Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance

Thomas Mandl
Information Science
University of Hildesheim
Marienburger Platz 22, D-31141 Hildesheim, Germany
E-mail: mandl@uni-hildesheim.de

Overview Paper

Keywords: information retrieval, evaluation

Received: July 25, 2007

The evaluation of information retrieval systems has gained considerable momentum in the last few years. Several evaluation initiatives are concerned with diverse retrieval applications, innovative usage scenarios and different aspects of system performance. These evaluation initiatives have led to a considerable increase in system performance. Data for evaluation efforts include multilingual corpora, structured data, scientific documents, Web pages as well as multimedia objects. This paper gives an overview of the current activities of the major evaluation initiatives. Special attention is given to the current tracks and developments within TREC, CLEF and NTCIR. The evaluation tasks and issues, as well as some results, will be presented.

Povzetek: Pregledni članek opisuje usmeritve v informacijskih povpraševalnih sistemih.

1 Information retrieval and its evaluation

Information retrieval is the key technology for knowledge management which guarantees access to large corpora of unstructured data. Very often, text collections need to be processed by retrieval systems. Information retrieval is the basic technology behind Web search engines and an everyday technology for many Web users.

Information retrieval deals with the storage and representation of knowledge and the retrieval of information relevant to a specific user problem. Information retrieval systems respond to queries which are typically composed of a few words taken from a natural language. The query is compared to document representations which were extracted during the indexing phase. The most similar documents are presented to the users who can evaluate the relevance with respect to their information needs and problems.

In the 1960s, automatic indexing methods for texts were developed. They implemented the bag-of-words approach at an early stage, and this still prevails today. Although automatic indexing is widely used today, many information providers and even internet services still rely on human information work.

In the 1970s, research shifted its interest to partial match retrieval models and proved superior compared to Boolean retrieval models. Vector space and later probabilistic retrieval models were developed. However, it took until the 1990s for partial match models to succeed in the market. The Internet accelerated this development. All Web search engines were based on partial match models and provided results as ranked lists rather than unordered sets of documents. Consumers got used to this kind of search system and eventually all big search engines included partial match functionality. However, there are many niches in which Boolean methods still dominate, e.g. patent retrieval. The growing amount of machine-readable documents available requires more powerful information retrieval systems for diverse applications and user needs.

The evaluation of information retrieval systems is a tedious task. Obviously, a good system should satisfy the needs of a user. However, the users' satisfaction requires good performance in several dimensions. The quality of the results with respect to the information need, system speed and the user interface are major dimensions. To make things more difficult, the most important dimension, the level to which the search result documents help the user to solve the information need, is very difficult to evaluate. User-oriented evaluation is extremely difficult and requires many resources. In order to evaluate the individual aspects of searches and the subjectivity of user judgments regarding the usefulness of searches, an impracticable effort would be necessary. As a consequence, information retrieval evaluation experiments try to evaluate only the system. The user is an abstraction and not a real user. In order to achieve that, the users are replaced by objective experts who judge the relevance of a document to one information need. This evaluation methodology is called the Cranfield paradigm, based on the first information

retrieval system evaluation in the 1960's (Cleverdon 1997). This is still the evaluation model for modern evaluation initiatives. The first main modern evaluation initiative was the Text Retrieval Conference (TREC). TREC had a huge impact on the field. The emphasis on evaluation in information retrieval research was strengthened. System development and the exchange of ideas was fostered by TREC and systems greatly improved in the first few years.

Recent evaluation efforts try to keep their work relevant for the real world and make their results interesting for practical applications. Yet, in order to cope with these new heterogeneous requirements and to account for the changing necessities of different domains and information needs, new approaches and tasks need to be established.

The remainder of the paper is organized as follows. The next section provides an introduction to the measures commonly used in information retrieval evaluation. Section 3 introduces the basic activities and the history of the three major evaluation initiatives. The following sections present challenges recently taken up in the scientific evaluation of information retrieval systems. They discuss how different document types, multimedia elements and large corpora are introduced. Section 5 points to new developments regarding evaluation methods.

2 Evaluation of information retrieval systems

The information retrieval process is inherently vague. In most systems, documents and queries consist of natural language. The content of the documents needs to be analyzed, which is a hard task for computers. Robust semantic analysis for large text collections or even multimedia objects has yet to be developed. Therefore, text documents are represented by natural language words mostly without syntactic or semantic context. This is often referred to as the bag-of-words approach. These keywords or terms can only imperfectly represent an object because their context and relations to other terms are lost in the indexing process.

