matching between components and assisting effective assemblies of components. In another alternative, the maturity of components in structuring and deploying software as well as possibilities of automatic assemblies of components encouraged the use of components as building units to construct agents. This form of combination appears is several works such as [6],[7] and [8]. Other forms of combination of agents and components may be found in the literature. Among the works where the components are used to build agents, we are particularly interested in the works of Siam et al [8] where software components are

used to build self-adaptable agents that form an organization in which agents are grouped into coalition groups.

The consideration of a social or organizational standpoint as framework for analysis and design presents an important key to develop MASs, notably open and adaptive ones. The organization is concerned with defining, managing and changing the relationships between agents. It corresponds to the concept linking the description of a society of agents and that of control and coordination activities. The concept of organization was the subject of numerous works giving birth to an even greater number of MAS organizational models. These works are often inspired from sociology, social psychology, economics, ethology or cognitive science. However, inspirations processes from these disciplines were not driven with formal approaches that allow for reasoning about inspired solutions as well as inspiration processes.

In this paper, we are interested in looking for solutions for the organization of MASs by taking inspiration from collective phenomena and organizations in human societies. First, we use the category theory CT concepts [9][10] to capture the logics of some collective phenomena and organizations in human societies. After that, the captured logics serve as the basis for modeling categorically adaptive organizations of MASs in which agents are built based on software components. Category theory presents a sophisticated mathematical toolbox. It provides instruments that facilitate the modeling of complex situations and which involves structured objects. Concepts in category theory are typically formalized in terms of the relationships that each object exhibits to the other objects in the universe of discourse. Particularly, agents based methods as well as component based development methods; typically model the universe as a society of interacting elements. Also, collective phenomena in human societies may be described via interactions between individuals. Category theory proves a great power of expression. This power is the result of the composition of very natural and comprehensible constructions. Although category theory is not based on complex concepts, it is extremely powerful and it derives its power from the idea of composition. Category definition itself contains composition. In addition, it is possible to build categories of categories. Categories are to be composed in order to generate more abstract structures. With the ideas of compositional thinking and the possibilities of diagrammatical representation, categories can be very useful in addressing a large range of problems and consequently that addressed in this paper.

We present an approach for modeling organizations of MASs and verifying their proprieties. In this approach, we propose constructing models of organizations of MASs using category theory on the basis of categorical models of social processes and organizations in human societies. The constructed models permit checking the organization proprieties such as stability. The stability of an organization of a multi agent system includes three important aspects. The first one concerns the degree to which social structures, established by the organization, resist disturbance. The second aspect concerns the convergence of a MAS to a valid organization when organizing, i.e., when a MAS decides to change the current organization for different reasons, does the process of organizing converges always to a valid organization? The third aspect of organization stability depends on the frequency of changing organization; if the changes are less frequent, the organization is more stable.

The paper is structured as follows: the second section presents important concepts related to multi-agent systems and organizations; the third section is devoted to present collective phenomena and organizations in human societies; the fourth one provides a view on the category theory and its important concepts; in the fifth section, we show how category theory can assist in the articulation and the modeling of collective phenomena and organizations in human societies primarily and how the logics of collective phenomena and organizations in human societies captured categorically serve as basis for modeling organizations of MAS; in the last section, we conclude the paper and define directions for future work.

2 Multi-agent systems and organization

Multi-agent systems present an important and very useful development paradigm through an armada of tools and methodologies for constructing complex, open, distributed systems that can autonomously adapt to their environments. The goal of research in multi-agent systems is to figure out how a society of autonomous entities called agents can organize themselves in order to solve problems and produce global phenomena that each agent cannot solve or/and produce individually. MASs use social metaphor of the so-called social insects when agents are reactive or of human organizations when agents are cognitive.

The organization of MASs is concerned with defining, managing and the change of relationship between agents. It presents a powerful technique to constrain autonomous agents to behave in such a way to meet overall goals. It corresponds to a concept linking the description of a society of agents and that of the control and coordination activities. The definition of the organization of MASs is not a subject of consensus. However, two significations may present extremities of the interval of organization definitions. In [11], an organization is defined as a collective entity with its own identity which is represented by a group of agents exhibiting highly formalized social structures. In [12], an organization is a scheme or a stable structure of joint activity that can constrain or affect the actions and interactions of agents for a specific goal. The MAS design guided by the organization makes later implementation questions and introduces concepts of very high level of abstraction such as roles and groups. MASs designed according to an organizational view may be equipped with mechanisms of reorganization or self-organization allow-

ing them to change their organizations by passing from one organization to another one in order to respond to environments' dynamics. When the organization is treated according to a self-organizational point of view, the organization is seen as a process where the degree of order increases with time, from less to more organized.

Several organizational models of MAS have been developed. Many of these models are inspired from organizations in human societies. A survey of such models can be found in [13].

The organization of multi-agent systems may be seen from an angle of emergent phenomena in complex systems [14]. Indeed, the design of such systems focuses mainly on agents that compose the system. Accordingly, agents are the engines of the organization; organizations exist only as emergent and observable phenomena. According to this agent-centered view [15], emphasis is put on the agents by specifying and designing local behaviors and peer interactions. The global functionality of the system is the result of complex and dynamic interactions in the agent's society. However, such design approaches often introduce unpredictable phenomena which cannot be checked in advance. The overall behavior is more than a simple juxtaposition of agents' behaviors. Several studies on emergent phenomena in multi agent systems and multi agent simulations such as [16],[17] and [18] may be found in literature.

In the opposite direction of the agent-centered point of view, organizations exist as explicit entities of system. Thus, the observer of the system can obtain a description of the organization. A pattern of cooperation is determined by the designer to define or constrain the behavior of agents. Works that are included into this direction may be categorized into two classes: approaches where the organization is specified and formalized but agents do not know about it and cannot reason on this subject. Examples of such approaches are agent-oriented methods focusing on organizational models as in INGENIAS [19] and MaSE [20]. The second class includes approaches where organizations are manipulated by both designer and agents. Agents have a representation of the organization and can perform organizational acts and potentially change the organization. AGR [21], ISLENDER [22] and the model proposed in [8] are examples of this class.

It is remarkable that most of existing organizations of MAS are focused on the concept of role. A role presents an important organizational concept. It permits an organizational structuring of MASs. The concept of role has various meanings depending on the MAS organizational model and the context of its use. We will see later that the concept of role is an important concept also for organizations in human societies.

3 Human societies and organizations, landmarks and inspirations

Collective phenomena and organizations in human societies present a great source of inspiration for artificial organizations and especially for MASs. Human societies have always produced collective behaviors patterns, allowing them to overcome the most of encountered problems and difficulties. Capturing logics of collective phenomena and organizations in human societies for the sake of reproducing them requires thorough observation and in-depth studies of human societies. Nevertheless, these observations and studies are the core of sociology. Sociological studies and the studies produced by psychologists, economists and AI searchers offer tremendous beneficial potential for exploration and exploitation to bring out landmarks susceptible to serve as basis for developing organization based multi-agent systems. We can distinguish two main elements in human societies, whose study would allow the discovery of several models that could be operationalized and reproduced for developing more effective and efficient MASs. The first element is social processes, and the second, organizations.

3.1 Social processes

A social process may be defined as the observable and repetitive patterns of social interaction that have a consistent direction or quality [23]. It may be seen as a way in which individuals and groups interact and establish relationships and models of behaviors. These relationships and models are permanently adjusted and modified through social interactions. This concept refers to some of the general and persistent forms that social interaction may assume, i.e., forms of social interactions or mutual activities that occur repeatedly [23].

We use a large-scale process, drawn from classical social theory to illustrate how category theory can assist, at first, in the articulation and the capturing of logics of complex social phenomena. In a second time, captured logics formalized in terms of categorical concepts serve as metaphors and starting points for categorical MASs modeling. The used process which is presented below is the circulation of elites [24]. In sociological literature, several social processes may be found along with various works aimed at using formal models to describe them. For examples, in [25] and [26] as well as in several other works, investigations have explored the possible applicability of mathematical chaos theory [27] to the social sciences. In [28], Attractor models have been used in the formulations of some social processes.

In the circulation of elite process, sociologist and economist Vilfredo Pareto divided social classes into mass class and elite class [24]. According to him, the elite distinguishes itself by its eminent qualities, its natural and

psychological superiority. It is made up of all those who exhibit exceptional qualities and eminent aptitudes within their sphere of activity. It contains both governmental and non-governmental elites. Elite individuals perpetually rise from the mass; the ruling elite have the choice to fight emerging non ruling elite or integrate it, until the ruling one is finally defeated or replaced. Pareto observes elite renewal and replacement, or, on the contrary, its closure. The formation of counter-elite at the moment of the appearance of degeneration symptoms of the ruling class appeared to him as the dynamic constituent of history.

This social process may be abstracted as the interaction of three elements. The first one is the gradual degradation of the ruling elite; the second consists of the inherent willingness of the most talented members of the non-elite to upgrade; the third element is the degree of facility for rising to a higher social position permitted by prevailing social structure imposed by the ruling elite. Facility for rising to a higher social or economic position is known in sociology as "upward mobility". Depending on whether this mobility is relatively open or closed, the elite is continually reconstituted by incorporating new talented members and eliminating unworthy ones; or, talented and aggressive non-ruling elites accumulate within the non-ruling elite causing instability.

3.2 Organization

An organization is a social entity in which the members thereof perform predefined functions. In human societies, the organization allows the coordination of complex behaviors. The changes in human societies, as the emergence of large organizations such as the army and the administration instead of artisan corporations, large families and other forms of organizations, demonstrates that human societies change its trends of thinking about social life. The replacement of the behaviors based on common values or social norms by behaviors directed by a more systematic search for efficiency is an instrumental logic which subordinates any action to objectives, purposes. Organizational theory [29] tries to query this instrumental rationality and its concrete expression in the organization in several directions such as: analysis of rationality and decision making, analysis of relationships of dependence and power and analysis of the dynamics of change.

The study of organizations in the human societies enables for the identification of several different organizations which can be classified and cataloged according to various points of view and criteria. For examples, the following kinds of organization may be identified in human societies: the single person organization which is the simplest organization, bureaucracy structure organizations [29], matrix structure organizations [29] and team structure organizations [29]. The organization refers at the same time to the process of organizing and the result of this process. The study of an organization (existing or a new one to define) consists in: analyzing how it works, identifying its main features, accentuating the pertinent characteristics of the members, pointing out the roles adopted by members and the member relationships, identifying norms and rules that oversee the organization as well as the chain of command, specifying how tasks are subdivided into independent and formalized subtasks, identifying which of subtasks have dependencies, defining how and by what means tasks may be grouped and pointing out where/when decisions have to be made.

We can notice that the issues discussed above present the same concerns to be addressed by a designer in the analysis and the design of an organization based multi-agent system added to the defining of environment proprieties to be considered such as openness and dynamicity as well as constraints related to the operationalization of organization computing.

In this paper, we present team and group structure as examples of organizations for the purpose of illustrating how category theory can help in capturing of functioning logics of these organizations and which are subsequently mapped to organization based multi-MASs. Group and team are not the same. We use the term crew to cover both groups and teams. A crew structure is an organic organization. Contrary to mechanical organizations where tasks are precisely defined and broken down into separated and specialized parts, in organic organizations, tasks are adjusted and redefined by means of collaborative work in crews. In a crew based organization, an organization consists of a set of crews. Crews represent a system with several actors that share the common goal of accomplishing the crew's global task. The global task of a crew is fractionalized into subtasks which are assigned to members of the crew vis-à-vis members' degrees of qualifications.

3.2.1 Group structure organizations

In human societies, a group may be defined as two or more individuals, interacting and interdependent, who have come together to achieve particular objectives [29]. Groups may be formal or informal. A formal group is defined by the organization's structure, with designated work assignments establishing tasks. In formal groups, the behaviors group members should engage in are stipulated by and directed toward organizational goals [29]. An informal group is a group that is neither formally structured nor organizationally determined; such a group appears in response to the need for social contact [29].

One of the key concepts behind the concept of group in human societies is what is called social identity. It is perspective that considers when and why individuals consider themselves members of groups [29]. It helps people reduce uncertainty about who they are and what they should do [29]. Several characteristics make a social identity important to a person. Mainly, similarity and distinctiveness; people who have the same values or characteristics as other members of their organization have higher levels of group identification; people are more likely to no-

tice identities that show how they are different from other groups [29]. According to [29], Groups generally pass through a sequence of five stages in their evolution. The first stage (forming stage) is characterized by uncertainty about the group's purpose, structure, and leadership. Members try to determine what types of behaviors are acceptable. This stage is completed when members have begun to think of themselves as part of a group. In the second stage called storming stage , members accept the existence of the group but resist the constraints it imposes on individuality. There is conflict over who will control the group. In the third stage, the group demonstrates cohesiveness with a strong sense of group identity. This stage is called norming stage, it is complete when the group structure solidifies and the group has assimilated a common set of expectations of what defines correct member behavior. For temporary groups, the stage four (performing stage) is for wrapping up activities and preparing to disband. Some group members are upbeat, basking in the group's accomplishments. In the last stage called adjourning stage, the structure is fully functional. Group energy has moved from getting to know and understand each other to performing the task at hand. For permanent work groups, this stage is the last one in development. Groups have properties that format members' behavior and allow the explaining of individual behavior within the group besides the performance of the group. These proprieties include [29] roles, norms, status, size, cohesiveness, and diversity.

According to [29], a role may be defined as a set of particular behavior patterns attributed to an actor occupying a given position in a social unit. Norms are acceptable standards of behavior shared by members of a group. Norms express what members ought and ought not to do under certain circumstances. When become a member of a group7, its norms influence members' behavior with a minimum of external controls. Members are susceptible to conforming to the group's norms. This conformity means the adjustment of one's behavior to align with the norms of the group. Different groups, communities, and societies have different norms, but they all have them. Status is a socially defined position given to groups or group members by others. Status can help to differentiate members in group and may be an important motivator when individuals perceive a disparity between what they believe their status is and what others perceive it to be. Status tends to derive from one of three sources: the power a person wields over others; the person's ability to contribute to a group's goals and the individual's personal characteristics. Each group is characterized by a size. It has influence on its overall behavior in the sense that individuals perform better in smaller groups which are faster at completing tasks than larger ones. Conversely, large groups consistently get better results than their smaller ones. It seems that individuals have tendency to expend less effort when working collectively than alone. The Cohesiveness of a group expresses the degree to which members are attracted to each other and motivated to stay in this one. Cohesiveness has influence on group's productivity. Finally, the diversity in a group expresses the degree to which its members are similar to, or different from, one another. This propriety appears to increase group conflict. One of the most important points when studying a group is the group decision making. Group decisions compared to those made by an individual alone have advantages in the sense that groups generate more complete information and knowledge; groups take more input and heterogeneity into the decision process as well as they offer a wider range of points of view. This allows obtaining more approaches and alternatives. However, Group decisions are time consuming; discussions in the group can be dominated by a few members; and group decisions suffer from ambiguous responsibility.

3.2.2 Team structure organizations

Work teams are different from work groups. In groups, members interact primarily to share information and make decisions to help each member perform within his area of responsibility [29]. The group performance is simply the sum of contributions of the group members. The group is not able to reach the advanced level of collaboration and cooperation that would create an overall level of performance superior than the sum of the inputs. Work teams are characterized by positive synergy that means cooperation of two or more substances, or to produce a combined effect greater than the sum of their separate effects. When an organization is oriented work teams rather than groups, it creates the potential for to generate more outputs without increasing inputs. In teams, skills are complementary, unlike groups where skills are random and varied. Teams are crews as well as groups; according to [29], effective teams have characteristics related to their key components organized into three general categories summarized in Table 1.

The most common types of teams in an organization are: problem-solving teams, self managed work teams, cross-functional teams, and virtual teams [29]. In problem-solving teams, members discuss ways of improving quality, efficiency, and the work environment. Generally, problem-solving teams have not authority to implement any of their suggestions, i.e., problem-solving teams make recommendations only. Self managed work teams make recommendations, propose solutions and implement them. Members of such teams perform: planning and scheduling work, assigning tasks to members, making operating decisions, taking action on problems, and working with suppliers and customers. Cross-functional teams are constituted of employees from the same hierarchical level but different work areas, who are regrouped together to accomplish a task. Virtual teams are teams that use computer technology to link together physically detached members in consideration of achieving a common goal.

Context.	Composition.	Process.
Adequate resources	Abilities of members	Common purpose
Leadership	Personality	Specific goals
Structure	Allocating roles	Team efficacy
Climate of trust	Diversity	Conflict levels
Performance	Size of teams	Social loafing
Evaluation	Member flexibility	
reward systems	Member preferences	

Table 1: Effective teams characteristics.

4 Category theory, key concepts, applicability and applications

The effective way of dealing with complexity in order to represent and reasoning about required information to build computer systems is via formal methods. Among the most effective methods and which is adequately capable of addressing structures is the category theory [9],[10].

Category theory is a sophisticated mathematical toolbox. It provides instruments that make easier the modeling of complex situations and that involve structured objects. Concepts in category theory are typically formalized in terms of the relationships that each object exhibits to the other objects in the universe of discourse. This way of thinking is totally different from set theory in which concepts are formalized extensionally in the sense that a set is defined by its elements. The focus on social aspect of objects lives is the reason of applying CT to software engineering as well as for synthesizing social theory. Particularly, agents based methods as well as component based development methods; typically model the universe as a society of interacting elements. Likewise, human societies are subject to the same vision.

Our objective is to develop organization based MASs by taking inspiration from social processes and organizations that have proven their success and effectiveness in human societies. For the reasons outlined in the introduction, agents that constitute such MASs are self adaptive component-based agents in accordance with the agent model presented in [8]. MASs development methods, component based development methods and the social theory present three different areas. In these areas, and particularly in the social theory, the perspective of representing collective phenomena and social processes at multiple scales and with any level of details is highly probable to produce complex and interlaced models. The lack of mathematical formalization can make the results of applying these models hard to validate and render them difficult to reproduce. One solution to managing the conceptual and computational complexity of social models is to move them toward a higher level of abstraction.

The CT may be involved to formalize these models and exploring their potential interaction. In one hand, category theory is able to integrate diverse areas and provides a common language which can be applied to deal with the diversity of social theories in the synthesis of social theory, as well as dealing with agents, MAS organizational concepts and software components with the same terms. Category theory not only allows these areas to be expressed in the same categorical terms, it also provides a basis for their possible integration. In the other hand, category theory supplements firm equality with weakened definitions of equivalence relations and/or classes. This propriety makes important contribution in studying areas in which concepts involve proximity assessments which is the case of social analysis as well as organizations in MASs. The potential is to address qualitative concepts with greater precision. Category theory can assist in the articulation of complex social phenomena and organizations in human societies; it helps to capture logics of validated phenomena from the domain of human societies and move them to the domain of organizations of MASs, categorically. The coherence of such categorical transformations makes a direct and important contribution to the process of validating MAS organizational models.

4.1 Some category theory concepts and notations

A category C consists of objects and morphisms with composition, identity, associativity and unit. Objects $a, b, c,$ ...etc are denoted $ob(C)$. Morphisms $f, g, h,$ ...etc denoted $hom(C)$where, for each morphism g there are given objects: $dom(g)$ and $cod(g)$ called the domain and the codomain of g. The notation $g : a \longrightarrow b$ indicates that $a = dom(g)$ and $b = cod(g)$. Given the objects a, b, c and d of C and morphisms $f : a \longrightarrow b$ and $g : b \longrightarrow c$, with $cod(f) = dom(g)$, there is a given morphism: $gof : a \longrightarrow c$, called the composite of $f$and g. For each object a there is a given morphim $ida : a \longrightarrow a$, called the identity morphism of a.

A category C is characterized by the associativity which means that for the morphisms $f : a \longrightarrow b, g : b \longrightarrow c$ and $h : c \longrightarrow d, ho(gof) = (hog)of$. The unit propriety means that $foida = f = idbof$, for all $f : a \longrightarrow b$. A discrete category is a category where the morphisms are only identity morphisms. An initial object in a category C is an object for which every other object of C is a codomain of a unique morphism with the initial object

as a domain. A terminal object has every object of C as the domain of a unique morphism where the terminal object is a codomain. Formally, in a category C an object is called initial object o if for any other object a in C, there is a unique morphism $o \longrightarrow a$. An object is called terminal object t if for any other object a in C, there is unique morphism $a \longrightarrow t$. In the case where objects of a category are categories, morphisms are called functor. A functor $F : C \longrightarrow D$ between categories C and D is a structure preserving mapping of objects to objects and morphisms to morphisms with: $F(f : A \longrightarrow B) = F(f) : F(A) \longrightarrow F(B); F(gof) = F(g)oF(f)$; and $F(idA) = idF(A)$.

This section offers a very brief overview of some category theory concepts. To obtain a broad or complete view on the concepts of this theory, works [9] or [10] can serve as references.

5 Categorical modeling of organizations of MASs on the basis of capturing logics of some collective phenomena in human societies

We use two examples to explain the ideas of looking for organizational solutions by studying collective phenomena and organizations in human societies when modeling a multi agent system for a given problem. For the first example, we assume that it is a question of proposing an organization for a multi-agent system comprising a number of heterogeneous agents. Each agent has abilities and skills that differ from other agents. The best agents in terms of abilities and skills decide the overall strategy of the system. For example, in the case of a multi-agent system for information retrieval [30], the agents best equipped with planning and researcher capabilities decide how or which information retrieval algorithm will be applied. Given that agents are self-adaptive and can acquire skills and qualities as the system progresses in its functioning, the set of agents that decide changes over time. The second example presents a case of MASs where agents must be structured into crews in which agents plays roles while respecting certain norms. For example, a distributed auction system in which three roles are played. The role Auctioneer played by agents that want to sell items; the role Seller played by agents that sell items on behalf of Auctioneers; and the role Bidder played by agents that make a bid on an item being auctioned [8].

After having explored the collective phenomena or the organizations which have proved their effectiveness and which seem to have a certain similarity with the problem to be solved, it is necessary to target a phenomenon or an organization to be the basis for the organization of the multi-agent system to be modeled. Once a phenomenon or an organization is targeted, we proceed to its categorical modeling. Such a modeling allows the study of the properties considered as important such as the stability presented in the introduction. A categorical modeling is to construct categories for which it is possible to find functors allowing the construction of categories that model a MAS. These functors are schematized in the figure 1 by the different connections between the concepts of the category theory and social concepts on the one hand, and the concepts of the domains agent, MAS and the component based development on the other. This path of reflection can be summarized as follows:

- Analysis of the problem for which we are seeking an organizational solution;
- Explore collective phenomena or the organizations which have proved their effectiveness and which seem to have a certain similarity with the problem;
- Modeling with categories the selected phenomenon or organization;
- The study of the properties of the categorical model (If the properties of the chosen model are not satisfactory, it is necessary to return to step 2);
- Use the correspondences schematized in the figure 1 between categorical and social concepts and agent and MASs concepts to obtain a categorical organizational model of MAS;
- The study of the properties of the categorical MAS model (If the properties of the chosen model are not satisfactory, it is necessary to return to step 5 or 2);
- Refine the categorical model of the organization of the MAS by introducing more details where objects of a category are modeled as new categories. For example an agent that is an object in a MAS category can be modeled as a category whose objects are software components;
- Make implementation choices.

When modeling a social process or a collective phenomenon in human societies categorically, it is possible to take social object as objects of categories, then, According to the behavior to be modeled, adapted sorts of morphisms can be identified. Social objects include several kinds of objects. Among these objects, social actors and Social aggregates are identified [31]. A Social actor may be defined as an active social entity that is considered at any scale, such as, individuals, Alliances, groups, teams... etc. Social actors make decisions and try actively to arouse coordination with internal and external actors. Social aggregates are passive collections of social actors such as populations, generation and aggregations. A social aggregate is determined by exogenous structures, events and process. It is changed through proximate adaptation and evolution.

After modeling a social process or a collective phenomenon, we use the same set of social mappings to

metaphorize the process of organizing multi agent systems. This categorical way of thinking is immensely important for how we understand organizations both in human societies and multi agent systems. Thereby, a MAS can be seen as a category which objects are agents and morphisms are interactions between agents. We can also build a category of MASs which objects are MASs and morphisms present reorganization processes. An agent can be seen as a category also. Its objects are different components that constitute the agent and morphisms are interactions between components or compositions. A functor between two categories agent A and agent B may present an adaptation of the agent A to become similar to B, as well as it may present all sorts of interactions between the agents A and B. These same principles are valid for components development concepts as depicted in figure 1 that illustrates mappings from Categories to social concepts, MAS and agents' concepts and components development concepts. For example, component based architecture can be seen as a category in which components present objects and interactions present morphisms. Concerning architectural styles, the choice of a specific category can be seen to reflect the choice of a specific architectural style.

It is clear that there is a sort of universality in the way that social theory, MAS, organizations, and component based development can be approached through categories. This universality supports the integration of different approaches.

5.1 Organizing a MAS on the basis of the circulation of elite process

Assuming that we want to propose a multi-agent organization-oriented solution for the problem presented earlier as a first example. In this class of problems, a subset of agents that show good capacities decide about the global strategy of the system. For example, in the case of a multi-agent system for information retrieval [30], the decision to choose the information retrieval strategy is taken by the most qualified agents. An organizational solution for this class of problems can be inspired from the process of circulation of elites. According to the process of circulation of elites presented above, a society may be represented by the set of all individuals noted Ω. At the time i, the society is given as the triplet (A_i, B_i, C_i) which is a partition of the set Ω where: A_i is the set of the mass individuals at the time i; B_i is the set of the non ruling elite individuals at the time i; C_i the set of the ruling elite individuals at the time i. Thus, $\Omega = A_i \cup B_i \cup C_i$.

At times $i(i = 1, 2, \ldots, n)$, a society may be described with the stats $M_1(A_1, B_1, C_1), M_2(A_2, B_2, C_2), ..., M_n(A_n, B_n, C_n)$. The transition of the society from a state M_j to another one M_k means that: (i) an element from the non ruling elite is integrated in the ruling elite; (ii) an element form the ruling elite is excluded from this one; or (iii) an element from the mass is integrated in the elite.

If we consider two successive states M_i and M_{i+1}, D_i is the transformation passing the society from M_i to M_{i+1}. For two non-successive stats M_j and M_k, there is a sequence of transformations $D : M_j \longrightarrow M_k$. If $H$is a transformation form M_k to M_{k+1}, the composition $H \circ D : M_j \longrightarrow M_{k+1}$, changes the society from the stat M_j to M_{k+1}. We are in front of a category of social transformations.

Each stat M_i may be described categorically. Let be the set E defined as: $E = \{(x, X_i)/x \in X_i, X_i = A_i \vee B_i \vee C_i\}$.

Let be the discrete category $CatE$ which objects are the elements of E and morphisms are identity morphisms. We define categories $CatE_m$ by recurrence. The objects of a category $CatE_n$ are the elements of the set E_n defined below. Morphisms of $CatE_n$ are identities.

$$E_n = \{(x, X_i^n)/x \in X_i^n, X_i^n = A_n \vee B_n \vee C_n\}.$$

With

$$A_0 = A, B_0 = B, C_0 = C$$

and

$$A_n = A_{(n-1)} - D_{(n-1)};$$

$$B_n = (B_{(n-1)} - F_{(n-1)}) \cup G_{(n-1)};$$

$$C_n = C_{(n-1)} - G_{(n-1)}$$

Where sets D_i, F_i and G_i are defined as follows:

$$D_i = \{x \in A_i \wedge T1(x) = Cte_1\}$$

$$F_i = \{x \in B_i \wedge T2(x) = Cte_2\}$$

$$G_i = \{x \in C_i \wedge T3(x) = Cte_3\}$$

$T1$, $T2$ and $T3$ are functions expressing different situations allowing an individual x from Ω to pass from one class to another at the time i as depicted in figure 2. The arrows to the symbol Φ express the disappearance of an individual.

We define now a category CAT of the categories $CatE_m$ which objects are categories $CatE_1$, $CatE_2$,..., $CatE_m$ and morphisms (functors) are inclusion maps (canonical injections) $\Im$ defined as:

$$\Im : (x, X_i) \longrightarrow (x, X_j), X_i \subseteq X_j$$

The formalization presented above stipulates that the transition of the society to the state n from the state $n-1$ means that: some elements from the mass class have become members of the class of non-governing elite $(A_n = A_(n-1) - D_(n-1))$; some elements from the non ruling elite class have become members of the class of the governing elite and vice-versa $(B_n = (B_(n-1) - F_(n-1)) \cup G_(n-1))$; or some elements from the ruling elite class have become members of the class of the non ruling elite $(C_n = C_(n-1) - G_(n-1))$.

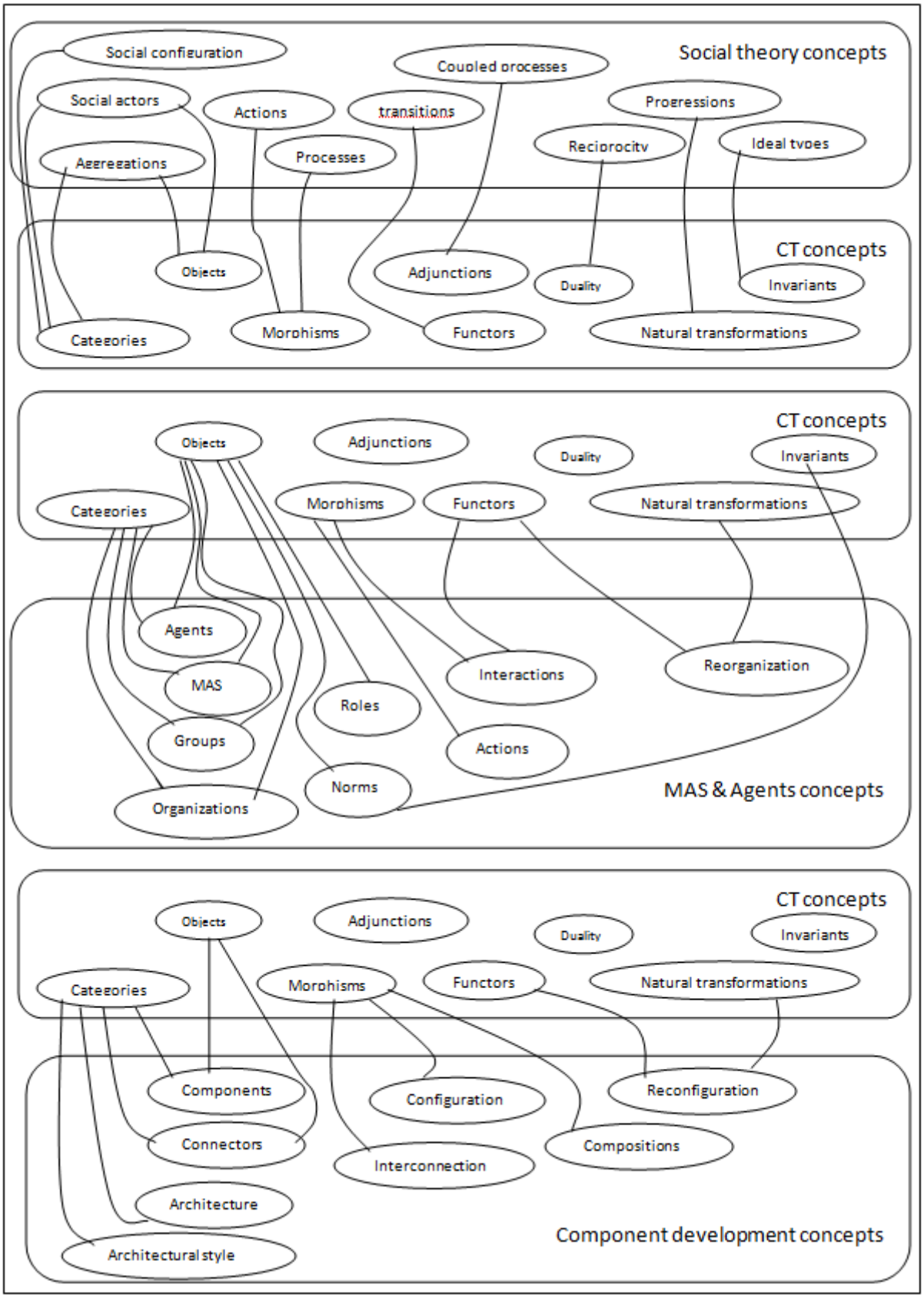


Figure 1: Correspondences between categorical concepts and social, MAS and component based development concepts.

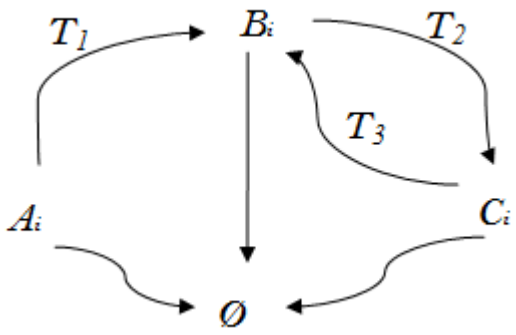


Figure 2: A schematic of passages of individuals between classes according to the circulation of elite process.

Once the categorical model is established, the study of all properties becomes possible. For example, the study of the stability of the circulation of elites process can be achieved through monitoring cardinalities of the sets A_i,B_i and C_i. With the progressive changes in a society, if the cardinality of one of the three sets becomes lower compared to a minimum limit or upper compared to a maximum limit, the instability is the result. At the time i, the number of elements of the population Ω is N_i. Elements of Ω are divided into the three subsets A_i,B_i and C_i which cardinalities are respectively ni_1, ni_2 and ni_3 with $N_i = ni_1 + ni_2 + ni_3$. For example, One of the factors of system stability is the comply with the condition formulated as:

$$\alpha_j \leq \frac{n_i j}{N_i} \leq \delta_j, \forall j = (1..3)$$

The functions $T1$, $T2$ and $T3$ and the values $Cte1$, $Cte2$ and $Cte3$ express conditions allowing an individual x from Ω to pass from one class to another. $T1$, $T2$ and $T3$ expressions may be determined in several ways. For example, for the defining of the functions $T1$, $T2$ and $T3$ that produce the probability that a member of a class moved to another class or disappear at the time i, it is possible to take inspiration from the Maxwell-Boltzmann law used in statistical physics to determine the distribution of particles between different energy levels [32]. At the basis of Maxwell-Boltzmann law and according to the case of one of its variants relative to bosons (Bose-Einstein) or fermions (Fermi-Dirac) taking into account the parameters giving the characteristic of elements. It is possible to correspond the level 0 to the elements that disappear; to the elements that remain in their class the level 1 and those who pass to another class the level 2. The number of elements in a given level takes a similar form to that presented in the previous paragraph by considering as parameters some characteristics of the elements and the characteristic of each pre-defined level.

The circulation of elite process categorically modeled can be the basis of modeling multi-agent systems. This social process may be reproduced in the modeling of a MAS for information retrieval as well as for a great class of similar problems. Agents of such system are partitioned into steering agents and executor ones. Steering agents produce plans and decide about strategies of information retrieval that executor agents execute. The Steering agents are the best agents in terms of planning capacities. The same categorical modeling may be used in such a way steering agents are the ruling elite, executor agents form the mass class. The agents are self adaptive and can acquire new skills. As the system progresses in its execution, agents acquire new skills. The best executor agents in terms of capacities, progress to join the steering agents passing by an intermediary status of a non-ruling elite (non ruling steering agents). The stability of such organization of MAS may be studied in the same way in which the stability of societies is treated i.e. via cardinalities of the groups of ruling steering agents, non ruling elite agents and executor agents.

For such system we use the model of agent presented in [8]. In this model, agents are composed of software components assembled automatically and dynamically by an assembly engine. The different capabilities of an agent are implemented based on software components. The architecture of the agent includes a control module. Among the tasks for which the control module is responsible is the decision to make changes in the internal composition of the agent. These changes consist in perform assemblies or reassemblies of components. By achieving changes in the composition of an agent, it acquires new skills. Agents are to present categorically in order to study their capacities. Each agent may be modeled as a category which objects are software components and morphisms are interactions between components. Agent's capacities may be studied as proprieties determined using the capacities fuzzy measure proposed by Siam et al [8]. The figure 3 offers a general view of such modeling in which the different status of the agents society constitutes a category in which objects are status and morphisms are the transitions of the society from one status to another one. A transition expresses the fact that an agent has changed class. A status which is an object of the category described above is also a category. Its objects are agents and morphims are interactions between agents. Agents in their turns are categories where objects are software components and morphisms are interactions and compositions of components.

5.2 Organization of a MAS based on organizations in human societies

It is possible to describe with category theory concepts, organizations in human societies and their organizational concepts. All concepts presented in the section 3 such as organizations, groups, roles, teams, norms...etc may be represented using categorical concepts. The figure 1 illustrates correspondences between categorical concepts and organizational ones.

