Semantic Searching of Biological Documents Using Gene Ontology

Marwa Mostafa[1], Enas M.F. El Houby[2] and Akram Salah[1]

[1]Faculty of computers and information, Cairo University, computer science department
5 Dr. Ahmed Zewail, Orman, Giza, Egypt
E-mail: m.mostafa_fci@yahoo.com, akram.salah@fci-cu.edu.eg

[2]Systems & Information Department, Engineering Division, National Research Centre, Dokki, Giza, Egypt
E-mail: em.fahmy@nrc.sci.eg

Keywords: information retrieval, biological ontology, ranking methodology, precision, recall

Received: September 21, 2013

Semantic information retrieval of biological documents is an information retrieval approach that utilizes semantics to improve the search recall and precision. This research presents a framework for a semantic biological retrieval system that effectively searches and retrieves meaningful results using Gene Ontology. The system takes two related biological terms as an input and retrieves relevant documents which contain these inputs. Since the user searches for the documents that contain two related biological terms, the system helps the user to know the hierarchical relationship between these two terms using Gene Ontology. The system utilizes the Gene Ontology to infer semantically related terms to the inputs. The inferred words may include synonyms, parents and grandparents of the input terms entered in the search query. The system uses these related inferred terms in expanding the user query to produce meaningful results since it retrieves the documents that contain the input terms and these inferred related terms. The system uses a ranking methodology to help in ordering the retrieved documents based on the rank values. The proposed technique improves the precision of the retrieved documents as well as the recall which saves researcher time and focus.

Povzetek: Razvita je metoda iskanja bioloških dokumentov z uporabo genskih ontologij.

1 Introduction

The biological repositories contain hundreds of thousands of electronic collections that often contain high quality information [1]. During the past years, the increase in scientific knowledge and the massive data production have caused an exponential growth in the number and size of biological databases and repositories. However, data size, which can reach hundreds of gigabytes, involves serious problems of data access through data storage in local disks. Other challenging issues associated to biological data are that much relevant information is spread out in different databases or repositories [2]. So the biological data is still locked in a large number of resources; remaining not computer-readable. In the current search engines when the user enters two terms it returns a lot of documents including unhelpful ones.

Keyword-based search is currently the most commonly employed search strategy in biomedical digital libraries. When users search by a few keywords, a large number of matched results could be returned. Users spend a significant amount of time to browse these results to find out those documents they are truly interested in because the publications returned may not be organized based on the user needs, forcing users to browse thousands of publications. In most cases, it is impossible for users to manually read every returned entry thus leads to loss of many truly relevant publications [3].

The goal of an information retrieval (IR) system is to rank documents optimally given a query to rate the relevance of documents. In order to achieve this goal, the system must be able to score documents so that the relevant document would ideally have a higher score than the irrelevant one [4].

Most of the current forms of web content are designed to be presented to humans; they are not understandable by computers. The semantic web aims at enhancing existing web content with semantic structure in order to make it meaningful to computers as well as to humans. Ontology plays a key role in the semantic web [5], [6] which offers an advanced approach for managing, retrieving information and processing it.

Ontology is a formal conceptualization of a particular domain into a human understandable, machine-readable format [7].One of the most important bio-ontology is Gene Ontology [8]. It organizes terms in a parent-child hierarchy.

Our first publication about this framework was "Ontology based Biological Information Retrieval System" (OBIRS) [9] which shows how we improved the efficiency of the method used in the system algorithm.

The proposed system presented in this paper uses Gene Ontology to infer semantically related terms to the input terms. The inferred terms may include synonyms which are useful in retrieving documents by authors who use different wording in reference to the same concept.

The system also infers related terms through parent-child relationship up to 2 levels (parents and grandparents) for each term of the input terms to expand the search query.

The proposed system helps the researchers to get more relevant and accurate retrieval of the documents. It allows the researchers to enter two related terms to get the documents that contain both of them. Also the system semantically retrieves the documents if they contain synonyms of the input terms inferred from the Gene Ontology even if these documents do not contain the exact phrase of the input terms. Also the system retrieves the documents that contain the input terms and/or synonyms with any combination of the other inferred terms (parents and grandparents).

The system searches for two related terms because its main idea is to retrieve documents that contain relation among related biological terms and we found that the least number of possible terms to find a relation between is two.

The system uses a ranking methodology that helps in ranking the retrieved documents to achieve the researchers satisfaction and save time and effort consumed by the researchers to rate the relevance of documents manually. The system groups the retrieved documents into five classes to save the time of the researchers. The system also extracts the relation between the input terms from the gene ontology to give the researchers the hierarchical relation between them.

The remainder of the paper is organized as follows: in section 2, an overview for the previous work related to our subject is presented. In section 3, the architecture of the proposed system is described. In section 4, ranking issues are explained. In section 5, an example is introduced to illustrate the proposed system. In section 6 relationship extraction is explained. In section 7, testing the system and the results are produced, before drawing conclusions and future work in section 8.

2 Related work

A lot of previous work was studied for the subject of semantic web and biological information retrieval. Sumithiradevi et al.[10]proposed one such tool called BIOMINING that is designed to eliminate anomalous and redundancy in biological web content. The authors use indexing and mining technology on biological databases to summarize the information of biological data in the document. Zhou et al. [11] designed a biological information retrieval and analysis system (BIRAS) based on the Internet. The system could send and receive information from the Entrez search and retrieval system maintained by the National Centre for Biotechnology Information (NCBI) in USA. Marta Bleda et al.[2]proposed the "CellBase" that provides a solution to the growing necessity of integration by easing the access to biological data. CellBase implemented a set of RESTful web services that query a centralized database containing the most relevant biological data sources. Minlie Huang et al.[12]proposed Ontology-based biological relation extraction system to automatically extract biological relations from a huge number of online MEDLINE abstracts. Authors then made Ontology-based semantic annotation of online biological documents. Anália Lourenço et al. [13] present BioDR which is a novel approach that allows the semantic indexing of the results of a query by identifying relevant terms in the documents. This system makes it possible to navigate semantically between documents and relevant terms, taking advantage of the rich contents of full-text.