Information retrieval systems can be implemented in many ways by selecting a model and specific language processing tools. They interact in a complex system and their performance for a specific data collection cannot be predicted. As a consequence, the empirical evaluation of performance is a central concern in information retrieval research (Baeza-Yates & Ribeiro-Neto 1999). Researchers are faced with the challenge to find measures which can be used to determine whether a system is better than another one (Bollmann 1984).

The most important basic measures are recall und precision. Recall indicates the ability of a system to find relevant documents, whereas precision measures show how good a system is in finding only relevant documents without ballast. Recall is calculated as the fraction of relevant documents found among all relevant documents, whereas precision is the fraction of relevant documents in the result set. The recall requires knowledge of all the relevant documents in a collection that could never be put together in any real world collection. The number of known relevant documents is usually used to calculate the value. Both measures are set oriented, However, most current systems present ranked results. In this case, a recall and precision value pair can be obtained for each position on the ranked list taking into account all documents from the top of the list down to that position. Plotting these values leads to the recall-precision graph. The average of precision values at certain levels of recall is calculated as the mean average precision (MAP), which expresses the quality of a system in one number.

Evaluation initiatives compare the quality of systems by determining the mean average precision for standardized collections and topics like descriptions of information needs. The relevant documents for the topics are assessed by humans who work through all the documents in a pool. The pool is constructed from the results of several systems and ultimately limits the number of relevant documents which can be encountered. Research on the pooling technique has shown that the results are reliable (Buckley & Voorhees 2005).

3 Major evaluation initiatives

The three major evaluation initiatives are historically connected. TREC was the first large effort, which started in 1992. Subsequently, CLEF and NTCIR adopted the TREC methodology and developed specific tasks for multilingual corpora, cross-lingual searches as well as for specific application scenarios.

3.1 Text Retrieval Conference (TREC)

TREC[1] was the first large-scale evaluation initiative and is now in 2008 in its 17th year. TREC is sponsored by the National Institute for Standards and Technology (NIST) in Gaithersburg, Maryland, USA where the annual TREC conference is held. TREC may be considered as the start of a new era in information retrieval research (Voorhees & Buckland 2006). For the first time in information science, TREC achieved a high level of comparability of system evaluations. In the first few years, the effectiveness of the systems approximately doubled. The initial TREC collections for ad-hoc retrieval, which were based on some statement expressing an information need, were newspaper and newswire articles. These test data and collections have stimulated considerable research and are still a valuable resource for development. The model of the user for the ad-hoc evaluation is that of a "dedicated searcher" who is willing to read through hundreds of documents. In the first few years, the topics were very elaborate and long. Starting with TREC 3, the topics became shorter.

TREC has organized the evaluation in 26 tracks which started and ended in different years (Voorhees 2007). Important tracks, apart from the ad-hoc track,

[1] http://trec.nist.gov

were Filtering, Question Answering, Web and Terabyte Track. Other tracks which ran over the last years were the following ones:

- The Question Answering (QA) track requires systems to find short answers to mostly fact-based questions from various domains. In addition to the identification of a relevant document, question-answering systems need to extract the snippet which serves as an answer to the question. In recent years, the QA track is also moving towards more difficult questions like list and definition questions.
- The track with the most participants in 2005 was the Genomics track which combines scientific text documents with factual data on gene sequences (Hersh et al. 2004, Hersh et al. 2006). These tasks attract researchers from the bio-informatics community as well as text retrieval specialists (see also section 3 for domain specific data). The Genomics track ran three times and ended in 2007.
- In the HARD track (High Accuracy Retrieval from Documents), the systems are provided with information on the user and the context of the search. This meta-data needs to be exploited during the retrieval (Allen 2004).
- The Robust Retrieval track applied new evaluation measures which focus on a stable performance over all topics instead of just rewarding systems with good mean average precision (see section 5).
- In the Spam track, which started in 2005, the documents are e-mail messages and the task is to identify spam and non-spam mail. One English and one Chinese corpus need to be filtered. In the Immediate Feedback Track, the system is given the correct class after each message classification and in the Delayed Feedback after a batch of mails. Both tasks simulate a user who gives feedback to a spam filter (Cormack 2006).
- The Blog track, which started in 2006, explores information behavior in large social computing data sets (see section 5.2).
- The Terabyte track can be seen as a continuation of the ad-hoc track and its successor, the Web track. The data collection of almost one terabyte comprises a large and recent crawl of the GOV domain containing information provided by US government agencies. Here, the participants need to scale information retrieval algorithms to large data sets.

```
   <top> <head> Tipster Topic Description
   <num> Number:  051
   <dom> Domain:  International Economics
   <title> Topic:  Airbus Subsidies
   <desc>  Description:  Document will
discuss government assistance to Airbus
Industrie, or mention a trade dispute
between Airbus and a U.S. aircraft
producer over the issue of subsidies.
   <smry> Summary: Document will discuss
government assistance to Airbus Industrie,
or mention a trade dispute between Airbus
and a U.S. aircraft producer over the
issue of subsidies.
   <narr> Narrative: A relevant document
will cite or discuss assistance to Airbus
Industrie by the French, German, British
or Spanish government(s), or will discuss
a trade dispute between Airbus or the
European governments and a U.S. aircraft
producer, most likely Boeing Co. or
McDonnell Douglas Corp., or the U.S.
government, over federal subsidies to
Airbus.
   <con> Concept(s):
   … </top>

   <top> <num> Number: 400
   <title> Amazon rain forest
   <desc> Description: What measures are
being taken by local South American
authorities to preserve the Amazon
tropical rain forest?
   <narr> Narrative: Relevant documents
may identify: the official organizations,
institutions, and individuals of the
countries included in the Amazon rain
forest; the measures being taken by them
to preserve the rain forest; and
indications of degrees of success in these
endeavors.
   </top>
```

Figure 1: Example Topics from TREC 1 (51) and TREC 7 (400)[2]

TREC continuously responded to ideas from the community and created new tracks. In 2008, the following five tracks are organized at TREC:

- In the Enterprise Track, the participants have to search through the data of one enterprise. The model for this track is intranet search, which is becoming increasingly important. This track started in 2005.
- The Legal track intends to develop effective techniques for legal experts. It was organized for the first time in 2006.
- The large amount of results and submission data has been analysed in many studies. The Million Query Track is a consequence of such evaluation research stimulated by TREC. It was organized for the first time in 2007. Some 10,000 queries from a search engine log were tested against the GOV Web collection (see section 6).[3]
- In 2008, a new Relevance Feedback Track was established.

TREC has greatly contributed to empirically-driven system development and it has improved retrieval systems considerably.

[2] http://trec.nist.gov/data/topics_eng
[3] http://ciir.cs.umass.edu/research/million/

3.2 Cross-language evaluation forum (CLEF)

CLEF[4] is based on the Cross-Language Track at TREC which was organized three years ago (Peters et al. 2005). In 2000, the evaluation of multilingual information retrieval systems moved to Europe and the first CLEF workshop took place. Since then, the ever-growing number of participants has proved that this was the right step. Different languages require other optimization methods in information retrieval. Each language has its own morphological rules for word creation and its words with specific meanings and synonyms. As a consequence, linguistic resources and retrieval algorithms need to be developed for each language. CLEF intends to foster this development.

CLEF closely followed the TREC model for the creation of an infrastructure for research and development. The infrastructure consisted of multi-lingual document collections comprised of national newspapers from the years 1994, 1995 and 2002. CLEF has been dedicated to include further languages. Document collections for the following languages have been offered over the years: English, French, Spanish, Italian, German, Dutch, Czech, Swedish, Russian, Finnish, Portuguese, Bulgarian and Hungarian.

All topics developed in one year are translated into all potential topic languages. Participants may start with the topics in one language and retrieve documents in another language. CLEF offers more topic languages than document languages. Some languages which attract less research from computational linguistics can be used as topic languages as well. These have included Amharic, Bengali, Oromo and Indonesian over the years.

The participating systems return their results, which are then intellectually assessed. These relevance assessments are always done by native speakers of the document languages (Braschler & Peters 2004). The results from CLEF have led to scientific progress as well as significant system improvement. For example, it could be shown that character n-grams can be used for representing text instead of stemmed terms (McNamee & Mayfield 2004).

Similar to TREC, a question-answering (QA) track has been established, which has attracted many participants. In addition to finding a short answer to a question, the system needs to cross a language border. The language of the query and the document collection are not identical in most cases. Like in the ad-hoc tasks, languages are continuously added. Furthermore, the types of questions are modified. Questions for which no answer can be found in the collection need to be handled properly as well. Temporally restricted questions have also been added (for example, "Where did a meteorite fall in 1908"). A document collection of eight languages has been established as the standard collection.