To propose an organizational modeling of a MAS allowing the implementation of the distributed auction system

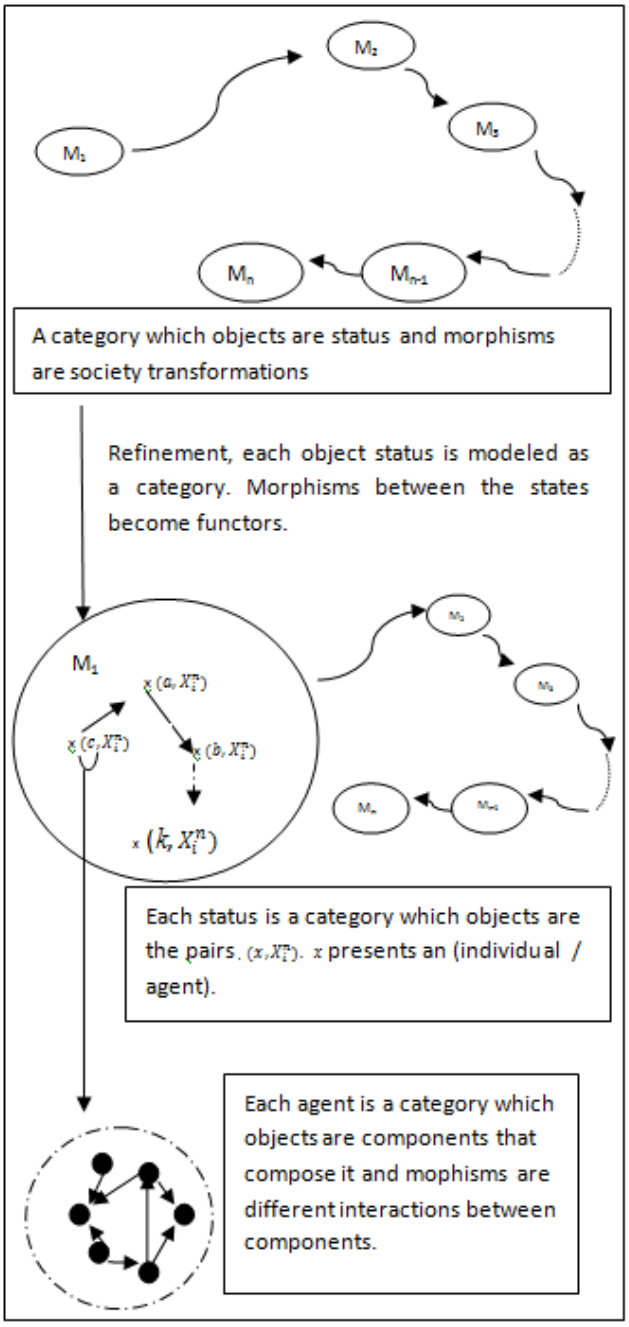


Figure 3: A categorical modeling on the basis of the circulation of elite process.

described above, all organizations in human societies in which a society is structured into groups or teams of members who play roles can serve as a source of inspiration. For example, it is possible to define a category ROL of roles. The objects of this category are triplets (R_i, S_i, P_i) where R_i is a set of roles; S_i are protocols that connect the roles R_i, P_i denote pair of roles connected by each protocol. Morphisms of ROL are relations Ψ between triplets (R_i, S_i, P_i). A relation between two triplets (R_1, S_1, P_1) and (R_2, S_2, P_2) may be defined as a function $f = (\alpha, \beta)$ with $\alpha : R_1 \longrightarrow R_2$ and $\beta : S1 \longrightarrow S2$such that $\forall s \in S1, P(s) = (r1, r2) \Longrightarrow \beta(s) = (\alpha(r1), \alpha(r2))$.

We can verify that the objects (R_i, S_i, P_i) with the relations Ψ are a category by checking the different proprieties of a category. Objects of the category are (R_i, S_i, P_i) and its morphisms are relations Ψ. The identity morphism is $id_\Psi = (id_R, id_S)$ such that id_R and id_S are identity functions; id_R maps each role to itself and id_S maps each protocol to itself. To verify the composition propriety, Let M_1, M_2, M_3 be three objects and $f_1 = (\alpha_1, \beta_1), f_2 = (\alpha_2, \beta_2)$ be two relations such that $f_1 : M_1 \longrightarrow M_2$ and $f_2 : M_2 \longrightarrow M_3$. The composition is defined as: $f_2 \circ f_1 = (\alpha_2 \circ \alpha 1, \beta 2 \circ \beta 1)$. Regarding the associativity propriety, the relations between triplet objects consist of functions between sets. Thence, the assiociativity is derived from the associativity of functions between sets.

In the same way, all organizational and social concepts such as norms, groups ...etc, may be modeled categorically. After modeling all concepts of an organization with categories C_i, the organization categorical model may be obtained by building a new category from the categories C_i. Several ways are available to build a new category from existing ones. For examples, by compositions, by adding structures, using sub categories, using product of categories...etc.

After modeling an organization with categories, it becomes possible to study formally its properties such as stability and adaptation by using advanced categorical concepts and notions such as natural transformations, Pushout...etc.

By using correspondences between categorical concepts and organizational ones and correspondences between categorical concepts and multi-agent organizational ones illustrated in figure1, the categorical model of an organization in a human society may be moved in a categorical model of an organization of MAS. The resulting model may be extended categorically in the same way with that presented in the first example. Once the organizational categorical model of MAS produced, studying organization proprieties becomes possible. Finally, an implementation of the MAS is to realize. Several ways to implement MASs are available; the use of MAS platforms [33] is one of the most effective.

6 Conclusion

The important property of category theory is that it allows studying structures and their proprieties in one domain and moving them to another one together with the capacities to deal with all concepts of all domains with the same categorical terms. In this paper, we presented an idea of studying and formalizing organizations and collective phenomena in human societies with category theory with the aim of capturing their logics into categorical models. Afterward, the captured models are mapped categorically to categorical MAS organizational models. This way of thinking allows taking inspiration from efficient organizations in human societies in order to develop efficient organizational models for MAS. Categories are compositional; in a categorical MAS organizational model, agents are often presented as the objects of certain categories. Such model may be refined and extended by considering each agent as a category which objects are software components and morphisms are interactions between components. So, agent's proprieties may also be studied. With this way of thinking, it becomes possible to study the properties of organizations in human societies such as stability and adaptation before taking them as landmarks for developing MAS organizational models. The inspirations from human societies are conducted with formal concepts. So, it is possible to reason about the inspiration processes themselves regardless of the source or inspired models.

The present work presents a nucleus around which several research projects can be initiated. These projects are to be managed according to three main axes. In the first one, it is necessary to study the maximum possible of collective phenomena and organizations in human societies and subsequently to model them with category theory concepts. A classification of the produced categorical models must be carried out. Each class of models responds to the specific needs of one or more types of problems. This classification requires a parameterization operation to output a signature that characterizes and allows classifying each model in one or more classes. Once reached at an advanced stage in the study, categorical modeling and classifying collective phenomena and organizations, a library of models is the result. To look for an organizational solution for a multi-agent system designed for a given problem, we have to do the parameterization of the problem. The result signature directs the search for a solution in the model library. The search for a solution in the model library may be done in an automatic way. Thus, the reuse of solutions is pushed very far. In addition, it is possible to evaluate the appropriateness of each chosen solution with the type of problem for which the solution is chosen. The traces of the different evaluations allow better directing the search for organizational solutions for each type of problems. This approach is similar to component-oriented development approaches where off-the-shelf components are ready to use. In the second axis, works must be driven to study on the basis of the categorical models the emergent phenomena in both MASs and human societies. In the third axis, sociology benefits from the possibilities of making simulations to study the impact of the variations that societies can undergo. .

References

[1] C. Szyperski (1998) *Component Software Beyond Object- Oriented Programming*, Addison-Wesley.

[2] B. Z Abraham, J. C Aguilar (2007) Software Component Selection Algorithm Using Intelligent Agents, *Agent and Multi-Agent Systems: Technologies and Applications, Lecture Notes in Computer Science 4496*, Springer-Verlag, pp. 82–91. https://doi.org/10.1007/978-3-540-72830-6_9

[3] P. Pelliccione, M. Tivolia, A. Bucchiaroneb and A. Polini (2008) An architectural approach to the correct and automatic assembly of evolving component-based systems, *Journal of Systems and Software*, sciencedirect, Vol 81, pp. 2237–2251. https://doi.org/10.1016/j.jss.2008.05.030

[4] N. Obeid, S. Al-Areqi (2013) Using Agents for Dynamic Components Redeployment and Replication in Distributed Systems, *Contemporary Challenges and Solutions in Applied Artificial Intelligence, Studies in Computational Intelligence*, Springer-Verlag, pp. 19–25. https://doi.org/10.1007/978-3-319-00651-2_3

[5] A. Siam , R. Maamri and Z. Sahnoun (2012) An approach based on software components and mobile agents for developing distributed applications with verification of validity criterion, *Proceedings of The 6th International Conference on Complex, Intelligent, and Software Intensive Systems*, IEEE Xplore, Palermo, pp. 407–413. https://doi.org/10.1109/CISIS.2012.97

[6] S. Leriche , JP. Arcangeli (2007) Adaptive Autonomous Agent Models for Open Distributed Systems, *Proceedings of the International Multi-Conference on Computing in the Global Information Technology*, IEEE Xplore, Gosier Guadaloupe, pp. 18–24. https://doi.org/10.1109/ICCGI.2007.10

[7] G. Eleftherakis , P. Kefalas and E. Kehris (2011)A methodology for developing component-based agent systems focusing on component quality, *Proceedings of the Federated Conference on Computer Science and Information Systems*, IEEE Xplore, Szczecin, Poland, pp. 561–568.

[8] A. Siam , R. Maamri and Z. Sahnoun Zaidi (2014) Organization of Self-Adaptive Agents Based On Software Components, *International Journal of Intelligent Information Technologies*, IGI Global, Vol 10, No 3, pp. 36–56. https://doi.org/10.4018/ijiit.2014070103

[9] A. J Macintyre , D. S Scott (2010) *Category theory, Second Edition*, Oxford University press.

[10] B. Michael, S. W Charle (1990) *Category Theory for Computing Science*, Prentice Hall.

[11] W. R Scott (1998) *Organizations: rational, natural and open systems, 4 edition*, Prentice Hall.

[12] O. Boissier, L. Coutinho and J.S. Sichman (2006) Organization oriented programming: from closed to open organizations ıEngineering Societies in the Agents World VII (ESAW'06), LNCS, Springer-Verlag, Berlin, Heidelberg, Vol.4457, pp.86–105. https://doi.org/10.1007/978-3-540-75524-1_5

[13] A. Estefania ,J. Vicente and B. Vicente (2006) Multi-Agent System Development Based on Organizations, *Electronic Notes in Theoretical Computer Science*, Vol 150, Elsevier, pp. 51–77. https://doi.org/10.1016/j.entcs.2006.03.005

[14] J. Deguet, L. Magnin, and Y. Demazeau (2006) Elements about the emergence issue: A survey of emergence definitions, *BIO INSPIRED METHODS*, ComPlexUs, pp. 24–31. https://doi.org/10.1159/000094185

[15] C. Lemaître, C. B Excelente (1998) Multiagent Organization Approach., *Proceedings 2nd Iberoamerican Workshop on Distributed Artificial Intelligence and Multi-Agent Systems* , Toledo, Spain, pp. 07–16.

[16] J. L Dessalles, J. P Müller and D Phan (2007) *Emergence in multi-agent systems: conceptual and methodological issues. In: Agent-based modelling and simulation in the social and human sciences*, Oxford : Bardwell Press.

[17] S. Bouarfa, H. AP Blom, R. Curran and M. HC Everdij (2013) Agent-based modeling and simulation of emergent behavior in air transportation, *Complex Adaptive Systems Modeling*, SpringerOpen, pp. 1–15. https://doi.org/10.1186/2194-3206-1-15

[18] Y. M Teo, B. L Luong and C. Szabo (2013) Formalization of emergence in multi-agent systems, *Proceedings of the 1st ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, ACM DL, NTU, Singapore, pp. 231–240. https://doi.org/10.1145/2486092.2486122

[19] J. Pavon, J. J Gomez-Sanz (2003) Agent oriented software engineering with INGENIAS, *Multi-Agent Systems and Applications III*, Springer Berlin Heidelberg, Vol. 2691 pp. 394–403. https://doi.org/10.1007/3-540-45023-8_38

[20] Author (2004) The MaSE Methodology, *In Methodologies and Software Engineering for Agent Systems*, SringerLink, pp. 107–126. https://doi.org/10.1007/1-4020-8058-1_8

[21] J. Ferber, O Gutknecht (1998) A meta-model for the analysis and design of organizations in multiagents systems, *Proceedings of the 3rd International Conference on Multi-Agent Systems*, IEEE Computer Society, Paris, France, pp. 128–135.

[22] M. Esteva, A. J. A Rodriguez-Aguiar, C. Sierra,P. Garcia and J. L Arcos (2001) On the formal specification of electronic institutions, *Notes in Computer Science*, Springer, Vol 191 pp. 126–147. https://doi.org/10.1007/3-540-44682-6_8

[23] D. B Panos (1979) Social Interaction and Social Processes, *Social Science*, JSTOR,Vol. 54, No. 3 pp. 147–167.

[24] V. Pareto (1968) *The rise and fall of the elites*, Totowa NJ Bedminster Press.

[25] J. K DeVree, J. C Dagevos (1994) The structure of action and interaction: The structural similarity of systems in social science, *Journal of Mathematical Sociology*, Taylor and francis, Vol. 19 pp. 91–127. https://doi.org/10.1080/0022250X.1994.9990138

[26] D. S Dendrinos (1996) Cities as spatial chaotic attractors, *Chaos Theory in the Social Sciences*, University of Michigan Press, pp. 237–269. https://doi.org/10.3998/mpub.14623

[27] P. Cvitanovic (1984) *Universality in Chaos, 2nd ed*, Bristol, Adam Hilger Ltd. https://doi.org/10.1002/bimj.4710270818

[28] D. L Sallach (2000) Classical social processes: Attractor and computational models, *The Journal of Mathematical Sociology*, Taylor and francis, Vol. 24, pp. 245–272. https://doi.org/10.1080/0022250X.2000.9990238

[29] S. P Robbins, A. Timothy (2012) *Organizational Behavior, 15 Edition*, Pearson.

[30] C. D Manning, P. Raghavan and H. Schütze (2009) *An Introduction to Information Retrieval*, Cambridge University Press.

[31] S. A Paul, K. Seok-Woo (2002) Social Capital: Prospects for a New Concept, *The Academy of Management Review*, JSTOR,Vol. 27 pp. 17–40. https://doi.org/10.2307/4134367

[32] F. Mandl (1991) *Statistical Physics 2nd Edition*, John Wiley and Sons.

[33] K. Kravari, N. Bassiliades (2015) A Survey of Agent Platforms , *Journal of Artificial Societies and Social Simulation*, Vol.18 pp. 11. https://doi.org/10.18564/jasss.2661

https://doi.org/10.31449/inf.v42i4.2027 Informatica **42** (2018) 577–585 **577**

Enhanced V-Model

Mustafa Seçkin Durmuş
Sadenco, Safe, Dependable Engineering and Consultancy, Antalya, Turkey
E-mail: msd@saden.co

İlker Üstoğlu
Yildiz Technical University, Department of Control and Automation Engineering, Istanbul, Turkey
E-mail: ustoglu@yildiz.edu.tr

Roman Yu. Tsarev
Siberian Federal University, Department of Informatics, Krasnoyarsk, Russia
E-mail: tsarev.sfu@mail.ru

Josef Börcsök
Kassel University, Computer Architecture and System Programming Department, Kassel, Germany
E-mail: j.boercsoek@uni-kassel.de

Keywords: Software development lifecycle, V-model, fault diagnosis, discrete event systems, EN 50128, fixed-block railway signaling systems

Received: November 16, 2017

Typically, software development processes are time consuming, expensive, and rigorous, particularly for safety-critical applications. Even if guidelines and recommendations are defined by sector-specific functional safety standards, development process may not be completed because of excessive costs or insufficient planning. The V-model is one of the most well-known software development lifecycle model. In this study, the V-model lifecycle is modified by adding an intermediate step. The proposed modification is realized by checking the fault diagnosability of each module. The proposed modification provides three advantages: (1) it checks whether the constructed model covers all software requirements related with faults; (2) it decreases costs by early detection of modeling deficiencies before the coding and testing phases; and (3) it enables code simplicity in decision of fault occurrence.

Povzetek: Osnovnemu modelu V razvoja programskih sistemov je dodana izboljšava na osnovi možnosti testiranja napak modulov.

1 Introduction

The concept known as Safety Integrity Level (SIL) is used to quantify safety. The SIL is a degree of safety system performance for a Safety Instrumented System (SIS), which is an automatic system used to avoid accidents and to reduce their impact both on humans and the environment. A SIS has to execute one or more Safety Instrumented Functions (SIFs) to maintain a safe state for the equipment under control [1]. Bear in mind that, a safe state is known as the state where the whole system is prevented from falling into a dangerous situation. A SIF has a designated SIL level depending on the ratio of risk that needs to be decreased. IEC 61508, the standard for functional safety of electrical/electronic/programmable-electronic Safety Related System (SRSs), mentions that a SIL should be designated to each SIF and defines the safety integrity as the probability of a SRS adequately performing the required safety functions under all the stated conditions within a given period of time from the lowest requirement level (SIL 1) to highest requirement level (SIL 4).

The third part of IEC 61508 applies to any software used to develop a safety-related system within the scope of first and second parts of the standard, and establishes the requirements for *safety lifecycle* phases. Industry and domain specific implementations of IEC 61508 include IEC 61511 for industrial processes, IEC 61513 for the nuclear industry, and IEC 62061 for machinery etc.

A lifecycle model is defined in [2] as a *model that describes stages in a software product development process*. The IEC 61508-4 standard discusses the term lifecycle in the context of both *safety lifecycle* and *software lifecycle*. The safety lifecycle includes the necessary activities involved in the implementation of SRSs [3]. IEC 61508 states that a *safety lifecycle for software development shall be selected* and specified during the safety-planning phase in accordance with Clause 6 of IEC 61508-1. The safety lifecycle includes the definition of scope, hazard and risk analysis, determination of safety requirements, installation, commissioning, validation, operation, maintenance, repair, and decommissioning. On the other hand, the software lifecycle includes the activities occurring from the conception of the software to the decommissioning of the software.

Numerous lifecycle models have been addressed in the literature, such as the waterfall, spiral, iterative development, and butterfly models [4-8]. However, despite the availability of many lifecycle alternatives, safety standards such as IEC 61508, EN 50126, EN 50128, and IEC 62278 recommend using the V-model for software development processes. The V-model lifecycle has been applied to various domains such as the automotive [9], aerospace [10], railways [11], and the nuclear industry [12].

In this study, a Discrete Event System (DES)-based fault diagnosis method is added to the V-model lifecycle as an intermediate step between the module design and the coding phases. A DES is a discrete-state, event-driven system in which the state evolution of the system depends totally on the occurrence of discrete events over time.

The main difference of the proposed enhancement is its simplicity, when compared with the existing model checking tools and techniques in the literature [13, 14]. Because the fault diagnoser is built from the software model itself and; since the modular approach is a must in the *Software Design Phase* of the V-model in EN 50128 (recommended as mandatory) there is no need for any tool to check the diagnosability of a simple software module (component) model [15]. The remainder of this paper is organized as follows. The V-model lifecycle and the modified V-model lifecycle are explained in Sections 2 and 3, respectively. DES-based fault diagnosis is introduced in Section 4, and conclusion section is given in Section 5.

2 V-model lifecycle

Paul Rook introduced the V-model lifecycle in 1986 as a guideline for software development processes [2]. The primary aim of the V-model is to improve both the efficiency of software development and the reliability of the produced software. The V-model offers a systematic roadmap from project initiation to product phase-out [2]. The V-model also defines the relationship between the development and test activities; it implements verification of each phase of the development process rather than testing at the end of the project. The V-model, as defined in IEC 61508-3, is shown in Figure 1.

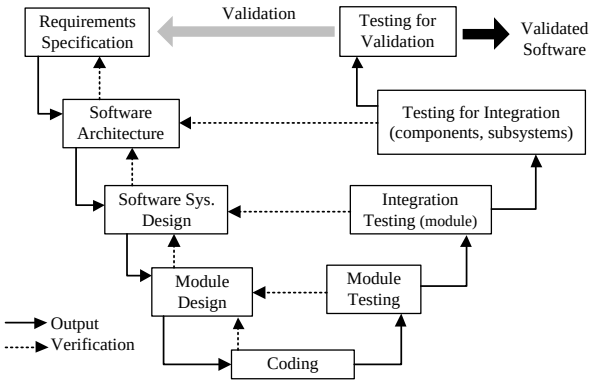


Figure 1: V-model software safety integrity and development lifecycle [16].

Before initializing a *software development process* according to the V-model, a *software planning phase* has to be realized, wherein a *software quality assurance plan*, *software verification* and *software validation plans*, and a *software maintenance plan* are fully defined. Later, the software requirements should be determined in cooperation with both the customer and the stakeholders. Using the selected software architectures (including modeling methods), software modules are developed by the designers. Each phase is verified immediately after completion. Note that, the left side of the V-model in Figure 1 represents the decomposition of the problem from the business world to the technical world [17]. After the coding phase, the right side of the V-model denotes the testing phase of the developed software.

The number of person may expand in the development process but this expansion shall be identified from the very beginning of the project. In [2], the change of the total number of software development teams are illustrated as given in Figure 2. Additionally, the cost of detection of faults in the different phases of V-model is given in Figure 3.

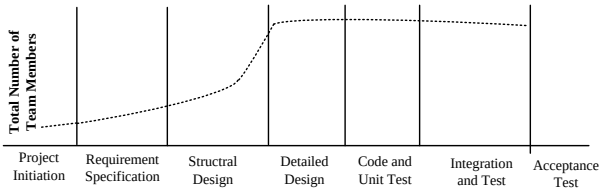


Figure 2: Software development teams [2].

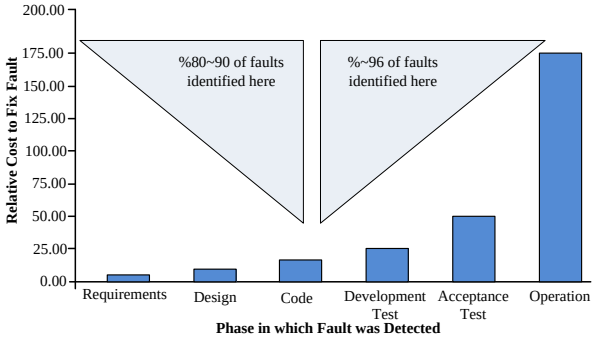


Figure 3: Software development teams [18].

The advantages and disadvantages of the V-model can be summarized as follows [4, 5]:

Advantages

1. Facilitates greater control due to the standardization of products in the process.
2. Cost estimation is relatively easy due to the repeatability of the process.
3. Each phase has specific products.
4. Greater likelihood of success because of the early development of test plans and documentation before coding.
5. Provides a simple roadmap for the software development process.

Disadvantages

1. Low flexibility, expensive, and difficult to change scope.
2. No early prototypes.
3. Addresses software development within a project rather than a whole organization.
4. Too simple to precisely reflect the software development process and may steer managers into a false sense of security.

3 Modified V-model lifecycle

As mentioned in [19], and [20], the required workforce and the cost of the development process of the software increases towards the end with respect to the initial phases of the development lifecycle. Therefore, the proposed modification is realized on the left side of the V-model.

In the usual software development process according to V-model, the fulfillment of the requirements are checked by realizing the module tests after coding. By checking the module diagnosability, one can decide if the module fully covers the software requirements related with faults or not (will be explained in section 4). This intermediate phase can be considered as time consuming and an extra workload. However, rather than turning back again from the module testing phase to the module design phase in the V-model, the proposed phase provides a final inspection of modules before proceeding to the coding and module testing phases. The proposed V-model is given in Figure 4.

The proposed modification in Figure 4 has three unique advantages:

a. *It checks whether the constructed model covers all software requirements related to faults*: If the developed software model (see Table A.2 and Table A.17 of [15]) is not diagnosable, then the software model does not contain all software requirements related with the faults.

b. *It decreases costs through early detection of modeling deficiencies before proceeding to coding and testing phases*: As can be seen from Figure 4, after proceeding to the coding phase, the designer can only go back to the module design phase at the end of the module tests. Many studies showed that, it is 5 times more expensive to fix a problem at the design stage than in the course of initial requirements, 10 times more expensive to fix it through the coding phase, 20 to 50 times more expensive to fix it at acceptance testing and, 100 to 200 times more expensive to fix that error in the course of actual operation [20-23].

c. *It enables designers to write simple and more readable code in decision of the faults*: This will be explained with a simple case study in the next section.

4 DES-based fault diagnosis

An *event* is defined as an encountered specific action, i.e., an unplanned incident that occurred naturally or due to numerous conditions that are encountered simultaneously [24]. Events are classified as observable or unobservable events in a DES.

A DES system is considered as diagnosable if it is possible to identify, within a finite delay, occurrences of precise unobservable events that are referred to as fault

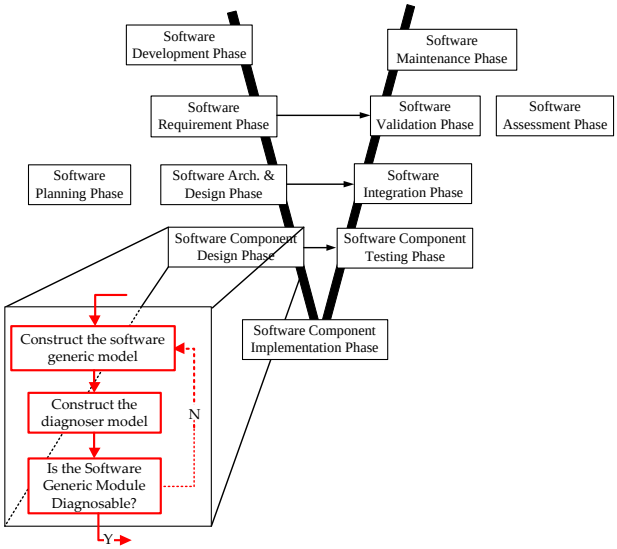


Figure 4: Enhanced V-model (Y-Yes, N-No).

events [25]. In other words, a system is diagnosable if the fault type is always identified within a uniformly bounded number of transition firings after the occurrence of the fault [26]. The diagnoser is obtained from the system model itself and carries out diagnostics to observe the system behavior. Diagnoser states involve fault information, and occurrences of faults are identified within a finite delay by examining these states [27].

Finite state machines and Petri nets are considered as DES-based modeling methods and, these methods are also highly recommended by functional safety standards (see [15]).

4.1 Basic petri net (*PN*) definitions

A Petri net [28] is defined as;

$$PN = (P, T, F, W, M_0) \tag{1}$$

where

- $P = \{p_1, p_2, \ldots, p_k\}$ is the finite set of places,
- $T = \{t_1, t_2, \ldots, t_z\}$ is the finite set of transitions,
- $F \subseteq (P \times T) \cup (T \times P)$ is the set of arcs,
- W: F $\rightarrow \{1, 2, 3, \ldots\}$ is the weight function,
- M_0: P $\rightarrow \{0, 1, 2, 3, \ldots\}$ is the initial marking,
- $P \cap T = \varnothing$ and $P \cup T \neq \varnothing$.

For a marking M, $M(p_i) = n$ represents the token number of the ith place where it is equal to n [28]. Representation of a marking $M : P \rightarrow \{1, 2, 3, ...\}$ can be realized by a k-element vector, where k denotes the total number of places.

Definition 1 [28]: If a *PN* has no self-loops, then it is considered as *pure* and when all arc weights of a *PN* are 1, then it is said to be *ordinary*.

Definition 2 [28]: $M_0[t_1 > M_1[t_2 > \ldots M_{k-1}[t_k > M_k$ means that, the marking M_k is reachable from the initial marking M_0 by the $t_1 t_2 \ldots t_k$ transitions sequence. Note

that, $R(M_0)$ denotes the set of all reachable markings from M_0.

Definition 3 [29]: A *PN* is said to be *free from deadlocks* if it is possible to find at least one enabled transition at every reachable marking.

The set of places, P is partitioned into a set of observable places and a set of unobservable places (P_o and P_{uo}). Likewise, the set of transitions, T is partitioned into a set of observable transitions and a set of unobservable transitions (T_o and T_{uo}). Thus, the partitioned sets for P and T can be expressed as

$$P = P_{uo} \cup P_o \text{ and } P_{uo} \cap P_o = \varnothing$$
$$T = T_{uo} \cup T_o \text{ and } T_{uo} \cap T_o = \varnothing \quad (2)$$

In addition, a subset T_F of T_{uo} represents a faulty transitions set. It is assumed that there are m different fault types. Here, $\Delta_F = \{F_1, F_2, \ldots, F_m\}$ is the set of fault types. T_F is expressed as $T_F = T_{F_1} \cup T_{F_2} \cup \cdots \cup T_{F_m}$, where $T_{F_i} \cap T_{F_j} = \varnothing$ (if $i \neq j$).

$\Delta = \{N\} \bigcup 2^{\Delta_F}$ is used to define the label set, where N is used to represent the label "normal," which specifies that all fired transitions are *not faulty*, and 2^{Δ_F} represents the power set of Δ_F. In the remainder of this paper, unobservable transitions and places are represented as shown in Figure 5.

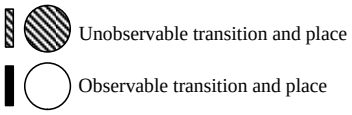


Figure 5: Representation of places and transitions.

4.2 Diagnosis of faults by using *PN* models

Since the system model contains unobservable places, it is not always possible to distinguish some markings. Thus, if $M_1(p_i) = M_2(p_i)$ for any $p_i \in P_o$, then it is denoted as $M_1 \equiv M_2$. That is to say, M_1 and M_2 markings have the same observations. As done in [30], the definition of the quotient set $\hat{R}(M_0)$ according to the equivalence relation $(\equiv)$ is useful; $\hat{R}(M_0) := R(M_0)/\equiv := \{\hat{M}_0, \ldots, \hat{M}_n, \ldots\}$ where $M_0 \in \hat{M}_0$. An observable marking or the observation of a marking is represented by each member of $\hat{R}(M_0)$. We assume the following two statements are true for simplicity.

Assumption [25, 26]: Only *deadlock free PN*s are considered and there does not exist an order of unobservable transitions whose firing produces a cycle of markings that have the same observation.

A diagnoser is given for a *PN* [26, 27] by

$$G_d = (Q_d, \Sigma_o, \delta_d, q_0). \quad (3)$$

The diagnoser given by (3) is an automaton where the set of states are represented by $Q_d \subseteq Q$, the set of events are represented by $\Sigma_o = \hat{R}(M_0) \cup T_o$, the notation $\delta_d : Q_d \times \Sigma_o \to Q_d$ represents the partial state transition function, and the initial state is denoted by $q_0 = \{(M_0, N)\}$. A diagnoser state q_d is given as $q_d = \{(M_1, l_1), (M_2, l_2), \ldots, (M_n, l_n)\}$, which involves pairs of a marking $M_i \in R(M_0)$ and a label $l_i \in \Delta$. The state set $Q_d \subseteq Q$ represents the reachable states from the initial state q_0 by using δ_d. Each observed event $\sigma_o \in \Sigma_o$ represents an observation of a marking in $\hat{R}(M_0)$ or an observable transition in T_o. The state transition function δ_d is defined with the use of the label propagation function and the range function.

The label propagation function associates a label (*faulty* or *normal*) over a sequence of transitions. If the sequence of transitions does not contain any faulty transition, the resulting marking is labeled as normal (N). Detailed explanation of the label propagation function, the range function and the state transition function can be seen from [25-27, 30, 31].

4.3 Obtaining diagnosability

A *PN* is diagnosable if, and only if, the states of the diagnoser given by (3) shall be F_m-certain or does not involve any F_m-indeterminate cycle for any fault type F_m. Due to page restriction, the reader is referred [25-27] for detailed explanation of DES-based fault diagnosis and the proof of this theorem.

4.4 Railway point example

Trains can move from one track to another by the help of railway *points* (rail switches or point machines) placed at necessary locations. Points have two position indications, i.e., Normal (*Nr*) and Reverse (*Rev*). At any railway point, three main faults may occur. These faults are identified in the V-model software requirements specification phase as follows:

- F_1: Point may not reach the desired position in a predefined time (e.g., 5 sec) while moving from *Nr* to *Rev*.
- F_2: Point may not reach the desired position in a predefined time while moving from *Rev* to *Nr*.
- F_3: Both position indications may be received simultaneously.

Examples of diagnosable and not diagnosable *PN* models of a railway point are given in Figure 6 and Figure 7, respectively. The meanings of the transitions and places of the models in Figure 6 and Figure 7 are given in Table 1 and Table 2, respectively. Note that the striped places and transitions represent unobservable places (P_{uo}) and transitions (T_{uo}), whereas the other places (P_o) and transitions (T_o) are observable. M_0 represents the initial marking of the *PN*. The underlined numbers in the diagnoser are used to represent the marking of an unobservable place.

Representation of the *PN* model in Figure 6 is as follows:

$$P_o = \{P_{PM\_1}, P_{PM\_2}, P_{PM\_5}, P_{PM\_6}, P_{PM\_8}, P_{PM\_10}, P_{PM\_12}\},$$
$$P_{uo} = \{P_{PM\_3}, P_{PM\_4}, P_{PM\_7}, P_{PM\_9}, P_{PM\_11}\},$$
$$T_o = \{t_{PM\_1}, t_{PM\_2}, t_{PM\_3}, t_{PM\_4}, t_{PM\_5}, t_{PM\_6}, t_{PM\_7}, t_{PM\_8}, t_{PM\_9}, t_{PM\_10}, t_{PM\_11}, t_{PM\_12}, t_{PM\_13}, t_{PM\_14}\}, \quad T_{uo} = \{t_{PM\_f1}, t_{PM\_f2}, t_{PM\_f3}\}, \quad (4)$$
$$M_0 = (M_0(P_{PM\_1}), M_0(P_{PM\_2}), M_0(P_{PM\_3}), M_0(P_{PM\_4}), M_0(P_{PM\_5}), M_0(P_{PM\_6}), M_0(P_{PM\_7}), \ldots$$
$$\ldots, M_0(P_{PM\_8}), M_0(P_{PM\_9}), M_0(P_{PM\_10}), M_0(P_{PM\_11}), M_0(P_{PM\_12})) = (0,0,\underline{0},\underline{0},1,0,\underline{0},0,\underline{0},0,\underline{1},0).$$

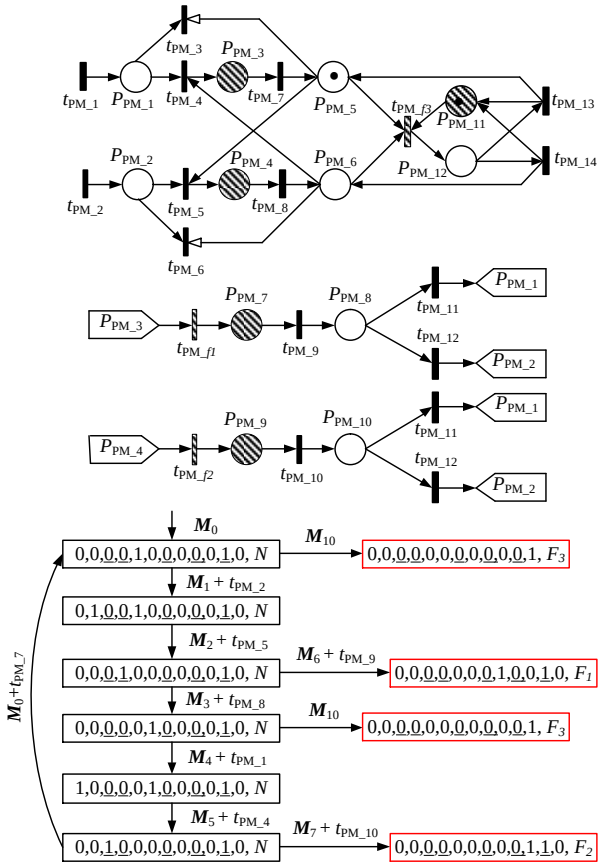


Figure 6: PN model of a railway point and its diagnoser (*diagnosable*).

Three different fault types given in Figure 6 are $\Delta_F = \{F_1, F_2, F_3\}$, where $T_{F_1} = \{t_{PM\_f1}\}$, $T_{F_2} = \{t_{PM\_f2}\}$, and $T_{F_3} = \{t_{PM\_f3}\}$. The rectangles are used to diminish the complexity of the *PN* model. Each rectangle represents the label of the related place.

The diagnoser illustrated in Figure 6 is built from the *PN* model of railway point. A rectangle is used to denote each state and each state contains a pair of place markings and an attached label, normal (*N*) or fault. In other words, in parts of the diagnoser, a marking immediately after an observed event is detected precisely.

In accordance with the definition of the diagnoser in (3), a label which represents an observable transition or the observation of a marking is attached to all diagnoser state transitions.

In this study, with a slight abuse of notation, labels containing the observation of a marking or a pair of the observation of a marking and an observable transition are attached to all state transitions of the diagnoser.

Place	Definition	Transition	Definition
P_{PM_1}	*Nr* position requested	t_{PM_1}	Point *Nr* position request
P_{PM_2}	*Rev* position requested	t_{PM_2}	Point *Rev* position request
P_{PM_3}	Point is moving to *Nr*	t_{PM_3} (t_{PM_6})	Request ignored
P_{PM_4}	Point is moving to *Rev*	t_{PM_4}	Point left *Rev*
P_{PM_5}	Point is in *Nr*	t_{PM_5}	Point left *Nr*
P_{PM_6}	Point is in *Rev*	t_{PM_7} (t_{PM_8})	Point reached to *Nr* (*Rev*)
P_{PM_7}	Fault type F_1 has occurred	t_{PM_9} (t_{PM_10})	Filter time has expired
P_{PM_8}	Point is faulty (F_1)	t_{PM_11} (t_{PM_12})	*Nr* (*Rev*) position request
P_{PM_9}	Fault type F_2 has occurred	t_{PM_13} (t_{PM_14})	Point moved to *Nr* (*Rev*) and the fault acknowledged
P_{PM_10}	Point is faulty (F_2)	t_{PM_f1}	Point indication fault
P_{PM_11}	Unobservable fault restriction	t_{PM_f2}	Point indication fault
P_{PM_12}	Point is faulty (F_3)	t_{PM_f3}	Point position fault

Table 1: Definition of transitions and places in the models given in Figure 6.

For example, at $\hat{M}_0$ in Figure 6, the event label $\hat{M}_{10}$ represents that the observable marking $\hat{M}_{10}$ is observed by firing the unobservable transition t_{PM_f3}. Similarly, the diagnoser state changes by firing the unobservable transition t_{PM_f3}. Similarly, the diagnoser state changes from $\{((0,0,\underline{0},\underline{0},1,0,\underline{0},0,\underline{0},0,\underline{1},0,),N)\}$, to $\{((0,1,\underline{0},\underline{0},1,0,\underline{0},0,\underline{0},0,\underline{1},0),N)\}$, as a function of firing the observable transition t_{PM_2} with the observation $\hat{M}_1$ of the resulting marking. According to the definition given in Section 4.3, since all states are F_m-certain and there is no F_m-indeterminate cycle in the diagnoser, the *PN* model is diagnosable.