Many other researchers [1], [14], [15], [16], [17], [18], [19] used ontologies, inverted list (different tech.) and query expansion to assist biological information retrieval search.

After reviewing several researches that support the retrieval of biomedical information it is our conclusion that the most similar to our system is[12]. However the previous reviewed researches aim to study the design of a biological information retrieval and analysis systems using the Internet. These systems are designed to eliminate anomalies and redundancy in biological web content, integrate biological database, retrieve biological information and extract relations. Our proposed system is a biological semantic retrieval system that tries to improve the recall and precision of the retrieved documents and helps the researchers to get the relevant documents that contain information and relationships between two related biological terms. The system retrieves documents that contain the terms as well as other semantically related terms inferred from the Gene Ontology. The system also ranks the retrieved documents based on their relevance to the input terms. The system retrieves the content of the document, not its address, unlike other retrieval systems. It is our assumption that retrieving relevant documents that contain information about two related biological terms entered from the researchers and ranking them should save the researchers time and effort.

3 Proposed system

Testing our previous system presented in [9] shows that there are many documents that satisfy the researcher's needs and have not been retrieved. Many biological terms have synonyms and it is possible to have a document that contains synonyms of the two terms entered in the search query or that contains one term and the synonym of the other term during the searching process. These documents have not been retrieved to the researchers in spite of its relevancy to the query. Also the retrieved documents have not been ranked appropriately. That was a motivation for improving the effectiveness of the previous system since there are many documents that semantically may satisfy the researchers needs and have not been retrieved.

The system (EOBIRS) presented in this paper is the Enhanced Biological Information Retrieval System which is the updated version of the previous system that highlights the importance of the synonyms of the searched terms and retrieves documents from corpus even if they have the synonyms only and doesn't contain the same wording of the terms entered in the search query because semantically they are reference to the

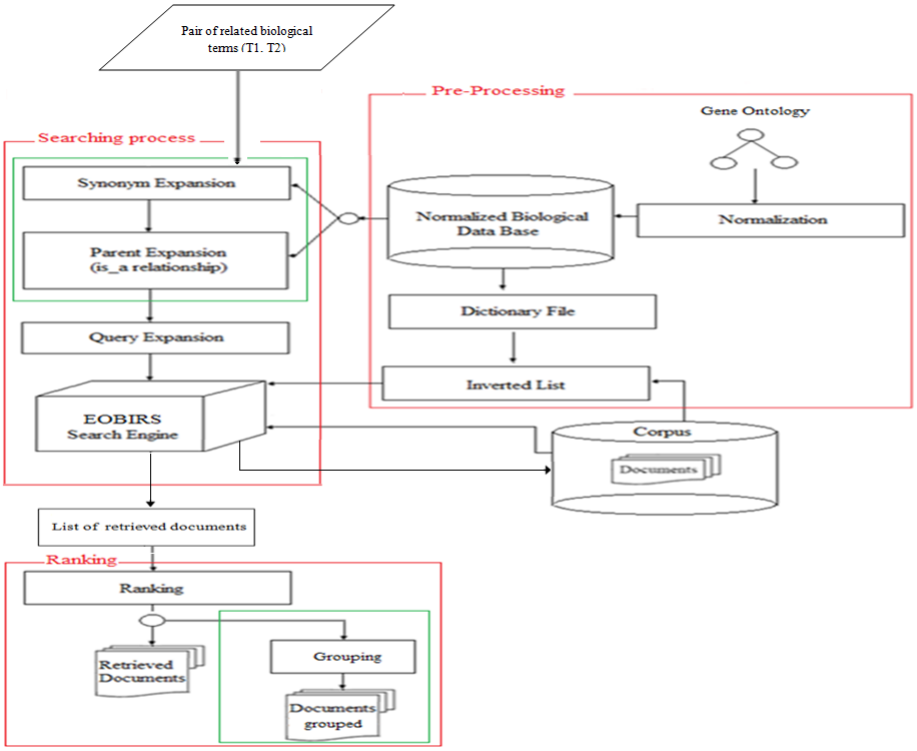

Figure 1: System architecture.

same concept. The system presented in this paper retrieves documents that contain the two entered terms, their synonyms and their parents up to two levels (parents and grandparents).

The system presented in this paper has the same system design like the previous system. It also has the same pre-processing instructions like [9] and differs in the searching process instructions, ranking criteria, grouping criteria and reordering the classes based on the balance value that adds another facility to minimize the time and effort of the researchers. The system architecture is shown in Figure 1.

3.1 Pre-processing

i. The system normalizes the Gene Ontology to a database named "DBGenes". The database "DBGenes" contains all genes that exist in the Gene Ontology with their attributes such as name, id, definition, synonymous, is_a and part_of.
ii. The system builds a dictionary file that contains all the biological terms exist in the normalized database.
iii. The system builds inverted list based on the biological terms only that exist in the corpus's documents. The system compares all terms exist in the corpus's documents with the biological terms exist in dictionary file, so a term added to the inverted list if it was found in the dictionary file. The terms were added to the inverted list with a list of the documents that contains these terms and the positions of the terms and the frequencies of their appearance in each document.