The number of questions answered correctly is the main evaluation measure. Over the last years, systems have considerably improved. In 2005, six systems reached an accuracy of 40% and two were even able to achieve 60% accuracy. Most experiments submitted are monolingual (61%), bi-lingual experiments reach an accuracy of 10% less than the monolingual. There is a tendency toward applying more elaborated linguistic technologies like deep parsing (Vallin et al. 2005, Leveling 2006).

This performance gap between mono- and cross-lingual retrieval is mainly due to translation errors which lead to non-relevant documents. On the other hand, there are some topics which benefit from the translation. In the target language there might be no synonyms for a topic word leading to a performance decrease in the initial language. Overall, it needs to be said that the variance between topics is typically much larger than the performance difference between systems (Mandl, Womser-Hacker et al. 2008).

In the ImageCLEF track, combined access to textual and graphic data is evaluated (see section 3). The Interactive task (iCLEF) is focused on the user interaction and the user interface. Participants need to explore the differences between a baseline and a contrastive system in a user test setting. The comparison is done only within the runs of one group. The heterogeneity of approaches does not allow for a comparison between groups. In 2004, the interactive track included question answering and in 2005, systems for image retrieval were evaluated in user tests. In the target search task, the user is presented with one image and needs to find it through a keyword search (Gonzalo et al. 2005). In the interactive setting, systems for question answering and image retrieval proved that they are mature enough to support real users in their information needs.

A Web track was installed in 2005 (see section 4) as well as the GeoCLEF track focusing on geographic retrieval (see section 5.1). The tracks Spoken Document Retrieval and Domain Specific are also mentioned in sections below.

For CLEF 2008, library catalogue records from The European Library (TEL) will form a new collection for ad-hoc retrieval. A new filtering task will be established. After a pre-test in 2007, the role of disambiguation in retrieval will be investigated in cooperation with the SemEval Workshop. Participants will receive disambiguated collections and topics and can experiment on ways to use that additional information successfully.

3.3 Asian language initiative NTCIR

NTCIR[5] is dedicated to the specific language technologies necessary for Asian languages and cross-lingual searches among these languages and English (Oyama et al. 2003, Kando & Evans 2007). The institution organizing the NTCIR evaluation is the National Institute for Informatics (NII) in Tokyo where the workshops have been held since the first campaign in 1997. NTCIR takes one and a half years to run an evaluation campaign. In December 2005, the fifth

[4] http://www.clef-campaign.org

[5] http://research.nii.ac.jp/ntcir/

workshop was organized with 102 participating groups from 15 countries. NTCIR is attracting more and more European research groups. NTCIR established a raw data archive which contains the submissions of participants. This will allow long-term research on the results.

The cross-lingual ad-hoc tasks include the three Asian languages Chinese, Japanese and Korean (CJK). Similar to TREC and CLEF, the basic document collections are newspaper corpora. Meanwhile the newspaper collection contains some 1.6 million documents. Overall, the results of the systems are satisfactory and comparable to the performance levels reached at TREC and CLEF; however, the performance between language pairs differs greatly. The fifth workshop emphasized named entity recognition, which is of special importance for Asian language retrieval. In Asian languages, word borders are not marked by blanks like in Western languages. Consequently, word segmentation is not trivial and the identification of named entities is more complicated than in Western languages.

Apart from the ad-hoc retrieval tasks, NTCIR has a patent, a Web and a question-answering track. The general model for the cross-language question-answering task from newspaper data is report writing. It requires processing series of questions belonging together. Patent retrieval focuses on invalidity search and text categorization. The collection has been extended from 3.5 to 7 million documents. Rejected claims from patent offices are used as topics for invalidity search. A set of 1200 such queries has been assembled. Patent search by non-experts based on newspaper articles is also required. A sub-task for passage retrieval aims at more precise retrieval within a document. The number of passages which need to be read until the relevant passage is encountered is the evaluation measure (Oyama et al. 2003).

The Web track comprises a collection of approximately one Terabyte of page data collected from the JP domain. The search task challenges developers to find named pages. This is called a navigational task because users often search for homepages or named pages in order to browse to other pages from there.