Representation of the *PN* model in Figure 7 is as follows:

$$
\begin{aligned}
&P_o = \{P_{PM\_1}, P_{PM\_2}, P_{PM\_3}, P_{PM\_4}, P_{PM\_5}, P_{PM\_6}\}, \quad P_{uo} = \{P_{PM\_7}\},\\
&T_o = \{t_{PM\_1}, t_{PM\_2}, t_{PM\_3}, t_{PM\_4}, t_{PM\_5}, t_{PM\_6}, t_{PM\_7}, t_{PM\_8}\}, \quad T_{uo} = \{t_{PM\_f1}, t_{PM\_f2}, t_{PM\_f3}\},\\
&M_0 = (M_0(P_{PM\_1}), M_0(P_{PM\_2}), M_0(P_{PM\_3}), M_0(P_{PM\_4}), M_0(P_{PM\_5}), M_0(P_{PM\_6}), M_0(P_{PM\_7}))\\
&\quad = (0,0,1,0,0,0,\underline{0}).
\end{aligned} \tag{5}
$$

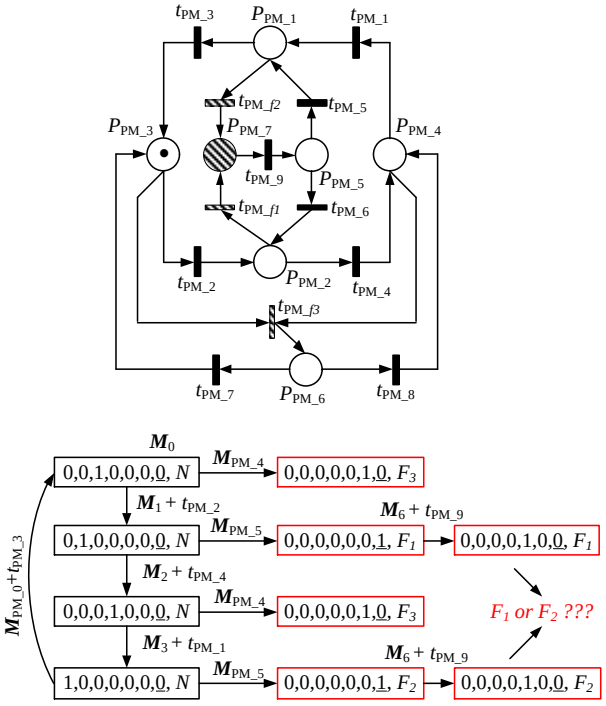


Figure 7: PN model of a railway point and its diagnoser (*not diagnosable*).

The diagnoser given in Figure 7 is not diagnosable because it is not possible to distinguish the fault type after observing the marking $\hat{M}_5$. In this case, the *PN* model will identify only one of the faults (F_1 or F_2) while the obtained code from this *PN* model is running. Therefore, the designers should revise the *PN* model before proceeding to the coding phase; otherwise, this deficiency will result in an unsuccessful test case in the module testing phase.

4.5 Railway signal example

An example *PN* model of a Two-Aspect Signal (TAS) and its diagnoser is given in Figure 8. TAS is generally used in railway depot areas and has two signal color indications (red means stop and green means proceed). The meanings of the transitions and places of the model in Figure 8 are given in Table 3. It is assumed that two different faults may occur in TAS which are F_1: Both signal aspects are lit at the same time; and F_2: No signals are lit.

Place	Definition	Transition	Definition
P_{PM_1}	Point is moving to *Nr*	t_{PM_1} (t_{PM_2})	Movement request is received for *Nr* (*Rev*) position
P_{PM_2}	Point is moving to *Rev*	t_{PM_3} (t_{PM_4})	Point reached to *Nr* (*Rev*)
P_{PM_3}	Point position is *Nr*	t_{PM_5} (t_{PM_6})	Point request to *Nr* (*Rev*)
P_{PM_4}	Point position is *Rev*	t_{PM_7} (t_{PM_8})	Point moved to *Nr* (*Rev*) and the fault acknowledged
P_{PM_5}	Fault type F_1 or F_2 has occurred	t_{PM_9}	Filter time has expired
P_{PM_6}	Point is faulty (F_3)	t_{PM_f1} (t_{PM_f2})	Point indication fault
P_{PM_7}	Point is moving from one position to another	t_{PM_f3}	Point position fault

Table 2: Definition of transitions and places in the models given in Figure 7.

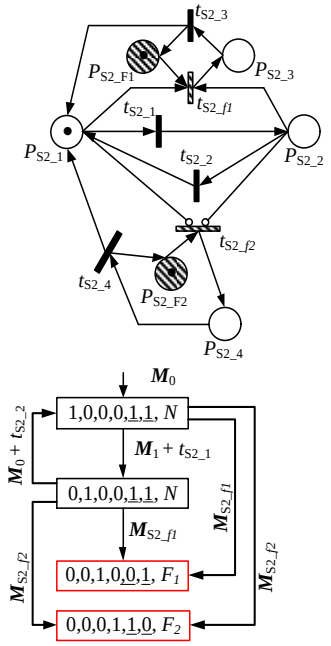


Figure 8: *PN* model of TAS and its diagnoser.

Place	Definition	Transition	Definition
P_{S2_1}	Signal is red	t_{S2_1}	Turn signal to green

P_{S2_2}	Signal is green	t_{S2_2}	Turn signal to red
P_{S2_3}	Fault type F_1 has occurred	t_{S2_3}	Signal turned to red and the fault acknowledged
P_{S2_4}	Fault type F_2 has occurred	t_{S2_4}	Signal turned to red and the fault acknowledged
P_{S2_F1}	Unobservable fault restriction	t_{S2_f1}	Point aspect fault
P_{S2_F2}	Unobservable fault restriction	t_{S2_f2}	Point indication fault

Table 3: Definition of transitions and places in the models given in Figure 8.

To compare the simplicity in decision of the faults with and without a diagnoser, an example Programmable Logic Controller (PLC) code snippet of TAS model is shown in Figure 9 and Figure 10, respectively.

As can be seen from Figure 9 and Figure 10, decision of fault occurrence with a diagnoser is simpler than without a diagnoser. For the PLC code given in Figure 9, the diagnoser compares the actual states of the *PN* model with predefined faulty states. When the faulty state of the diagnoser is fully matched with the marking of the actual *PN* states, the diagnoser sets the corresponding fault label.

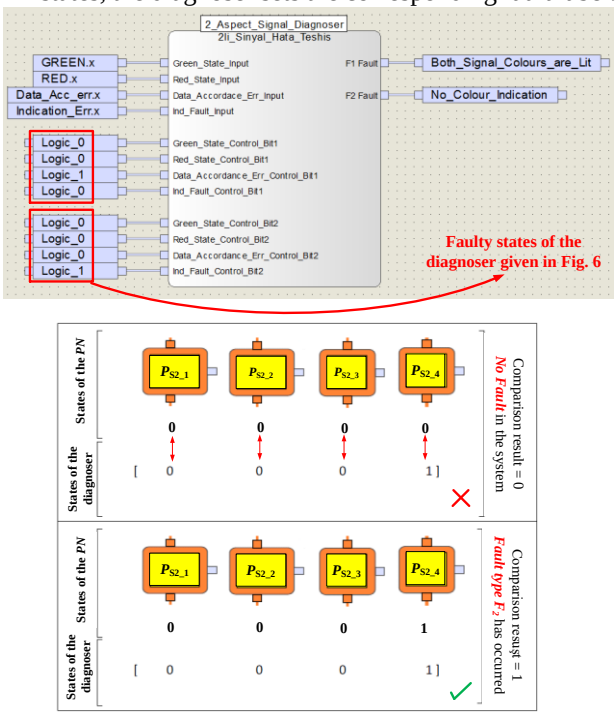


Figure 9: Diagnoser block of TAS and decision of fault occurrence.

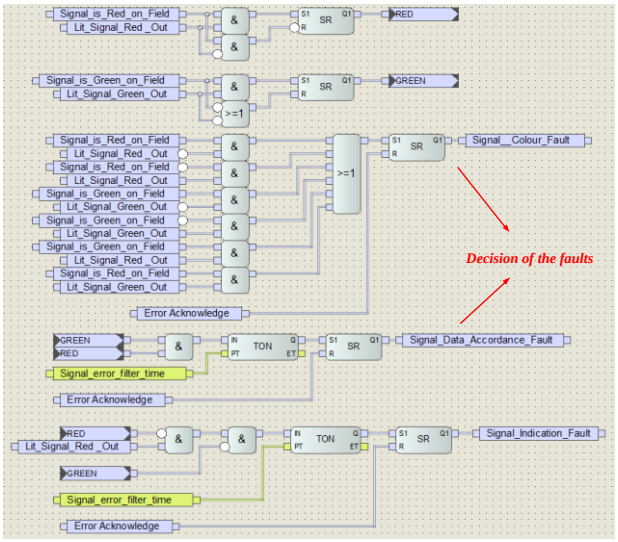


Figure 10: Decision of fault for TAS without diagnoser.

Moreover, since the V-model is modified by adding an additional step, we also defined a new task for the organizational structure of the software development team. The Diagnoser designer (DDes) is added to the preferred organizational structure of the enhanced V-model as given in Figure 11 (PM: Project Manager, RQM: Requirement Manager, Des: Designer, IMP: Implementer, VER: Verifier, VAL: Validator, DDes: Diagnoser Designer, ASR: Assessor). The original organizational structure can be seen from EN 50128 [15].

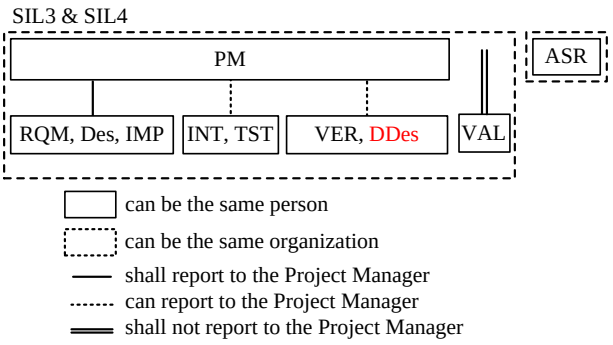


Figure 11: Recommended organizational structure for the enhanced V-model for SIL3&SIL4 software.

5 Conclusion

Faults in a safety-critical system may cause severe harm to humans. Therefore, the development steps of software for such safety-critical systems must be executed very carefully. Designers, developers, and engineers must consider the recommendations of both the international safety standards and the national rules to satisfy the required safety level and fulfill requirements.

Although enhancing the V-model with DES-based fault diagnosis is time consuming, however, the advantages of this intermediate step are threefold: (1) it checks whether the developed model fulfills all software requirements related to the faults; (2) decision of faults with a diagnoser is simpler than without a diagnoser; and (3) an early check of the models is possible before

proceeding to the coding and testing phase because the V-model leads developers from the module testing phase to the module design phase rather than the coding phase.

On the other hand, when costs and work hours are considered, adding such an intermediate step to the V-model can result in considerable benefits to both project management and product development departments.

Acknowledgement

The authors are thankful to Enago (www.enago.com) for the review of the English language of the paper.

References

[1] IEC61508 (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems, Parts 1–7*. International Electrotechnical Commission.

[2] Rook P (1986). Controlling Software Projects. *Software Engineering Journal*, 1, pp. 7-16. https://doi.org/10.1049/sej.1986.0003

[3] IEC 61508-4 (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems, Part 4: Definitions and Abbreviations*. International Electrotechnical Commission.

[4] Munassar NM, Govardhan A (2010). A Comparison Between Five Models of Software Engineering. *International Journal of Computer Science Issues*, 7, pp. 94-101.

[5] Krishna ST, Sreekanth S, Perumal K, Kumar Reddy KR (2012). Explore 10 Different Types of Software Development Process Models. *International Journal of Computer Science and Information Technologies*, 3:4580-4584.

[6] Royce WW (1970). Managing the Development of Large Software Systems: Concepts ad Techniques. *Proceedings Wescon*, pp. 1-9.

[7] Boehm BW (1988). A Spiral Model of Software Development and Enhancement. *Computer*, 21, pp. 61-72. https://doi.org/10.1109/2.59

[8] Lehman MM (1980). Programs, Life Cycles, and Laws of Software Evolution. *Proceedings of the IEEE*, 68, pp. 1060-1076. https://doi.org/10.1109/PROC.1980.11805

[9] Rahman RA, Pulm U, Stetter R (2007). Systematic Mechatronic Design of a Piezo-Electric Brake. *16th International Conference on Engineering Design*, 28-31 July, Paris, France, pp. 1-12.

[10] Märtin L, Schatalov M, Hagner M, Goltz U, Maibaum O (2013). A Methodology for Model-Based Development and Automated Verification of Software for Aerospace Systems. *IEEE Aerospace Conference*, 2-9 March, Big Sky, MT, USA, pp. 1-19. https://doi.org/10.1109/AERO.2013.6496950

[11] Scippacercola F, Pietrantuono R, Russo R, Zentai A (2015). Model-Driven Engineering of a Railway Interlocking System. *3rd Int Conf on Model-Driven Eng and Soft Development*, 2-9 September, Angers, France, pp pp. 509-519. https://doi.org/10.1007/978-3-319-27869-8_22

[12] SSG-39 (2016). *Design of Instrumentation and Control Systems for Nuclear Power Plants*. IAEA Safety Standards Series.

[13] Kwiatkowska M, Norman G, Parker D (2002). PRISM: Probabilistic Symbolic Model Checker. Field T, Harrison PG, Bradley J, Harder U (ed) *Computer Performance Evaluation: Modeling Techniques and Tools, Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, pp. 200-204. https://doi.org/10.1007/3-540-46029-2_13

[14] Holzmann GJ (2003). *Spin model checker, the: primer and reference manual*. Addison-Wesley.

[15] BS EN 50128 (2011). *Railway Applications-Communication, Signalling and processing systems: Software for railway control and protection systems*. International Electrotechnical Commission.

[16] IEC 61508-3 (2010). *Functional safety of electrical/electronic/programmable electronic safety-related systems, Part 3: Software Requirements*. International Electrotechnical Commission.

[17] Ratcliffe A (2011). SAS Software Development with the V-Model. *3SAS Global Forum, Coder's Corner*, 4-7 April, Las Vegas, Nevada, USA, pp. 1-9.

[18] Brat GP (2017). Reducing V&V Cost of Flight Critical Systems: Myth or Reality? *AIAA Information Systems, AIAA SciTech Forum*, American Institute of Aeronautics and Astronautics, 9-13 January, Grapevine, Texas, USA, pp. 1-10.

[19] Boehm BW (1984). Verifying and Validating Software Requirements and Design Specifications. *IEEE Software*, 1, pp. 75-88. https://doi.org/10.1109/MS.1984.233702

[20] Boehm BW (1984). Software Engineering Economics. *IEEE Transactions on Software Engineering*, SE-10, pp. 4-21. https://doi.org/10.1109/TSE.1984.5010193

[21] Boehm BW (1987). Industrial Software Metrics: A Top Ten List. *IEEE Software*, 4, pp. 264-271.

[22] Haskins B, Stecklein J, Dick B, Moroney G, Lovell R, Dabney J (2004). Error Cost Escalation Through the Project Life Cycle. *14th Annual Int Symp, Int Council on Systems Engineering*, 19-24 June, Toulouse, France, pp. 1723-1737. https://doi.org/10.1002/j.2334-5837.2004.tb00608.x

[23] Schneider GM, Martin J, Tsai WT (1992). An Experimental Study of Fault Detection in User Requirements Documents. *IACM Transactions on Software Engineering and Methodology*, 1, pp. 188-204. https://doi.org/10.1145/128894.128897

[24] Cassandras CG, Lafortune S (2008). *Introduction to Discrete Event Systems*. Springer, New York. https://doi.org/10.1007/978-0-387-68612-7

[25] Sampath M, Sengupta R, Lafortune S, Sinnamohideen K, Teneketzis D (1995). Diagnosability of discrete-event systems. *IEEE Trans on Automatic Control*, 40, pp. 1555-1575.

https://doi.org/10.1109/9.412626

[26] Ushio T, Onishi I, Okuda K (1998). Fault detection based on Petri net models with faulty behaviours. *International Conference on Systems, Man, and Cybernetics*, 11-14 October, San Diego, CA, USA, pp. 113-118.

[27] Sampath M, Sengupta R, Lafortune S, Sinnamohideen K, Teneketzis D (1996). Failure diagnosis using discrete-event models. *IEEE Transactions on Control Systems Technology*, 4, pp. 105-124. https://doi.org/10.1109/87.486338

[28] Murata T (1989). Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77, pp. 541-580. https://doi.org/10.1109/5.24143

[29] Li ZW, Zhou MC, Wu NQ (2008). A survey and comparison of Petri net-based deadlock prevention policies for flexible manufacturing systems. *IEEE Trans on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 38, pp. 173–188. https://doi.org/10.1109/TSMCC.2007.913920

[30] Chung SL (2005). Diagnosing PN-based models with partial observable transitions. *International Journal of Computer Integrated Manufacturing*, 18, pp. 158-169. https://doi.org/10.1080/0951192052000288206

[31] Durmuş MS, Takai S, Söylemez MT (2014). Fault Diagnosis in Fixed-Block Railway Signaling Systems: A Discrete Event Systems Approach. *IEEJ Transactions on Electrical and Electronic Engineering*, 9, pp. 523-531. https://doi.org/10.1002/tee.22001

https://doi.org/10.31449/inf.v42i4.2003 Informatica **42** (2018) 587–594 **587**

Integrated Speaker and Speech Recognition for Wheel Chair Movement using Artificial Intelligence

Gurpreet Kaur
Research Scholar, I.K Gujral Punjab Technical University, Kapurthala-144603, India
Assistant Professor, University Institute of Engineering & Technology, Panjab University,Chandigarh-160025, India
E-mail: regs4gurpreet@yahoo.co.in

Mohit Srivastava
Professor, Chandigarh Engineering College, Landran, Mohali-140307, India
E-mail: mohitsrivastava.78@gmail.com

Amod Kumar
Scientist, Central Scientific Instruments Organisation, Chandigarh-160030, India
E-mail: csioamod@yahoo.com

Keywords: speaker recognition, speech recognition, mel frequency cepstral coefficients, artificial bee colony algorithm, feed forward back propagation neural network

Received: November 10, 2017

Abstract: A speech signal is a result of the constrictions of vocal tract and different sounds can be generated by different vocal tract constrictions. A speech signal carries two things i.e. speaker's identity and meaning. For a specific applications of speaker and speech recognition like voice operated wheel chair, both the speaker and speech is to be recognized for movement of wheel chair. Automation in wheelchair is, today's requirement as the numbers of people are increasing with disabilities like injuries in spine, amputation, impairments in hands etc. They need assistance for moving their wheel chair. Voice operated wheel chair is one of the solution. The intention of this study is to use a speaker and speech dependent system to control the wheelchair and minimize the risk of unwanted accident. We have proposed a system in which speaker (patient) as well as speech (commands) is recognized based upon the acoustic features like Mel Frequency Cepstral Coefficients (MFCC). Optimization of the features using Artificial Bee Algorithm (ABC) is done to gain good accuracy with artificial intelligence technique as a classifier. We have tested our system on standard dataset (TIDIGITS) and our own prepared dataset. Also, validation of proposed work is done by generating control signal to actuate the wheel chair in real time scenario.

Povzetek: Predstavljen je nadzor invalidskega vozička s pomočjo govora in metod umetne inteligence.

1 Introduction

In the recent years, speaker and speech recognition has become a major domain of research because of various applications in real world from home to health care services. Speaker and speech recognition is jointly used in voice operated wheel chair. Wheel chair should move only when specific person (speaker recognition) should give commands (speech recognition). To guarantee good accuracy and less learning time, feature extraction process is very important for any recognition system. We have done detailed analysis on various feature extraction methods [1-2]. MFCC are the most used features for the speaker and speech recognition systems. Various optimization algorithms like Genetic algorithms with convolutional neural networks have been used for various applications like human action recognition [3-4]. Weights can be optimized using Genetic Algorithm (GA) framework. Dervis Karaboga proposed ABC algorithm for optimizing numerical problems[5]. This algorithm can be used in various fields like data mining, designing of filters, speech/speaker recognition etc. [6]. Initially DNN based system are used on phone recognition task. After that DNNs are used at large scale on big vocabulary continuous speech [7-10].When DNN based systems are compared with other systems like dynamic time warping (DTW), Hidden Markov Models (HMM), Gaussian mixture models (GMM) based systems, then recognition accuracy is more for DNN based systems [11].There is fast learning in deep neural network by parameterization of weight matrix by using periodic functions. In this way training time is reduced and classification accuracy improves. Different configurations of layers in neural network results in different results in terms of accuracy by considering complexity and memory requirements of the system [12-13]. Feed forward back propagation network (FFBPN) and recurrent neural network (RNN) are most used networks type [14-15]. This paper presents the integrated speaker and speech recognition system with optimized

MFCC features using ABC algorithm and FFBPN is used for classification. The main aim of proposed work is to combined the features of speech according to the speaker and create a speech and speaker dependent system to recognize the command which is given by the speaker to operate the wheelchair. If the wheelchair is operated only on commands without recognizing speaker then anybody can provide the wrong commands. So we need to combine the speech with speaker for the movement of wheel chair without any error. Operating the wheelchair according to the command of speaker is a big task and the feature extraction technique plays an important role. If the extracted features of speech signal are not unique then the accuracy of recognition system is not acceptable and the chances of error will increase. So the selection of a feature extraction technique is major issue in the speech recognition system. MFCC is a better option to find out the features from the speech signals. But the chances of unwanted features are still there, so optimization is needed and we have used the artificial bee colony algorithm with a novel objective function. So in the proposed work, an efficient speech and speaker recognition system is presented using the artificial intelligence technique along with the ABC optimization algorithm.

2 Related work

There are many existing speech and speaker recognition system using different types of techniques.

Many researchers have developed smart wheel chairs with voice commands using commercially available recognition systems. Simpson et al. [16] developed a wheelchair with joystick as well as voice controlled module. They have used commercially available Verbex speech commander recognition system with nine commands. With the development of many algorithms in artificial intelligence domain, researchers have moved to this field. Pacnik et al. [17] proposed a voice operated intelligent wheel chair (VOIC) using LPC cepstral analysis with neural network. They have analyzed that recognition of speech is possible with neural network, giving 4% error. Another author Jabardi [18] developed a wheelchair based upon artificial neural network with 86% of accuracy for known speakers and 76% of accuracy for unknown speakers. Speech recognition field has gain more advancement in terms of recognition rate, environmental conditions, speaker variability etc. with the development of deep neural networks[19-24].

From the above survey we have decided to present an integrated speaker and speech recognition system using ABC algorithm with artificial intelligence technique for the movement of wheel chair.

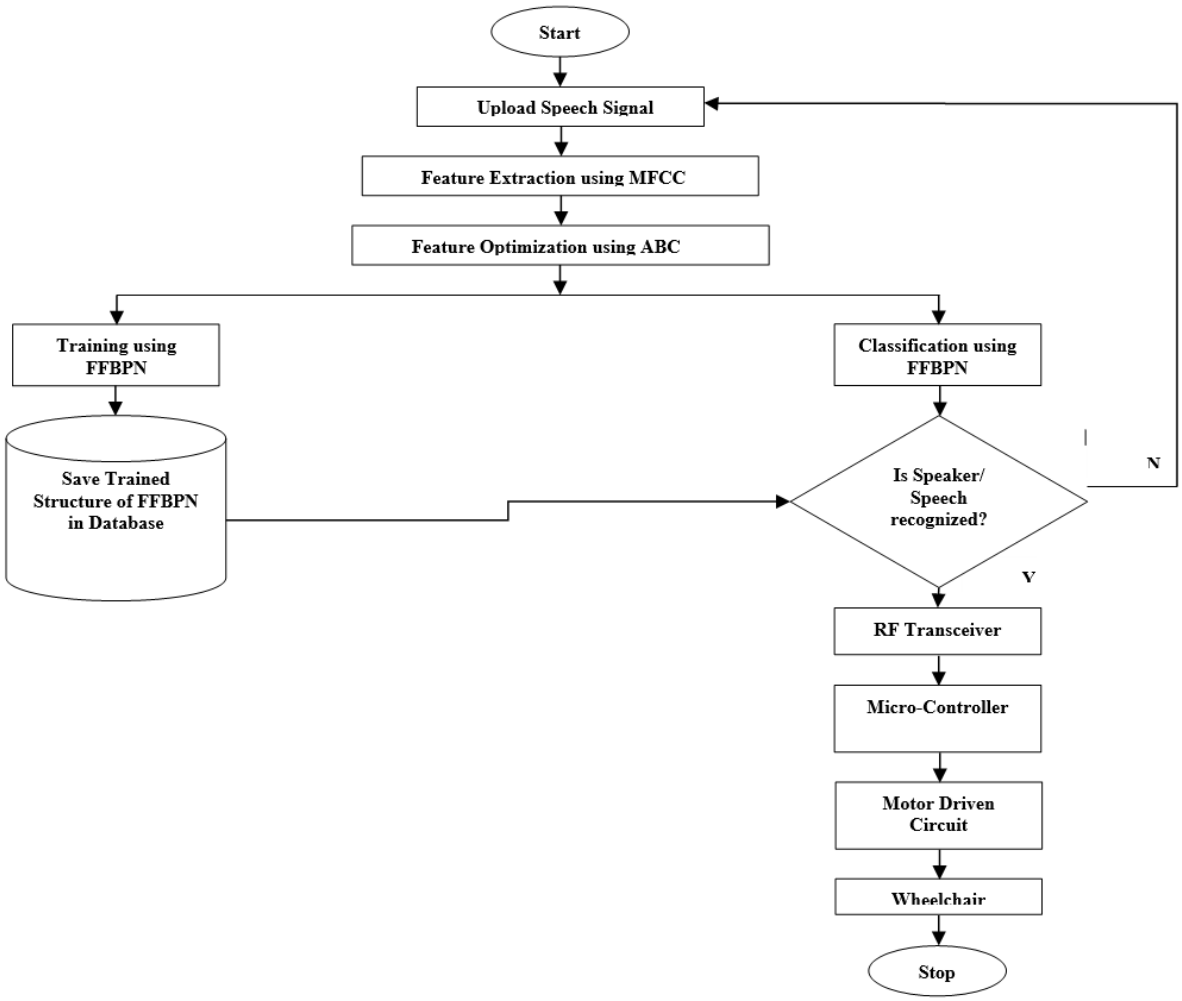


Figure 1: Integrated Speaker and Speech Recognition System for movement of wheel chair.

3 Proposed system

We have proposed an integrated speaker and speech recognition system as shown in figure 1.

Above figure represent the flow diagram of integrated speaker and speech recognition system based on the artificial intelligence concept. In the speaker and speech recognition system, firstly features are extracted from the speech signal. These features should be robust to noise and efficient enough so that classification can discriminate between the speakers and words. We have used MFCC features and to gain in accuracy, optimization of these features is done by ABC algorithm. Then classification is done with feed forward back propagation neural network. The FFBPN output is what is recognized i.e. who is speaking and what he or she is speaking. The command word is given through the RF transceiver to the microcontroller. The MCU (ATMEGA 8) interprets the commands received and accordingly motor is controlled through driver circuit (L293D) to move the wheel chair.

3.1 Feature extraction using MFCC algorithm

In this section, we have described the MFCC algorithm which is used to find out the feature set from the speech signal with respect to the speaker. The algorithm of MFCC is given below.

Firstly Initialized parameters
Tw = 25(analysis frame duration (ms))
Ts = 10 (analysis frame shift (ms))
Alpha = 0.97(pre emphasis coefficient)
R = [300 3700] (frequency range to consider)
M = 20 (number of filter bank channels)
C = 13 (number of cepstral coefficients)
L = 22 (cepstral sine lifter parameter)
Hamming=((N)(0.54-0.46*cos(2*pi*[0:N-1].'/(N-1))))

$$\mathrm{MFCC}_{\mathrm{features\{i\}}} = \sum_{i=1}^{n} \mathrm{MFCC}(\ \mathrm{Signal, fs, Tw, Ts, Alpha, Hamming, R, M, C, L})$$

Where Signal is the speech data which is uploaded by user and MFCC_features are the extracted feature set from the uploaded speech data.

3.2 Optimization with ABC algorithm

The probability of unwanted signals are more in extracted features and due to the unwanted signal the accuracy of work is degraded. So we need to enhance the features set by removing the unwanted signal using the artificial bee colony algorithm as an optimization technique. To optimize the features set, we have defined a novel objective function and fitness function of ABC algorithm as shown in equation 1.

$$ABC_{ff} = \begin{cases} Bee_{current} if Bee_{current} > Bee_{onlooker} \\ Bee_{onlooker} else \end{cases} \quad ..(1)$$

Where $\mathrm{ABC_{ff}}$ is the output of fitness function and $\mathrm{Bee_{current}}$is the total bee which is called MFCC feature and $\mathrm{Bee_{onlooker}}$ is the threshold value of feature set. The steps of ABC algorithm is given in below;

Upload dataset for Training
Select Case (B, F, L, R and S)
Choose Noise Type
A: Without Noise
B: White Gaussian Noise (WGN)
C: Adaptive WGN
If user=1 (Without Noise)
Speech_signal=load (Speech Data)

$$\mathrm{Speech\_MFCC_{features\{i\}}} = \sum_{i=1}^{n} \mathrm{mfcc}(\ \mathrm{Speech\_signal})$$

Initialize ABC Algorithm
Define - Employed bee
- Onlookers bee and
- Scouts bee
Set objective function:

$$ABC_{ff} = \begin{cases} Bee_{current} & if \quad Bee_{current} > Bee_{onlooker} \\ Bee_{onlooker} & else \end{cases}$$

$$\mathrm{Opitmized\_MFCC\{i\}} = \sum_{i=1}^{r}\sum_{\mathrm{j}=1}^{\mathrm{c}} \mathrm{ABC}(\ \mathrm{MFCC_{features}, fs, ft})$$

where fs is selected value and ft is threshold value
else if user=2 (WGN Noise)
Speech_signal_WGN=load (Speech Data)

$$\mathrm{Speech\_WGN\_MFCC_{features\{i\}}} = \sum_{i=1}^{n} \mathrm{mfcc}(\ \mathrm{Speech\_signal\_WGN})$$

$$\mathrm{Opitmized\_WGN\_MFCC\{i\}} = \sum_{i=1}^{r}\sum_{\mathrm{j}=1}^{\mathrm{c}} \mathrm{ABC}(\ \mathrm{MFCC_{features}, fs, ft})$$

where fs is selected value and ft is threshold value

else if user=3 (AWGN Noise)
Speech_signal_AWGN=load (Speech Data)

$$\mathrm{Speech\_AWGN\_MFCC_{features\{i\}}} = \sum_{i=1}^{n} \mathrm{mfcc}(\ \mathrm{Speech\_signal\_AWGN})$$

$$\mathrm{Opitmized\_AWGN\_MFCC\{i\}} = \sum_{i=1}^{r}\sum_{\mathrm{j}=1}^{\mathrm{c}} \mathrm{ABC}(\ \mathrm{MFCC_{features}, fs, ft})$$

Where fs is selected value which is called current bee and ft is threshold value which is called onlooker bee. The best optimized is called scout bee which is the optimal feature from the MFCC feature sets.

end

Where speech_signal is the speech data which is uploaded by users and Opitmized_AWGN_MFCC is the

optimized features set which is used in the training of proposed system as an input of FFBPN algorithm.

3.3 FFBPN algorithm

A Feed forward back propagation neural network (FFBPN) is an authoritative machine learning technique from the field of deep learning. FFBPNs are trained using large collections of optimized features set. From these large collections, FFBPNs can learn prosperous feature representations for a wide range of features. The used algorithm of FFBPN is given as;

Load Opitmized_MFCC_Data
Trainingdata = Opitmized_MFCC_Data
Initialize FFBPN
Generate group of data = group
Set iteration = 1000
for i = 1: iteration
Weight = Opitmized_MFCC_Data(i)
Hidden_Layer = [25, 25, 25] (tansig)
Net_algo = trainrp
Generate Net structure of FBPNN (net)
Net_Output = train (net, Trainingdata, group)
end

We have saved the Net_Output as a training data and simulated with test data and appropriate results are calculated with feed forward back propagation neural network. The Net_Output depends on the training data of network and it contains the categories of data which is used in the classification stage of proposed work. In the training phase we have considererd the 25 neurons in each hidden layers with tan sigmoid transfer function. This is used as a carrier of signal from one layer to another layer of FFBPN. Each layer of FFBPN produces a response, or activation, to an input feature. However, there are only a few layers within a FFBPN that are suitable for feature training. Here we have set the 1000 iteration for the training of input data based on the performance criteria of FFBPN. In the each iteration FFBPN adjust the weight of input feature and create a structure of output according to the defined group at the time of initialization of network.

4 Proposed system

We have proposed an integrated speaker and speech recognition system as shown in figure 1.

Above figure represent the flow diagram of integrated speaker and speech recognition system based on the artificial intelligence concept. In the speaker and speech recognition system, firstly features are extracted from the speech signal. These features should be robust to noise and efficient enough so that classification can discriminate between the speakers and words. We have used MFCC features and to gain in accuracy, optimization of these features is done by ABC algorithm. Then classification is done with feed forward back propagation neural network. The FFBPN output is what is recognized i.e. who is speaking and what he or she is speaking. The command word is given through the RF transceiver to the microcontroller. The MCU (ATMEGA 8) interprets the commands received and accordingly motor is controlled through driver circuit (L293D) to move the wheel chair.

4.1 Feature extraction using MFCC algorithm

In this section, we have described the MFCC algorithm which is used to find out the feature set from the speech signal with respect to the speaker. The algorithm of MFCC is given below.

Firstly Initialized parameters
Tw = 25(analysis frame duration (ms))
Ts = 10 (analysis frame shift (ms))
Alpha = 0.97(pre emphasis coefficient)
R = [300 3700] (frequency range to consider)
M = 20 (number of filter bank channels)
C = 13 (number of cepstral coefficients)
L = 22 (cepstral sine lifter parameter)
Hamming=((N)(0.54-0.46*cos(2*pi*[0:N-1].'/(N-1))))

$$\mathrm{MFCC}_{\mathrm{features\{i\}}} = \sum_{i=1}^{n} \mathrm{MFCC}(\ \mathrm{Signal, fs, Tw, Ts, Alpha, Hamming, R, M, C, L})$$

Where Signal is the speech data which is uploaded by user and MFCC_features are the extracted feature set from the uploaded speech data.

4.2 Optimization with ABC algorithm

The probability of unwanted signals are more in extracted features and due to the unwanted signal the accuracy of work is degraded. So we need to enhance the features set by removing the unwanted signal using the artificial bee colony algorithm as an optimization technique. To optimize the features set, we have defined a novel objective function and fitness function of ABC algorithm as shown in equation 1.

$$ABC_{ff} = \begin{cases} Bee_{current} if\, Bee_{current} > Bee_{onlooker} \\ Bee_{onlooker} else \end{cases} \quad ..(1)$$

Where ABC_{ff} is the output of fitness function and $Bee_{current}$is the total bee which is called MFCC feature and $Bee_{onlooker}$ is the threshold value of feature set. The steps of ABC algorithm is given in below;

Upload dataset for Training
Select Case (B, F, L, R and S)
Choose Noise Type
A: Without Noise
B: White Gaussian Noise (WGN)
C: Adaptive WGN
If user=1 (Without Noise)
Speech_signal=load (Speech Data)

$$\text{Speech\_MFCC}_{\text{features\{i\}}} = \sum_{i=1}^{n} \text{mfcc(Speech\_signal)}$$

Initialize ABC Algorithm
Define - Employed bee
- Onlookers bee and
- Scouts bee
Set objective function:

$$ABC_{ff} = \begin{cases} Bee_{current} & if \quad Bee_{current} > Bee_{onlooker} \\ Bee_{onlooker} & else \end{cases}$$

$$\text{Opitmized\_MFCC\{i\}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \text{ABC(MFCC}_{\text{features}}, \text{fs, ft)}$$

where fs is selected value and ft is threshold value
else if user=2 (WGN Noise)
Speech_signal_WGN=load (Speech Data)

$$\text{Speech\_WGN\_MFCC}_{\text{features\{i\}}} = \sum_{i=1}^{n} \text{mfcc(Speech\_signal\_WGN)}$$

$$\text{Opitmized\_WGN\_MFCC\{i\}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \text{ABC(MFCC}_{\text{features}}, \text{fs, ft)}$$

where fs is selected value and ft is threshold value

else if user=3 (AWGN Noise)
Speech_signal_AWGN=load (Speech Data)

$$\text{Speech\_AWGN\_MFCC}_{\text{features\{i\}}} = \sum_{i=1}^{n} \text{mfcc(Speech\_signal\_AWGN)}$$

$$\text{Opitmized\_AWGN\_MFCC\{i\}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \text{ABC(MFCC}_{\text{features}}, \text{fs, ft)}$$

Where fs is selected value which is called current bee and ft is threshold value which is called onlooker bee. The best optimized is called scout bee which is the optimal feature from the MFCC feature sets.
end
Where speech_signal is the speech data which is uploaded by users and Opitmized_AWGN_MFCC is the optimized features set which is used in the training of proposed system as an input of FFBPN algorithm.