3.2 Searching process

a. The researcher enters the two related biological terms that he/she wants to search for. Where the system searches for unique identifiers for biological terms, the system begins to check if the "DBGenes" contains these terms or not. The search process starts if the two terms exist in the "DBGenes".
b. The system gets all the synonymous for both terms from the normalized database "DBGenes".
c. The system gets all the parents up to two levels (parents and grandparents) for both terms from the normalized database "DBGenes".
d. The system expands the query "term1 AND term2" using synonymous provided from the Ontology as well as parents and grandparents using "is_a"relation that describes the parent-child relationship. The query will expanded as follow:

If we assume that the two related biological terms entered to the retrieval system are G_1 and G_2. The set of synonyms are later called a synset. If the two synsets are $GS_1=\{gs_{11}, gs_{12},\ldots,gs_{1m}\}$ and $GS_2=\{gs_{21},gs_{22},\ldots,gs_{2n}\}$, and if the gene parents are $GP_1=\{gp_{11}, gp_{12},\ldots,gp_{1i}\}$ and $GP_2=\{gp_{21},gp_{22},\ldots,gp_{2j}\}$, and if the gene grandparents are $GGP_1=\{ggp_{11}, ggp_{12},\ldots,ggp_{1k}\}$ and $GGP_2=\{ggp_{21},ggp_{22},\ldots,ggp_{2l}\}$, the query will expanded into these queries:

Q_1 retrieves all the documents that contain the two related biological terms and/or the synonyms and their parents and their grandparents.

$$Q_1=[((G_1)OR(gs_{11}ORgs_{12}OR\ldots gs_{1m}))AND(gp_{11}OR\ gp_{12}OR\ldots,gp_{1i})\ AND(ggp_{11}OR\ ggp_{12},OR\ldots,ggp_{1k})]\ \underline{AND}\ [((G_2)\ OR\ (gs_{21}ORgs_{22}OR\ldots gs_{2n}))AND(gp_{21}OR\ gp_{22}OR\ldots,gp_{2j})AND\ (ggp_{21}ORggp_{22}OR\ldots,ggp_{2l})]$$

Q_2 retrieves all the documents that contain the two biological terms and/or the synonyms with their parents or grandparents.

$$Q_2=[((G_1)OR(gs_{11}ORgs_{12}OR\ldots gs_{1m}))AND((gp_{11}O\ Rgp_{12}OR\ldots,gp_{1i})\ OR\ (ggp_{11}\ ORggp_{12}\ OR\ldots,ggp_{1k}))]$$

AND [((G2) OR (gs_{21}ORgs_{22}OR…gs_{2n}))AND((gp_{21}OR gp_{22}OR…gp_{2j}) OR (ggp_{21} OR ggp_{22} OR …,ggp_{2l}))]

Q3 retrieves all the documents that contain the two terms or their synonyms or one term and the synonym of other term.

Q3= [(G1) OR (gs_{11}ORgs_{12}OR…gs_{1m})] AND [(G2) OR (gs_{21}ORgs_{22}OR…gs_{2n})]

The expanded query will be:

$Q = Q_1$ OR Q_2 OR Q_3

e. The system uses the inverted list to get the list of the documents that satisfy the query Q. This list will contain the document's names that contain the two terms (G1 and G2) and/or any combination of the related terms inferred from the Gene Ontology.
f. The system calculates the rank value of each document which used to order the retrieved documents. The system ranks the documents under a certain criteria:
   - The initial value of ranking of the document is the count of occurrence of the two terms multiplied by weight W_1.
   - Finding synonyms of any of the two terms increases the value of ranking by adding W_2 of the number of their occurrence.
   - Finding a parent or grandparent of any of the two terms increases the value of ranking by adding W_3 of the number of their occurrence.

The rank value will be calculated as follow:

Rank value =

$$[(F(T_1)+F(T_2))*W_1]+[(F(ST_1)+F(ST_2))*W_2]+ [(F(PT_1)+F(PT_2)+ F(GPT_1)+F(GPT_2)) * W_3] \quad (1)$$

Where T_1 and T_2 are the input terms ST_1, ST_2 are the synonyms of the input terms, PT_1, PT_2 are the parents of the input terms and GPT_1 and GPT_2 are the grandparents of the input terms. F is to count the number of occurrence of the terms.

Supposed that: $W_1 > W_2 > W_3$

We supposed that W_1 to be greater than W_2 because we assume that the weight of input terms must be greater than that of synonyms this is due to that we should give a strong concern to the input terms entered by the researcher. We think that the researcher is more concern about the retrieved documents that contain the exact wording of the input terms than the documents that contain the synonyms of the input terms. We choose W_2 to be greater than W_3 because the existence of the synonyms means the existence of the input terms so we assume that the weight of finding a parent or grandparent must be less than the weight of finding a synonym. The existence of a parent or grandparent adds another prove that the retrieved document talks about the input terms but in the same time it still doesn't represent the same meaning of the input terms so we cannot give it a weight equal to the synonyms.

g. The system retrieves from corpus the documents resulted from the query Q ranked by the system ranking values.
h. The system calculates the value of the precision and recall of the retrieved documents.

$$\text{Precision} = \frac{\text{Number of relevant documents}}{\text{Number of retrieved documents}} \quad (2)$$

$$\text{Recall} = \frac{\text{Number of relevant documents retrieved}}{\text{Number of relevant documents}} \quad (3)$$

i. The system extracts the relation between the two related terms from the Gene Ontology and presents it to the user as additional information about hierarchical relation of the two terms in addition to that mentioned in the documents.
j. The user can open any of the retrieved documents and notice the two terms that he/she searches for are highlighted.

4 Ranking Issues

During testing the system issue has been released "what about if the user wants to get specific documents as the first outcome in the list of the retrieved documents for example the documents that contain terms and their parents only". Because of this issue we have added a grouping option that the system provides to the user in addition to a list of the whole documents. The list is grouped into the following classes:

Class one: Provides all the documents that each one contains the two related biological terms and/or their synonyms and their parents and grandparents.