4 Document types

TREC started to develop collections for retrieval evaluation based on newspaper and news agency documents. This approach has been adopted by CLEF and NTCIR because newspaper articles are easily available in electronic formats, they are homogeneous, no domain experts are necessary for relevance assessments and parallel corpora from different newspapers, dealing with the same stories, can be assembled. Nevertheless, this approach has often been criticized because it was not clear how the results gained from newspaper data would generalize to other kinds of data. Especially domain-specific texts have other features than newspaper data and the vocabularies used are quite different across domains. The focus on newspapers seemed to make evaluation results less reliable and relevant for other realistic settings.

As a consequence, many other collections and document types have been integrated into evaluation collections throughout recent years. These include structured documents and multimedia data which are discussed in the following sections.

An important step was the establishment of the domain specific track at CLEF where systems can be evaluated for domain specific data in mono- and multi-lingual settings for German, English and Russian. The collection is based on the GIRT (German Indexing and Retrieval Test Database) corpus from the social sciences, containing English and German abstracts of scientific papers (Kluck 2004). At TREC, the demand for bio-informatics led to the integration of the Genomics track where genome sequences and text data are combined. The new legal track at TREC is also dedicated to domain experts. The patent retrieval task at NTCIR requires the optimization for the text type patent. For all these domain specific tasks, the special vocabularies and other characteristics need to be considered in order to achieve good results.

4.1 Structured documents

Newspaper stories have a rather simple structure. They contain a headline, an abstract and the text. In many applications, far more numerous and complex document structures need to be processed by information retrieval systems.

The inclusion of Web documents into evaluation campaigns has been a first step to integrate structure. Web documents written in HTML have very heterogeneous structures and only a small portion is typically exploited by retrieval systems. The HTML tag title is most often used for specific indexing, but headlines and links texts are used, too.

One initiative is specifically dedicated to the retrieval from documents structured with XML. INEX[6] (Initiative for the Evaluation of XML Retrieval) started in 2002 and is annually organized by the University of Duisburg-Essen in Germany. The topics are based on information needs and as such, cannot be solved merely by XML database retrieval. The challenge for the participants lies in tuning their systems such that they do not only retrieve relevant document parts, but the smallest XML element which fully satisfies the information need (Fuhr 2003). The need to exploit structure has attracted many database research groups to INEX. The test collection within INEX includes several computer science bibliography and paper collections as well as the Lonely Planet travel guide books which exhibit a rich structure and even contain pictures. Based on these pictures, a multimedia track has been established at INEX.

[6] http://inex.is.informatik.uni-duisburg.de/

4.2 Multimedia data

Multimedia data is becoming very important and most search engines already provide some preliminary form of image retrieval. Research has been exploring the algorithms for content based multimedia retrieval but is still struggling with the so-called “semantic gap” (Mittal 2006). Systems cannot yet make the step from atomic features of an image, like the color of a pixel, to the level of an object which a human would recognize. Evaluation campaigns are integrating multimedia data in various forms into their efforts.

The track ImageCLEF began in 2003 and explores the combination of visual and textual features in cross-lingual settings. Images seem to be language independent, but they often have associated text (e.g. captions, annotations). ImageCLEF assembled collections of historic photographs and medical images (radiographs, photographs, power-point slides). For the historic photographs, ad-hoc retrieval is performed and the topics are motivated by a log-file analysis from an actual image retrieval system (for example “waves breaking on the beach”, “a sitting dog”). Visual as well as textual topics were developed and some topics contain both text and images. In contrast to other tasks at CLEF, where usually binary assessments are required, ternary relevance assessment is carried out by three assessors at ImageCLEF. The best systems reach some 0.4 mean average precision; however, performance varies greatly among languages (Clough et al. 2005).

For the medical images, retrieval and annotation is required. Medical doctors judged the relevance of the images for the information need. For the automatic annotation task, images needed to be classified into 57 classes identifying, for example, the body region and image type.

In addition, ImageCLEF introduced an interactive image retrieval task in cooperation with the Interactive track to investigate the interaction issues of image retrieval. It could be shown that relevance feedback improved results similarly to ad-hoc retrieval.

In 2001, a video track started up and ran again in 2002. Starting in 2003, the evaluation for video retrieval established an independent evaluation campaign called TRECVid[7] (TREC Video Retrieval Evaluation). In 2005, TRECVid concentrated on four tasks:

- Shot boundary determination: systems need to detect meaningful parts within video data.
- Low-level feature extraction: systems need to recognize whether camera movement appears in a scene (pan, tilt or zoom)
- High-level feature extraction: ten features from a Large Scale Concept Ontology for Multimedia (LSCOM) were selected and systems need to identify their presence in video scenes. The ontology includes cars, explosions and sports.
- Search tasks include interactive, manual, and automatic retrieval. Examples of topics are: "Find shots of fingers striking the keys on a keyboard which is at least partially visible" and "Find shots of Boris Yeltsin".