4.3 FFBPN algorithm

A Feed forward back propagation neural network (FFBPN) is an authoritative machine learning technique from the field of deep learning. FFBPNs are trained using large collections of optimized features set. From these large collections, FFBPNs can learn prosperous feature representations for a wide range of features. The used algorithm of FFBPN is given as;

```
Load Opitmized_MFCC_Data
Trainingdata = Opitmized_MFCC_Data
Initialize FFBPN
Generate group of data = group
Set iteration = 1000
for i = 1: iteration
}Weight = Opitmized_MFCC_Data(i)
}Hidden_Layer = [25, 25, 25] (tansig)
Net_algo = trainrp
Generate Net structure of FBPNN (net)
Net_Output = train (net, Trainingdata, group)
end
```

We have saved the Net_Output as a training data and simulated with test data and appropriate results are calculated with feed forward back propagation neural network. The Net_Output depends on the training data of network and it contains the categories of data which is used in the classification stage of proposed work. In the training phase we have considererd the 25 neurons in each hidden layers with tan sigmoid transfer function. This is used as a carrier of signal from one layer to another layer of FFBPN. Each layer of FFBPN produces a response, or activation, to an input feature. However, there are only a few layers within a FFBPN that are suitable for feature training. Here we have set the 1000 iteration for the training of input data based on the performance criteria of FFBPN. In the each iteration FFBPN adjust the weight of input feature and create a structure of output according to the defined group at the time of initialization of network.

5 Experiments and results

In this section, the simulation results and analysis of proposed work is described. The speech is acquired by sound recorder with the help of headphone at 16 KHz frequency at room environment in mono format. In our case, database is prepared for four speakers of age 27-34, two females (F1, F2) and two males (M1, M2). The words recorded are 'Forward, Backward, Left, Right, Stop'. Each word is recorded 80 times and hence 400 words are recorded for each speaker creating a database of 1600 words. It is much more difficult to recognize speech in presence of noise. Proposed work is tested on various types of noises like White Gaussian Noise (WGN), Adaptive White Gaussian Noise (AWGN) etc. We have tested our system on TIDIGITS database and our own created database. For the all types of signal, we have extracted the features and then optimized them to enhance the features set. After the optimization, we have trained the features with FFBPN.

In the training phase, we have used the set of 25 neurons in each hidden layers with tan sigmoid transfer function to train the input feature data. After the training, we have tested the simulation with a test speech signal and process is repeated for testing phase.

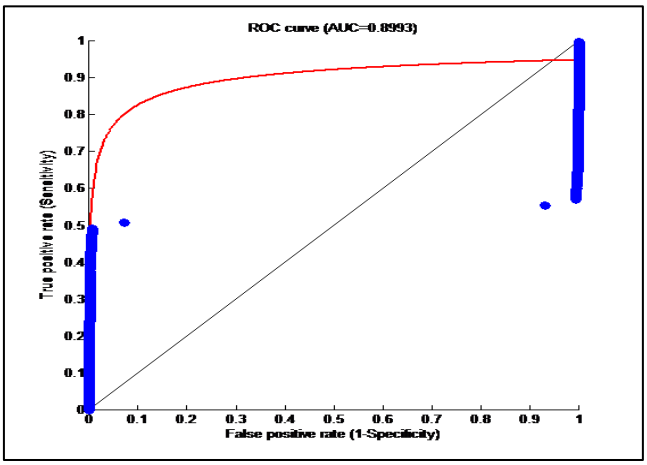


Figure 2: ROC Curve for Proposed Work.

Figure 2 shows the receiver operating characteristics (ROC) curve of proposed speech and speaker recognition system. It is a graphical method for comparing two empirical distributions where x-axis denotes the false positive rate and y-axis denotes the true positive rate. On the basis of ROC curve we have calculated the probability of recognition accuracy using the area under curve (AUC), which varies from 0 to 1. Form the figure 2, the AUC value is 0.8993 which indicates the training to system is good, therefore better classification rate. Table 1 shows the recognition accuracy for four persons including 2 men and 2 women. We have compiled the results for four persons using TIDIGITS dataset.

Speech Signal (Words)	Man 1	Man 2	Woman 1	Woman 2
One	96.88	97.11	97.35	96.20
Two	95.38	94.27	95.00	96.24
Three	99.42	98.30	97.91	99.35
Four	98.32	99.14	98.10	98.65
Five	99.16	97.88	99.02	99.35
Six	96.88	97.11	97.35	96.26
Seven	95.38	94.27	95.01	96.28
Eight	99.42	98.30	97.91	99.35
Nine	98.32	99.14	98.13	98.65
Zero	99.16	97.88	99.02	99.35

Table 1: Accuracy of integrated speaker and speech recognition for different isolated words (Clean Environment).

Figure 3 shows the accuracy of integrated speaker and speech recognition work for the digit database in clean environment. The average accuracy is more than 97% in clean environment.

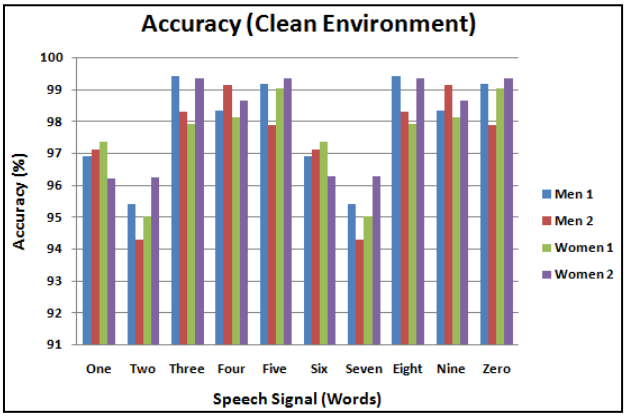


Figure 3: Accuracy of integrated speaker and speech recognition in clean environment.

Speech Signal (Words)	Man 1	Man 2	Woman 1	Woman 2
One	88.76	90.93	91.17	90.08
Two	89.26	88.09	88.82	90.03
Three	93.26	92.12	91.79	93.17
Four	92.19	92.96	91.92	92.47
Five	92.91	91.74	92.85	93.84
Six	90.75	90.93	91.19	90.56
Seven	89.21	88.09	88.87	90.07
Eight	93.24	92.15	91.72	93.17
Nine	92.14	92.99	91.92	92.45
Zero	92.98	91.72	92.84	93.17

Table 2: Accuracy of integrated speaker and speech recognition for different isolated words (Noisy Environment).

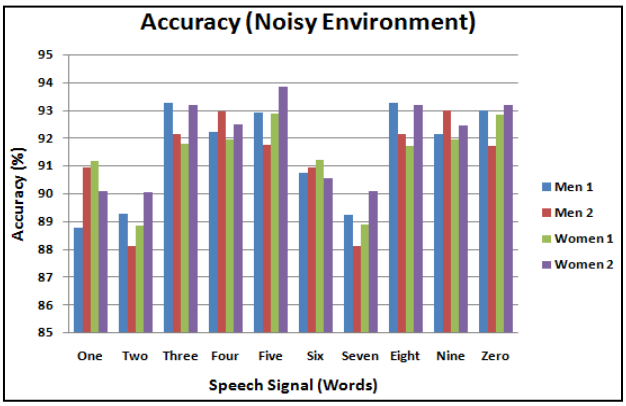


Figure 4: Accuracy of integrated speaker and speech recognition in noisy environment.

Figure 4 and table 2 shows the achieved accuracy in noisy environment. Speech signal is corrupted by adding White Gaussian Noise and then accuracy is measured. The average accuracy is more than 91% in noisy environment.

Table 3 shows the comparison of two methods: One with MFCC only and another is with MFCC and ABC

algorithm on our own created algorithm using FFBPN as a classifier.

No. of Iterations	Proposed work using MFCC	Proposed work using MFCC with ABC Algorithm
1	92.85	97.64
2	94.06	98.94
3	90.68	97.47
4	89.84	95.94
5	91.64	97.69
Average (%)	**91.82**	**97.53**

Table 3: Accuracy of integrated speaker and speech recognition for own created database.

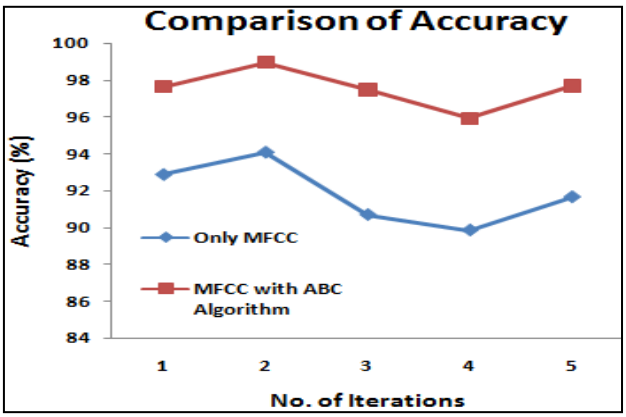


Figure 5: Comparison of accuracy.

In the figure 5 the comparison of accuracy for proposed work using optimization and without optimization is given. The accuracy is better for optimization case. So in integrated speaker and speech recognition, optimization is a better tool to create a unique feature set.

Further, we have tested our system in real time scenario for the movement of wheelchair. The command from MATLAB software is received using RF data modem. It works on 2.4 GHz frequency with adjustable baud rates of 9600 /115200 for direct interfacing with MCU. The MCU (ATMEGA 8) interprets the commands received and accordingly motor is controlled through driver circuit (L293D). Programming for MCU (ATMEGA 8) is done on ARDUINO Compiler. We have achieved average 87.4% accuracy for five isolated words in different environments like lab, canteen, office etc.

6 Conclusion

In proposed work, we have presented that speaker as well as speech recognition system with MFCC, ABC and FFBPN is helpful in achieving more accuracy. To be specific, we have found that optimization and feature extraction are very important as well as difficult steps in any pattern recognition system. In proposed work, we have extracted more useful feature set from speech signal using MFCC technique, feature optimization using ABC optimization algorithm and for the training and classification of data, FFBPN is used. The experimental results analyzed that proposed method using MFCC with ABC algorithm provides good results with 97% of accuracy and it is 6% more than without using optimization technique. In real time scenario, average accuracy achieved is 87.4%.

7 References

[1] Cutajar M., Micallef J., Casha O., Grech I., and Gatt E. Comparative study of automatic speech recognition techniques, IET Signal Processing, 7(1): 25–46, 2013. http://dx.doi.org/10.1049/iet-spr.2012.0151

[2] Kaur, G., Srivastava, M., and Kumar, A. Analysis of feature extraction methods for speaker dependent speech recognition, International journal of engineering and technology innovation, 7(2):78–88, 2017.

[3] Ijjina E. P. and Mohan C. K. Human action recognition using genetic algorithms and convolutional neural networks, Pattern recognition, 59: 199–212, 2016. http://dx.doi.org/10.1016/j.patcog.2016.01.012

[4] Ijjina E. P. and Mohan C. K. Hybrid deep neural network model for human action recognition, Applied soft computing, 46: 936–952, 2015. https://doi.org/10.1016/j.asoc.2015.08.025

[5] Karaboga D., and Akay B. A comparative study of Artificial Bee Colony algorithm, Applied mathematics and computation, 214(1): 108–132, 2009. https://doi.org/10.1016/j.amc.2009.03.090

[6] Bolaji A. L., Khader A. T., Al-Betar M. A., and Awadallah M. A. Artificial bee colony algorithm, its variants and applications: A survey, Journal of theoretical and applied information technology, 47(2): 434–459, 2013. https://doi.org/10.1504/IJAIP.2013.054681

[7] Chandra B., and Sharma R. K. Fast learning in deep neural networks, Neuro-computing, 171: 1205–1215, 2016. https://doi.org/10.1016/j.neucom.2015.07.093

[8] Li K., Wu X., and Meng H. Intonation classification for L2 English speech using multi-distribution deep neural networks, computer speech & language, 43: 18–33, 2017. https://doi.org/10.1016/j.csl.2016.11.006

[9] Richardson F., Member S., Reynolds D., and Dehak N. Deep neural network approaches to speaker and language recognition, IEEE signal processing letters, 22(10): 1671–1675, 2015. https://doi.org/10.1109/LSP.2015.2420092

[10] Dahl G. E., Yu D., Deng L., and Acero A., Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, 20(1): 30–42, 2012. https://doi.org/10.1109/TASL.2011.2134090

[11] Solera-Urena R. and Garcia-Moral A. I.Real-time robust automatic speech recognition using compact

support vector machines, Audio speech and language processing, 20(4): 1347–1361, 2012. https://doi.org/10.1109/TASL.2011.2178597

[12] Mohamad D., and Salleh S. Malay isolated speech recognition using neural network : a work in finding number of hidden nodes and learning parameters, International Arab journal of information technology, 8(4): 364–371, 2011.

[13] Desai Vijayendra A., and Thakar V. K. Neural network based Guajarati speech recognition for dataset collected by in-ear microphone, Procedia computer science, 93: 668–675, 2016. https://doi.org/10.1016/j.procs.2016.07.259

[14] Abdalla O. A., Zakaria M. N., Sulaiman S., and Ahmad W. F. W. A comparison of feed-forward back-propagation and radial basis artificial neural networks: A Monte Carlo study," Proceedings 2010 International Symposium on Information Technology , 2: 994–998, 2010. https://doi.org/10.1109/ITSIM.2010.5561599

[15] Chen X., Liu X., Wang Y., Gales M. J. F., and Woodland P. C. Efficient training and evaluation of recurrent neural network language models for automatic speech recognition, IEEE/ACM Transactions on audio speech and language processing, 24(11): 2146–2157, 2016. https://doi.org/10.1109/TASLP.2016.2598304

[16] Simpson, R.C. et al. NavChair : An assistive wheelchair navigation system with automatic adaptation, Assistive technology and artificial intelligence, 1458: 235–255, 1998

[17] Pacnik, G., Benkic, K. and Brecko, B. Voice operated intelligent wheelchair - VOIC," IEEE international symposium on industrial electronics, 1221–1226, 2005. https://doi.org/10.1109/ISIE.2005.1529099

[18] Jabardi, M.H., "Voice controlled smart electric-powered wheelchair based on artificial neural network," International journal of advanced research in computer science, 8(5): 31–37, 2017. https://doi.org/10.26483/ijarcs.v8i5.3650

[19] Siniscalchi S. M., Svendsen T., and Lee C.-H. An artificial neural network approach to automatic speech processing, Neurocomputing, 140: 326–338, 2014. https://doi.org/10.1016/j.neucom.2014.03.005

[20] Hossain A., Rahman M., Prodhan U. K., and Khan F. Implementation of back-propagation neural network for isolated Bangla speech recognition, International journal of information sciences and techniques, 3(4): 1–9, 2013.

[21] Mansour A. H., Zen G., Salh A., Hayder H., and Alabdeen Z. Voice recognition using back propagation algorithm in neural networks," International journal of computer trends and technology, 23(3): 132–139, 2015.

[22] Qian Y., Tan T., and Yu D. Neural network based multi-factor aware joint training for robust speech recognition, IEEE/ACM transactions on audio speech and language processing, 24(12): 2231–2240, 2016. https://doi.org/10.1109/TASLP.2016.2598308

[23] Dede G. and Sazlı M. H. Speech recognition with artificial neural networks, Proceedings of the annual conference of the international speech communication association, INTERSPEECH, 20(3), 763–768, 2015. https://doi.org/10.1016/j.dsp.2009.10.004

[24] Shahamiri S. R. and Binti Salim S. S. Real-time frequency-based noise-robust automatic speech recognition using multi-nets artificial neural networks: A multi-views multi-learners approach, Neurocomputing, 129(5), 1053–1063, 2014. https://doi.org/10.1016/j.neucom.2013.09.040

https://doi.org/10.31449/inf.v42i4.1914 Informatica **42** (2018) 595–605 **595**

Blur Invariant Features for Exposing Region Duplication Forgery Using ANMS and Local Phase Quantization

Diaa Mohammed Uliyan
Middle East University, Faculty of Information Technology, Amman, Jordan
E-mail: diaa_uliyan@hotmail.com

Mohammad A. M. Abushariah
Computer Information Systems Department, King Abdullah II School of Information Technology
The University of Jordan, Amman, Jordan
E-mail: m.abushariah@ju.edu.jo

Ahmad M. Altamimi
Applied Science Private University, Faculty of Information Technology, Amman, Jordan
E-mail: a_altamimi@asu.edu.jo

Keywords: copy-move forgery, image forgery detection, image forensics, local interest points, region duplication, segmented regions

Received: October 17, 2017

In digital image forensics, local interest points can be employed to faithfully detect region duplication forgery. Authentic images may be abused by copy-move forgery to fully contained duplicated regions such as objects. R*ecent existing local interest point forgery detection methods fail to detect this type of forgery in the retouched regions by some geometric transformations. To solve this challenge, local interest points should be detected which cover all the regions with high primitives like corners and edges. These primitives represent the internal structure of any object in the image which makes them have a discriminating property under geometric transformations such as scale and rotation operation. They can be exposed based on Scale-Invariant Features Transform (SIFT) algorithm. Here, we provide an image forgery detection technique by using local interest points. First, the image is segmented based on fuzzy C means to divide the image into homogenous regions that have the same texture. Second, local interest points are exposed by extracting Adaptive non-maximal suppression (ANMS) from dividing blocks in the segmented image to detect such corners of objects. We also demonstrate that ANMS Keypoints can be effectively utilized to detect blurred and scaled forged regions. The ANMS features of the image are shown to exhibit the internal structure of copy moved region. We provide a new texture descriptor called local phase Quantization (LPQ) that is robust to image blurring and also to eliminate the false positives of duplicated regions. Experimental results show that our scheme has the ability to reveal region duplication forgeries under scaling, rotation and blur manipulation of JPEG images on MICC-F220 and CASIA v. 2 Image Datasets.*

Povzetek: Predstavljena je izvirna metoda za odkrivanje ponarejenih področij v sliki.

1 Introduction

In the digital era, it is quite popular for expert users of image editing tools to manipulate images easily. Nowadays, we are facing the abuse of digital image tools, image forgery has begun to crumble the trustworthiness of visual images [12], that seeing is no longer believing. Image forgery has inspired researchers [20] to investigate and check the authenticity of digital images due to its effect to the judgment of the truth of suspected images in many sectors, such as digital newspapers, law evidence, medical documents, etc. Region duplication forgery is one of the most common image editing tools to abuse image. It is a simple operation that gives high visual impact to suspected images. Furthermore, it is known as Copy-move, cloning or region duplication. Copy-move forgery duplicates a region of an image and moves it to another location within the same image. This type of forgery has a good effect which conveys misleading information in order to support an individual agenda.

Some Existing methods are developed to examine and locate Copy-moved regions in a forged image [9, 2]. Some can detect duplicate regions [27, 41, 44] and another can locate multiple duplicated regions [47]. The region duplication forgery detection methods have been categorized and evaluated based on their sensitivity towards two types of attacks: a) Geometrical manipulation attacks and b) Post-processing attacks. For a geometrical attacks, the copy-move detection methods are resilient against spatial domain changes such as rotation [40], scaling [14, 11]. Conversely, some scientific papers have examined the robustness against the retouching or blending tools which hide visual editing artifacts in the

image through some post-processing attacks. Such attacks include: blurring [43, 46], additive noise [38] and JPEG compression [19, 42] impacts are obtained after applying geometrical transformation operations. Hence, this type of forgery is a challenging problem that motivates us to investigate forged images against scale, rotation and blur attacks. As blurring could transform the features of any region in the image, further inspection of this attack should consider [43]. The blur transformation in the image features may also make the standard copy-move forgery detection methods struggle to detect the blurred duplicated regions. The proposed method starts a forensic job by collecting images that contain simple transformation attacks and blur attacks. The original images are collected from the Dataset MICC-F220 [4] and CASIA v2.0 [33]. Then, the proposed method is implemented to combine the Scale Invariant Feature with LPQ matching technique. We then compare the performance of the proposed method by F-scores with state-of-the-art methods: [4, 25, 39] and block-based methods: [3, 24].

The paper is organized into five sections. Section 2 highlights Related Works on copy-move forgery detection per some attacks included. Section 3 introduces the proposed method. In Section 4, it will discuss the experimental results and performance evaluation. In Section 5, the conclusion and future works are summrized.

2 Related works

The common flowchart of most copy-move forgery detection methods has six steps as shown in Figure 1. These steps are: 1) image preprocessing, 2) image division, 3) feature extraction, 4) building descriptor 5) matching and 6) show detection results. The first step is optional, which tries to improve the image content by defeating undesired noise. The most frequent preprocessing step is image color conversion be converting an RGB color image into grayscale image [32] by using the Eq. 1.

$$Grayscale = 0.228\,R + 0.587\,G + 0.114\,B \quad (1)$$

Where R,G and B channels represent the Red, Green and Blue channels as pixel information in the image.

Rafsanjany et al. [17] converts the input RGB image to Gray scale and Lab color space. Then, they divided it into square blocks to extract features. Their method achieved about 90% F-measure for JPEG images with size 512x512. Another color conversion is used such as YCbCr color system to give the luminance information Y or chrominance information Cb and Cr [26]. Shinfeng et al. [21] used YCbCr color system for image conversion and divide it into blocks, for each block, DCT coefficients are extracted to produce 64 bit feature vector. Later, they computed the probability of each block by identifying the period of the it's histogram.

The main goal of the image conversion is to achieve the dimensionality reduction of the image features and extract the distinctive local interest points or visual features. This could help on performance the proposed copy-move forgery detection methods in the aspect of time complexity [13]. Similarly, Hue saturation Value (HSV) color space is used in method [31], which help to detect intense dark duplicated regions or bright regions with around 7.22 % false positive rate.

Based on the way of dividing the image on the second stage of copy-move forgery detection, these techniques are classified into three classes: block-based schemes [34], segmented regions-based schemes [41] and local keypoints based schemes [38]. In the block-based, the image is divided into a number of sub-blocks either square blocking or circle blocking. Similarly, segmented-based method tries to segment the image into different regions that fully covered the forged objects in the image based on color, texture and property palette properties. Conversely, the Keypoint based method detects local interest points to find primitive features in the image. The benefit of this stage is that can minimize the time complexity for matching step in order to search the similar feature vectors of building descriptor in an image compared to exhaustive search.

After image division, the feature extraction can help to choose the relevant data that exhibit the internal structure and its properties in the image. These features are saved into a feature vector. Finally, matching between two feature vectors is employed using the distance of the nearest neighbours from all points in the feature space to show forged regions.

Based on Copy-move forgery detection steps, common schemes focused on image division and feature extraction steps that exhibit invariant features against

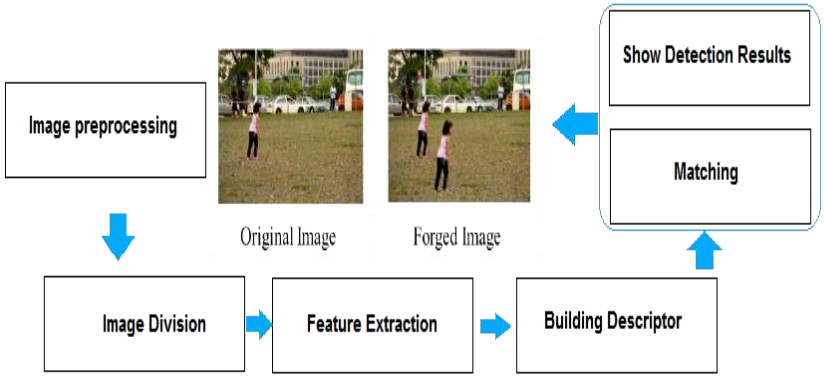


Figure1: The basic flowchart of standard copy move forgery detection schemes [38].

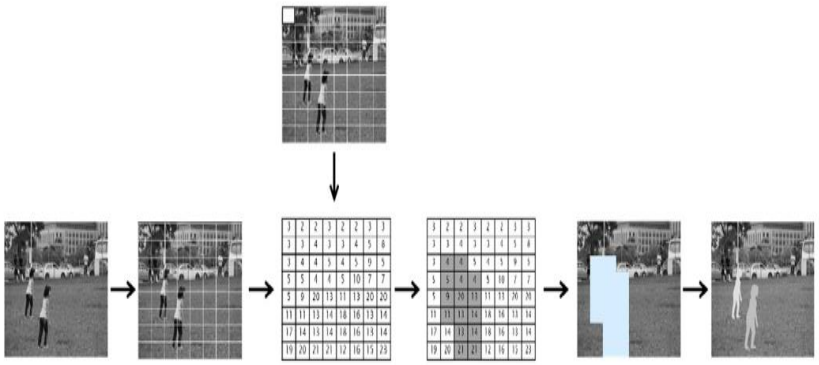


Figure 2: The image is divided into 8x8 blocks, features are highlighted and saved for matching process.

geometric transformation and post-processing attacks. These schemes are introduced in details [37] as follows:

I. **Block-Based Methods** divide the image into square or circle blocks to extract features from these blocks as shown in Figure 2. The main advantage of this approach is that give high detection accuracy for the textured forged regions. But, it still gives high computational complexity due to exhaustive search between divided blocks in the image [34].
II. **Segmented-Based Methods** Segment the input image into homogenous regions based on color or texture. This approach works well in the forged images that have duplicated objects [10].
III. **Keypoint-Based Methods** discard block division step and use local interest point detectors to extract features. These features are distinctive to represent corners, edges or blobs in the image. Then, a robust texture descriptor is built to increase a reliability against geometric transformation attacks [37].

Different types of attacks have been considered in existing methods for detecting region duplication forgery. These methods are called Passive methods due to detecting image forgery without requiring explicit prior information. The main goal is to analyze the history of the image tampering blindly by examining pixel-level correlations [35].

In this article, popular feature extraction methods in copy-move forgery detection methods were covered for various geometric transformations and post-processing attacks. The robustness of detection methods depends on invariant features to possible attacks as pointed in [9]. Copy-move forgery detection methods based on type of features are classified into two classes: Frequency transform methods [16], Texture and intensity based methods [42].

A. Frequency Transform Methods convert the image pixel information into frequency domain to extract high frequency coefficients form the image. This approach is robust to JPEG compression and can detect duplicated regions with a large size 128 x 128 pixel. The limitations are the high computational complexity and struggle to detect duplicated regions with scale and rotation attacks. The frequency features are: Discrete cosine transform (DCT) [47], Fourier Transform (FT) [37], Discrete wavelet transform (DWT) [27], Curvelet Transform (CT)[1] and Wiener Filter. The limitation of this approach is that features are sensitive to blur attack.

B. Texture And Intensity Based Methods extract features that exhibit image texture regions with the smoothness property. Various features have been used to detect textured duplicated regions in copy-move forgery detection methods for instance, Local binary Patterns (LBP), Histogram of Gradient (HOG), Zernike moments (Zm) [36] which is robust to rotation, log polar transform [28] that detects rotated duplicated regions, Principle component analysis (PCA) and Singular value decomposition (SVD) that reduce the size of feature vector to enhance the time complexity.

All of these methods that utilize frequency and texture features were employed in block-based methods and did not suppose that forged regions may be geometrically transformed. Another direction has been discovered to detect duplicated regions against scaling and rotations.

This can be done by keypoint-based approach for instance, Scale invariant transform features (SIFT), speed up robust features (SURF) [5] and Harris features. These features are slightly blur invariant. This motivates us to develop a blur invariant detection method to detect blurred forged duplicated regions in the suspected images.

Blurring is made effectively through image forgery process suing averaging of neighbor pixels in a square block [49]. The blur is commonly applied by Gaussian, defocus and motion blurs. In practice, the Gaussian blur filter is well known by users that do tampering in the image due to it's simplicity. If the duplicated region is retouched by blur, then the main features of the blurred region are minimized and details cannot be seen.

Blurring on forged regions aims to manipulate region's information and assists hiding retouch and blending artifacts. As a result, blurring allow the duplicated region to be consistent with its surrounding area. The scope of locating tampered regions attacked by blurring artifact is even smaller. Only few related papers have been discovered that deal with blur attack [5, 15, 49, 46, 23, 18].

The first attempt was made by [23] to detect burred duplicated region forgery. The extracted blur invariant moments from image blocks. Then, principal Component

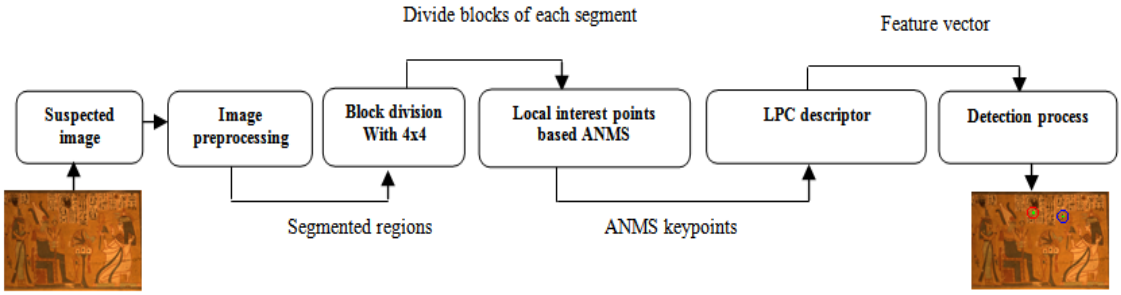


Figure 3: The Flowchart of the proposed forensic detection scheme.

Analysis was employed to achieve the dimensionality reduction of feature vectors, finally, they, used a kd tree to locate the duplicated regions. The weakeness of their method is that struggle to detect uniform duplicated regions and also gives high false positives. Another blur detection method is developed by Zhou et al. [49] for revealing blurred edges in the duplicated regions. Their method starts by preprocessing step to convert the image into binary one. Then, the method applied edge preserving–smoothing filters, followed by a mathematical morphology operation using the erosion filter to expose forged duplicated area with malicious blurred edges. The average accuracy rate about 89.26% in images with blurred edges manually attacked by the Gaussian noise filter. Zheng et al. [48] located tampered regions with blur attack via wavelet homomorphic filtering to represent pretty high frequency edges. Then, erosion operation was applied to expose blurred edges in forged reigon from normal regions which effectively reduced the false positive rates. Wang et al. [5] used non sub sampled contourlet transform (NCST) to examine manually blurred edges from duplicating regions. The detection of forged duplicated regions is done using support vector machine (SVM). In [46], blur artifacts were explored in forged regions by using combined blur and affine transform moments. The relative detection error was employed to estimate the stability of local invariant features deformed by Gaussian and motions blurs. The method achieved high accuracy rate with small feature vector. Guzin et al. [45] applied Object Removal operation from Uniform Background Forgery by adapting accelerated diffusion filter (AKAZE). The Local binary difference descriptor was built in AKAZE features which are scale invariant features. The size of feature vector is 486 bits. The performance of their method in terms of TPR is 85.74%, 71.35% and 76.73% against Gaussian blurring, rotation and JPG compression respectively.

The paper proposed a region duplication forgery detection scheme based on ANMS features and LPQ texture descriptor. In this paper, a part of the authentic image is copied and pasted to another area to mislead the semantic visual meaning of the image. While copy-move operation is applied, the duplicated region may be post-processed using rotation, scaling, blurring to create better forgery. The common pipeline of the proposed method is, first the input image is segmented-based on color features. Fuzzy C-means method is used to cluster and label the segments in the image. The centroid of each segment is located in the image. We assume that forgery is made by for small regions. These regions can be detected by calculating the least frequent occurrence of labeled segments in the image. For each candidate segment, ANMS local interest points are extracted. ANMS features are scale invariant to represent the structure of segmented region. Second, each segment is split into 4 blocks, the size of the block is 4 x 4. The distribution of ANMS points the blocks of each segment contributes to detect duplicated regions against rotation. Third, blur invariant LPQ descriptor is built to the approximation of the ANMS points in each segment. Finally, the closest local keypoint search of features between two segments is employed by Generalized Nearest neighbor (G2NN) to improve the performance of our method in terms of True positive rate (TPR) and false positive rate (FPR).

3 Proposed method

In this section, we introduce in details the flowchart of the proposed method for exposing the copy-move forgery, with scaling and blurring of the cloned region. Our contribution is proposing a forensic keypoint-based method for blur and scale invariant copy-move forgery detection in digital images. A diagram representing the workflow of the proposed technique is shown in Figure 3.

3.1 Image preprocessing: color image segmentation

Image segmentation is the one of the most important techniques for image analysis and object detection [8]. The main aim of Segmentation of our method is to perform an efficient search strategy to detect duplicated regions such objects in the image. It starts from coarse search to quickly split an image into homogeneous objects based on discontinuity and similarity of image intensity values. Then a feature extraction is applied to these query regions to improve the TPR of copy-move forgery detection. The proposed color segmentation approach, followed by Fuzzy C-means clustering (FCM) is introduced in [7]. The fuzzy C-means is an unsupervised technique which estimate the RGB channel of every pixel in the image and compare it with the centroid of the cluster. It makes a decision about which category the pixel should relate to. Each pixel in the image should be in [0-1], which the value describes how much pixel value relates to its cluster. A fuzzy

membership criterion denotes that the sum of the membership value of a pixel to all clusters equals 1. The FCM clustering is an iterative optimization that minimizes the cost function which is described as follows:

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{m} |p_i - v_k|^2 \quad (2)$$

Where, an image I with n pixels to be partitioned into c clusters, p_i represents the i^{th} image pixels. μ_i is the fuzzy membership value with fuzziness factor k >1. Here, the membership function μ_i with the centroid of K^{th} cluster v_k is defined as follows:

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left(\frac{|p_i - v_k|}{|p_i - v_l|}\right)^{2/m-1}} \quad (3)$$

$$v_k = \frac{\sum_{i=1}^{n} \mu_{ik}^{m} p_i}{\sum_{i=1}^{n} \mu_{ik}^{m}} \quad (4)$$

Here, v_k denotes to the centroid of the k^{th} cluster and $|p_i - v_k|$ refers to the Euclidean distance between two points: p_i and v_k. By using the cluster information (c=5, maximum number of iterations=10) and the pixel information pi from the forged image I with size 512 x 512, the homogeneous regions including copy-moved regions can be extracted as shown in Figure 4.

Consequently, each segment is split into 4 non overlapping blocks of b x b pixels, where $b = 4$ as shown in Figure 4. We introduce below, the process of extracting features from these blocks to exhibit the internal structures of segments and achieve rotation invariance.

3.2 Adaptive Non Maxima Suppression (ANMS) features

Keypoint-based methods are significantly helpful in detecting visual objects in the image. While the block-based shemes split the image into blocks, keypoint-based schemes identify and highlight only regions with high entropy, called the local interest points or keypoints. However, keypoints such as SIFT are robust against geometric transformations such as scaling. Hence, the major drawback is that keypoints may be insufficient or even none in the forged region of uniform texture. To avoid the drawback in SIFT based methods, we adopt the ANMS method which is an effective approach suggested by Brown, Szeliski, & Winder [6] to select uniformly distributed interest points for instance,

$K = \{K1, K2, \ldots, K_m | K \in (\mu_{K_m}, V_{K_m})\}$ in image and provide the stability and good performance in scale and rotation through detection of duplicated regions. The principal of ANMS is to select $K_m \in K$, K_m is the maximum neighborhood of region of interest with radius r pixels. K are generated from Harris corners can be described in Equation 5:

$$E(\mu, v)|_{(x,y)} = \sum w(x,y)[I(x+u, y+v) - I(x,y)]^2 \quad (5)$$

Where w(x,y) denotes a Gaussian kernel defined below and (u,v) is the minimal Euclidean distance.

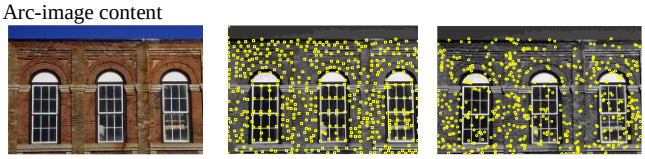
$$w(x,y) = \exp\left(-\frac{1}{2}(u^2 + v^2) \Big/ \sigma^2\right) \quad (6)$$

Where σ is the Standard Deviation. Then, Taylor series expansion is employed to the Equation of $E(\mu, v)$ to eliminate the weak interest points as follows:

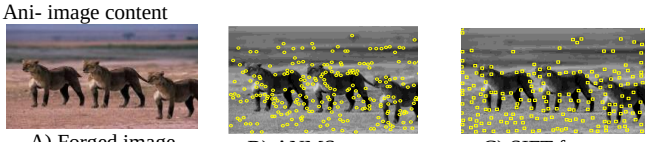
A B C

Figure 4: A) Original image, B) suspected image with duplicated regions and C) Segmented image using the FCM algorithm.

Arc-image content



Ani- image content



A) Forged image B) ANMS C) SIFT features

Figure 5: Keypoints detected from Forged images in column (A) by B) ANMS method and C) SIFT method.

$$A = w.I_{x^2}, B = w.I_{y^2}, C = w.I_x \quad (7)$$

Here, . denotes the image convolution operator. I_x, I_y are the horizontal and vertical directions in the image *I*. a corner response measure is defined as follows:

$Z = \det(V) - \alpha \times tr^2(V)$,

where $V = \begin{bmatrix} A & C \\ C & B \end{bmatrix}$ (8)

V is a matrix has two eigenvalues. tr is the trace of a matrix and $\alpha = 0.06$ in our method. Figure 5 shows the results obtained by the ANMS compared with the SIFT based method [22]. ANMS points are much better distributed in the image and represent the structure of windows object by local interest points such as corners. In Figure 5, two types of images are regarded: a)Arc - architecture content and b) Ani - animal content.

3.3 Local Phase Quantization (LPQ) descriptor

Ojansivu et al. [30] proposed a blur invariant method to extract phase information in the Fourier transform domain and consider only the best energy of sampling low frequencies varying with blur changes. The blurring process in LPQ is applied by convolving the image with a Point Spread Function (PSF) which is defined as follows:

$$g(x,y) = (f * h)(x,y) + n(x,y) \quad (9)$$

Where, where g(x, y) denotes blurred image, f(x, y) represents the original image, h(x, y) is the PSF of blur and n(x, y) is the additive noise. Here * is the image convolution operator. In terms of frequency domain, the Equation 9 is converted to:

$$G(u,v) = (F * H)(u,v) + N(u,v) \quad (10)$$

Where G(u,v), F(u,v) and H(u,v) dentote to the discrete Fourier transforms (DFT) of the blurred PSF image g(x,y), the original image f(x,y) and the PSF h(x,y), respectively. u,v are frequency coefficients in the blurred image. After

applying the Fourier transform, the image will have two parts: the real part $Re(u,v)$ and imaginary part $Im(u,v)$. Only real valued will be kept as follows:

$$G(u,v) = |Re\{F(u,v)\}| + |Im\{F(u,v)\} \quad (11)$$

Real valued parts are quantized based on scalar quantizer as follows:

$$q_i = \begin{cases} 1, & if\ Re_i(u,v) \geq 0 \\ 0, & otherwise \end{cases} \quad (12)$$

Here $q_i$is the i[th] component of $Re(u,v)$. The quantized coefficients are integer values between 0-255.