Class two: Provides all the documents that each one contains the two related biological terms and/or their synonyms and their parents.

Class three: Provides all the documents that each one contains the two related biological terms and/or their synonyms and their grandparents.

Class four: Provides all the documents that each one contains the two related biological terms only.

Class five: Provides all the documents that each one contains the synonyms of the two related biological terms only or one term and the synonym of other term.

Each class can be ordered based on the frequencies of the two related biological terms under search with concern of the balancing between the frequency of term 1 and the frequency of term 2 to ensure that the documents contain material that tackle the relation between the two terms. The system calculates the absolute value of the difference between the frequency of term1 and the frequency of the term2 in the document. It orders the documents based on the balance value, the document is ordered first if it has low balance value. If there are two or more documents having the same balance value then they will be ranked based on the summation of the "term frequency" value of term 1 and the "term frequency"

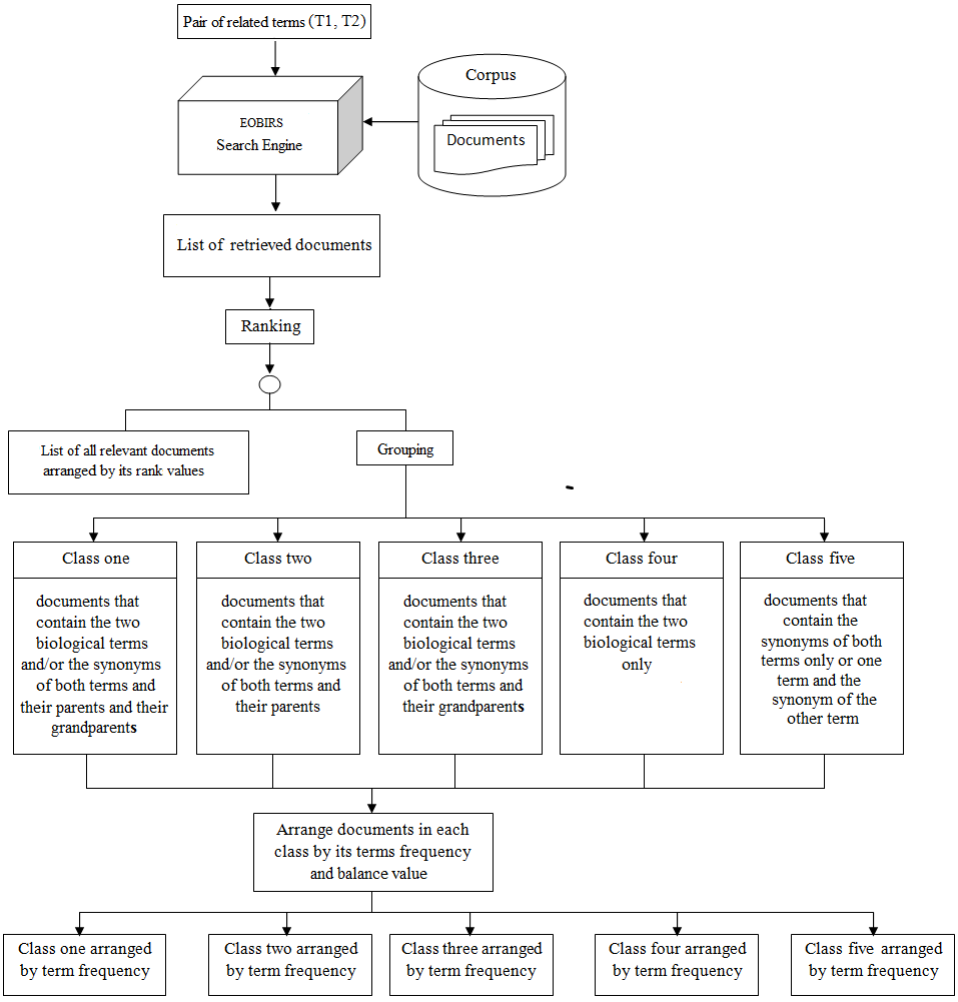

Figure 2: Ranking workflow.

value of term2. The ranking workflow is shown in Figure 2.

For illustration:
If we have a list of relevant retrieved documents from class four that contains Term1 and Term 2 as shown in Table 1.

Table 1: "term frequency" values of the two terms and the calculation of the balance value for each document.

Document number	Term 1	Term 2	Balance value (absolute value of the difference)
D1	2	2	0
D2	4	4	0
D3	2	1	1
D4	6	5	1
D5	4	1	3
D6	5	19	14

The system will calculate the balance for this class as shown in Table1. The list of the documents will be ordered based on balance value as shown in Figure 3.

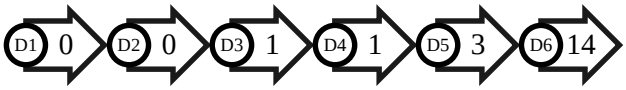

Figure 3: List of documents ordered based on balance values.

Then the system reorders the documents that have the same balance values based on the summation of the "term frequency" values of both term1 and term2.

So for D1 and D2:

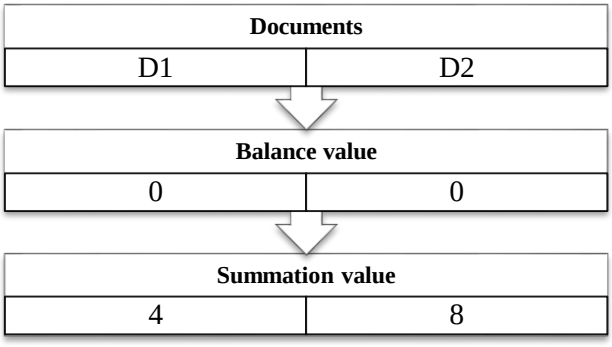

Figure 4: The comparison between term frequency values of D1 and D2.