The data collection includes 170 hours of television news in three languages (English, Arabic and Chinese) from November 2004 collected by the Linguistic Data Consortium[8] (LDC), some hours of NASA educational programs and 50 hours of BBC rushes on vacation spots (Smeaton 2005). Considerable success has been achieved by applying speech recognition to the audio track of a video and by running standard text retrieval techniques to the result. On the other hand, content-based techniques for the visual data still require much research to bridge the semantic gap.

Apart from visual data, retrieval of audio data has also attracted considerable research. At CLEF, a Cross-Language Spoken Document Retrieval (CL-SR) track has been running since 2003. In 2005, the experiments were based on the recordings of interviews with Holocaust survivors (Malach collection). The interviews last 750 hours and are provided as audio files and as transcripts of an automatic speech recognition (ASR) system. Participants may base the retrieval on their own ASR or use the transcript provided. The data was tagged by humans who added geographical and other terms. For the retrieval test, interviews in Czech and English are provided. The retrieval systems need to be optimized for the partially incorrect output of the ASR (Oard et al. 2006).

Even for music retrieval, an evaluation campaign has been established. The Music Information Retrieval Evaluation eXchange (MIREX[9]) focuses on content-based music data processing. The tasks include query by humming, melody extraction and music similarity (Downie 2003, Downie et al. 2005).

5 Specific user needs

Focusing on very specific user needs makes evaluation more real-world-oriented and increases its value for that specific application area. Each application has its own particular character. While some users work on a recall oriented basis (patent attorneys), others focus on precision (web users). Many users want all aspects of a topic to be represented in the result set independent of the number of retrieved relevant documents. This aspect has been evaluated in the Genomics Track (Hersh et al. 2006) and has previously been researched in the Novelty Track.

5.1 Geographic information retrieval

In GeoCLEF[10], systems need to retrieve news stories with a geographical focus. GeoCLEF is a modified ad hoc retrieval task, involving both spatial and multilingual aspects based on newspaper collections previously offered at CLEF (Gey et al. 2007, Mandl, Gey, et al.

[7] http://www.itl.nist.gov/iaui/894.02/projects/trecvid/

[8] http://www.ldc.upenn.edu/

[9] http://www.music-ir.org/mirex2006/

[10] http://www.uni-hildesheim.de/geoclef/

2008). Examples of topics are "shark attacks off California and Australia" or "wind power on Scottish Islands". In order to master the last topic, the systems need knowledge of what the Scottish Islands are. For other topics, it is necessary to include symbolic knowledge about the inclusion of one geographical region within another. Participants applied named entity identification for geographical names and used geographical knowledge sources like ontologies and gazetteers. However, standard approaches outperformed specific geographical tools in the first two editions 2005 and 2006 and still perform similarly in 2007. This might be due to the fact that standard approaches like blind relevance feedback lead to results similar to geographical reasoning systems.

For 2007, the topics were developed to include more challenging aspects. Ambiguity, vaguely defined geographic regions (Near East) and more complex geographical relations were emphasized (Mandl, Gey, et al. 2008).

5.2 Opinion retrieval

Social Software applications allow users to create and modify Web pages very easily. Such systems enable users to quickly publish content and share it with other users. The success of social systems encouraged millions of users to become members of social networks and led to the creation of a large amount of user-generated content.

Users create huge amounts of text in blogs which can be simplistically described as online diaries with comments and discussions. Many blogs contain personal information; others are dedicated to specific topics. The huge interest in blogs has also led to blog spam. In order to explore searching in blogs, TREC initiated a blog track in 2006. A collection was created by crawling well-known blog locations on the Web. More than 3.2 million documents, in this case blog entries from more than 100,000 blogs, were collected (Ounis et al. 2006).

One of the most interesting and blog-specific issues is the subjective nature of the content. It is very likely to find opinions on topics. Companies e.g. are beginning to exploit blogs by looking for opinions on their products. Consequently, a very natural retrieval task regarding blogs is the retrieval of opinions on a given topic.