Finally, LPQ descriptor, which is similar to Local binary pattern (LBP) [42] and is calculated as follows:

$$LPQ(x,y) = \sum_{j=1}^{j=8} q_i(x,y)^{\,2^{\,j-1}} \quad (13)$$

In Figure 6, an example of the computing LPQ for sample images from CASIA dataset and the duplicated regions are clearly recognized.

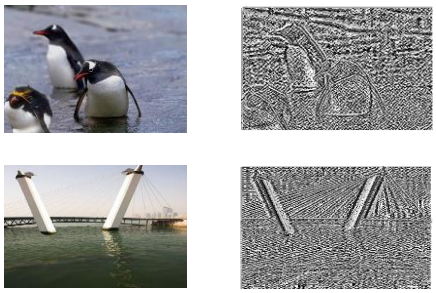


Figure 6: LPQ descriptor of sample images.

3.4 Forgery localization process

As discussed above, keypoints for each segmented region are extracted by ANMS. The LPC descriptor for each segment in the image was calculated to do matching between keypoints and discover the duplicated regions. The best matching between keypoints is founded by generalized nearest neighbor (G2NN) [4]. In G2NN, a ratio between closest keypoint $d_i$with the second nearest neighbor d_{i+1} is calculated as follows:

$$d = \frac{d_i}{d_{i+1}} \leq T, \quad T \in [0,1] \quad (14)$$

Where d is Euclidean Distance, T is threshold value=0.89 in our experiments. x denotes the value on which the iterative procedure G2NN stops, then every keypoint related to a calculated distance in $\{d1, d2, d3, d4 \ldots\ldots dx\}$ satisfies $1 \leq x < n$, is regarded to be matched for keypoint. However, to search the similarity between two local keypoints, simply the proposed method evaluates the distance between two descriptors with respect to a global threshold *T*.

4 Experimental results

The performance of the blur invariant detection method was examined through a set of forged images were collected from two standard datasets, namely MICC-F220 and CASIA v2. Firstly, we introduce the experimental setup of our method and performance evaluation metric where used on detecting duplicated regions. These regions have repetitive texture patterns which are required to make a convinced forgery via post-processing operation such as blurring and scaling. Then, the proposed method is evaluated with existing methods developed in [4], [10] and [39]. The details of the experiments are discussed below.

4.1 Evaluation metric

Our method is developed by MATLAB R2014a on Intel Core i5 processor, with 16 GB memory. The forged images under copy move forgery were collected from the first Dataset MICC-F220 which are produced by a well-known copy-move forgery detection method [4]. It consists of digital images from the Columbia photographic image repository [29] and their personal collection. MICC-F220 includes of 220 images with various sizes from 722 x 480 to 800 x 600 pixels. The size of the duplicated regions conceal about 1.2% of the whole image. The second Dataset (CASIA v2) has about 5123 forged images in JPEG Format with various quality factors. The image resolutions is varying from 240×160 to 900×600. A duplicated region on these images was copied and moved with considering the post-processing after copy move operation to finish the fake image generation; simple post-processing attacks comprising scaling, rotation, blurring, JPEG compression and additive noise.

Here, A Gaussian blur filter is applied in duplicated pattern regions. The similarity threshold is set experimentally to $T = 0.8$ which give a high detection rate. The performance of the proposed detection scheme is evaluated via True Positive Rate (T_{PR}) and False Positive Rate (F_{PR}). The evaluation metric is defined to include others: True positive (TP), True negatives (TN), False positives (FP), False negatives (FN) and F-score calculated as follows:

$$F_{score} = \frac{2Tp}{2Tp+FN+FP} \quad (15)$$

$$T_{PR} = \frac{\text{No.of detected images as forged being forged}}{\text{No.of forged images}} \quad (16)$$

$$F_{PR} = \frac{\text{No.of detected images as forged being original}}{\text{No.of original images}} \quad (17)$$

Where TP is the number of exposed forged images, FN is undetected forged images and FP is incorrectly detected original images.

4.2 Region duplication Forgery detection without attacks

Normal forgery is defined as creating a forged image without applying any attacks to the original part or on the whole image. In Figure 7, the small car has been copied and pasted to another area of the image without applying any attack on the original part, as results illustrate our method has better detection results compared with SIFT based method [4]. This is due to number of local keypoints detected by the ANMS directly improving the detection rate in the image. Here, the number of keypoints detect by our method in the Car image is 70 while other method detects 50 keypoints only. More keypoints are selected means better performance in terms of T_{PR}. However, it will

spend much time than Sift based method. The average detection time of the proposed method is about 13.8 seconds.

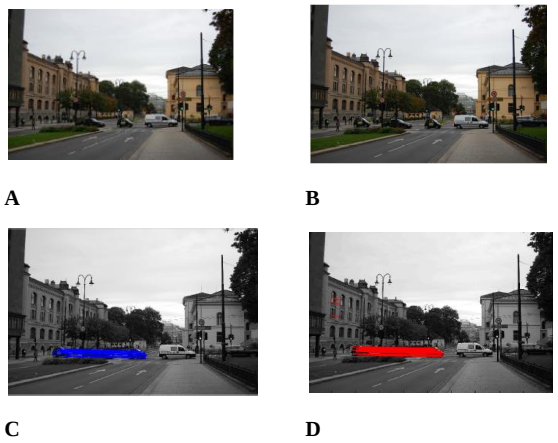


Figure 7: (A) Original image, (B) Forged image with Normal forgery, (C) Detection result of our method with T_{PR}=96%, (D) Detection result of SFIT based method with T_{PR}=94% and F_{PR}=7%.

4.3 Scale attacks

To examine the proposed method under scaling attack, Various scaling transformations with scaling Factors(SF=0.5,0.7,1,1.5) have been applied to images (A-D) in the dataset: MICC-F220, where Sx and Sy are scale factors applied to the x and y axis of the image part as shown in Figure 8.

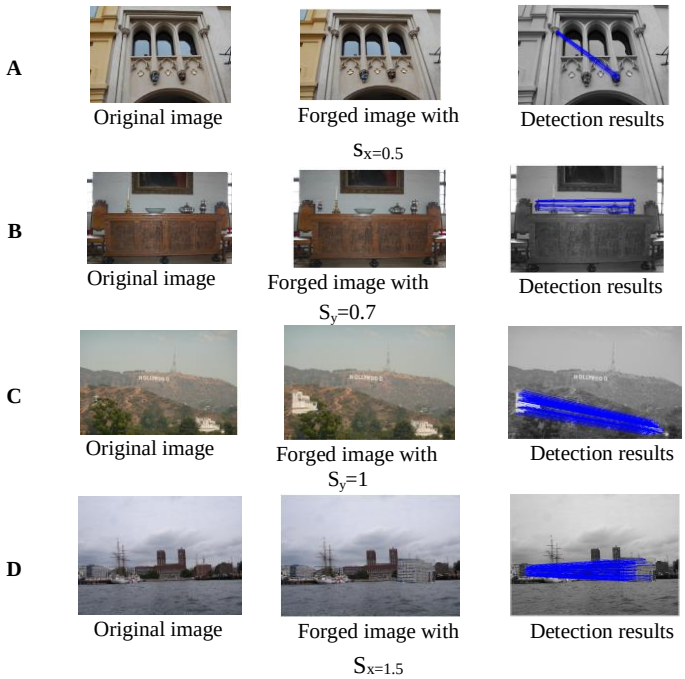


Figure 8: Detection of duplicated regions with horizontal and vertical scaling attacks.

Furthermore, the proposed method is examined to identify the optimal threshold T in the detection step to achieve the best detection rate for scaling attack. Table 1 shows that the value of 80% is identified as the best threshold value where the best true positive rate (TPR) and false positive rate (FPR) results are achieved. The goal of our method is achieved the lowest FPR which means only a few percent of all images didn't authenticate correctly; the TPR value is about 96% which means the majority of images in a dataset are authenticated correctly.

Threshold Value	Average T_{PR}%	Average F_{PR}%
0.1	75%	20%
0.3	80%	36%
0.5	90%	10%
0.7	92%	12%
0.8	96%	7%

Table 1: Threshold estimation for images in MICC-F220 under scale attack with scaling Factors (SF=0.5, 0.7, 1, 1.5).

4.4 JPEG compression

Some experiments for JPEG compressions are addressed.

The performance of our method is evaluated on a set of images compressed with various quality factors (QF=80, 70 and 50) as shown in Figure 9. The ROC curve in Figure 10 shows that the TPR and FPR of the proposed method are 90%, 4% respectively for JPEG quality factors up to 40.

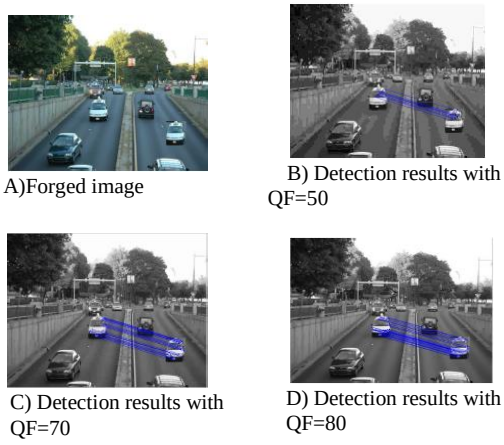


Figure 9: The ability of our method to detect duplicated regions via various JPEG factors.

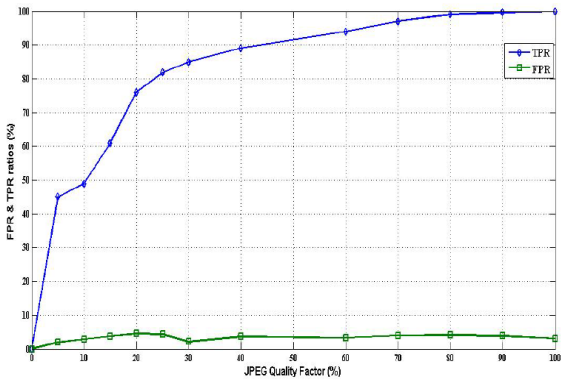


Figure 10: ROC curve in terms of TPR and FPR based on MICC-F220.

As shown in Figure 10, it can be concluded that the

proposed method is still reliable and robust against JPEG compression even with a low quality factor such as Q=50.

4.5 Forgery with different block sizes

100 original images from CASIA v2 image are selected. For each original image and each duplicate region with a block size 32x32 pixels, 64x 64pixels and 96 x96 pixels, four forged images are created with the additive noise duplicated regions by SNRs (dB=10, 15,20,30). This results in 400 forged images in total. The detection performances of duplicated regions for each block size with additive noise are presented in Table 2. It shows the efficiency of the system in case of very high signal-to noise ratios.

SNR (dB)	Block size					
	32 x 32		64 x 64		96 x 96	
	T_{PR}	F_{PR}	T_{PR}	F_{PR}	T_{PR}	F_{PR}
10	96%	6%	95%	6%	97%	3%
15	96%	8%	94%	8%	96%	8%
20	95%	8%	93%	8%	95%	15%
30	94%	10%	93%	10%	95%	15%

Table 2: The detection performance of region duplication forgery with different block size from images in CASIA v2.

4.6 Blurring Attack

Some experiments of detecting region duplication forgery under blur with their corresponding descriptors constructed by our method. Here, we use Gaussian blurs with radius varying from 0.5 to 2. The details are shown in Figure 11. Comparative study

As shown in Table 3, the proposed method is examined with a well known state of art methods such as keypoint-based methods: [4], [25], [39] and block-based methods: [3], [24]. These methods focused on detecting region duplication forgery with different post-processing attacks for instance, scaling and blurring.

Table 3 shows that, the proposed scheme gives a TPR=97%, which is better than TPRs in the methods: [25] and [39] due to the robustness of ANMS features against scale and blur attacks compared with SURF features. [4] method gives high FPR due to the weakness of SIFT method to detect local keypoints of duplicated regions when the textures of some forged regions are almost in uniform, since the local extrema may not exist in such region. The FPR is about 3% which is less than FPR of [25] method due to G2NN clustering technique to find best matching. The proposed method extract local phase quantized coefficients from divided regions 4 x4 in the image. LPQ texture descriptor is insensitive to blurring manipulations which gives a high F-score=97% for detecting this type of forgery compared with [3] method and [24] method.

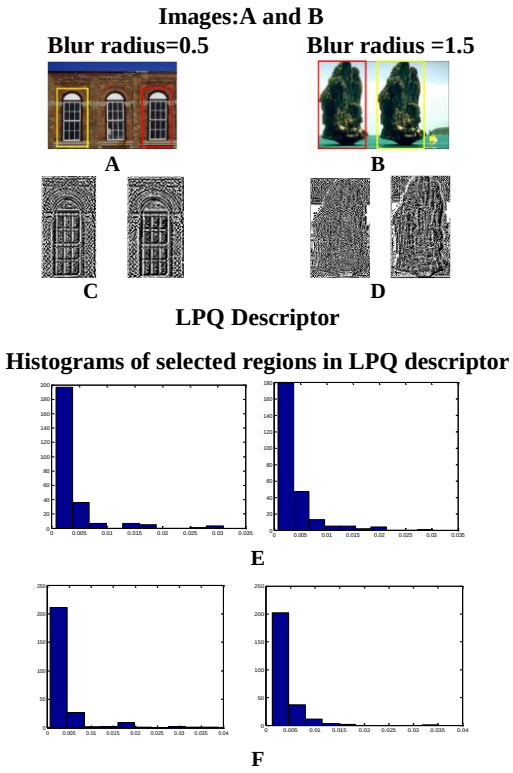


Figure 11: Illustrating region duplication forgery detection by local phase quantized coefficients from images on CASIA v2. (A) Image "window" has blurred duplicated region with (Gaussian blur radius = 0.5) which highlighted by the red rectangle. Image (B) has blurred duplicated region with (Gaussian blur radius = 1.5). (C) and (D) are LPQ image maps of (A) and (B) to extract a significant features of internal structure of foreground objects. (E) and (F) The histograms of selected regions in LPQ descriptor show the similarity of features between blurred region and Normal region.

Methods	TPR%	FPR%	Fscore%	Features	Block size	Time(s)
Amerini et al. 2011 [4]	100	8	81.40	SIFT	NA	4.94
Mishra et al. 2013 [25]	73.6	3.64	NA	SURF and HAC	4 x4	2.58
Silva et al. 2015 [39]	94.08	1.70	NA	SURF on HSV color features	Circle block with radii=4	18.81
Alkawaz et al. 2016 [3]	96.579	NA	75.166	DCT	4 x4	296.74
Mahmood et al. 2017 [24]	96.606	NA	96.05	Stationary wavelet transforms (SWT)	4 x4	NA
The proposed method	97	3	97.05	ANMS and LPQ	4 x4	13.80

Table 3: The overall performance of the proposed compared with the state of the art methods on MICC-F220.

5 Conclusion

In this paper, robust features such as local interest points play an important rule to expose copy move forgery on images. ANMS keypoints and LPQ texture descriptor have been proposed. The use of image preprocessing like color segmentation has reduced the FPR in the suspected image. Clustering segmented regions in the image based on fuzzy C means will increase the TPR of matching duplicated regions over ANMS keypoints. From the suspected forged images, the proposed method can find the duplicated regions, even if they are post-processed by some transformations like scaling or blurring. Future works will focus on image forgery with reflections and illumination changes.

Acknowledgment

The authors are grateful to the Middle East University, Amman, Jordan for the financial support granted to cover the publication fee of this research article.

References

[1] Al-Hammadi, M. H., G. Muhammad, M. Hussain and G. Bebis (2013). Curvelet transform and local texture based image forgery detection. International Symposium on Visual Computing, Springer, DOI: https://doi.org/10.1007/978-3-642-41939-3_49

[2] Al-Qershi, O. M. and B. E. Khoo (2013). Passive detection of copy-move forgery in digital images: State-of-the-art. Forensic science international 231(1): 284-295, DOI: https://doi.org/10.1016/j.forsciint.2013.05.027

[3] Alkawaz, M. H., G. Sulong, T. Saba and A. Rehman (2016). Detection of copy-move image forgery based on discrete cosine transform. Neural Computing and Applications: 1-10, DOI: https://doi.org/10.1007/s00521-016-2663-3

[4] Amerini, I., L. Ballan, R. Caldelli, A. Del Bimbo and G. Serra (2011). A sift-based forensic method for copy–move attack detection and transformation recovery. Information Forensics and Security, IEEE Transactions on 6(3): 1099-1110, DOI: 10.1109/TIFS.2011.2129512

[5] Bo, X., W. Junwen, L. Guangjie and D. Yuewei (2010). Image copy-move forgery detection based on SURF. Multimedia Information Networking and Security (MINES), 2010 International Conference on, IEEE, DOI: 10.1109/MINES.2010.189

[6] Brown, M., R. Szeliski and S. Winder (2005). Multi-image matching using multi-scale oriented patches. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE, DOI: 10.1.1.124.3167

[7] Chen, M. and S. A. Ludwig (2017). Color Image Segmentation Using Fuzzy C-Regression Model. Advances in Fuzzy Systems 2017, DOI: https://doi.org/10.1155/2017/4582948

[8] Cheng, H.-D., X. H. Jiang, Y. Sun and J. Wang (2001). Color image segmentation: advances and prospects. Pattern recognition 34(12): 2259-2281, DOI: https://doi.org/10.1016/S0031-3203(00)00149-7

[9] Christlein, V., C. Riess, J. Jordan and E. Angelopoulou (2012). An evaluation of popular copy-move forgery detection approaches. IEEE Transactions on Information Forensics and Security 7(6): 1841-1851, DOI: 10.1109/TIFS.2012.2218597

[10] Cozzolino, D., G. Poggi and L. Verdoliva (2015). Efficient dense-field copy–move forgery detection. IEEE Transactions on Information Forensics and Security 10(11): 2284-2297, DOI: 10.1109/TIFS.2015.2455334

[11] Dadkhah, S., M. Köppen, H. A. Jalab, S. Sadeghi, A. A. Manaf and D. M. Uliyan (2017). Electromagnetismlike Mechanism Descriptor with Fourier Transform for a Passive Copy-move Forgery Detection in Digital Image Forensics. ICPRAM, DOI: 10.5220/0006232206120619

[12] Farid, H. (2008). Digital image forensics. Scientific American 298(6): 66-71, DOI: doi:10.1038/scientificamerican0608-66

[13] Gan, Y. and J. Zhong (2014). Image copy-move tamper blind detection algorithm based on integrated feature vectors. Journal of Chemical and Pharmaceutical Research 6(6): 1584-1590, DOI:

[14] Guo, J.-M., Y.-F. Liu and Z.-J. Wu (2013). Duplication forgery detection using improved DAISY descriptor. Expert Systems with Applications 40(2): 707-714, DOI: https://doi.org/10.1016/j.eswa.2012.08.002

[15] Hsiao, D.-Y. and S.-C. Pei (2005). Detecting digital tampering by blur estimation. Systematic Approaches to Digital Forensic Engineering, 2005. First International Workshop on, IEEE, DOI: 10.1109/SADFE.2005.8

[16] Huang, Y., W. Lu, W. Sun and D. Long (2011). Improved DCT-based detection of copy-move forgery in images. Forensic science international 206(1): 178-184, DOI: https://doi.org/10.1016/j.forsciint.2010.08.001

[17] Kushol, R., M. S. Salekin, M. H. Kabir and A. A. Khan (2016). Copy-Move Forgery Detection Using Color Space and Moment Invariants-Based Features. Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on, IEEE, DOI: 10.1109/DICTA.2016.7797027

[18] Li, H. and J. Zheng (2012). Blind Detection of Digital Forgery Image Based on the Edge Width. Intelligent Science and Intelligent Data Engineering. Y. Zhang, Z.-H. Zhou, C. Zhang and Y. Li, Springer Berlin Heidelberg. 7202: 546-553, DOI: 10.1007/978-3-642-31919-8_70

[19] Li, X.-h., Y.-q. Zhao, M. Liao, F. Shih and Y. Shi (2012). Passive detection of copy-paste forgery between JPEG images. Journal of Central South University 19(10): 2839-2851, DOI: https://doi.org/10.1007/s11771-012-1350-5

[20] Li, Y. (2012). Image copy-move forgery detection based on polar cosine transform and approximate nearest neighbor searching. Forensic science international 224(1-3): 59-67, DOI: https://doi.org/10.1016/j.forsciint.2012.10.031

[21] Lin, S. D. and T. Wu (2011). An integrated technique for splicing and copy-move forgery image detection. Image and Signal Processing (CISP), 2011 4th International Congress on, IEEE, DOI: 10.1109/CISP.2011.6100366

[22] Lowe, D. G. (1999). Object recognition from local scale-invariant features. Computer vision, 1999. The proceedings of the seventh IEEE international conference on, Ieee, DOI: 10.1109/ICCV.1999.790410

[23] Mahdian, B. and S. Saic (2007). Detection of copy–move forgery using a method based on blur moment invariants. Forensic science international 171(2): 180-189, DOI: https://doi.org/10.1016/j.forsciint.2006.11.002

[24] Mahmood, T., Z. Mehmood, M. Shah and Z. Khan (2017). An efficient forensic technique for exposing region duplication forgery in digital images. Applied Intelligence: 1-11, DOI: https://doi.org/10.1007/s10489-017-1038-5

[25] Mishra, P., N. Mishra, S. Sharma and R. Patel (2013). Region duplication forgery detection technique based on SURF and HAC. The Scientific World Journal 2013, DOI: http://dx.doi.org/10.1155/2013/267691

[26] Muhammad, G., M. H. Al-Hammadi, M. Hussain, A. M. Mirza and G. Bebis (2013). Copy move image forgery detection method using steerable pyramid transform and texture descriptor. EUROCON, 2013, IEEE, DOI: 10.1109/EUROCON.2013.6625188

[27] Muhammad, G., M. Hussain and G. Bebis (2012). Passive copy move image forgery detection using undecimated dyadic wavelet transform. Digital Investigation 9(1): 49-57, DOI: https://doi.org/10.1016/j.diin.2012.04.004

[28] Myrna, A., M. Venkateshmurthy and C. Patil (2007). Detection of region duplication forgery in digital images using wavelets and log-polar mapping. Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on, IEEE, DOI: 10.1109/ICCIMA.2007.271

[29] Ng, T.-T., S.-F. Chang, J. Hsu and M. Pepeljugoski (2005). Columbia photographic images and photorealistic computer graphics dataset. DOI:

[30] Ojansivu, V. and J. Heikkilä (2008). Blur insensitive texture classification using local phase quantization. International conference on image and signal processing, Springer, DOI: https://doi.org/10.1007/978-3-540-69905-7_27

[31] Panzade, P. P., C. S. Prakash and S. Maheshkar (2016). Copy-move forgery detection by using HSV preprocessing and keypoint extraction. Parallel, Distributed and Grid Computing (PDGC), 2016 Fourth International Conference on, IEEE, DOI: 10.1109/PDGC.2016.7913156

[32] Peng, F., Y.-y. Nie and M. Long (2011). A complete passive blind image copy-move forensics scheme based on compound statistics features. Forensic science international 212(1): e21-e25, DOI: https://doi.org/10.1016/j.forsciint.2011.06.011

[33] Peng Gao, H. Z., Ruier Guo, Jingli Liu, Lihu Ma, Jin Zhang and Qian He. (2009). CASIA Image Tempering Detection Evaluation Database (CAISA TIDE) V2.0. Retrieved 23, May 2016, from http://forensics.idealtest.org/casiav2/.

[34] Qazi, T., K. Hayat, S. U. Khan, S. A. Madani, I. A. Khan, J. Kołodziej, H. Li, W. Lin, K. C. Yow and C.-Z. Xu (2013). Survey on blind image forgery detection. IET Image Processing 7(7): 660-670, DOI: 10.1049/iet-ipr.2012.0388

[35] Redi, J. A., W. Taktak and J.-L. Dugelay (2011). Digital image forensics: a booklet for beginners. Multimedia Tools and Applications 51(1): 133-162, DOI:

[36] Ryu, S.-J., M.-J. Lee and H.-K. Lee (2010). Detection of copy-rotate-move forgery using Zernike moments. Information Hiding, Springer, DOI: https://doi.org/10.1007/978-3-642-16435-4_5

[37] Sadeghi, S., S. Dadkhah, H. A. Jalab, G. Mazzola and D. Uliyan (2017). State of the art in passive digital image forgery detection: copy-move image forgery. Pattern Analysis and Applications: 1-16, DOI: 10.1007/s10044-017-0678-8

[38] Sadeghi, S., H. A. Jalab, K. Wong, D. Uliyan and S. Dadkhah (2017). Keypoint based authentication and localization of copy-move forgery in digital image. Malaysian Journal of Computer Science 30(2): 117-133, DOI: https://doi.org/10.22452/mjcs.vol30no2.4

[39] Silva, E., T. Carvalho, A. Ferreira and A. Rocha (2015). Going deeper into copy-move forgery detection: Exploring image telltales via multi-scale analysis and voting processes. Journal of Visual Communication and Image Representation 29: 16-32, DOI: https://doi.org/10.1016/j.jvcir.2015.01.016

[40] Uliyan, D. M., H. A. Jalab, A. W. Abdul Wahab and S. Sadeghi (2016). Image Region Duplication Forgery Detection Based on Angular Radial Partitioning and Harris Key-Points. Symmetry 8(7): 62, DOI: https://doi.org/10.3390/sym8070062

[41] Uliyan, D. M., H. A. Jalab, A. Abuarqoub and M. Abuhashim (2017). Segmented-Based Region Duplication Forgery Detection Using MOD Keypoints and Texture Descriptor. Proceedings of the International Conference on Future Networks and Distributed Systems, ACM, DOI: 10.1145/3102304.3102310

[42] Uliyan, D. M., H. A. Jalab and A. W. A. Wahab (2015). Copy move image forgery detection using Hessian and center symmetric local binary pattern. Open Systems (ICOS), 2015 IEEE Confernece on, IEEE, DOI: 10.1109/ICOS.2015.7377269

[43] Uliyan, D. M., H. A. Jalab, A. W. A. Wahab, P. Shivakumara and S. Sadeghi (2016). A novel forged blurred region detection system for image forensic applications. Expert Systems with Applications 64: 1-10,

DOI: https://doi.org/10.1016/j.eswa.2016.07.026

[44] Uliyan, D. M. H. (2016). Region Duplication Forgery Detection Technique Based on Keypoint Matching, Fakulti Sains Komputer dan Teknologi Maklumat, Universiti Malaya.

[45] Ulutas, G. and G. Muzaffer (2016). A New Copy Move Forgery Detection Method Resistant to Object Removal with Uniform Background Forgery. Mathematical Problems in Engineering 2016, DOI: http://dx.doi.org/10.1155/2016/3215162

[46] Wang, T., J. Tang and B. Luo (2013). Blind detection of region duplication forgery by merging blur and affine moment invariants. Image and Graphics (ICIG), 2013 Seventh International Conference on, Qingdao, China, IEEE, DOI: 10.1109/ICIG.2013.61

[47] Zhao, J. and J. Guo (2013). Passive forensics for copy-move image forgery using a method based on DCT and SVD. Forensic science international 233(1): 158-166, DOI: 10.1016/j.forsciint.2013.09.013

[48] Zheng, J. and M. Liu (2009). A digital forgery image detection algorithm based on wavelet homomorphic filtering. Digital Watermarking, Springer: 152-160, DOI: https://doi.org/10.1007/978-3-642-04438-0_13

[49] Zhou, L., D. Wang, Y. Guo and J. Zhang (2007). Blur detection of digital forgery using mathematical morphology. Agent and Multi-Agent Systems: Technologies and Applications, Springer: 990-998, DOI: https://doi.org/10.1007/978-3-540-72830-6_105

https://doi.org/10.31449/inf.v42i4.2297 Informatica **42** (2018) 607–616 **607**

The Impact of Online Indexing in Improving Arabic Information Retrieval Systems

Tahar Dilekh,
Computer Science Department, University of Batna 2, Batna 05078, Algeria
E-mail: tahar.dilekh@univ-batna2.dz

Saber Benharzallah
Computer Science Department, University of Batna 2, Batna 05078, Algeria
LINFI Laboratory, University of Biskra, Biskra 07000, Algeria
E-mail: s.benharzallah@univ-batna2.dz

Ali Behloul
LaSTIC Laboratory, University of Batna 2, Batna 05078, Algeria
E-mail: a.behloul@univ-batna2.dz

Keywords: online indexing, offline indexing, semi-automatic indexing, Arabic keywords extraction, Arabic information retrieval system

Received: April 17, 2018

This paper suggests a new type of indexing Arabic Language text that contribute to improving the quality of IRS. The proposed method of indexing belongs to semi-automatic category of indexing and consists of two types. The first type conducts an online indexing and the output of this process give a rise to a Partial index. The second type – under this method- is an offline indexing and the output of this process leads to a General index. We illustrate application and the performance of this new method of indexing using an Arabic text editor and Information Retrieval tool developed and designed for this purpose. We also illustrate the process of building a new form of Arabic corpus appropriate to conduct the necessary experiments. Our findings show that the online indexing model successfully identifies the descriptors most relevant to the document. In addition, this model is more efficient as it helps minimizing index storage size, consequently, improving the response time of the different requests. Finally, the paper proposes a solution to issues and deficiencies Arabic language processing suffers from, especially regarding corpora building and information retrieval evaluation systems.

Povzetek: V prispevku je predlagan nov način indeksiranja arabskih besedil z namenom izboljšanja jezikovno-računalniških operacij.

1 Introduction

Recent developments in the internet technology made information abundant, which made it highly available to users. On the other hand, the vast availability of information made it particularly challenging for users to obtain and find relevant and useful information. In this context, Information Retrieval Systems (IRS) have emerged as a tool to address this problem.

IRS consists of two stages: the 'indexing' and the 'search' stages. In the first stage, the descriptors are extracted from documents and prepared to facilitate and accelerate the search process in the second stage. In general, the indexing stage consists of three types. First, manual indexing, in which the descriptors selection process is performed by a human expert. Second, the automatic indexing where the descriptors are automatically extracted from documents, and finally, the semi-automatic indexing (or supervised indexing). This latter provides automated assistance to the expert.

Currently, IRS benefit from the indexing processes, most of which remains under-performing in the extraction of accurate descriptors that contribute to improving the quality of these systems including extracting the semantic of these descriptors. This remains a challenging task of automatic indexing that often requires human intervention to choose the appropriate descriptors. This is because of several reasons including the ambiguity of language, the power of language to transfer thoughts from one mind to another and the dynamic nature of language.

While the literature consists of many studies concerning various natural languages, there are relatively fewer studies on Arabic language, where the complex grammatical and morphological features of this language make the task of automatic processing even more challenging. Thus, this paper suggests a new type of indexing to contribute to improving the quality of IRS. The proposed method of indexing belongs to semi-automatic category of indexing and consists of two types. The first type conducts an online indexing where one document is the indexing unit. This type of indexing refers to the indexing process that begins directly after the

writing of each unit ends, which allows to assist human expert (author of text) to select Arabic appropriate descriptors to improve the search results. The output of this process give a rise to a Partial index. The second type – under this method- is an offline indexing, which refers to the process of indexing based on the collection of textual documents available from different corpora. The output of this process leads to a General index.

We also illustrate implementing and the performance of this new method of indexing using an Arabic text editor developed and designed to allow for an online semi-automatic indexing system and Information Retrieval tool that contains an offline automatic indexing system. We also illustrate the process of building a new form of Arabic corpus appropriate to conduct the necessary experiments.

Thus, this study contributes to two key areas of the literature. First, it offers applications of some tools such as SIRAT[1] and OIRDA[2] that have been developed to show the extent to which the integration of online semi-automatic indexer into text editors is effective in improving indexing, and thus improving the precision of IRS. Second, the study is conducted on Arabic texts, which contributes to the enrichment and development of Arabic language processing tools.

The remainder of the paper is organized as follows. Section 2 offers an account of the main developments and recent advances of Arabic documents indexing literature. Section 3 identifies the main characteristics of Arabic language followed by an illustration of the proposed semi-automatic system in Section 4. Section 5 and 6 illustrate implemented applications and analyze the results of the conducted experiments respectively. Section 7 concludes.

2 Literature review

We being with a review the main literature of Arabic documents indexing, and identify the challenges facing this research area. We categorize the literature according to the most commonly used approach. We then present some work related to the automatic Arabic keyword extraction, which helps to improve the quality of Arabic indexing systems.

2.1 Arabic documents indexing

Various studies have proposed different methods for Arabic documents indexing. However, to the best of our knowledge, all of these studies focused on manual and automatic indexing. This prevented us from comparing the existing methods to that proposed in this paper. This paper proposes various automatic indexing techniques according to the following approaches: linguistic, statistical, semantic, and hybrid.

2.1.1 The linguistic approach

The linguistic approaches consist of a morphological and syntactic analysis of the document based on the grammatical rules and relationships between the different textual units. The methods of this approach are widely used in Arabic natural language processing due to the reliability of syntactic and semantic recognition algorithms. Saadi et al. [1] proposed knowledge extraction systems, based on a deep linguistic analysis and using a domain ontology to extract the semantic content, they have achieved promising results, but reveal other problems in need of careful investigation.

Mansour et al.[2] proposed a method mainly based on morphological analysis and on a technique for assigning weights to words. The morphological analysis uses a number of grammatical rules to extract candidate index words. The weight assignment technique computes weights for these words relative to the container document. The weights are based on how spread are the words in a document and not only on their rate of occurrence. The experimental results carried out for a number of texts have demonstrated the advantage of their auto-indexing method.

Al Molijy et al. [3] proposed and implemented a method to create and index for books written in Arabic language using the syntactic analysis. The process depends largely on text summarization and abstraction processes to collect main topics and statements in the book automatically.

This approach offers good results in specific situations, such as determining the exact meaning of a vague word as expressed in the sentence; the name is gold, but the verb is gone, but remains less able to match other approaches, given the complexity of the Arabic language.

2.1.2 The statistical approach

Statistical approaches are mostly based on statistical techniques. A variety of these approaches have been developed to extract descriptors (terms) and study their occurrence in a document, or even in the corpus.

The frequency distribution of words has been a key object of study in statistical approach for the past decades. This distribution approximately follows a simple mathematical form known as Zipf's law. According to this law, words occur according to a systematic frequency distribution such that there are few very high-frequency words that account for most of the text and many low-frequency words. We very briefly mention some of the places where this law affects research in our study:

- Zipf's Law tells us how much text we have to look at and how precise our statistics have to be to achieve what level of expected error [4].
- Zipf's Law also provides a base-line model for expected occurrence of target terms and the answers

[1] The Arabic text editor SIRAT (Semantic Information Retrieval of Arabic Texts) is an application that we have developed to conduct experiments on semantic Arabic information retrieval domain.

[2] It is an indexing and retrieval program for Arabic texts, we have developed in Java. OIRDA is abbreviation of the French sentence (Outil d'Indexation et de Recherche dans les Documents Arabes) i.e. Indexing and retrieval tool for Arabic documents.

to certain questions may provide considerable information about its role in the corpus [5]: what does it mean to ask if a word is significant in a corpus, beyond mere occurrence or relative probability? What is the range of the semantic influence of a word in a corpus? What does the pattern of occurrences contribute to our assessment of its relevance in the corpus? [6]

- Zipf's Law provides a basis for evaluating parsers and taggers [7]. Again we summarize the potential role in the form of a series of questions: How does a language model developed on one corpus transfer to another? How do we translate performance estimates on a few test corpora to estimates for the language as a whole? How do differences in register, genre and medium affect the utility of a system, and how do we compensate for these differences? [6]

The Term Frequency–Inverse Document Frequency (TF-IDF) method is also one of the statistical approaches that provides a good representation of the weight of corpora words whose document size is homogeneous. Several alternatives have been proposed for the TF-IDF method, which has become the subject of many comparative studies.

The feasibility of this approach also depends on the process of extracting the root/stem of each word, according to root-based approach; or stem-based approach; in order to overcome the polymorphism of the word.

Several studies have shown that the process of stemming of the word from its prefixes and suffixes is more useful for Arabic information retrieval systems than in other approaches.

Researchers adopted various statistical methods and techniques in the indexing process [8] [9] [10] [11] [12] [13] [14] [15] [16].

In conclusion, these methods, considered as simple to implement, are efficient and perfectly tolerant of large masses of documentary. On the other hand, the hypothesis considering the words as independent units generates a loss of semantic information. The resulting indexes may generate polysemy problems and deviate from the general context of the document [17].

2.1.3 The semantic approach

This approach aims, on the one hand, to reduce the ambiguity of the words meaning and, on the other hand, allows to extract the semantic relations between these words. Thus, texts are represented focuses on the unit of meaning rather than simple words. Semantic relationships can also be calculated using methods that evaluate the amount of information between words.

Researchers [18] have integrated semantic process into an Internet search engine and used several techniques (Harman, Croft, and Okapi) to evaluate the performance of this engine. In a recent study [19] [20] have exploited the lexical base of Arabic WordNet in an IRS in order to index the collection of documents and query of the user. Others[21], introduced a query expansion approach using an ontology built from Wikipedia pages in addition to other thesaurus to improve search accuracy for Arabic language.

This approach provides the best semantic cover for the documents due to relies on semantic resources (dictionaries, anthologies or others). However, it remains restricted by the type of resource used and its ability to describe the words of the text being processed.

2.1.4 The hybrid approach

Several researchers [22] [23] [24] [25] have experimented with different combinations of linguistic, statistical, and semantic methods, taking the advantages of each method in an attempt to overcome their shortcomings and to improve the process of indexing by extracting hidden information in a document. These approaches often led to better results than those obtained through the use of standard methods.