Based on the calculation in Figure 4 the system will rank D2 higher than D1 because the summation of the "term frequency" values of term 1 and term2 in D2 is greater than their summation in D1.

For D3 and D4:

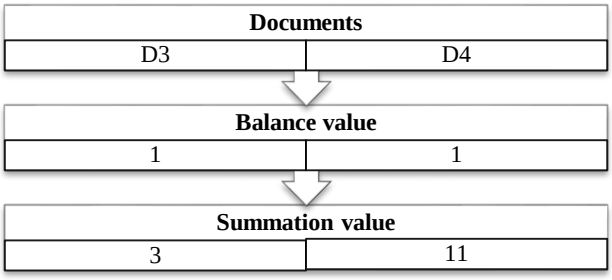

Figure 5: Comparing term frequency values of D3 and D4.

Based on the calculation in Figure5 the system will rank D4 higher than D3 because the summation of the "term frequency" values of term1 and term2 in D4 is greater than their summation in D3.

The system will present the documents for the user as shown in Figure 6.

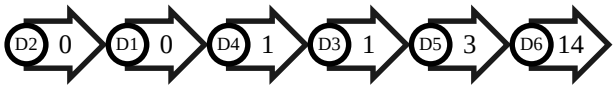

Figure 6: List of documents ordered by term's frequency values.

5 An example

To show how the system searches and orders the documents we present the following example, supposing that W_1=1, W_2= 0.8, W_3=0.25.

If the researcher searches for two terms, Term1:"regulation of DNA recombination"and Term 2:"mitochondrion inheritance"and the corpus contains a list of the following documents:

D1: mitochondrion inheritance and regulation of DNA recombination are biological_process in the Gene Ontology.

D2: mitochondrion inheritance has synonyms and regulation of DNA recombination doesn't have synonyms. Mitochondrial inheritance is a synonym of mitochondrion inheritance and organelle inheritance is a parent for it.

D3:mitochondrion inheritance and regulation of DNA recombination have parents. Regulation of DNA metabolic process is a parent of regulation of DNA recombination. Mitochondrial inheritance is a synonym of mitochondrion inheritance and organelle organization is a grandparent for it.

D4: regulation of DNA recombination is a biological process. Mitochondrion inheritance and regulation of DNA recombination have parents. Organelle organization is a grandparent of mitochondrion inheritance.

D5: Gene Ontology contains regulation of DNA recombination and mitochondrion inheritance. Regulation of DNA recombination is a biological_process. Regulation of DNA recombination is any process that modulates the frequency, rate or extent of DNA recombination. Regulation of DNA recombination is a subset of gosubset_prok. Regulation of DNA recombination has only one parent. Recombination regulates has a relationship with regulation of DNA recombination. Regulation of DNA recombination has "intersection_of" relation with biological regulation and DNA recombination regulates. We can find regulation of DNA recombination in Gene Ontology version1.2.The id of regulation of DNA recombination is GO:0000018 in the Gene Ontology. Mitochondrion inheritance is a biological_process. Mitochondrion inheritance is the distribution of mitochondria, including the mitochondrial genome, into daughter cells after mitosis or meiosis, mediated by interactions between mitochondria and the cytoskeleton.

D6: this document talks about regulation of DNA recombination and mitochondrion inheritance. Regulation of DNA recombination is any process that modulates the frequency, rate or extent of DNA recombination.

D7: Gene Ontology contains genes.

D8: mitochondrial inheritance is a biological_process in the Gene Ontology.

D9: Organelle inheritance is a biological_process in the Gene Ontology.

Table2: The calculations of documents rank values for the presented example.

Document number	Term frequency			Total	Parent	Grandparent	Document Rank value	Balance value (absolute value of the difference)
	Term1	Term2	Synonymous					
D1	1	1	0	2	-	-	2	0
D2	1	2	1	3.8	1	-	4.05	1
D3	2	2	1	4.8	1	1	5.3	0
D4	2	2	0	4	-	1	4.25	0
D5	9	3	0	12	-	-	12	6
D6	2	1	0	3	-	-	3	1
D7	0	0	0	0	-	-	0	0
D8	0	0	1	0.8	-	-	0	0
D9	0	0	0	0	1	-	0	0

Table2 presents the rank value of each document in the corpus. As shown document number 7 does not contain any of the two terms or their synonyms so it has a rank value equal to 0. Also the document number 8 has a value equal to 0 because it contains the synonymous of one term only. A document that contains synonyms of both terms will be ordered based on the total number of synonyms found in it. Also the table shows that document number 9 has a rank value equal to 0 because it contains parents only and does not contain any of the two terms or their synonyms so it will not be retrieved for the user.

The system calculates the balance values to order the documents within the class. In Figure 7 we show how the system retrieves the relevant documents based on our example.

6 Relation Extraction

Since our system objective is to retrieve the documents that relate two biological terms the system extracts the hierarchical relationship between the two terms from the Gene Ontology as additional information for the researcher in addition to that mentioned in the documents. The relationship shows the kinship between term 1 and term 2. The system determines four relationships between terms; these relations are sibling, cousin, child and uncle.

Sibling relationship:

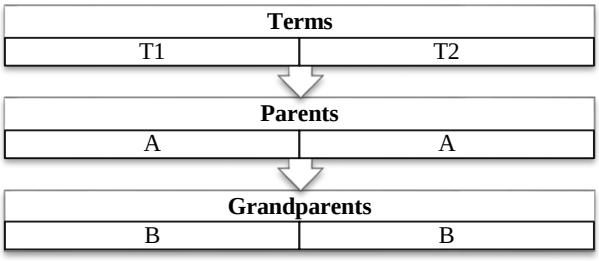

Figure 8: Sibling relationship.