Typical approaches for opinion retrieval include list-based and machine learning approaches. List-based methods rely on large lists of words of a subjective nature. Their occurrence in a text is seen as an indicator of opinionated writing. Machine learning methods are trained on typically objective texts like online lexical documents and on subjective texts like product review sites. Systems learn to identify texts with opinions based on features like individual words, the number of pronouns or adjectives.

The opinion retrieval task in the blog track at TREC was based on relevance assessment at several levels. The typical relevant and non-relevant judgments were supplemented by explicitly negative, explicitly positive and both positive and negative judgments. The subjective documents were well balanced in the pool. The document pool contained 2% spam blog posts, showing that spam is a problem. The variance among topics is very large. However, systems managed to retrieve spam documents more likely at later ranks rather than on earlier ranks. Interestingly, opinion finding and relevance scores of the systems correlate substantially. The opinion finding scores are higher than the topic relevance scores overall (Ounis et al. 2006).

The idea of opinion analysis is considered at NTCIR as well. For NTCIR-7, a track for Cross-Language Information Retrieval for Blogs (CLIRB) and a track for Multilingual Opinion Analysis Task (MOAT) are envisioned.

User-created content is also becoming a subject for CLEF. The interactive track at CLEF 2008 intends to investigate how users use the picture-sharing platform FlickR to search for images in a multilingual way.

6 Large corpora

Information Retrieval is faced with new challenges on the Web. The mere size of the Web forces search engines to apply heuristics, in order to find a balance between efficiency and effectiveness. One example for a heuristic would be to only index significant parts of each document. The dynamic nature of the Web makes frequent crawling necessary and creates the need for efficient index update procedures. One of the most significant challenges of the Web is the heterogeneity of the documents in several respects. Web pages vary greatly in length, document format, intention, design and language. These issues have been dealt with in evaluation initiatives.

The Web Track at TREC ran from 1999 until 2004. In its last edition, it attracted 74 runs (Craswell & Hawking 2004). The Web corpus used at TREC had a size of 18 GB and was created by a crawl of the GOV domain, containing US government information. This track is focused on retrieval of Web pages in English. Similarly, a Chinese Web Evaluation Initiative organized by Beijing University is focused on the Chinese Web . The document collection crawled from the CN domain contains some 100 Gigabyte . The tasks for the systems are named page finding, home page finding and an information ad-hoc task based on topics selected from a search engine log. NTCIR also includes a Web collection for retrieval evaluation, based on a collection of one Terabyte of document data from the Japanese Web.

The task design for Web retrieval evaluation in evaluation initiatives is oriented towards navigational information needs (Broder 2002) and known item finding tasks. As such, these evaluations differ from ad-hoc retrieval, where an informational need is the model for the topics developed. The main difference between the two search types is that the navigational information needs aims for one specific Web page (homepage or another page) which the user might even have visited before. In contrast, the informational task aims at finding pages on a certain topic to satisfy a certain information need. In these cases, it is not known how many potential

target pages exist. The pooling technique is not necessary for navigational information needs. On the contrary, for informational search tasks, the quality of the pooling technique needs to be re-evaluated. The quality and depth of the pool from which the relevant documents will be extracted by human assessors cannot be judged. The effect of this fact on the evaluation results needs to be assessed. Consequently, most evaluation tracks for Web retrieval remain restricted to navigational information needs.

Most Web retrieval tracks include mainly navigational information needs. This may be partially due to the need for many resources to create relevance judgments for informational Web search tasks.

The results of the TREC Web track indicate that the use of Web-specific knowledge of document structure and anchor text positively affects retrieval quality. The contribution of link structure and URL length is less obvious. Typical information retrieval techniques like stemming do not seem to be necessary for Web retrieval (Craswell & Hawking 2004).

At TREC 2002, a navigational task as well as a topic distillation task were offered. Both led to different results. For navigational tasks, link analysis led to better results whereas link analysis could not improve topic distillation (Craswell & Hawking 2003:6).

```
   <title> highway safety
   <desc> Description:
   Find documents related to improving
highway safety in the U.S.
   <narr> Narrative:
Relevant documents include those related
to the improvement of safety of all
vehicles driven on highways, including
cars, trucks, vans, and tractor trailers.
Ways to reduce accidents through
legislation, vehicle checks, and drivers'
education programs are all relevant.
```

Figure 2: Example for a Topic of the TREC Web Track 2002 (Craswell & Hawking 2003)

A new Terabyte Track at TREC is based on a crawl of the domain GOV and contains more than 400 Gigabyte of document data. The topics are developed from ad-hoc type information needs. The goal of this track is scaling the systems and evaluation methodology to a large-size collection. It is expected that the pool and the relevance assessments will be dramatically less complete than for newspaper collections for ad-hoc retrieval. The effects of this problem for evaluation methods are being investigated (Clarke et al. 2004).