Despite the positive results of this approach, it suffers from the problem of complexity, depending on the integration of other approaches.

2.2 Extraction Arabic keywords

Keywords (descriptors) are a subset of words or phrases that can describe the meaning of a document, where several natural language processing applications can benefit from keywords. Unfortunately, most documents do not contain these words. On the other hand, adding high-quality keywords manually is costly, time-consuming, and error-prone. Therefore, this domain has emerged to develop novel algorithms and systems designed to extract keywords automatically.

[26] presented the KP-Miner (Keyphrases-Miner) system to extract keyphrases from both English and Arabic documents of varied length. This system does not need to be trained on a particular document set in order to achieve its task (i.e. unsupervised learning). It also has the advantage of being configurable as the rules and heuristics adopted by the system are related to the general nature of documents and keyphrases. In general, Experiments and comparison studies with widely used systems suggest that KP-Miner is effective and efficient.

[27] introduced AKEA, a keyphrase extraction -unsupervised- algorithm for single Arabic documents. They relied on heuristics that collaborate linguistic patterns based on Part-Of-Speech (POS) tags, statistical knowledge and the internal structural pattern of terms. They employed the usage of Arabic Wikipedia to improve the ranking of candidate keyphrases by adding a confidence score if the candidate exists as an indexed Wikipedia concept. Experimental results have shown that the performance of AKEA outperforms other unsupervised algorithms as it has reported higher precision values.

[28] presented a keyword extraction system for Arabic documents using term co-occurrence statistical information. In case the co-occurrence of a term is in the biasness degree, then the term is important and it is likely to be a keyword. The biasness degree of the terms and the set of frequent terms are measured using $\chi 2$. Therefore, terms with high $\chi 2$ values are likely to be keywords. This

technique showed an acceptable performance compared to other techniques.

[29] presented a supervised learning technique for extracting keyphrases of Arabic documents. The extractor is supplied with linguistic knowledge to enhance its efficiency instead of relying only on statistical information such as term frequency and distance. An annotated Arabic corpus is used to extract the required lexical features of the document words. The knowledge also includes syntactic rules based on part of speech tags and allowed word sequences to extract the candidate keyphrases. The experiments carried out show the effectiveness of this method to extract Arabic keyphrases.

[30] presented a framework for extracting keyphrases from Arabic news documents. It relies on supervised learning, Naïve Bayes in particular, to extract keyphrases. The final set of keyphrases is chosen from the set of phrases that have high probabilities of being keyphrases.

Various experiments have shown the effectiveness of these methods to extract Arabic keywords in varying percentages. However, while supervised techniques are costly and limited by the type of language resources used, unsupervised techniques suffer from the best semantic cover for the documents.

3 Characteristics of the Arabic language

The complex grammatical and morphological features of the Arabic language make the task of automatically processing more difficult. Among these features, we highlight the following:

- Arabic scripts have diacritics to represent the short vowels, which are marks above or below the letters. However, these diacritics have been disappearing in most contemporary writings, and readers are expected to fill in the missing diacritics through their knowledge of the language. The absence of diacritics from contemporary Arabic texts makes the automatic processing a difficult task.
- Morphological analysis is a complex procedure because Arabic is an agglutinative language. For example, the word "أفاستسقيناكموها" (*did we ask you- plural- for water to her (it)*) is one of the longest words in the Arabic language dictionaries. It consists of 15 letters and 9 diacritics. Its root is the verb "سقى" (*to water*). We add to the word the prefix "است" to become "استسقى" (*he asked for water*). Adding a subject pronoun, the word becomes "استسقينا" (*we asked for water*). Then we add the indirect object pronoun to become "استسقيناكم" (*we asked you – plural- for water*), and we add the direct object to become "استسقيناكموها" (*we asked you – plural- for water for her (it)*) Next, we add "F" of appeal (ف الاستئناف) and "A" of question (أ الاستفهام) to become a fully-meaningful phrase: "أفاستسقيناكموها؟" (*did we ask you- plural- for water to her (it)?*).
- Arabic is a highly inflectional and derivational language where many of the nouns and verbs are derived from the same root. This latter is based on more than 150 patterns, which makes them more complex and difficult to handle.

4 Semi-automatic indexing system

As emphasised in the introduction above, we have designed and developed a semi-automatic indexing system that is based on:

1. An Online semi-automatic indexing of Arabic documents (Figure 1).
2. An Offline automatic indexing of Arabic corpus (Figure 5).

4.1 Online semi-automatic indexing system

This system consists of three units: a unit for automatic indexing, a unit for the automatic extraction of keywords and a unit for updating partial index of a document after the intervention of the human expert to select the relevant keywords.

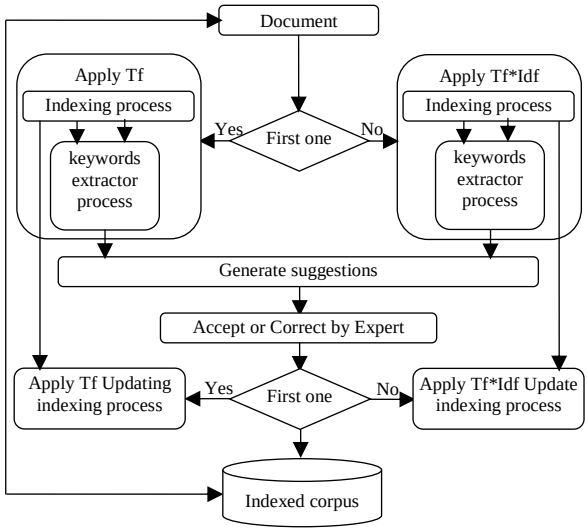


Figure 1: Online semi-automatic indexing system of Arabic documents.

In addition, we integrated our online indexing system to an Arabic text editor (Figure 6) that we developed for the purpose of testing and running experiments. We also created an Arabic corpus in a new format (Figure 7) that allows us running the necessary experiments.

4.1.1 Automatic indexing unit

Indexing is the process of representing the given text into the list of informative terms, which reflects its content in order to optimize speed and performance in finding relevant documents for a search query.

The automatic indexing of Arabic texts had dominated most of the research literature in Arabic text retrieval. In our study, we followed the approach due to [25] to create the index with some modifications, which we discuss in the next section. This method has proved to be effective in improving the process of indexing Arabic documents.

4.1.1.1 Encoding

The corpus and queries can be encoded differently, making them incomparable. In order to standardize the documents with the queries, we must reuse converting tools between different encodings systems. Thus, everything would be converted into UTF-16 encoding in our case, because it allows the representation of letters and symbols in a wide range of languages, including Arabic.

4.1.1.2 Normalization

Normalization involves the following steps:

- Remove punctuation;
- Remove diacritics (primarily weak vowels);
- Remove the Tatweel '-'.
- Replace the 'إ' or the 'أ' initial by Alif nu 'ا';
- Replace the 'آ' by the 'ا';
- Replace the 'ىء' of order by the 'ئ';
- Replace the 'ى' final by the 'ي';
- Replace the 'ة' final by the 'ه'.

4.1.1.3 Removing stop words

The removal of stop words has the advantage of reducing the number of indexing terms and may reduce the recall rate (i.e. the proportion of relevant documents returned by the system to all relevant documents). We use a list of stop words to remove stop words.

4.1.1.4 Stemming

We used a hybrid method, as proposed by [25], to extract the roots of the words and use them as index terms. This combines the application of three previously used techniques, which deal with three key issues related to Arabic stemming including affix removal proposed by [31], dictionaries [32] and morphological analysis[33]. This method has been found to be effective in indexing process compared to other methods.

4.1.1.5 Term frequency and weighting

Several statistical measure are available to assign weights to words of a document in a corpus. Currently, TF-IDF is one of the most popular term-weighting procedure. TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

In our study, we used TF-IDF that combine the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document. The TF-IDF weighting procedure assigns a weight to term t in document d given by

$$tf - idf_{t,d} = tf_{i,j} * idf_i$$

where

- $tf_{i,j}$: the number of times that term *i* occurs in document *j*.
- $idf_i = \log \frac{|D|}{|\{d_i : t_j \in d_j\}|}$
- $|D|$: total number of documents in the corpus.
- $|\{d_i : t_j \in d_j\}|$: number of documents where the term t appears (i.e., tf $(t, d) \neq 0$).

Our automatic indexing unit deals differently with the first document added to the corpus (Figure 2). Since there are no documents available prior to the first document to compute $tf - idf_{t,d}$, we only count a $tf_{i,j}$ value.

The automatic indexing unit constructs a *partial index* for every document of every corpus. The output of this unit is a *partial index* for each document (Figure 2). The main motivation behind constructing *partial indexes* is to allow the expert intervention in the creation of index later.

Indexing function pseudo code
Input:
Document d_i ϵ corpus
Output:
$Index_i$ // partial index
Algorithm
For each token in d_i loop
Encoding ();
Normalize ();
Removing_stop_words ();
Stemming ();
If (tf_type = tf) then
Weighting(tf)
Else
Weighting(tf-idf);
End
Stored tf for the term = token
End loop.
Add d_i to $Index_i$.

Figure 2: Automatic indexing algorithm.

4.1.2 Automatic keyword extraction unit

We have adopted a simple method of extracting keywords as long as the human expert is responsible for the final decision making regarding the acceptation or modification of the appropriate keywords for the document being processed (see the example in figure 3).

Instructions	Execute?	If no, why?
1. ***Input:*** في الأمم المتحدة أن ... (In the united nations that …)	-	
2. ***Selected word from the result of the indexing module*** ... في الامم المتحدة أن ...	Yes	
3. ***Add 1st right word*** ... في الامم المتحدة أن ...	No	Stop word
4. ***Add 1st left word*** ... في الامم المتحدة أن ...	Yes	
5. ***Add 2nd left word*** ... في الامم المتحدة أن ...	No	Stop word
6. ***Output:*** الأمم المتحدة (The united nations)		

Figure 3: Automatic keyword extraction example.

The automatic keyword extraction unit (Figure 4) proposes the list of candidate words. This list is limited to twelve keywords, each consisting of at most five words. These words are extracted in two stages:

In the first stage, we adopt the results of the automatic indexing unit, where we retrieve the index words with the highest weights. Then, we add, if possible, to each index word, from original text, two nearest neighbor words on the right and two others on the left while ensuring that this five-word string does not contain Arabic punctuation marks in between words. Otherwise, we just take the number of words between two punctuations. We also give priority to a noun phrase or nominal phrase by setting terms for the candidate words in the following order:

- Words that begin with "ال" letters and end with " ,"ة ي"" or "ء" letters.
- Words that begin with "ال" letters.
- Words that end with "ي" ,"ة" or "ء" letters.
- Ordinary words.

In the second stage, we propose to the human expert twelve key words arranged in descending order, after which the human expert would accept or modify the suggestions generated by the automatic keyword extraction unit.

Keywords_Extract function pseudo code
Input:
Document $d_i \in$ corpus
Output :
Keywords [ ]
Algorithm
For j = 1 to 12 loop
word ← Paratial_Index.canditat_word[j];
word ← From_Original_text (word);
if (Setting_terms (*fst_right_word*))
word ← word + fst_right_word;
if (Setting_terms (*snd_right_word*))
word ← word + snd_right_word;
if (Setting_terms (*fst_leftt_word*))
word ← fst_left_word + word;
if (Setting_terms (*snd_lest_word*))
word ← snd_left_word + word;
Keywords [i] ← word
End loop.

Figure 4: Automatic keyword extraction algorithm.

4.1.3 Unit of updating partial index

The role of this unit is to update a partial index of a document. The expert's opinions are accounted for by updating the weights of the selected index words and assigning to them higher values. This phase concludes with the integration of this partial index into the document, and saving it to an object file in order to exploit it later.

4.2 Offline indexing system of building and updating general index

The role of this system is to build and update a general index based on partial indexes of several corpora (Figure 5).

It retrieves all documents indexes (partial indexes) that created by the online semi-automatic indexing system, and merges them into a single general index. It also updates this index whenever necessary.

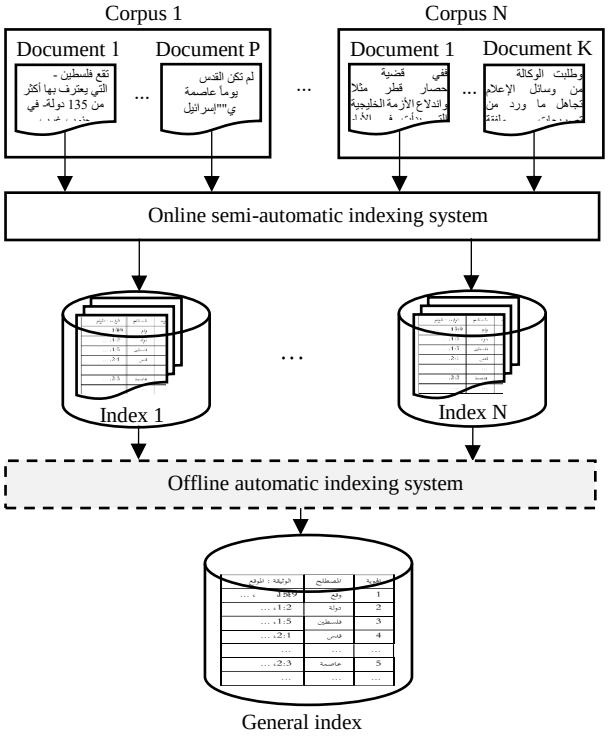


Figure 5: Offline automatic indexing system.

5 Implemented applications

To implement the online semi-automatic indexing system that we designed, we developed an Arabic text editor that contains an online document indexing system. In addition, we worked on building a suitable new form of Arabic corpus, which contains keywords proposed by a human expert, to conduct the necessary experiments. We also used OIRDA application for general indexing and information retrieval and equipped by an offline automatic indexing system of building and updating general index.

5.1 Arabic text editor

We first developed an Arabic text editor (Figure 6), which -in addition to the regular functions as text editor-, is provided with the automatic indexing option to editor's users. We have adopted the design of online semi-automatic indexing system described above (Figure 1) to add this option.

As discussed above, we deal differently with the first document added to the corpus, where there are no other documents, so it only counts a $tf_{i,j}$ value. We then integrate the keyword extraction unit, which is based on the results obtained from the automatic indexing unit prompting some keywords suggestions to the expert indexers, giving them the opportunity to modify these proposed words. Finally, the index is updated. The output

of this editor is an object file that contains the processed text and the generated partial index.

Figure 6: Arabic Text Editor "SIRAT".

5.2 New Arabic corpus form

To study the efficiency of the proposed system, it was necessary to obtain a test corpus consisting of a set of Arabic documents that would meet a set of necessary and sufficient features for testing.

We have developed a program to build an Arabic corpus, through the organization of a number of web pages of Al Jazeera's website[3], in a new corpus form that is different from the usual ones, by appending keywords suggested by the human expert (Al Jazeera journalists) to the end of documents (Figure 7). This allows evaluating the performance of the automatic keywords extraction unit. In addition, we have taken into account the set of rules used globally in the building of such corpus, especially those provided by (TREC) [34].

<DOC>
<DOCNO> ALJAZEERA-511 </DOCNO>
<HEADLINE> ماذا تعرف عن الجيوش الإلكترونية؟ </HEADLINE>
<DATELINE> المصدر: الجزيرة + وكالات, مواقع إلكترونية </DATELINE>
<TEXT>
الجيوش الإلكترونية مجموعات مدربة تعمل وفق أجندة خاصة هدفها اختراق مواقع الخصوم، والترويج لوجهة نظر معينة عبر مختلف منصات الإنترنت، وإسكات وتشويه سمعة المناوئين، إلى جانب ترويج الإشاعات والأكاذيب وخلق البلبلة، وقد بدأت الدول في إنشاء وحدات إلكترونية داخل أجهزتها العسكرية والأمنية لحماية أمنها القومي. خدمة رسمية الجيوش الإلكترونية مجموعة من الأشخاص وقراصنة الإنترنت (هاكرز) تعمل لصالح أجهزة المخابرات والأمن في الغالب، تسعى لاختراق المواقع الإلكترونية الخاصة بالشخصيات والمؤسسات والدول، ولا تكاد تترك منتديات أو نقاشات أو تعليقات على مواقع التواصل الاجتماعي وغيرها من المواقع الإلكترونية إلا ودخلت إليها للدفاع عن وجهة النظر الرسمية، ونشر الإشاعات والأكاذيب التي تربك رؤية الناس وتوجههم باتجاه معين.
...
وفي عصر المعلومة والحروب الإلكترونية، بدأت دول عديدة سياسة إنشاء "جيوش إلكترونية" نظامية لها ميزانيتها الخاصة، وتسعى للدفاع عن البلاد ضد الهجمات الإلكترونية التي لا تكاد تنتهي حتى تبدأ. وقد أعلنت وزيرة الدفاع الألمانية أورسولا فون دير لاين يوم 6 أبريل/نيسان 2017 عن تكوين جيش إلكتروني كوحدة مستقلة داخل الجيش الألماني إلى جانب القوات البرية والبحرية والجوية، حيث يمارس مهام دفاعية وهجومية على شبكة الإنترنت. وقالت الوزيرة إن عمل الجيش الإلكتروني لن يقتصر على صد هجمات القرصنة، بل سيرد عليها أيضا في ساحة المعركة، وهي الإنترنت. وأضافت "في حال تعرض شبكات الجيش الألماني للهجوم فمن حقنا أيضا أن نرد".
</TEXT>
<KEYWORDS> الجيش الإلكتروني؛ إنترنت؛ حروب إلكترونية؛ فيروسات؛ الجيش الإلكتروني السوري؛ موسوعة الجزيرة؛ </KEYWORDS>
</DOC>

Figure 7: New Arabic corpus form.

Thus we were able to obtain an Arabic corpus containing 2416 documents and 25 requests. The vocabulary number of this corpus is 1475148 words, of which 133474 different words (i.e. 9.03% of the total words).

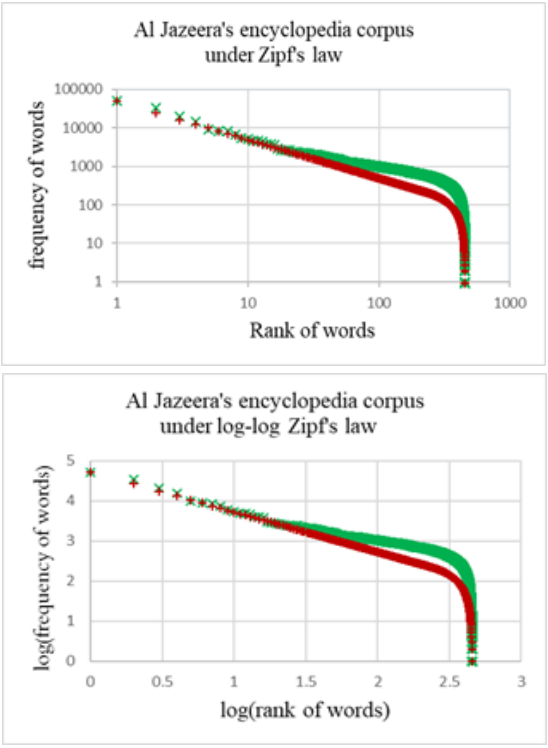


Figure 8: Curve Al Jazeera site corpus according to Zipf's law curve.

According to Zipf's Law, which is concerned with the distribution of words across documents, the range, and highlight the importance of the corpus words. Figure 8 illustrates the Al Jazeera's corpus curve (in red color represented by symbols (+)) and Zipf's curve (in green color represented by symbols (x)). The Figure suggests that Al Jazeera's corpus curve is very close to Zipf's curve. Furthermore, according to some other criteria [31], our new form corpus is very rich and qualified to use as a test collection for IR system quality.

This new format enables us to benefit from, among other things:

- The Contribution to building a system for IRSs evaluation, which enables researchers to test the effectiveness of their applications. In addition to the quality and quantity of the documents considered in this corpus, we have created two types of requests set and their relevant documents. The first is a brief and simple; while the second is extensive and complex, based on the corpus keywords, for example: " الحرب الالكترونية التي يقودها الجيش السوري" (Electronic warfare led by the Syrian Army).
- The Contribution to building a system for keywords systems evaluation, where we have been able to perform extracting experiments using the corpus documents, compare the results of these systems

[3] http://www.aljazeera.net/encyclopedia. Uploaded on November 16, 2017.

with the available keywords and calculate the precision and recall scores.

6 Analysis and results

The aim of our experiments is to evaluate different methods of indexing performance in Arabic information retrieval. A series of experiments was conducted to show the effect of each method of indexing in retrieval performance.

We conducted several experiments using the OIRDA application and endowed it with an offline indexing system for general indexing and information retrieval.

We first compare the following two indexing models:

- *Keyword-based indexing*: the index is composed only of keywords approved by the expert.
- *Indexing without keyword-based or normal indexing*: the index is generated by automatic indexing unit without the intervention of the expert.

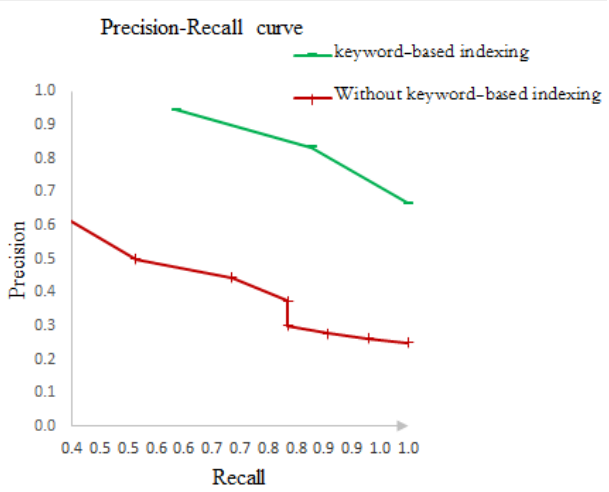


Figure 9: Experiment 1.

Figure 9 represents a comparison between these two models based on their recall-precision curves. The results show that the model of keyword-based indexing, curve in red color represented by symbols (-), is more efficient than the model of indexing without keyword-based, curve in green color represented by symbols (+), on all points of recall and precision.

Then, we compare the two following models of indexing

- *Hybrid*: different combinations of keyword-based indexing and indexing without keyword-based, in a way they token the advantages from each of them.
- *keyword-based indexing*.

In the series of our experiments, the results show that the keyword-based model, curve in green color represented by symbols (-), is more efficient than hybrid model, curve in red color represented by symbols (+). One can observe this behavior in (Figure 10); the curve keyword-based indexing representing the precision based on points of recall is above the hybrid curve.

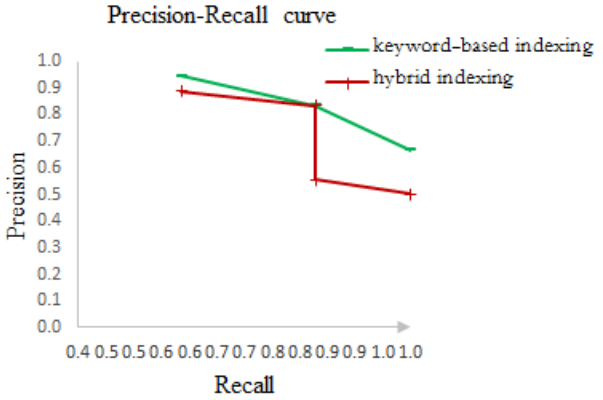


Figure 10: Experiment 2.

The results, further, show that the keyword-based indexing model is the best approach as it is more successful in identifying the descriptors relevant the most to the document. This is primarily due to the intervention of the human expert in keywords identification, especially with ambiguous queries that include polysemy, compound words, etc., which are in need for an accurate semantic processing.

In addition, this model also proved effective to help minimizing index storage size, and thus, improving the response time of the different requests.

The keyword-based indexing model suffers from problems, especially in the case where the expert cannot identify the descriptors that are relevant the most to the document, the aspect this model must improve and find a viable solution to.

7 Conclusion

The main objective of this study is to show the effects of online indexing, which require the semi-automatic indexing, on information retrieval system performance. In addition, this model proved to be effective to help minimize index storage size, and thus, improving the response time of different requests. Therefore, we recommend integrating this model into word processing tools in order to allow the editor to contribute effectively to build a high quality indexes while accounting for the drawbacks and shortcomings of this model. This study also proposes a solution to problems and deficiencies that Arabic language processing suffers from, especially regarding corpus building by developing an application framework for the building and development of corpora. In addition, the paper suggests a solution to reduce deficiencies information retrieval evaluation systems suffer from, which enable researchers to test their indexing and retrieval algorithms and complete systems on common tasks and datasets.

References

[1] S. Bessou, A. Saadi, and M. Touahria, “Un système d’indexation et de recherche des textes en arabe (SITRA).” 1er séminaire national sur le langage naturel et l’intelligence artificielle (LANIA),

Université HAssiba ben Bouali, Département d'Informatique, Chlef (DZ), 2007.

[2] N. Mansour, R. A. Haraty, W. Daher, and M. Houri, "An auto-indexing method for Arabic text," Inf. Process. Manag., vol. 44, no. 4, pp. 1538–1545, 2008. https://doi.org/10.1016/j.ipm.2007.12.007

[3] A. Al Molijy, I. Hmeidi, and I. Alsmadi, "Indexing of Arabic documents automatically based on lexical analysis," Int. J. Nat. Lang. Comput., vol. 1, no. 1, pp. 1–8, 2012.

[4] S. Finch, "Finding structure in language." University of Edinburgh, 1993.

[5] R. Steele and D. Powers, "Evolution and evaluation of document retrieval queries," in Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, 1998, pp. 163–164. https://doi.org/10.3115/1603899.1603927

[6] D. M. W. Powers, "Applications and explanations of Zipf's law," in Proceedings of the joint conferences on new methods in language processing and computational natural language learning, 1998, pp. 151–160. https://doi.org/10.3115/1603899.1603924

[7] J. Entwisle and D. M. W. Powers, "The present use of statistics in the evaluation of NLP parsers," in Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, 1998, pp. 215–224. https://doi.org/10.3115/1603899.1603935

[8] R. El-Khoribi and M. Ismael, "An intelligent system based on statistical learning for searching in arabic text," ICGST Int. J. Artif. Intell. Mach. Learn. AIML, vol. 6, pp. 41–47, 2006.

[9] L. Khreisat, "Arabic text classification using N-gram frequency statistics a comparative study," Conf. Data Mining| DMIN'06, vol. 2006, pp. 78–82, 2006.

[10] A. M. El-Halees, "Arabic text classification using maximum entropy," IUG J. Nat. Stud., vol. 15, no. 1, 2015.

[11] F. Thabtah, "VSMs with K-Nearest Neighbour to categorise Arabic text data," Proc. World Congr. Eng. Comput. Sci., no. WCECS 2008, October 22-24, 2008, San Francisco, USA, pp. 22–25, 2008.

[12] S. Al-Harbi, A. Almuhareb, and A. Al-Thubaity, "Automatic Arabic text classification," 9es Journées Int. Anal. Stat. des Données Textuelles, pp. 77–84, 2008.

[13] F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi, "Naïve Bayesian based on Chi Square to categorize Arabic data," in proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt, 2009, pp. 4–6.

[14] T. F. Gharib, M. B. Habib, and Z. T. Fayed, "Arabic Text Classification Using Support Vector Machines," Int. J. Comput. Their Appl., vol. 16, no. 4, pp. 192–199, 2009.

[15] R. Al-Shalabi, G. Kanaan, and M. H. Gharaibeh, "Arabic Text Categorization Using kNN Algorithm," in Proceedings of The 4th International Multiconference on Computer Science and Information Technology, 2006, vol. 4, pp. 5–7.

[16] S. Raheel and J. Dichy, "An empirical study on the feature's type effect on the automatic classification of Arabic documents," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, vol. 6008 LNCS, pp. 673–686.

[17] E. Bazzi, M. Salim, T. Zaki, D. Mammass, and A. Ennaji, "Indexation automatique des textes arabes: état de l'art.," E-Ti E-Review Technol. Inf., no. 9, 2016.

[18] N. Tazit, S. S. El Hossin Bouyakhf, A. Yousfi, and K. Bouzouba, "Semantic internet search engine with focus on Arabic language," 2007.

[19] M. A. Abderrahim, M. Dib, M. E. A. Abderrahim, and M. A. Chikh, "Semantic indexing of Arabic texts for information retrieval system," Int. J. Speech Technol., vol. 19, no. 2, pp. 229–236, 2016. https://doi.org/10.1007/s10772-015-9307-3

[20] M. A. Abderrahim, M. E. A. Abderrahim, and M. A. Chikh, "Using Arabic wordnet for semantic indexation in information retrieval system," arXiv Prepr. arXiv1306.2499, 2013.

[21] A. Mahgoub, M. Rashwan, H. Raafat, M. Zahran, and M. Fayek, "Semantic query expansion for Arabic information retrieval," in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, pp. 87–92. https://doi.org/10.3115/v1/W14-3611

[22] F. Harrag, E. El-Qawasmah, and A. M. S. Al-Salman, "Stemming as a feature reduction technique for Arabic text categorization," in Proceedings of the 10th International Symposium on Programming and Systems, ISPS' 2011, 2011, pp. 128–133. https://doi.org/10.1109/ISPS.2011.5898874

[23] R. Mohamed and J. Watada, "An evidential reasoning based LSA approach to document classification for knowledge acquisition," in IEEM2010 - IEEE International Conference on Industrial Engineering and Engineering Management, 2010, pp. 1092–1096. https://doi.org/10.1109/IEEM.2010.5674188

[24] F. S. Al-Anzi and D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing," J. King Saud Univ. Inf. Sci., vol. 29, no. 2, pp. 189–195, 2017. https://doi.org/10.1016/j.jksuci.2016.04.001

[25] T. Dilekh and A. Behloul, "Implementation of a New Hybrid Method for Stemming of Arabic Text," Analysis, vol. 46, no. 8, pp. 14–19, 2012.

[26] S. R. El-Beltagy and A. Rafea, "KP-Miner: A keyphrase extraction system for English and Arabic documents," Inf. Syst., vol. 34, no. 1, pp. 132–144, 2009. https://doi.org/10.1016/j.is.2008.05.002

[27] E. Amer and K. Foad, “Akea: an Arabic keyphrase extraction algorithm,” in International Conference on Advanced Intelligent Systems and Informatics, 2016, pp. 137–146.

[28] M. Al-Kabi, H. Al-Belaili, B. Abul-Huda, and A. Wahbeh, “Keyword extraction based on word co-occurrence statistical information for arabic text,” Abhath Al-Yarmouk" Basic Sci. Eng., vol. 22, no. 1, pp. 75–95, 2013.

[29] T. El-Shishtawy and A. Al-sammak, “Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques,” ReCALL, pp. 1–8, 2012.

[30] R. Duwairi and M. Hedaya, “Automatic keyphrase extraction for Arabic news documents based on KEA system,” J. Intell. Fuzzy Syst., vol. 30, no. 4, pp. 2101–2110, 2016. https://doi.org/10.3233/IFS-151923

[31] Y. Kadri and J. Y. Nie, “Effective stemming for Arabic information retrieval,” Proc. Chall. Arab. nLP/mt, Int. conf. Br. Comput. Soc., pp. 68–74, 2006.

[32] I. A. Al-Kharashi and M. W. Evens, “Comparing words, stems, and roots as index terms in an Arabic information retrieval system,” J. Am. Soc. Inf. Sci., vol. 45, no. 8, p. 548, 1994. https://doi.org/10.1002/(SICI)1097-4571(199409)45:8<548::AID-ASI3>3.0.CO;2-X

[33] K. Beesley, “Arabic morphological analysis on the Internet,” in Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing, 1998.

[34] L. S. Larkey and M. E. Connell, “Arabic Information Retrieval at UMass in TREC-10.,” in TREC, 2001.

[35] E. M. Voorhees, “Overview of TREC 2003.,” in Trec, 2003, pp. 1–13.

Entropy, Distance and Similarity Measures under Interval-Valued Intuitionistic Fuzzy Environment

Pratiksha Tiwari and Priti Gupta
Delhi Institute of Advanced Studies, Plot No. 6, Sector- 25, Rohini, Delhi, India
E-mail: parth12003@yahoo.co.in

Keywords: entropy measure, distance, similarity measure, interval valued intuitionistic fuzzy sets

Received: July 21, 2016

This paper presents new axiomatic definitions of entropy measure using concept of probability and distance for interval-valued intuitionistic fuzzy sets (IvIFSs) by considering degree of hesitancy which is consistent with the definition of entropy given by De Luca and Termini. Thereafter, we propose some entropy measures and also derived relation between distance, entropy and similarity measures for IvIFSs. Further, we checked the performance of proposed entropy and similarity measures on the basis of intuition and compared with the existing entropy and similarity measures using numerical examples. Lastly, proposed similarity measures are used to solve problems in the field of pattern recognition and medical diagnoses.

Povzetek: V prispevku so predstavljene nove aksiomatske definicije entropijske mere za intervalno intuicionistične mehke množice.

1 Introduction

Fuzzy set theory (Zadeh, 1965) is tool that can handle uncertainty and imprecision effortlessly. Interval-valued, intuitionistic, interval-valued intuitionistic fuzzy sets (Zadeh (1975), Atanassov (1986), Atanassov & Gargov (1989)), vague sets (Gau & Buehrer, 1993) and R-fuzzy sets (Yang, Hinde, 2010) are various generalizations of Fuzzy sets (FSs). From all these generalizations IvIFSs and intuitionistic fuzzy sets (IFSs) are two conventional extensions of FSs. IvIFSs are more practical and flexible than IFSs as they are characterized by membership and non-membership degree range instead of real numbers. It makes IvIFSs more useful in dealing with real world complexities which arises due to insufficient information, lack of data, imprecise knowledge and human nature wherein range is provided instead of real numbers. Distance, entropy and similarity measures are the central arenas that are investigated by various researchers under intuitionistic and interval-valued fuzzy environment (IFE and IvFE). These measures identify the similarity or dissimilarity between two FSs. Till date, vivid entropy, distance or similarity measures are presented by various investigators. Some of these research findings are mentioned as follows: Xu (2007 a, b) introduced the concept of similarity between IvIFSs along with some distance measure. Zang et al. (2009) defined a entropy axiomatically for interval-valued fuzzy sets (IvFSs) and discussed relation between entropy and similarity measures. Xu and Yager (2009) studied preference relation and defined similarity measure under IvFE and interval-valued intuitionistic environment (IvIFE). Wei et al (2011) derived a generalized measure of entropy for IvIFSs. Cosine similarity measures for IvIFSs are defined by both Ye (2012) and Singh (2012). Sun & Liu (2012), Hu & Li (2013), Zhang et al. (2014) proposed entropy and similarity measure along with their relationship for IvIFSs. Applications of the aforesaid entropy, distance and similarity measures are for recognition of patterns, medical diagnoses, and decision making with multiple criteria and expert systems problems. However, most of these distance, similarity or entropy measures do not consider hesitancy index between IvIFSs. Hesitance index play a very important role when membership and non-membership degree do not differ much for two IvIFSs but their hesitant index does. Some of the authors, Xu (2007), Xu & Xia (2011), Wu et al. (2014) considered hesitancy into the measure of distance, similarity and entropy developed by them. Since hesitancy index also has a vital role in any decision making as it outclasses the existing methods and deals with decision process in a better way. Dammak et al. (2016) studies some possibility measures in multi-criteria decision making under $IvIFE$. Tiwari & Gupta(in press) proposed generalized entropy and similarity measure for $IvIFS$ with application in decision making. Zang et al. (2016) defined some operations on $IvIFSs$ and proposed some aggregation operators for $IvIFSs$ w.r.t. the restricted interval Shapley function with application in multi-criteria decision making. In this paper we have developed some of the distance, entropy and similarity measures by taking all the three degrees in account and applied it to pattern recognition and medical diagnoses under $IvIFE$.

This work is organized in various sections. Section 2 has basic definition and operations on $IvIFSs$. Section 3, presents the relationship between distance and entropy measures along with example to check the performance of entropy measures on the basis of intuition. A relation between measure of entropy and similarity measure is proposed in Section 4. Further, comparison of new similarity measures with the few existing one in done.

Thereafter in section 5 we applied new similarity measures to recognition of patterns and medical diagnoses. Lastly conclusion is drawn in Section 6.

2 IvIFSs along with its distance and similarity measures

This section has definitions and concepts for IvIFSs. In this paper $\Omega = \{x_1, \ldots, x_n\}$ denotes the universe of discourse;$\mathbb{C}(\Omega)$ and IvIFSs(Ω) denote all crisp sets and $IvIFSs$ respectively in Ω.

Definition 1 (Atanassov & Gargov,1989): An IvIFS A in the finite universe Ω is defined by a triplet $\langle x_i, MV_A(x_i), NV_A(x_i), HV_A(x_i)\rangle$ as $x_i \in \Omega$ where $MV_A(x_i) = [MV_{AL}(x_i), MV_{AU}(x_i)]$ is called membership value interval ,$NV_A(x_i) = [NV_{AL}(x_i), NV_{AU}(x_i)]$ is called non-membership value interval and $HV_A(x_i) = [HV_{AL}(x_i), HV_{AU}(x_i)], HV_{AL}(x_i) = 1 - MV_{AU}(x_i) - NV_{AU}(x_i)$, $H_{AU}(x_i) = 1 - MV_{AL}(x_i) - NV_{AL}(x_i)$ such that $0 \le MV_{AU}(x_i) + NV_{AU}(x_i) \le 1$ for each $x_i \in A$.