As shown in Figure 8 if the parent of T1 is the same as T2 and the grandparent of T1 is the same as T2 then T1 and T2 are sibling.

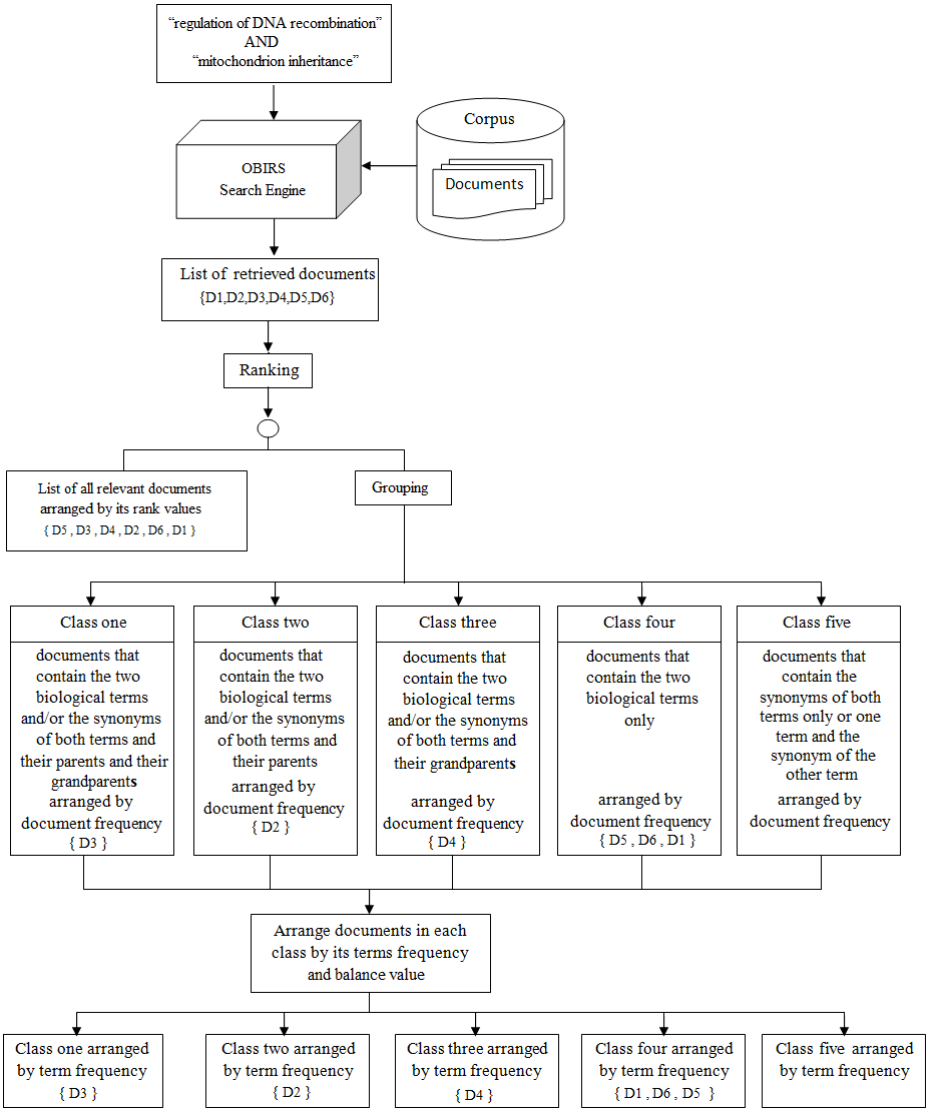

Figure 7: System workflow based on our example.

"Cousin" relationship:

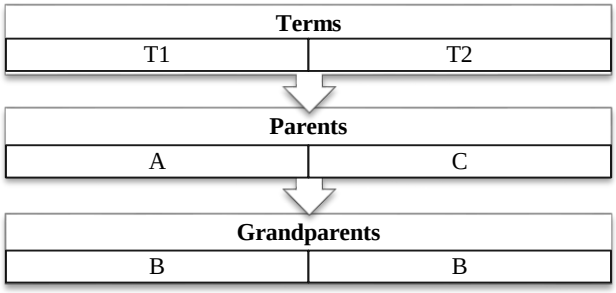

Figure 9: "Cousin" relationship.

As shown in Figure 9 if the parent of T1 is not the same as the parent of T2 and the grandparent of T1 is the same as T2 then T1 and T2 are cousins.

Child relationship:

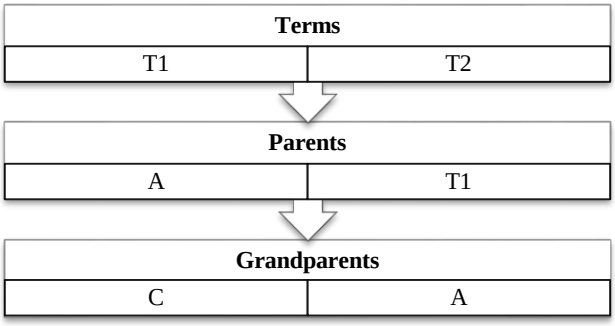

Figure 10: Child relationship.

As shown in Figure 10 if the parent of T1 is a grandparent of T2 and the parent of T2 is T1 then T2 is the child of T1.