Liu (1992) defined distance and similarity measures for IvIFSs axiomatically which are given as follows:

Definition 2: For any two IvIFSs A and B, a real valued function $\mathrm{D}: \mathrm{IvIFSs}(\Omega) \times \mathrm{IvIFSs}(\Omega) \longrightarrow [0,1]$ is termed as a distance measure of IvIFSs on Ω, if it satisfies the below mentioned axioms:

1. For any crisp set A, we have $\mathrm{D}(\mathrm{A}, \overline{\mathrm{A}}) = 1$.
2. Distance between any two IvIFSs A and B is zero iff $\mathrm{A} = \mathrm{B}$.
3. Distance measure is symmetrical w.r.t to any two IvIFSs A and B.
4. For any three IvIFSs A, B and C such that $\mathrm{A} \subseteq \mathrm{B} \subseteq \mathrm{C}$, we have $\mathrm{D}(\mathrm{A}, \mathrm{C}) \ge \mathrm{D}(\mathrm{A}, \mathrm{B})$ and $\mathrm{D}(\mathrm{A}, \mathrm{C}) \ge \mathrm{D}(\mathrm{B}, \mathrm{C})$.

Distance between FSs was presented by (Kacprzyk, 1997). Then its extension was proposed by Atanassov in 1999 as two dimensional distances whereas third parameter hesitancy degree in distance was introduced by Szmidt and Kacprzyk (2000) for intuitionistic fuzzy sets. Yang & Chiclana (2012) proved three dimensional distance consistency over two dimensional distances. Grzegorzewski (2004) and Park et al. (2007) gave distance measure for IvFSs and IvIFSs respectively. Here we extend the distance measures by considering hesitancy degree for IvIFSs. For any two IvIFSs A and B, we define the following measures of distance:

1) Normalized Euclidean Distance

$$\mathrm{D}_1(\mathrm{A},\mathrm{B}) = \left\{\frac{1}{12n}\sum_{i=1}^{n}\left[\left(MV_{AL}(x_i) - MV_{BL}(x_i)\right)^2 + \left(MV_{AU}(x_i) - MV_{BU}(x_i)\right)^2 + \left(NV_{AL}(x_i) - NV_{BL}(x_i)\right)^2 + \left(NV_{AU}(x_i) - NV_{BU}(x_i)\right)^2 + \left(HV_{AL}(x_i) - HV_{BL}(x_i)\right)^2 + \left(HV_{AU}(x_i) - HV_{BU}(x_i)\right)^2\right]\right\}^{1/2} \quad \ldots(1)$$

2) Normalized Hamming Distance

$$\mathbf{D_2(A,B)} = \mathbf{\frac{1}{8n}\sum_{i=1}^{n}[|MV_{AL}(x_i) - MV_{BL}(x_i)| + |MV_{AU}(x_i) - MV_{BU}(x_i)| + |NV_{AL}(x_i) - NV_{BL}(x_i)| + |NV_{AU}(x_i) - NV_{BU}(x_i)| + |HV_{AL}(x_i) - HV_{BL}(x_i)| + |HV_{AU}(x_i) - HV_{BU}(x_i)|]} \quad \ldots(2)$$

3) Hamming Hausdorff Normalized Distance

$$\mathrm{D}_3(\mathrm{A},\mathrm{B}) = \frac{1}{4n}\sum_{i=1}^{n}[|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)| + |NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)| + |HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|] \quad \ldots(3)$$

4) Hausdorff Normalized Hamming Distance

$$\mathrm{D}_4(\mathrm{A},\mathrm{B}) = \frac{1}{4n}\sum_{i=1}^{n} max\left\{\frac{|MV_{AL}(x_i)-MV_{BL}(x_i)|+|MV_{AU}(x_i)-MV_{BU}(x_i)|}{2}, \frac{|NV_{AL}(x_i)-NV_{BL}(x_i)|+|NV_{AU}(x_i)-NV_{BU}(x_i)|}{2}, \frac{|HV_{AL}(x_i)-HV_{BL}(x_i)|+|HV_{AU}(x_i)-HV_{BU}(x_i)|}{2}\right\} \quad \ldots(4)$$

5) Averaged fifth Distance Measure

$$\mathrm{D}_5(\mathrm{A},\mathrm{B}) = \frac{1}{2n}\sum_{i=1}^{n}\left\{\frac{\begin{bmatrix}|MV_{AL}(x_i)-MV_{BL}(x_i)|+|MV_{AU}(x_i)-MV_{BU}(x_i)| \\ +|NV_{AL}(x_i)-NV_{BL}(x_i)|+|NV_{AU}(x_i)-NV_{BU}(x_i)| \\ +|HV_{AL}(x_i)-HV_{BL}(x_i)|+|HV_{AU}(x_i)-HV_{BU}(x_i)|\end{bmatrix}}{8} + \frac{max\begin{pmatrix}|MV_{AL}(x_i)-MV_{BL}(x_i)|+|MV_{AU}(x_i)-MV_{BU}(x_i)|, \\ |NV_{AL}(x_i)-NV_{BL}(x_i)|+|NV_{AU}(x_i)-NV_{BU}(x_i)|, \\ |HV_{AL}(x_i)-HV_{BL}(x_i)|+|HV_{AU}(x_i)-HV_{BU}(x_i)|\end{pmatrix}}{4}\right\} \quad \ldots(5)$$

6) Generalized Measure of Distance , for $\mathrm{p} \ge 2$,

$$\mathrm{D}_6(\mathrm{A},\mathrm{B}) = \left\{\frac{1}{12n}\sum_{i=1}^{n}(|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|)^p + (|NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|)^p + (|HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)^p\right\}^{1/\mathrm{p}} \quad \ldots(6)$$

Definition 3: Let A and B be any two IvIFSs, a real valued function $\mathrm{S}: \mathrm{IvIFSs}(\Omega) \times \mathrm{IvIFSs}(\Omega) \longrightarrow [0,1]$ is defined as a measure of similarity for IvIFSs on Ω, if it satisfies axioms mentioned below:

1. For any crisp set A, we have $\mathrm{S}(\mathrm{A}, \overline{\mathrm{A}}) = 0$
2. Measure of similarity between any two IvIFSs is 1 iff $\mathrm{A} = \mathrm{B}$.
3. Measure of similarity is symmetric w.r.t. any two IvIFSs.
4. For any three IvIFSs A, B and C such that $\mathrm{A} \subseteq \mathrm{B} \subseteq \mathrm{C}$. We have $\mathrm{S}(\mathrm{A}, \mathrm{C}) \le \mathrm{S}(\mathrm{A}, \mathrm{B})$ and $\mathrm{S}(\mathrm{A}, \mathrm{C}) \le \mathrm{S}(\mathrm{B}, \mathrm{C})$.

From axiomatic definition of distance and similarity measures it is clear that $S(A,B) = 1 - D(A,B)$ where A and B are IvIFSs , D and S are distance and similarity measure for IvIFSs respectively.

2.1 Entropy measure for IvIFSs

In 1972, De Luca and Termini defined measure of entropy for FSs. Hung & Yang (2006) extended definition for IFSs considering hesitancy degree. The following definition for entropy is an extension of definition of entropy proposed by Hung & Yang (2006) for IvIFSs.

Definition 4: A real valued function $E: IvIFSs(\Omega) \to [0,1]$ is termed as measure of entropy under IvIFE, if below mentioned axioms are satisfied:

1. $E(A) = 0, \forall A \in \mathbb{C}(\Omega)$;
2. $E(A) = 1$, iff $M_A(x_i) = N_A(x_i) = H_A(x_i) = \left[\frac{1}{3}, \frac{1}{3}\right], \ \forall\, x_i \in \Omega$;
3. $E(A) \leq E(B)$, if A is less fuzzy than B;
4. $E(A) = E(\bar{A})$, where $\bar{A}$ is complement of A, where $A, B \in IvIFSs(\Omega)$.

Above definition is steady with description of measure of entropy given by De Luca & Termini (1972). As it is known that complete description of an IvIFS $A \in \Omega$ has three degrees membership, non-membership and hesitancy with $MV_{AU}(x_i) + NV_{AU}(x_i) + HV_{AL}(x_i) = 1$ and $MV_{AL}(x_i) + NV_{AL}(x_i) + HV_{AU}(x_i) = 1$ with $0 \leq MV_{AU}(x_i), NV_{AU}(x_i), HV_{AL}(x_i),$ $MV_{AL}(x_i), NV_{AL}(x_i), HV_{AU}(x_i) \leq 1$. By taking all the three in to consideration we may assume them as probability measure. Therefore the entropy is maximum when all the variables are equal (i.e. $MV_{AU}(x_i) = NV_{AU}(x_i) = HV_{AL}(x_i) = \frac{1}{3}$ and $MV_{AL}(x_i) = NV_{AL}(x_i) = HV_{AU}(x_i) = \frac{1}{3}$) and zero (minimum) when only one variable is exists (i.e. $MV_{AL}(x_i) = MV_{AU}(x_i) = 1, NV_{AL}(x_i) = NV_{AU}(x_i) = 0, HV_{AL}(x_i) = HV_{AU}(x_i) = 0$ or $MV_{AL}(x_i) = MV_{AU}(x_i) = 0, NV_{AL}(x_i) = NV_{AU}(x_i) = 1, HV_{AL}(x_i) = HV_{AU}(x_i) = 1$ or $MV_{AL}(x_i) = MV_{AU}(x_i) = 0, NV_{AL}(x_i) = NV_{AU}(x_i) = 0, HV_{AL}(x_i) = HV_{AU}(x_i) = 1$).
Again we extend the definition of entropy given by Zang et al. (2014) based on distance for IvIFSs. In following definition we have considered degree of hesitation, which is not considered by other definitions of entropy.

Definition 5: A real-valued function $E: IvIFSs(\Omega) \to [0,1]$ is termed as measure of entropy under IvIFE, if the following axioms are satisfied:

1. $E(A) = 0, \forall A \in \mathbb{C}(\Omega)$;
2. $E(A) = 1$, iff all the three description of IvIFSs intervals satifies $MV_A(x_i) = NV_A(x_i) = HV_A(x_i) = \left[\frac{1}{3}, \frac{1}{3}\right], \ \forall\, x_i \in \Omega$
3. If $D\left(A, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right) \geq D\left(B, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right)$, then $E(A) \leq E(B)$, $\forall A, B \in IvIFSs(\Omega)$, where D is measure of distance.
4. $E(A) = E(\bar{A})$, where $\bar{A}$ is complement of A, where $A, B \in IvIFSs(\Omega)$.

In the next section, we derive a relation which relates measure of distance and entropy for IvIFSs, which satisfies all the axioms of the definition of entropy.

3 Relation between measure of distance and entropy

Here, we develop a technique which obtains entropy measure for IvIFSs which satisfies the aforementioned properties.

Theorem 2: Let D_j, $j = 1, \ldots, 6$ be the above-mentioned six distance measure equations (1)-(6) between IvIFSs, then, $E_j(A) = 1 - 3D_j\left(A, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right)$, $j = 1, \ldots, 6$ for any $A \in IvIFSs(\Omega)$ are measure of entropy of IvIFSs.

Proof: We prove that $E_j(A)$, for $j = 1, \ldots, 6$ satisfies conditions given by definition 5.

Property1): If $A \in \mathbb{C}(\Omega) \Rightarrow A(x_i) = \langle[1,1], [0,0], [0,0]\rangle$ or $A(x_i) = \langle[0,0], [1,1], [0,0]\rangle, \forall\, x_i \in \Omega$, then for $j = 1, \ldots, 6$

$$D_j\left(A, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right) = \frac{1}{3}.$$

Thus, $E_j(A) = 0$

Property 2): For all $j = 1, \ldots, 6$, $E_j(A) = 1$

$$\Leftrightarrow 1 - 3D_j\left(A, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right) = 1$$

$$\Leftrightarrow 3D_j\left(A, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right) = 0$$

$$\Leftrightarrow A = \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle$$

Property3): Let A and B be any two IvIFSs and $D_j\left(A, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right) \geq D_j\left(B, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right)$ then

$$1 - 3D_j\left(A, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right) \geq 1 - 3D_j\left(B, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right)$$

$\Rightarrow E_j(A) \leq E_j(B)$, for all $j = 1, \ldots, 6$

Property 4) : Let A be any $IvIFS$ then $\bar{A} = \{\langle x_i, [NV_{AL}(x_i), NV_{AU}(x_i)], [MV_{AL}(x_i), MV_{AU}(x_i)]\rangle / x_i \in \Omega\}$

$$\Rightarrow D_j\left(A, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right) = D_j\left(\bar{A}, \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right)$$

Thus, $E_j(A) = E_j(\bar{A})$, for all $j = 1, \ldots, 6$. ■

From theorem 2 and various distance formulas' mentioned (equation (1) to (6)), we get corresponding entropy formulas as follows:

$$E_1(A) = 1 - 3\left\{\frac{1}{12n}\sum_{i=1}^{n}\left[\left(MV_{AL}(x_i) - \frac{1}{3}\right)^2 + \left(MV_{AU}(x_i) - \frac{1}{3}\right)^2 + \left(NV_{AL}(x_i) - \frac{1}{3}\right)^2 + \left(NV_{AU}(x_i) - \frac{1}{3}\right)^2 + \left(HV_{AL}(x_i) - \frac{1}{3}\right)^2 + \left(HV_{AU}(x_i) - \frac{1}{3}\right)^2\right]\right\}^{1/2}$$

$$E_2(A) = 1 - \frac{3}{8n}\sum_{i=1}^{n}\left[\left|MV_{AL}(x_i) - \frac{1}{3}\right| + \left|MV_{AU}(x_i) - \frac{1}{3}\right| + \left|NV_{AL}(x_i) - \frac{1}{3}\right| + \left|NV_{AU}(x_i) - \frac{1}{3}\right| + \left|HV_{AL}(x_i) - \frac{1}{3}\right| + \left|HV_{AU}(x_i) - \frac{1}{3}\right|\right]$$

$$E_3(A) = 1 - \frac{3}{4n}\sum_{i=1}^{n}\left[\left|MV_{AL}(x_i) - \frac{1}{3}\right| \vee \left|MV_{AU}(x_i) - \frac{1}{3}\right| + \left|NV_{AL}(x_i) - \frac{1}{3}\right| \vee \left|NV_{AU}(x_i) - \frac{1}{3}\right| + \left|HV_{AL}(x_i) - \frac{1}{3}\right| \vee \left|HV_{AU}(x_i) - \frac{1}{3}\right|\right]$$

$$E_4(A) = 1 - \frac{3}{2n}\sum_{i=1}^{n} max\left\{\frac{|MV_{AL}(x_i)-1/3|+|MV_{AU}(x_i)-1/3|}{2}, \frac{|NV_{AL}(x_i)-1/3|+|NV_{AU}(x_i)-1/3|}{2}, \frac{|HV_{AL}(x_i)-1/3|+|HV_{AU}(x_i)-1/3|}{2}\right\}$$

$$\mathrm{E}_5(\mathrm{A},\mathrm{B}) = 1 - \frac{3}{2\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}\left\{\frac{\begin{bmatrix}|MV_{AL}(x_i) - 1/3| + |MV_{AU}(x_i) - 1/3| \\ +|NV_{AL}(x_i) - 1/3| + |NV_{AU}(x_i) - 1/3| \\ +|HV_{AL}(x_i) - 1/3| + |HV_{AU}(x_i) - 1/3|\end{bmatrix}}{8} + \frac{\max\begin{pmatrix}|MV_{AL}(x_i) - 1/3| + |MV_{AU}(x_i) - 1/3|, \\ |NV_{AL}(x_i) - 1/3| + |NV_{AU}(x_i) - 1/3|, \\ |HV_{AL}(x_i) - 1/3| + |HV_{AU}(x_i) - 1/3|\end{pmatrix}}{2}\right\}$$

$$E_6(A,B) = 1 - 3\sum_{i=1}^{n}\left[\frac{1}{12n}\left[\left(\left|MV_{AL}(x_i) - \frac{1}{3}\right| \vee \left|MV_{AU}(x_i) - \frac{1}{3}\right|\right)^p + \left(\left|NV_{AL}(x_i) - \frac{1}{3}\right| \vee \left|NV_{AU}(x_i) - \frac{1}{3}\right|\right)^p + \left(\left|HV_{AL}(x_i) - \frac{1}{3}\right| \vee \left|HV_{AU}(x_i) - \frac{1}{3}\right|\right)^p\right]\right]^{1/p}$$

To check the consistency of proposed entropy measures with the intuitionist beliefs we have used the following example.

Example: Consider two IvIFSs $A = \{x, \langle[0.2,0.2], [0.2,0.3], [0.5,0.6]\rangle, x \in \Omega\}$ and $B = \{x, \langle[0.2,0.3], [0.4,0.6], [0.1,0.4]\rangle, x \in \Omega\}$, clearly we can see that A is more fuzzy than B. Then $\mathrm{E}_j(\mathrm{A})$ and $\mathrm{E}_j(\mathrm{B})$ are given in Table 1:

Entropies	A	B
E_1	0.5570	0.65866
E_2	0.675	0.7
E_3	0.6	0.525
E_4	0.675	0.75
E_5	0.5125	0.6
E_6	0.7171	0.672

Table 1: Entropies.

Since $\mathrm{E}_3(\mathrm{A}) > \mathrm{E}_3(\mathrm{B})$ and $\mathrm{E}_6(\mathrm{A}) > \mathrm{E}_6(\mathrm{B})$ which indicates that $\mathrm{E}_3$and $\mathrm{E}_6$are consistent with the intuition.

3.1 Comparison of existing entropy measure with proposed entropy measures

We compared performance of existing entropy measures with the proposed measures with the help of an example. Let A be an IvIFS, then

$$E_{ZJ}(A) = \frac{1}{n}\sum_{=1}^{n}\frac{min(MV_{AL}(x_i),NV_{AL}(x_i)) + min(MV_{AU}(x_i),NV_{AU}(x_i))}{max(MV_{AL}(x_i),NV_{AL}(x_i)) + max(MV_{AU}(x_i),NV_{AU}(x_i))},$$

$$E_{WW}(A) = \frac{1}{\mathrm{n}}\sum_{=1}^{\mathrm{n}}\frac{\left[\begin{matrix}\min(\mathrm{MV_{AL}(x_i)},\mathrm{NV_{AL}(x_i)}) \\ +\min(\mathrm{MV_{AU}(x_i)},\mathrm{NV_{AU}(x_i)}) + \mathrm{HV_{AL}(x_i)} + \mathrm{HV_{AU}(x_i)}\end{matrix}\right]}{\left[\begin{matrix}\max(\mathrm{MV_{AL}(x_i)},\mathrm{NV_{AL}(x_i)}) \\ +\max(\mathrm{MV_{AU}(x_i)},\mathrm{NV_{AU}(x_i)}) + + \mathrm{HV_{AL}(x_i)} + \mathrm{HV_{AU}(x_i)}\end{matrix}\right]},$$

$$E_{ZM}(A) = \frac{1}{n}\sum_{i=1}^{n}\left[1 - \left(\overline{MV}_A(x_i) + \overline{NV}_A(x_i)\right)e^{1-\left(\overline{MV}_A(x_i) + \overline{NV}_A(x_i)\right)}\right],$$

where $\overline{MV}_A(x_i) = MV_{AL}(x_i) + \tau\Delta MV_A(x_i)$, $\overline{NV}_A(x_i) = NV_{AL}(x_i) + \tau\Delta NV_A(x_i)$ *and* $\Delta MV_A(x_i) = MV_{AU}(x_i) - MV_{AL}(x_i)$ *and* $\Delta NV_A(x_i) = NV_{AU}(x_i) - NV_{AL}(x_i), \tau \in [0,1]$

are the measures of entropies given by Zang et al. (2010), Wei et al.(2011) and Zang et al. (2011) respectively.

Example: Consider the example from Sun & Liu (2012) to review the entropies for an IvIFSs A_i which are as follows:

"$A_1 = \{x, \langle[0.1,0.2], [0.2,0.4], [0.4,0.7]\rangle, x \in \Omega\}$,

$A_2 = \{x, \langle[0.2,0.2], [0.3,0.5], [0.3,0.5]\rangle, x \in \Omega\}$,

$A_3 = \{x, \langle[0.2,0.4], [0,0], [0.6,0.8]\rangle, x \in \Omega\}$,

$A_4 = \{x, \langle[0.3,0.4], [0,0.142857], [0.4571,0.7]\rangle, x \in \Omega\}$,

$A_5 = \{x, \langle[0.1,0.1], [0,0.2], [0.6,0.9]\rangle, x \in \Omega\}$,

$A_6 = \{x, \langle[0,0.2], [0,0.2], [0.4,1.0]\rangle, x \in \Omega\}$"

	$E_1(A_i)$	$E_2(A_i)$	$E_3(A_i)$	$E_4(A_i)$	$E_5(A_i)$	$E_6(A_i)$
A_1	0.58167	0.625	0.45	0.675	0.4875	0.6063
A_2	0.735425	0.75	0.65	0.8	0.675	0.765479
A_3	0.367544	0.4	0.3	0.45	0.15	0.490098
A_4	0.523512	0.582143	0.425	0.607143	0.398214	0.566987
A_5	0.27889	0.3	0.15	0.3	0.05	0.395848
A_6	0.271989	0.375	0.01	0.45	0.1375	0.292893

Table 2: Comparison of entropies.

The values of $E_{ZJ}(A_1) = 0.5 = E_{ZJ}(A_2)$, $E_{WW}(A_3) = 0.7 = E_{WW}(A_4)$ and $E_{ZM}(A_5) = 0.5 = E_{ZM}(A_6)$. Thus, the entropies $E_{ZJ}(A_i)$, $E_{WW}(A_i)$ and $E_{ZM}(A_i)$ are unreasonable. Proposed entropies $\mathrm{E_j}$, $\mathrm{j} = 1,2,..6$ can discriminate the fuzziness of all the IvIFSs $A_i, i = 1, \dots, 6$ given as follows and give the reasonable results given in Table 2.

Here we have proposed some entropy measures and evaluated its performance on the basis of intuitionistic belief and comparison with existing measures. In next section we propose relation between measures of entropy and similarity under IvIFE. Also, we have defined new measures of similarity and have evaluated their performance by comparing them with some existing measures.

4 Relations between measure of entropy and similarity together with new similarity measures

In this section contains definition of new measures of similarity under IvIFE and determined an important relation between entropy and similarity measure IvIFSs which is discussed as follows.

Theorem 3: Let S_j, for $j = 1,..,6$ be measure of similarity of IvIFSs w.r.t. the measure of distance D_j, for $j = 1,..,6$ respectively, and A be any IvIFS. Then $E_j(A) = 3S_j\left(A, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) - 2$, for $j = 1, \dots, 6$ are measures of entropy for IvIFSs.

Proof: We prove that $\mathrm{E_j(A)}$, for $\mathrm{j} = 1, \dots, 6$ satisfies conditions given by definition 5.

Property 1): If $\mathrm{A} \in \mathbb{C}(\Omega) \Rightarrow A(x_i) = \langle[1,1],[0,0],[0,0]\rangle$ or $A(x_i) = \langle[0,0],[1,1],[0,0]\rangle$, $\forall\, x_i \in \Omega$, then for $j = 1, \dots, 6$

$$\mathrm{S_j}\left(\mathrm{A}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) = 1 - \mathrm{D_j}\left(\mathrm{A}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) = \frac{2}{3}.$$

Thus, $\mathrm{E_j(A)} = 0$

Property 2): For all $j = 1, \dots, 6, E_j(A) = 1$

$$\Leftrightarrow 3\mathrm{S_j}\left(\mathrm{A}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) - 2 = 1$$

$$\Leftrightarrow \mathrm{S_j}\left(\mathrm{A}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) = 1$$

$$\Leftrightarrow \mathrm{A} = \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle$$

Property3): Let A and B be any two IvIFSs and $\mathrm{D_j}\left(\mathrm{A}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) \geq \mathrm{D_j}\left(\mathrm{B}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right)$ then

$$1 - \mathrm{D}_j\left(\mathrm{A}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) \leq 1 - \mathrm{D}_j\left(\mathrm{B}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right)$$

$$\Leftrightarrow \mathrm{S_j}\left(\mathrm{A}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) \leq \mathrm{S}_j\left(\mathrm{B}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right)$$

$$\Leftrightarrow \mathrm{E}_j(\mathrm{A}) \leq \mathrm{E_j(B)}, \text{ for all } j = 1, \dots, 6$$

Property 4) : Let A be any IvIFS then $\overline{\mathrm{A}} = \{\langle x_i, [NV_{AL}(x_i), NV_{AU}(x_i)], [MV_{AL}(x_i), MV_{AU}(x_i)]\rangle / x_i \in \Omega\}$

$$\Rightarrow D_j\left(\mathrm{A}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) = \mathrm{D_j}\left(\overline{\mathrm{A}}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right)$$

$$\Rightarrow \mathrm{S}_j\left(\mathrm{A}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) = \mathrm{S}_j\left(\overline{\mathrm{A}}, \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right)$$

Thus, $\mathrm{E_j(A)} = \mathrm{E_j}(\overline{\mathrm{A}})$, for all $\mathrm{j} = 1, \dots, 6$. ■

Next, we present a conversion technique to define similarity measures established by entropy measure for IvIFSs.

Definition 6 : For any two IvIFSs A and B in Ω, such that both A and B are defined by the triplet $\langle x_i, [MV_{AL}(x_i), MV_{AU}(x_i)], [NV_{AL}(x_i), NV_{AU}(x_i)]\rangle$ and $\langle x_i, [MV_{BL}(x_i), MV_{BU}(x_i)], [NV_{BL}(x_i), NV_{BU}(x_i)]\rangle$ respectively. we define an IvIFSs $\emptyset(\mathrm{A}, \mathrm{B})$using A and B as given below:

$MV_{\emptyset(A,B)L}(x_i) = \frac{1}{3}\{1 - [max(|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|, |NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|, |HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)]^{1/2}\}$;

$MV_{\emptyset(A,B)U}(x_i) = \frac{1}{3}\{1 - [max(|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|, |NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |N_{AU}(x_i) - N_{BU}(x_i)|, |H_{AL}(x_i) - H_{BL}(x_i)| \vee |H_{AU}(x_i) - H_{BU}(x_i)|)]\}$;

$NV_{\emptyset(A,B)L}(x_i) = \frac{1}{3}\{1 + [min(|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|, |NV_{AL}(x_i) - NV_{BL}(x_i)| \vee$

$|NV_{AU}(x_i) - NV_{BU}(x_i)|, |HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)]^2\}$;
$NV_{\emptyset(A,B)U}(x_i) = \frac{1}{3}\{1 + [min(|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|, |NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|, |HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)]\}$.

Theorem 4: $E(\emptyset(A, B))$ be a similarity measure for IvIFSs Aand B, where E is an entropy.

Proof: To prove that $E(\emptyset(A, B))$ is a measure of similarity, we need to prove property given by definition 3 holds .

Property 1): If $A \in \mathbb{C}(\Omega) \Rightarrow A(x_i) = \langle[1,1],[0,0],[0,0]\rangle$ *or* $A(x_i) = \langle[0,0],[1,1],[0,0]\rangle$, for any $x_i \in \Omega$, then $MV_{\emptyset(A,\bar{A})L}(x_i) = 0 = MV_{\emptyset(A,\bar{A})U}(x_i)$;
And $NV_{\emptyset(A,\bar{A})L}(x_i) = 1 = NV_{\emptyset(A,\bar{A})U}(x_i)$
Thus, $\emptyset(A, \bar{A}) = \{\langle x_i, [0,0], [1,1], [0,0]\rangle / x_i \in \Omega\}$
$\Rightarrow S(A, \bar{A}) = E(\emptyset(A, \bar{A})) = 0$

Property 2): Assume that $S(A, B) = 1 \Rightarrow E(\emptyset(A, B)) = 1$
$\Leftrightarrow MV_{\emptyset(A,B)}(x_i) = NV_{\emptyset(A,B)}(x_i) = HV_{\emptyset(A,B)}(x_i) = \left[\frac{1}{3}, \frac{1}{3}\right]$
$\Leftrightarrow max(|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|, |NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|, |HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|) = 0$
and $min(|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|, |NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|, |HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|) = 0$
$\Leftrightarrow |MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)| = 0$,
$|NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)| = 0$,
and $|HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)| = 0$.
$\Leftrightarrow MV_{AL}(x_i) = MV_{BL}(x_i), MV_{AU}(x_i) = MV_{BU}(x_i)$,
$NV_{AL}(x_i) = NV_{BL}(x_i), NV_{AU}(x_i) = NV_{BU}(x_i)$
and $HV_{AL}(x_i) = HV_{BL}(x_i), HV_{AU}(x_i) = HV_{BU}(x_i)$.
$\Leftrightarrow A = B$.

Property 3): $\emptyset(A, B) = \emptyset(B, A)$ by definition of $MV_{\emptyset(A,B)L}(x_i), MV_{\emptyset(A,B)U}(x_i), NV_{\emptyset(A,B)L}(x_i)$, $NV_{\emptyset(A,B)U}(x_i)$ for any $x_i \in \Omega$
$\Rightarrow E(\emptyset(A, B)) = E(\emptyset(B, A))$
$\Leftrightarrow S(A, B) = S(B, A)$

Property 4): Let A, B and C be any three IvIFSs such that $A \subseteq B \subseteq C$ for any $x_i \in \Omega$, we have $MV_A(x_i) \leq MV_B(x_i) \leq MV_C(x_i), NV_A(x_i) \geq NV_B(x_i) \geq NV_C(x_i)$ or $MV_{AL}(x_i) \leq MV_{BL}(x_i) \leq MV_{CL}(x_i), NV_{AL}(x_i) \geq NV_{BL}(x_i) \geq NV_{CL}(x_i)$ and $MV_{AU}(x_i) \leq MV_{BU}(x_i) \leq MV_{CU}(x_i), NV_{AU}(x_i) \geq NV_{BU}(x_i) \geq NV_{CU}(x_i)$.
$\Rightarrow |MV_{AL}(x_i) - MV_{CL}(x_i)| \geq |MV_{AL}(x_i) - MV_{BL}(x_i)|$, $|MV_{AU}(x_i) - MV_{CU}(x_i)| \geq |MV_{AU}(x_i) - MV_{BU}(x_i)|$;
$|NV_{AL}(x_i) - NV_{CL}(x_i)| \geq |NV_{AL}(x_i) - NV_{BL}(x_i)|$, $|NV_{AU}(x_i) - NV_{CU}(x_i)| \geq |NV_{AU}(x_i) - NV_{BU}(x_i)|$;
and $|HV_{AL}(x_i) - HV_{CL}(x_i)| = |2(MV_{CL}(x_i) - MV_{AL}(x_i)) + 2(NV_{CL}(x_i) - NV_{AL}(x_i))| \geq |2(MV_{BL}(x_i) - MV_{AL}(x_i)) + 2(NV_{BL}(x_i) - NV_{AL}(x_i))| = |NV_{AL}(x_i) - NV_{BL}(x_i)|$,
Similarly, we have $|NV_{AU}(x_i) - NV_{CU}(x_i)| \geq |NV_{AU}(x_i) - NV_{BU}(x_i)|$
So, we have
$MV_{\emptyset(A,B)}(x_i) \leq MV_{\emptyset(A,C)}(x_i) \leq \left[\frac{1}{3}, \frac{1}{3}\right]$ and $NV_{\emptyset(A,B)}(x_i) \geq NV_{\emptyset(A,C)}(x_i) \geq \left[\frac{1}{3}, \frac{1}{3}\right]$ for any $x_i \in \Omega$. $\Rightarrow \emptyset(A, C) \subseteq \emptyset(A, B) \subseteq \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle$
Similarly, we have $\Rightarrow \emptyset(A, C) \subseteq \emptyset(B, C) \subseteq \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle$. Thus
$D\left(\emptyset(A, B), \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right) \leq D\left(\emptyset(A, C), \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right)$
and $D\left(\emptyset(B, C), \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right) \leq D\left(\emptyset(A, C), \langle\left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right], \left[\frac{1}{3}, \frac{1}{3}\right]\rangle\right)$.
So form definition of entropy corresponding to distance function, we get $E(\emptyset(A, C)) \leq E(\emptyset(A, B))$ and $E(\emptyset(A, C)) \leq E(\emptyset(B, C))$
or $S(\emptyset(A, C)) \leq S(\emptyset(A, B))$ and $S(\emptyset(A, C)) \leq S(\emptyset(B, C))$. ■

Corollary 1: Let E be an entropy measure for IvIFSs and $\emptyset(A, B)$be an IvIFS defined on two IvIFSs A and B acaccording to definition 6, then $E(\overline{\emptyset(A, B)})$ measure of similarity for $A, B \in IvIFSs(\Omega)$.

Proof: Proof followed from the definition of complement interval-valued intuitionistic fuzzy sets and theorem 4. ■

Definition 7: Let $A, B \in IvIFSs(\Omega)$, we can define $IvIFS$ $\eta(A, B)$using A, B as follows:
$MV_{\eta(A,B)L}(x_i) = \frac{1}{3}\{1 + [min((|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|^{\alpha}), (|NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|)^{\alpha}, (|HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)^{\alpha})]^2\}$;
$MV_{\eta(A,B)U}(x_i) = \frac{1}{3}\{1 + [min((|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|^{\alpha}), (|NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|)^{\alpha}, (|HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)^{\alpha})]\}$;
$NV_{\eta(A,B)L}(x_i) = \frac{1}{3}\{1 - [max((|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|^{\alpha}), (|NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|)^{\alpha}, (|HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)^{\alpha})]^{1/2}\}$;
$NV_{\eta(A,B)U}(x_i) = \frac{1}{3}\{1 - [max((|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|^{\alpha}), (|NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|)^{\alpha}, (|HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)^{\alpha})]\}$,
Where $\alpha \in [1, \infty[$ and $x_i \in \Omega$.

Theorem 5: For any two $IvIFSs$ A and B, $\mathrm{E}(\eta(\mathrm{A},\mathrm{B}))$ is a measure of similarity, where E is an entropy measure.

Proof: To prove that $\mathrm{E}(\eta(\mathrm{A},\mathrm{B}))$ is a measure of similarity, we need to prove property given by definition 3 holds .

Property 1): If $A \in \mathbb{C}(\Omega) \Rightarrow A(x_i) = \langle[1,1],[0,0],[0,0]\rangle$ *or* $A(x_i) = \langle[0,0],[1,1],[0,0]\rangle$, for any $x_i \in \Omega$, then $MV_{\emptyset(A,\bar{A})L}(x_i) = 1 = MV_{\emptyset(A,\bar{A})U}(x_i)$;
and $NV_{\emptyset(A,\bar{A})L}(x_i) = 0 = NV_{\emptyset(A,\bar{A})U}(x_i)$
Thus, $\emptyset(\mathrm{A},\bar{\mathrm{A}}) = \{\langle x_i,[1,1],[0,0],[0,0]\rangle / x_i \in \Omega\}$
$\Rightarrow \mathrm{S}(\mathrm{A},\bar{\mathrm{A}}) = \mathrm{E}(\emptyset(\mathrm{A},\bar{\mathrm{A}})) = 0$

Property 2): Assume that $\mathrm{S}(\mathrm{A},\mathrm{B}) = 1$
Then $\mathrm{E}(\eta(\mathrm{A},\mathrm{B})) = 1$

$$\Leftrightarrow MV_{\eta(A,B)}(x_i) = \left[\frac{1}{3},\frac{1}{3}\right] = NV_{\eta(A,B)}(x_i) = HV_{\eta(A,B)}(x_i)$$

$$\Leftrightarrow min\begin{pmatrix}(|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|)^{\alpha} \\ (|NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|)^{\alpha} \\ (|HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)^{\alpha}\end{pmatrix} = 0$$

and

$$max\begin{pmatrix}(|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|)^{\alpha}, \\ (|NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|)^{\alpha}, \\ (|HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)^{\alpha}\end{pmatrix} = 0$$

$\Leftrightarrow (|MV_{AL}(x_i) - MV_{BL}(x_i)| \vee |MV_{AU}(x_i) - MV_{BU}(x_i)|)^{\alpha} = 0,$
$(|NV_{AL}(x_i) - NV_{BL}(x_i)| \vee |NV_{AU}(x_i) - NV_{BU}(x_i)|)^{\alpha} = 0,$
and $(|HV_{AL}(x_i) - HV_{BL}(x_i)| \vee |HV_{AU}(x_i) - HV_{BU}(x_i)|)^{\alpha} = 0.$
$\Leftrightarrow MV_{AL}(x_i) = MV_{BL}(x_i), MV_{AU}(x_i) = MV_{BU}(x_i), NV_{AL}(x_i) = NV_{BL}(x_i), NV_{AU}(x_i) = NV_{BU}(x_i)$
and $HV_{AL}(x_i) = HV_{BL}(x_i), HV_{AU}(x_i) = HV_{BU}(x_i).$
$\Leftrightarrow \mathrm{A} = \mathrm{B}.$

Property 3): $\eta(\mathrm{A},\mathrm{B}) = \eta(\mathrm{B},\mathrm{A})$ by definition of $MV_{\eta(A,B)L}(x_i), MV_{\eta(A,B)U}(x_i), NV_{\eta(A,B)L}(x_i), NV_{\eta(A,B)U}(x_i)$ for any $x_i \in \Omega$
$\Rightarrow \mathrm{E}(\emptyset(\mathrm{A},\mathrm{B})) = \mathrm{E}(\emptyset(\mathrm{B},\mathrm{A}))$
$\Leftrightarrow \mathrm{S}(\mathrm{A},\mathrm{B}) = \mathrm{S}(\mathrm{B},\mathrm{A})$

Property 4): Let A, B and C be any three IvIFSs such that $\mathrm{A} \subseteq \mathrm{B} \subseteq \mathrm{C}$, then for any $x_i \in \Omega$, we have $MV_A(x_i) \leq MV_B(x_i) \leq MV_C(x_i), NV_A(x_i) \geq NV_B(x_i) \geq NV_C(x_i)$ *or* $MV_{AL}(x_i) \leq MV_{BL}(x_i) \leq MV_{CL}(x_i), NV_{AL}(x_i) \geq NV_{BL}(x_i) \geq NV_{CL}(x_i)$ and $MV_{AU}(x_i) \leq MV_{BU}(x_i) \leq MV_{CU}(x_i), NV_{AU}(x_i) \geq NV_{BU}(x_i) \geq NV_{CU}(x_i)$.
$\Rightarrow |MV_{AL}(x_i) - MV_{CL}(x_i)| \geq |MV_{AL}(x_i) - MV_{BL}(x_i)|,$
$|MV_{AU}(x_i) - MV_{CU}(x_i)| \geq |MV_{AU}(x_i) - MV_{BU}(x_i)|;$
$|NV_{AL}(x_i) - NV_{CL}(x_i)| \geq |NV_{AL}(x_i) - NV_{BL}(x_i)|,$
$|NV_{AU}(x_i) - NV_{CU}(x_i)| \geq |NV_{AU}(x_i) - NV_{BU}(x_i)|;$
and $|HV_{AL}(x_i) - HV_{CL}(x_i)| = |2(MV_{CL}(x_i) - MV_{AL}(x_i)) + 2(NV_{CL}(x_i) - NV_{AL}(x_i))| \geq |2(MV_{BL}(x_i) - MV_{AL}(x_i)) + 2(NV_{BL}(x_i) - NV_{AL}(x_i))| = |NV_{AL}(x_i) - NV_{BL}(x_i)|,$
Similarly, we have $|NV_{AU}(x_i) - NV_{CU}(x_i)| \geq |NV_{AU}(x_i) - NV_{BU}(x_i)|$
So, we have from definition 7
$MV_{\eta(A,C)}(x_i) \geq MV_{\eta(A,B)}(x_i) \geq \left[\frac{1}{3},\frac{1}{3}\right]$ and $NV_{\eta(A,C)}(x_i) \leq NV_{\eta(A,B)}(x_i) \leq \left[\frac{1}{3},\frac{1}{3}\right]$ for any $x_i \in \Omega$. $\Rightarrow \eta(\mathrm{A},\mathrm{C}) \supseteq \eta(\mathrm{A},\mathrm{B}) \supseteq \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle$
Similarly, we have $\Rightarrow \eta(\mathrm{A},\mathrm{C}) \supseteq \eta(\mathrm{B},\mathrm{C}) \supseteq \langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle$. Thus

$$\mathrm{D}\left(\eta(\mathrm{A},\mathrm{B}),\langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) \leq \mathrm{D}\left(\eta(\mathrm{A},\mathrm{C}),\langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right)$$

and $\mathrm{D}\left(\eta(\mathrm{B},\mathrm{C}),\langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right) \leq \mathrm{D}\left(\eta(\mathrm{A},\mathrm{C}),\langle\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right],\left[\frac{1}{3},\frac{1}{3}\right]\rangle\right)$.
So from definition of entropy corresponding to distance function, we get $\mathrm{E}(\emptyset(\mathrm{A},\mathrm{C})) \leq \mathrm{E}(\emptyset(\mathrm{A},\mathrm{B}))$ and $\mathrm{E}(\eta(\mathrm{A},\mathrm{C})) \leq \mathrm{E}(\eta(\mathrm{B},\mathrm{C}))$
or $\mathrm{S}(\eta(\mathrm{A},\mathrm{C})) \leq \mathrm{S}(\eta(\mathrm{A},\mathrm{B}))$ and $\mathrm{S}(\eta(\mathrm{A},\mathrm{C})) \leq \mathrm{S}(\eta(\mathrm{B},\mathrm{C}))$. ■

Corollary 2: Let E be an entropy for $IvIFSs$ and $\eta(\mathrm{A},\mathrm{B})$ be an $IvIFS$ defined on $A, B \in IvIFSs(\Omega)$ as defined in definition 7, then $\mathrm{E}(\overline{\eta(\mathrm{A},\mathrm{B})})$ is measure of similarity for $A, B \in IvIFSs(\Omega)$.
Proof: Proof followed from the definition of complement of $IvIFSs$ and theorem 5. ■

4.1 Weighted similarity measure

Let $\mathrm{w} = (\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_n)^{\mathrm{T}}$ be the weights provided to each element $x_i \in \Omega, i = 1,2,\ldots,n$. Then the weighted similarity measure based on the aforesaid similarity measures are defined as $S(A,B) = \sum_{i=1}^{n} w_i\, S(A(x_i), B(x_i))$, where $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$.