"Uncle" relationship:

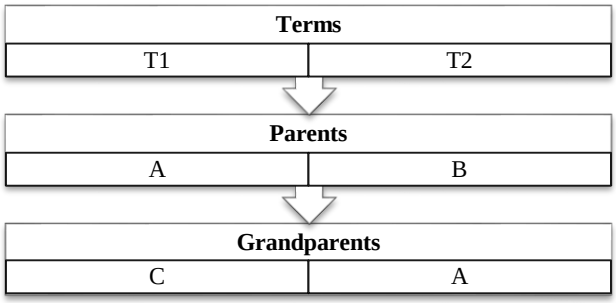

Figure 11: "Uncle" relationship.

As shown in Figure 11 if the parent of T1 is a grandparent of T2 and the parent of T2 is not T1 then T1 is uncle of T2.

7 System Evaluation

Extensive experiments are preformed to study the effectiveness of our algorithm. The system was tested using corpus named craft [20] and the Gene Ontology version 1.2 [8].

The performance of the system is improved since we retrieve the documents that contain the two related terms and the related inferred terms (synonyms, parents and grandparents). Our system retrieves the documents with a certain criteria of ranking that helps the research to find the document that he/she searches for. The following are screenshots from the system that represent how does the system work.

Pre processing:

The two steps represented in Figure 12 invoked once at the beginning of the system

In step 1, the system builds the dictionary file from the normalized database "DBGenes". In step 2, the system builds the inverted list that helps in retrieving the desired documents.

Searching process:

After building the inverted list the user can make any number of the search queries he/she wants. Figure 13 shows the search query request from user, Figure 14 shows the retrieved relevant documents in two alternative methods for ranking.

System testing:

"DNA" and "RNA"have been entered as two

Figure 12: Screen shot of preprocessing.

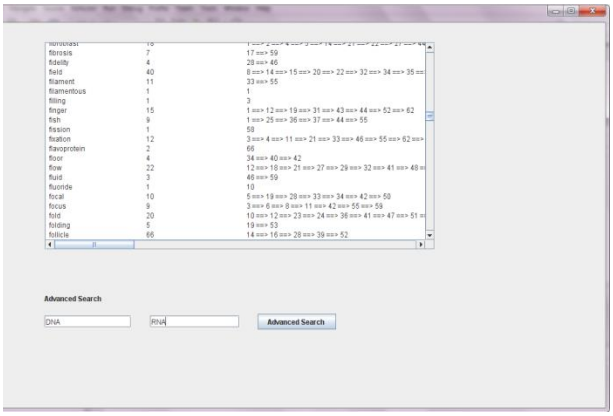

Figure 13: Search query request.

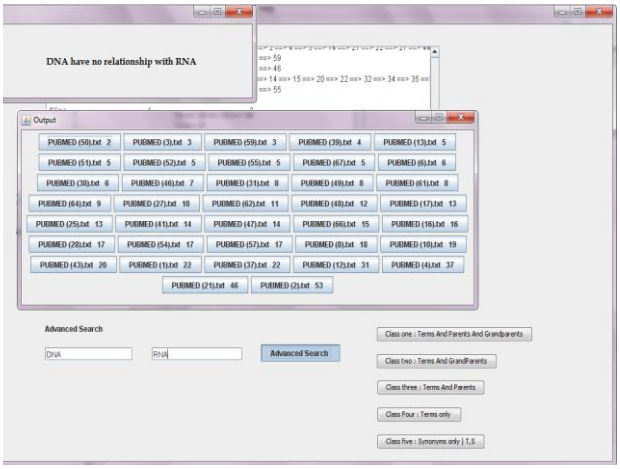

Figure 14: The set of retrieved relevant documents.

biological terms to the system and wanted to get all relevant documents that contain both terms from the craft corpus. The terms "DNA" and "RNA" have been added to the database although they have been removed from Gene Ontology since 2003 and they have been chosen to be searched for because they are very common in corpus and important in the search. The two terms found in the corpus as shown in Table 3.

Table 3: DNA and RNA statistics in craft 1.0.

Terms	Number of Documents that contain the term	Number of Documents that does not contain the term	Number of documents that contain one term and the synonym of the second term
DNA	55	12	18
RNA	44	23	7
Number of relevant documents (contain both term)	37		

The documents to be retrieved must have two terms entered by the user. In the previous experiment the system retrieves all the documents that contain both terms which are 37 documents.

Our assumptions insure that all the retrieved documents will be relevant documents as they contain the two biological terms entered by the user. After several experiments we calculated the precision and recall of the system and got a precision equal to 100% and recall equal to 100%.

The system gives these results because of the following points:

- The advantage of the "exact matching" for the query keywords and non-existence of the concept "partial matching" in the standard Boolean model. So documents can be retrieved if it contains the entered keywords otherwise it will not be retrieved.
- Biological keywords are unique. The "Polysemy" problem is absent, so there is no chance to have multiple words with the same meaning.

8 Conclusion and future work

This paper presents a semantic retrieval system that retrieves relevant documents with high performance. The system improves the performance of semantic information retrieving method since we use the Gene Ontology to infer related biological terms such as synonyms, parents and grandparents of the two related terms entered in the query to retrieve all relevant documents that contain these terms with any combination of the inferred terms. The system extracts the relations between the two related terms entered in the search query to give the researchers additional information about these terms.

In the system we used a ranking methodology to help in ordering the retrieved documents based on the rank values. The system groups the retrieved documents into five classes, each class can be ordered based on the frequencies of the input terms with concern of the balancing between the frequencies of input terms.

The system shows improvements in the percentage of the precision and recalls since it retrieves documents that actually contain needed information so all the retrieved documents are relevant ones.