4.2 Comparison with some select measures of similarity

Here, we compare the performance of proposed measure of similarity with some of the existing similarity measures as follows.

For any two IvIFSs A and B, then some existing similarity measures are given as follows:

- $S_W(A,B) = \frac{1}{n}\sum_{i=1}^{n} \frac{4-(MV_L(x_i)+MV_U(x_i)+NV_L(x_i)+NV_U(x_i))+(HV_L(x_i)+HV_U(x_i))}{4+(MV_L(x_i)+MV_U(x_i)+NV_L(x_i)+NV_U(x_i))+(HV_L(x_i)+HV_U(x_i))}$,

where $MV_L(x_i) = |MV_{AL}(x_i) - MV_{BL}(x_i)|$, $MV_U(x_i) = |MV_{AU}(x_i) - MV_{BU}(x_i)|$, $NV_L(x_i) = |NV_{AL}(x_i) - NV_{BL}(x_i)|$, $NV_U(x_i) = |NV_{AU}(x_i) - NV_{BU}(x_i)|$, $HV_L(x_i) = HV_{AL}(x_i) + HV_{BL}(x_i)$ and $HV_U(x_i) = HV_{AU}(x_i) + HV_{BU}(x_i)$ is proposed by Wu et al.(2014).

- $S_{HL}(A,B) = 1 - \frac{1}{4n}\sum_{i=1}^{n}|MV_{AL}(x_i) - MV_{BL}(x_i)| + |MV_{AU}(x_i) - MV_{BU}(x_i)| + |NV_{AL}(x_i) - NV_{BL}(x_i)| + |NV_{AU}(x_i) - NV_{BU}(x_i)|$is given by Hu and Li (2013).

- $S_S(A,B) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left[\begin{matrix}(MV_{AL}(x_i)+MV_{AU}(x_i))(MV_{BL}(x_i)+MV_{BU}(x_i))\\+(NV_{AL}(x_i)+NV_{AU}(x_i))(NV_{BL}(x_i)+NV_{BU}(x_i))\end{matrix}\right]}{\left[\begin{matrix}\sqrt{(MV_{AL}(x_i)+MV_{AU}(x_i))^2+(NV_{AL}(x_i)+NV_{AU}(x_i))^2}\\\sqrt{(MV_{BL}(x_i)+MV_{BU}(x_i))^2+(NV_{BL}(x_i)+NV_{BU}(x_i))^2}\end{matrix}\right]}$is introduced by Singh(2012)

- $S_{Su}(A,B) = \frac{1}{n}\sum_{i=1}^{n}\frac{\begin{matrix}|MV_{AL}(x_i)-MV_{BL}(x_i)|\vee|NV_{AL}(x_i)-NV_{BL}(x_i)|\\+|MV_{AU}(x_i)-MV_{BU}(x_i)|\vee|NV_{AU}(x_i)-NV_{BU}(x_i)|\end{matrix}}{3-min\left\{\begin{matrix}|MV_{AL}(x_i)-MV_{BL}(x_i)|\vee|NV_{AL}(x_i)-NV_{BL}(x_i)|,\\|MV_{AU}(x_i)-MV_{BU}(x_i)|\vee|NV_{AU}(x_i)-NV_{BU}(x_i)|\end{matrix}\right\}}$

is proposed by Sun & Liu (2012).

To review the performance of similarity measures let us consider an example. Consider the following IvIFSs
$A = \{x_i,\ \langle[0.5,0.5],[0.5,0.5],[0,0]\rangle, x_i \in \Omega\}$,
$B = \{x_i,\ \langle[0.3,0.4],[0.4,0.5],[0.1,0.3]\rangle, x_i \in \Omega\}$,
$C = \{x_i,\ \langle[0.3,0.3],[0.3,0.3],[0.4,0.4]\rangle, x_i \in \Omega\}$,
$D = \{x_i,\ \langle[0.6,0.6],[0.4,0.4],[0,0]\rangle, x_i \in \Omega\}$

Intuitively, it is clear that A is more similar to D than B and C. The result corresponding to measure of similarity measures given in Table 3:

	A B	A C	A D
S_W	0.83333	0.69308	0.8181
S_{HL}	0.9	0.8	0.9
S_S	0.99227	0.9	0.98058
S_{Su}	0.89655	0.8571	0.9310
$S_1(\phi)$	0.7450	0.73722	0.8646
$S_2(\phi)$	0.78806	0.761886	0.89594
$S_3(\phi)$	0.72614	0.68377	0.84188
$S_4(\phi)$	0.78806	0.74188	0.89594
$S_5(\phi)$	0.68210	0.62283	0.84391
$S_6(\phi)$ for p=2	0.77639	0.77141	0.87090
$S_1(\eta)$	0.93205	0.89426	0.98708
$S_2(\eta)$	0.95217	0.92075	0.99184
$S_3(\eta)$	0.91759	0.85853	0.98418
$S_4(\eta)$	0.95217	0.92075	0.99184
$S_5(\eta)$	0.92825	0.88113	0.98776
$S_6(\eta)$ for p=2	0.93292	0.89590	0.98709

Table 3: Comparison of Similarity Measures.

From the similarity measures listed in table 3, we can see that
S_W, S_{HL} and S_S are inconsistent with intuition where as S_{Su}, $S_j(\phi)$ and $S_j(\eta), j = 1, \dots, 6$.

In this section we have derived a relation between entropy and similarity measure. Then we defined some similarity measures, compared its performance with existing similarity measures. In section 5 we applied proposed similarity measures to draw conclusion in pattern recognition and medical diagnoses.

5 Applications of proposed similarity measures

Here the proposed similarity measures are applied to some of the situation that deals with imperfect information.

5.1 Pattern recognition

Here we use an example of pattern recognition considered by Xu (2007) and adapted by Wei et al. (2011) and Wu et al.(2014) for classification of building material.

Example: There are four types of building materials $A_i, i = 1,2,3,4$ and an anonymous building material B, which is characterized by the IvIFSs defined on $X = \{x_1, x_2, \dots, x_{12}\}$ with weighted vector

$$w = \left(\begin{matrix}0.1, 0.05, 0.08, 0.06, 0.03, 0.07,\\0.09, 0.12, 0.15, 0.07, 0.13, 0.05\end{matrix}\right)^T$$

and we have the data given as follows by Xu (2007).

"A_1

$$= \left\{\begin{pmatrix}\langle x_1,[0.1,0.2],[0.5,0.6]\rangle, \langle x_2,[0.1,0.2],[0.7,0.8]\rangle,\\ \langle x_3,[0.5,0.6],[0.3,0.4]\rangle, \langle x_4,[0.8,0.9],[0.0,0.1]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_5,[0.4,0.5],[0.3,0.4]\rangle, \langle x_6,[0.0,0.1],[0.8,0.9]\rangle,\\ \langle x_7,[0.3,0.4],[0.5,0.6]\rangle, \langle x_8,[1.0,1.0],[0.0,0.0]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_9,[0.2,0.3],[0.6,0.7]\rangle, \langle x_{10},[0.4,0.5],[0.4,0.5]\rangle,\\ \langle x_{11},[0.7,0.8],[0.1,0.2]\rangle, \langle x_{12},[0.4,0.5],[0.4,0.5]\rangle\end{pmatrix}\right\}$$

A_2

$$= \left\{\begin{pmatrix}\langle x_1,[0.5,0.6],[0.3,0.4]\rangle, \langle x_2,[0.6,0.7],[0.1,0.2]\rangle,\\ \langle x_3,[1.0,1.0],[0.0,0.0]\rangle, \langle x_4,[0.1,0.2],[0.6,0.7]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_5,[0.0,0.1],[0.8,0.9]\rangle, \langle x_6,[0.7,0.8],[0.1,0.2]\rangle,\\ \langle x_7,[0.5,0.6],[0.3,0.4]\rangle, \langle x_8,[0.6,0.7],[0.2,0.3]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_9,[1.0,1.0],[0.0,0.0]\rangle, \langle x_{10},[0.1,0.2],[0.7,0.8]\rangle,\\ \langle x_{11},[0.0,0.1],[0.8,0.9]\rangle, \langle x_{12},[0.7,0.8],[0.1,0.2]\rangle\end{pmatrix}\right\}$$

A_3

$$= \left\{\begin{pmatrix}\langle x_1,[0.4,0.5],[0.3,0.4]\rangle, \langle x_2,[0.6,0.7],[0.2,0.3]\rangle,\\ \langle x_3,[0.9,1.0],[0.0,0.0]\rangle, \langle x_4,[0.0,0.1],[0.8,0.9]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_5,[0.0,0.1],[0.8,0.9]\rangle, \langle x_6,[0.6,0.7],[0.2,0.3]\rangle,\\ \langle x_7,[0.1,0.2],[0.7,0.8]\rangle, \langle x_8,[0.2,0.3],[0.6,0.7]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_9,[0.5,0.6],[0.2,0.4]\rangle, \langle x_{10},[1.0,1.0],[0.0,0.0]\rangle,\\ \langle x_{11},[0.3,0.4],[0.4,0.5]\rangle, \langle x_{12},[0.0,0.1],[0.8,0.9]\rangle\end{pmatrix}\right\}$$

A_4

$$= \left\{\begin{pmatrix}\langle x_1,[1.0,1.0],[0.0,0.0]\rangle, \langle x_2,[1.0,1.0],[0.0,0.0]\rangle,\\ \langle x_3,[0.8,0.9],[0.0,0.1]\rangle, \langle x_4,[0.7,0.8],[0.1,0.2]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_5,[0.0,0.1],[0.7,0.9]\rangle, \langle x_6,[0.0,0.1],[0.8,0.9]\rangle,\\ \langle x_7,[0.1,0.2],[0.7,0.8]\rangle, \langle x_8,[0.1,0.2],[0.7,0.8]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_9,[0.4,0.5],[0.3,0.4]\rangle, \langle x_{10},[1.0,1.0],[0.0,0.0]\rangle,\\ \langle x_{11},[0.3,0.4],[0.4,0.5]\rangle, \langle x_{12},[0.0,0.1],[0.8,0.9]\rangle\end{pmatrix}\right\}$$

$B =$

$$\left\{\begin{pmatrix}\langle x_1,[0.9,1.0],[0.0,0.0]\rangle, \langle x_2,[0.9,1.0],[0.0,0.0]\rangle,\\ \langle x_3,[0.7,0.8],[0.1,0.2]\rangle, \langle x_4,[0.6,0.7],[0.1,0.2]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_5,[0.0,0.1],[0.8,0.9]\rangle, \langle x_6,[0.1,0.2],[0.7,0.8]\rangle,\\ \langle x_7,[0.1,0.2],[0.7,0.8]\rangle, \langle x_8,[0.1,0.2],[0.7,0.8]\rangle,\end{pmatrix}\right.$$
$$\left.\begin{pmatrix}\langle x_9,[0.4,0.5],[0.3,0.4]\rangle, \langle x_{10},[1.0,1.0],[0.0,0.0]\rangle,\\ \langle x_{11},[0.3,0.4],[0.4,0.5]\rangle, \langle x_{12},[0.0,0.1],[0.7,0.9]\rangle\end{pmatrix}\right\}$$

"

We need to identify which pattern is most similar to B using the maximum degree principle of measures of similarity between IvIFSs. Using the anticipated similarity measures defined in this paper, we get the following results given in table 4:

	A_1 B	A_2 B	A_3 B	A_4 B
$S_1(\phi)$	0.688569	0.803061	0.84671	0.936548
$S_2(\phi)$	0.725141	0.854747	0.868809	0.95075
$S_3(\phi)$	0.654357	0.742627	0.818343	0.9265
$S_4(\phi)$	0.725141	0.861997	0.868809	0.95075
$S_5(\phi)$	0.587711	0.789371	0.803214	0.926125
$S_6(\phi)$ for p=2	0.738413	0.80768	0.863582	0.939988
$S_1(\eta)$	0.780631	0.773645	0.768612	0.979413
$S_2(\eta)$	0.816725	0.811175	0.804125	0.98595
$S_3(\eta)$	0.767465	0.749444	0.76081	0.975
$S_4(\eta)$	0.816725	0.811175	0.804125	0.98595
$S_5(\eta)$	0.725088	0.716763	0.706188	0.978925
$S_6(\eta)$ for p=2	0.814325	0.801194	0.810194	0.979588

Table 4: Application to pattern recognition.

From the above values it is clear that B is most similar to A_4 as the value corresponding to each similarity measure is highest for A_4. So, we can conclude that A_4building material consistent with the specification and this result is consistent with the results presented by Wu et al.(2014).

5.2 Medical diagnoses

Many authors Wei et al. (2011), Wu et al. (2014), Singh (2012) employed IvIFSs to execute medical diagnosis in their works. Here we use the data used by Singh(2012) to do medical diagnosis using the proposed measure of similarity

Example: Let A and B be the set that represent the set of diagnoses and symptoms respectively given as $A = \{\langle A_1, \text{Viral fever}\rangle, \langle A_2, \text{Malaria}\rangle, \langle A_3, \text{Typhoid}\rangle\}$ and $= \{\langle B_1, \text{Temperature}\rangle, \langle B_2, \text{Headache}\rangle, \langle B_3, \text{Cough}\rangle\}$. Assume the patient is represented by

$$P = \begin{Bmatrix} \langle B_1, [0.6,0.8], [0.1,0.2]\rangle, \langle B_2, [0.3,0.7], [0.2,0.3]\rangle, \\ \langle B_3, [0.6,0.8], [0.1,0.2]\rangle \end{Bmatrix}$$

and the weights corresponding to each attribute is equal and each diagnosis is given by the following $IvIFSs$

$$A_1 = \begin{Bmatrix} \langle B_1, [0.4,0.5], [0.3,0.4]\rangle, \langle B_2, [0.4,0.6], [0.2,0.4]\rangle, \\ \langle B_3, [0.4,0.8], [0.1,0.2]\rangle \end{Bmatrix}$$

$$A_2 = \begin{Bmatrix} \langle B_1, [0.3,0.6], [0.3,0.4]\rangle, \langle B_2, [0.5,0.6], [0.3,0.4]\rangle, \\ \langle B_3, [0.4,0.5], [0.1,0.3]\rangle \end{Bmatrix}$$

$$A_3 = \begin{Bmatrix} \langle B_1, [0.7,0.8], [0.1,0.2]\rangle, \langle B_2, [0.6,0.7], [0.1,0.3]\rangle, \\ \langle B_3, [0.3,0.4], [0.1,0.2]\rangle \end{Bmatrix}$$

Using the proposed similarity measure we classify the patient P in one of the diagnoses A_1, A_2, A_3. The results are as follows in Table 5.

	A_1 P	A_2 P	A_3 P
$S_1(\phi)$	0.80666	0.753483	0.770859
$S_2(\phi)$	0.840736	0.788069	0.808633
$S_3(\phi)$	0.766473	0.703639	0.743099
$S_4(\phi)$	0.840736	0.788069	0.808633
$S_5(\phi)$	0.761105	0.682104	0.712949
$S_6(\phi)$ for p=2	0.821222	0.776552	0.796418
$S_1(\eta)$	0.916133	0.872675	0.885256
$S_2(\eta)$	0.938333	0.9025	0.911667
$S_3(\eta)$	0.89835	0.847525	0.865842
$S_4(\eta)$	0.938333	0.9025	0.911667
$S_5(\eta)$	0.9075	0.85375	0.8675
$S_6(\eta)$ for p=2	0.918319	0.877512	0.891129

Table 5: Application to Medical diagnoses.

From the above table 5 it is clear that patient P can be diagnosed with viral fever.

6 Conclusion

Entropy, distance and similarity measure are significant research area in fuzzy information theory as they are efficient tools to deal with uncertain and insufficient information. Here we have derived new definition of entropy based on distance measure by considering degree of hesitancy in to account and derived relation between distance, entropy and similarity measures under IvIFE. Further, we have compared the derived similarity measures with some of the existing similarity measure and instances are used to show that the derived measures are able to draw conclusion when existing measures give the same result. Thereafter, proposed measures of similarity are applied to recognition of patterns and medical diagnoses.

References

[1] Atanassov, K.(1986), Intuitionistic fuzzy sets, *Fuzzy Sets and Systems*, 20 (1), 87–96. https://doi.org/10.1016/S0165-0114(86)80034-3

[2] Atanassov, K. T. (1999). *Intuitionistic Fuzzy Sets*. Heidelberg, New York: Physica-Verlag. https://doi.org/10.1007/978-3-7908-1870-3

[3] Atanassov, K. , & Gargov, G.(1989) . Interval-valued intuitionistic fuzzy sets, *Fuzzy Sets and Systems*, 31 (3) , 343–349. https://doi.org/10.1016/0165-0114(89)90205-4

[4] Dammak, F. Baccour, L and Alimi, A.M.(2016), An Exhaustive Study of Possibility Measures of Interval-valued Intuitionistic Fuzzy Sets abd Application to Multicriteria Decision Making, *Advances in Fuzzy Systems*, Vol 2016, Article ID 9185706, 10 pages. http://dx.doi.org/10.1155/2016/9185706

[5] De Luca A.,& Termini S. (1972). A definition of a non-probabilistic entropy in setting of fuzzy sets. *Information and Control*, 20, 30.

https://doi.org/10.1016/S0019-9958(72)90199-41-312.

[6] Gau, W.L. ,& Buehrer, D.J.(1993) , Vague sets, IEEE Transactions on systems, *Man and Cybernetics*, 23 (2), 610–614. https://doi.org/10.1109/21.229476

[7] Grzegorzewski, P. (2004). Distances between intuitionistic fuzzy sets and/or interval-valued fuzzy sets based on the hausdorff metric. *Fuzzy Sets and Systems*, 48, 319–328. https://doi.org/10.1016/j.fss.2003.08.005

[8] Hu, K. & Li, J. (2013), The entropy and similarity measure of interval valued intuitionistic fuzzy sets and their relationship, *International Journal of Fuzzy Systems*, Vol. 15(3), 279-288.

[9] Hung, W., & Yang, M., (2006), Fuzzy entropy in intuitionistic fuzzy sets, *International Journal of Intelligent Systems*, Vol.21, 443-451. https://doi.org/10.1002/int.20131

[10] Kacprzyk, J. (1997). Multistage Fuzzy Control. Chichester: Wiley.

[11] Liu, X.C.(1992), Entropy, distance measure and similarity measure of fuzzy sets and their relations, *Fuzzy Sets and Systems*, 52, 305–318. https://doi.org/10.1016/0165-0114(92)90239-Z

[12] Park, J. H., Lin, K.M., Park, J.S. & Kwun, Y.C. (2007), Distance between interval-valued intuitionistic fuzzy sets, *Journal of Physics: Conference Series* 96. https://doi.org/10.1088/1742-6596/96/1/012089

[13] Singh, P.(2012), A new method on measure of similarity between interval-valued intuitionistic fuzzy sets for pattern recognition, *Journal of Applied & Computational Mathematics*, Vol.1(1), 1-5. https://doi.org/10.4172/2168-9679.1000101

[14] Sun, M. & Liu, J.,(2012), New entropy and similarity measures for interval-valued intuitionistic fuzzy set, Journal of information & *Computational Sciences*, vol.9(18), 5799-5806.

[15] Szmidt, E., & Kacprzyk, J. (2000). Distance between intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 114(3), 505–518. https://doi.org/10.1016/S0165-0114(98)00244-9

[16] Tiwari, P. & Gupta, P. (in press), Generalized Interval-valued Intuitionistic Fuzzy Entropy with some Similarity Measures, International Journal of Computing Science and Mathematics.

[17] Wei, C.P. , Wang, P. , & Zhang, Y.Z. (2011), Entropy, similarity measures of interval-valued intuitionistic fuzzy sets and their applications, *Inform. Sci*., 181, 4273–4286. https://doi.org/10.1016/j.ins.2011.06.001

[18] Wu, C., Luo, P., Li, Y. & Ren, X. (2014), A new similarity measure of interval-valued intuitionistic fuzzy sets considering its hesitancy degree and applications in expert systems, *Mathematical Problems in Engineering*. http://dx.doi.org/10.1155/2014/359214

[19] Xia, M. &Xu, Z. (2010), Some new similarity measures for intuitionistic fuzzy values and their application in group decision making , *Journal of Systems Science and System Engineering*, Vol. 19(4), 430-452. https://doi.org/10.1007/s11518-010-5151-9

[20] Xu, Z. (2007a), On similarity measures of interval-valued intuitionistic fuzzy sets and their application to pattern recognitions, *Journal of Southeast University (English Edition)* 23 (1), 139–143.

[21] Xu, Z. (2007 b), Some similarity measures of intuitionistic fuzzy sets and their application to multiple attribute decision making, *Fuzzy Optimization and Decision Making*, Vol. 6(2), 109-121. https://doi.org/10.1007/s10700-007-9004-z

[22] Xu, Z.S. &Yager, R. R., (2009), Intuitionistic and interval valued intuitionistic fuzzy preference relations and their measures of similarity for the evaluation of agreement within a group, *Fuzzy Optimization and Decision Making*, Vol. 8(2), 123-139. https://doi.org/10.1007/s10700-009-9056-3

[23] Yang, Y., & Chiclana, F. (2012), Consistency of 2D and 3D distances of intuitionistic fuzzy sets, *Expert Systems with Applications*, Vol. 39(10), 8665–8670. https://doi.org/10.1016/j.eswa.2012.01.199

[24] Yang, Y.J. & Hinde, C. (2010), A new extension of fuzzy sets using rough sets: R-fuzzy sets, *Information Sciences*, 180, 354–365. https://doi.org/10.1016/j.ins.2009.10.004

[25] Ye, J. (2012), Multicriteria decision making method using the Dice similarity measure based on the reduct intuitionistic fuzzy sets of interval-valued intuitionistic fuzzy sets, *Applied Mathematical Modelling*, vol.36(9), 4466-4472. https://doi.org/10.1016/j.apm.2011.11.075

[26] Zadeh, L. A.(1965), Fuzzy sets, Information and Control 8 (3), 338–356. https://doi.org/10.1016/S0019-9958(65)90241-X

[27] Zadeh, L.A.(1975), The concept of a linguistic variable and its application to approximate reasoning-I, *Information Science*, 8, 199–249. https://doi.org/10.1016/0020-0255(75)90036-5

[28] Zang, Q.S., Jiang, S.Y., Jia, B. G., & Luo, S. H., (2010), Some information measures for interval-valued intuitionistic fuzzy sets, *Information Sciences*, Vol.180, 5130-5145. https://doi.org/10.1016/j.ins.2010.08.038

[29] Zhang, Y.J., Ma, P.J., Su, H.,& Zhang, C.P.(2011), Entropy on interval-valued intuitionistic fuzzy sets and its application in multi-attribute decision making, *2011 proceedings of 14th International Conference on Fusion (FUSION)*,1-7.

[30] Zhang, S. Li, X. & Meng, F.(2016), An Approach to Multi-criteria Decision Making under Interval-valued Intuitionistic Fuzzy Vales

and Interval Fuzzy Measures , Journal of Industrial and Production Engineering, Vol 33(4).
https://doi.org/10.1080/21681015.2016.1146362

[31] Zhang, Q., Xing, H., Liu, F., Ye, J. & Tang, P. (2014), Some entropy measures for interval – valued intuitionistic fuzzy sets based on distances and their relationships with similarity and inclusion measures, *Information Sciences*, 283, 55-69.
https://doi.org/10.1016/j.ins.2014.06.012

[32] Zhang, H., Zhang, W., & Mei, C. (2009), Entropy of interval-valued fuzzy sets based on distance and its relationship with similarity measure, *Knowledge-Based Systems*, Vol. 22, 449-454.
https://doi.org/10.1016/j.knosys.2009.06.007

CONTENTS OF *Informatica* Volume 42 (2018) pp. 1–632

Papers

Editorials

JOŽEF STEFAN INSTITUTE

Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 900 staff, has 700 researchers, about 250 of whom are postgraduates, around 500 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of S**love**nia (or S♡nia). The capital today is considered a crossroad between East, West and Mediterranean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel.:+386 1 4773 900, Fax.:+386 1 251 93 85
WWW: http://www.ijs.si
E-mail: matjaz.gams@ijs.si
Public relations: Polona Strnad

INFORMATICA

AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS

INVITATION, COOPERATION

Submissions and Refereeing

Please register as an author and submit a manuscript at: http://www.informatica.si. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica LaTeX format and figures in .eps format. Style and examples of papers can be obtained from http://www.informatica.si. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

SUBSCRIPTION

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: drago.torkar@ijs.si

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than twentyfour years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering intelligent systems in the European computer science, informatics and cognitive community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica web edition is free of charge and accessible at http://www.informatica.si.
Informatica print edition is free of charge for major scientific, educational and governmental institutions. Others should subscribe.

Informatica WWW:

http://www.informatica.si/

Referees from 2008 on:

A. Abraham, S. Abraham, R. Accornero, A. Adhikari, R. Ahmad, G. Alvarez, N. Anciaux, R. Arora, I. Awan, J. Azimi, C. Badica, Z. Balogh, S. Banerjee, G. Barbier, A. Baruzzo, B. Batagelj, T. Beaubouef, N. Beaulieu, M. ter Beek, P. Bellavista, K. Bilal, S. Bishop, J. Bodlaj, M. Bohanec, D. Bolme, Z. Bonikowski, B. Bošković, M. Botta, P. Brazdil, J. Brest, J. Brichau, A. Brodnik, D. Brown, I. Bruha, M. Bruynooghe, W. Buntine, D.D. Burdescu, J. Buys, X. Cai, Y. Cai, J.C. Cano, T. Cao, J.-V. Capella-Hernández, N. Carver, M. Cavazza, R. Ceylan, A. Chebotko, I. Chekalov, J. Chen, L.-M. Cheng, G. Chiola, Y.-C. Chiou, I. Chorbev, S.R. Choudhary, S.S.M. Chow, K.R. Chowdhury, V. Christlein, W. Chu, L. Chung, M. Ciglarič, J.-N. Colin, V. Cortellessa, J. Cui, P. Cui, Z. Cui, D. Cutting, A. Cuzzocrea, V. Cvjetkovic, J. Cypryjanski, L. Čehovin, D. Čerepnalkoski, I. Čosić, G. Daniele, G. Danoy, M. Dash, S. Datt, A. Datta, M.-Y. Day, F. Debili, C.J. Debono, J. Dedič, P. Degano, A. Dekdouk, H. Demirel, B. Demoen, S. Dendamrongvit, T. Deng, A. Derezinska, J. Dezert, G. Dias, I. Dimitrovski, S. Dobrišek, Q. Dou, J. Doumen, E. Dovgan, B. Dragovich, D. Drajic, O. Drbohlav, M. Drole, J. Dujmović, O. Ebers, J. Eder, S. Elaluf-Calderwood, E. Engström, U. riza Erturk, A. Farago, C. Fei, L. Feng, Y.X. Feng, B. Filipič, I. Fister, I. Fister Jr., D. Fišer, A. Flores, V.A. Fomichov, S. Forli, A. Freitas, J. Fridrich, S. Friedman, C. Fu, X. Fu, T. Fujimoto, G. Fung, S. Gabrielli, D. Galindo, A. Gambarara, M. Gams, M. Ganzha, J. Garbajosa, R. Gennari, G. Georgeson, N. Gligorić, S. Goel, G.H. Gonnet, D.S. Goodsell, S. Gordillo, J. Gore, M. Grčar, M. Grgurović, D. Grosse, Z.-H. Guan, D. Gubiani, M. Guid, C. Guo, B. Gupta, M. Gusev, M. Hahsler, Z. Haiping, A. Hameed, C. Hamzaçebi, Q.-L. Han, H. Hanping, T. Härder, J.N. Hatzopoulos, S. Hazelhurst, K. Hempstalk, J.M.G. Hidalgo, J. Hodgson, M. Holbl, M.P. Hong, G. Howells, M. Hu, J. Hyvärinen, D. Ienco, B. Ionescu, R. Irfan, N. Jaisankar, D. Jakobović, K. Jassem, I. Jawhar, Y. Jia, T. Jin, I. Jureta, Đ. Juričić, S. K, S. Kalajdziski, Y. Kalantidis, B. Kaluža, D. Kanellopoulos, R. Kapoor, D. Karapetyan, A. Kassler, D.S. Katz, A. Kaveh, S.U. Khan, M. Khattak, V. Khomenko, E.S. Khorasani, I. Kitanovski, D. Kocev, J. Kocijan, J. Kollár, A. Kontostathis, P. Korošec, A. Koschmider, D. Košir, J. Kovač, A. Krajnc, M. Krevs, J. Krogstie, P. Krsek, M. Kubat, M. Kukar, A. Kulis, A.P.S. Kumar, H. Kwaśnicka, W.K. Lai, C.-S. Laih, K.-Y. Lam, N. Landwehr, J. Lanir, A. Lavrov, M. Layouni, G. Leban, A. Lee, Y.-C. Lee, U. Legat, A. Leonardis, G. Li, G.-Z. Li, J. Li, X. Li, X. Li, Y. Li, Y. Li, S. Lian, L. Liao, C. Lim, J.-C. Lin, H. Liu, J. Liu, P. Liu, X. Liu, X. Liu, F. Logist, S. Loskovska, H. Lu, Z. Lu, X. Luo, M. Luštrek, I.V. Lyustig, S.A. Madani, M. Mahoney, S.U.R. Malik, Y. Marinakis, D. Marinčič, J. Marques-Silva, A. Martin, D. Marwede, M. Matijašević, T. Matsui, L. McMillan, A. McPherson, A. McPherson, Z. Meng, M.C. Mihaescu, V. Milea, N. Min-Allah, E. Minisci, V. Mišić, A.-H. Mogos, P. Mohapatra, D.D. Monica, A. Montanari, A. Moroni, J. Mosegaard, M. Moškon, L. de M. Mourelle, H. Moustafa, M. Možina, M. Mrak, Y. Mu, J. Mula, D. Nagamalai, M. Di Natale, A. Navarra, P. Navrat, N. Nedjah, R. Nejabati, W. Ng, Z. Ni, E.S. Nielsen, O. Nouali, F. Novak, B. Novikov, P. Nurmi, D. Obrul, B. Oliboni, X. Pan, M. Pančur, W. Pang, G. Papa, M. Paprzycki, M. Paralič, B.-K. Park, P. Patel, T.B. Pedersen, Z. Peng, R.G. Pensa, J. Perš, D. Petcu, B. Petelin, M. Petkovšek, D. Pevec, M. Pičulin, R. Piltaver, E. Pirogova, V. Podpečan, M. Polo, V. Pomponiu, E. Popescu, D. Poshyvanyk, B. Potočnik, R.J. Povinelli, S.R.M. Prasanna, K. Pripužić, G. Puppis, H. Qian, Y. Qian, L. Qiao, C. Qin, J. Que, J.-J. Quisquater, C. Rafe, S. Rahimi, V. Rajkovič, D. Raković, J. Ramaekers, J. Ramon, R. Ravnik, Y. Reddy, W. Reimche, H. Rezankova, D. Rispoli, B. Ristevski, B. Robič, J.A. Rodriguez-Aguilar, P. Rohatgi, W. Rossak, I. Rožanc, J. Rupnik, S.B. Sadkhan, K. Saeed, M. Saeki, K.S.M. Sahari, C. Sakharwade, E. Sakkopoulos, P. Sala, M.H. Samadzadeh, J.S. Sandhu, P. Scaglioso, V. Schau, W. Schempp, J. Seberry, A. Senanayake, M. Senobari, T.C. Seong, S. Shamala, c. shi, Z. Shi, L. Shiguo, N. Shilov, Z.-E.H. Slimane, F. Smith, H. Sneed, P. Sokolowski, T. Song, A. Soppera, A. Sorniotti, M. Stajdohar, L. Stanescu, D. Strnad, X. Sun, L. Šajn, R. Šenkeřík, M.R. Šikonja, J. Šilc, I. Škrjanc, T. Štajner, B. Šter, V. Štruc, H. Takizawa, C. Talcott, N. Tomasev, D. Torkar, S. Torrente, M. Trampuš, C. Tranoris, K. Trojacanec, M. Tschierschke, F. De Turck, J. Twycross, N. Tziritas, W. Vanhoof, P. Vateekul, L.A. Vese, A. Visconti, B. Vlaovič, V. Vojisavljević, M. Vozalis, P. Vračar, V. Vranić, C.-H. Wang, H. Wang, H. Wang, H. Wang, S. Wang, X.-F. Wang, X. Wang, Y. Wang, A. Wasilewska, S. Wenzel, V. Wickramasinghe, J. Wong, S. Wrobel, K. Wrona, B. Wu, L. Xiang, Y. Xiang, D. Xiao, F. Xie, L. Xie, Z. Xing, H. Yang, X. Yang, N.Y. Yen, C. Yong-Sheng, J.J. You, G. Yu, X. Zabulis, A. Zainal, A. Zamuda, M. Zand, Z. Zhang, Z. Zhao, D. Zheng, J. Zheng, X. Zheng, Z.-H. Zhou, F. Zhuang, A. Zimmermann, M.J. Zuo, B. Zupan, M. Zuqiang, B. Žalik, J. Žižka,

Informatica

An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: http://www.informatica.si.

Subscription Information Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Litostrojska cesta 54, 1000 Ljubljana, Slovenia.
The subscription rate for 2018 (Volume 42) is
– 60 EUR for institutions,
– 30 EUR for individuals, and
– 15 EUR for students
Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.
Printing: ABO grafika d.o.o., Ob železnici 16, 1000 Ljubljana.

Orders may be placed by email (drago.torkar@ijs.si), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):
Slovene Society for Pattern Recognition (Vitomir Štruc)
Slovenian Artificial Intelligence Society (Mitja Luštrek)
Cognitive Science Society (Olga Markič)
Slovenian Society of Mathematicians, Physicists and Astronomers (Marej Brešar)
Automatic Control Society of Slovenia (Nenad Muškinja)
Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Mark Pleško)
ACM Slovenia (Borut Žalik)

Informatica is financially supported by the Slovenian research agency from the Call for co-financing of scientific periodical publications.

Informatica is surveyed by: ACM Digital Library, Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math

Volume 42 Number 4 December 2018 **ISSN 0350-5596**

Informatica

An International Journal of Computing and Informatics

Informatica **42** (2018) Number 4, pp. 485–632