The proposed system can be generalized to other domain specific fields. The authors use JAVA as a programming language to implement the system. JAVA has a limitation that affects the building of inverted list since it allows reading only 750 documents from the corpus. As a future work we aim to increase the number of documents read from the corpus by enhancing the index built in the system to be a multi-index that allows the system to read and store more terms from the documents and organizes the terms by other way. The presented system semantically expands the user query by parents and grandparents up to two levels in the Gene Ontology. As an improvement the system can use more than two levels from the Gene Ontology to enhance the semantic acting of the system. The system ranking issues can be changed and another ranking methodology can be used to get much close to the researchers' needs. The system ranking issues can be enhanced based on the researchers' feedback. The system grouping criteria's can be differed based on the application domain and can be decomposed based on domain requirements. The system extraction process can be enhanced to extract the relations between biological terms from the documents instead of the Ontology. Also other additional relations that are not mentioned in this research can be extracted. The system is based on two related terms and can be enhanced to use more than two terms.

References

[1] Parul Gupta and Dr. A.K.Sharma, *"Context based Indexing in Search Engines using Ontology"*, International Journal of Computer Applications (0975 – 8887), Volume 1 – No. 14, 2010.

[2] Marta Bleda, Joaquin Tarraga, Alejandro de Maria, Francisco Salavert, Luz Garcia-Alonso, MatildeCelma, Ainoha Martin, Joaquin Dopazo and Ignacio Medina, *"CellBase : a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources"*, Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), 46012 Valencia, Spain, Nucleic Acids Res,2012 Jul;40(Web Server issue):W609-14. doi: 10.1093/nar/gks575. Epub 2012 Jun 12, 2012.

[3] Meng Hu and Jiong Yang, *"A System of User-Guided Biological Literature Search Engine"*, EECS, Case Western Reserve University, 2005.

[4] ChengXiangZhai, *"Statistical Language Models for Information Retrieval A Critical Review"*, University of Illinois at Urbana-Champaign, 201 N. Goodwin, Urbana, IL, Foundations and Trends in Information Retrieval, Vol. 2, No. 3, 137–213, 2008.

[5] Mohammad Mustafa Taye, *"Understanding Semantic Web and Ontologies: Theory and Application, Journal of Computing"*, Volume 2, Issue 6, June Mohammad Mustafa Taye, ISSN 2151-9617, 2010.

[6] Thomas Eiter, GiovambattistaIanni, Thomas Krennwallner and Axel Polleres (2008) *"Rules and Ontologies for the Semantic Web"*, Reasoning Web, pp. 1-53.

[7] David Jin and Sally Lin, *"Advances in Computer Science"*, Intelligent Systems and Environment: Vol.1, Advances in Intelligent and Soft Computing, Springer - 1st Edition 1:134, 2011.

[8] The Gene Ontology Consortium, *"Gene Ontology: tool for theunification of biology"*. Nat. Genet., 25, 25–29, 2000.

[9] http://www.gene ontology.org/.

[10] MarwaMostafaMostafa,Enas M.F. El Houby and Akram Salah, *"Ontology-based Biological Information Retrieval System"*, Australian Journal of Basic and Applied Sciences, No-9177-AJBAS, 540-545 August 2012.

[11] C.Sumithiradevi, Dr.M.Punithavalli and S.Suresh, *"Biomining:-An Efficient Data Retrieval Tool for Bioinformatics to Avoid Redundant and Irrelevant Data Retrieval from Biological Databases"*, Global Journal of Computer Science and Technology, Volume XI Issue I Version I, 2011.

[12] Qi Zhou, Hong Zhang, Meiying Geng and Chenggang Zhang, *"A Real-Time and Dynamic Biological Information Retrieval and Analysis System (BIRAS)"*, Beijing Polytechnic University, Beijing, China, 2003.

[13] Minlie Huang, Xiaoyan Zhu, Shilin Ding, Hao Yu and Ming Li, *"ONBIRES:Ontology-based Biological Relation Extraction System"*, In Proceedings of the Fourth Asia Pacific Bioinformatics Conference, 2006.

[14] AnáliaLourenço, Rafael Carreira, Daniel Glez-Peña, José R. Méndez, SóniaCarneiro, Luis M. Rocha, Fernando Díaz, Eugénio C. Ferreira, Isabel Rocha, FlorentinoFdez-Riverola, Miguel Rocha, *"BioDR: Semantic indexing networks for biomedical document retrieval Expert Systems with Applications"*, 37(4), 3444-3453, 2010.

[15] Cui Tao, *"Information Extraction and Integration from Heterogeneous Biological Data Sources"*, Department of Computer Science, Brigham Young University, Provo, Utah 84602, U.S.A, 2006.

[16] Jiewen Wu, IhabIlyas and Grant Weddell, *"A Study of Ontology-based Query Expansion"*, Cheriton School of Computer Science, University of Waterloo, CS-2011-04, 2011.

[17] Philippe Thomas, Johannes Starlinger, Alexander Vowinkel, Sebastian Arzt and Ulf Leser, *"GeneView: a comprehensive semantic search engine for PubMed"*, Nucleic Acids Research, 40(Web Server issue):W585-91, 2012.

[18] M.C. Díaz-Galiano, M.T Martín-Valdivia and L.A. Ureña-López, *"Query expansion with a medical ontology to improve a multimodal information retrieval system"*, journal of Computers in Biology and Medicine, Elsevier Science, 2007.

[19] C. Pasquier, *"Biological data integration using Semantic Web technologies"*, Biochimie, 90(4), 584-594, 2008.

[20] Yungang Xu, MaozuGuo, Wenli Shi, Xiaoyan Liu, Chunyu Wang, *"A novel insight into Gene Ontology semantic similarity"*, Genomics, 101(6), 368-375, 2013.

[21] Craft1.0 http://bionlp-corpora.sourceforge.net/CRAFT/index.shtml