Another solution to this problem lies in the development of more topics. A statistical analysis of TREC results modified the number of topics and used different amounts of the relevance assessment available (Sanderson & Zobel 2005). It revealed that more topics and shallow pools led to more reliable results than deep relevance assessments for fewer topics. Fewer relevance judgments could diminish the cost of evaluation campaigns drastically. A new step in this direction is the so-called Million Query for TREC 2007 where this finding will be exploited. Some 10,000 queries from a search engine log were tested against the GOV Web collection. Relevance assessment will focus on a subset of a few hundred queries and it will consider 40 or more documents per topic.

At NTCIR-4, an informational retrieval task was organized which has been dropped at NTCIR 5. A navigational task was part of NTCIR 3 through NTCIR-5 (Eguchi et al. 2004). For the informational retrieval task, the pooling problem was addressed at NTCIR. Shallow vs. deep pooling was compared. For all topics, pooling with the top ten documents was carried out and for a subset, the top 100 documents of pooled runs were used and additional techniques were used to extend the pool (Eguchi et al. 2004). The pooling levels were mapped to different user models in Web search. The results varied between the two methods to a considerable extent. Another Web specific parameter in the evaluation was the document model behind the relevance assessment. The information unit can be the page itself or pages to which it directly links. This document model considers the hub function of pages, which is often highly valuable for informational search tasks.

The Web is obviously a very natural environment for multilingual retrieval. Users have many different native languages and for each user, most of the information on the Web is not in his or her native language. In 2005, a new Web track was established at CLEF focusing on the challenges of multi-linguality. Similar to the corpus used at TREC, European government sites were crawled and included in the collection. Unlike the TREC GOV collection, which is mainly English, and the NTCIR collection, which is English and Japanese, the EuroGOV collection contains pages in more than 25 languages (Sigurbjörnsson et al. 2005)[11]. Many pages are even multilingual (Artemenko et al. 2006). The multilingual corpus of Internet pages was engineered by the University of Amsterdam. The web crawl collected pages of official institutions (mainly ministries) in European countries. It covers 27 domains and contains 3.6 million Web pages. The documents are written in some 25 languages. The size of the corpus is some 100 Gigabyte. Together with the participants, the track organizers were able to create 575 topics for homepage and named page finding in the first year. The tasks offered were mixed-mono-lingual (many queries of different languages being submitted to one search engine), bi-lingual (retrieve English documents based on Spanish queries) and truly multi-lingual where the language of the target was not specified (Sigurbjörnsson et al. 2006).

The performance for the mixed-mono task is comparable to mono-lingual ad-hoc results; however, the performance for both cross-lingual tasks lacks far behind. There is a great need for further research. The first year of work on the Web task led to surprising results. Whereas the automatic translation of topics is the main approach to bridge the language gap in ad-hoc retrieval, translation harmed performance for the Web topics. This

[11] http://ilps.science.uva.nl/WebCLEF/

may be due to the reason that the Web task is focused on homepage and named page finding.

7 Evaluation measures

The initiatives adhered to the traditional evaluation measures mentioned in the second section of this paper. They assumed that there is a valid concept of the quality of a system, which can be assessed by several strongly correlating measures. However, the large-scale evaluations themselves have stirred interest in these basic questions of evaluation.

It has often been pointed out that the variance between queries is larger than the variance between systems. There are often very difficult queries. Few systems solve these well and they lead to very bad results for most systems (Harman & Buckley 2004). Thorough failure analysis can lead to substantial improvement. For example, the absence of named entities is a factor which can generally make queries more difficult (Mandl & Womser-Hacker 2004). It is also understood that the requests which are answered poorly will strongly contribute toward any negative feelings of the user.

As a consequence, a new evaluation track for robust retrieval has been established by the Text Retrieval Conference (TREC). Robustness can be seen as the capacity of a system to perform well under heterogeneous conditions. The robust track not only measures the average precision over all queries, but also emphasizes the performance of the systems for difficult queries. In order to perform well in this track, it is more important for the systems to retrieve at least a few documents for difficult queries than to improve the performance on average (Voorhees 2005). To allow for a system evaluation based on robustness, more queries are necessary than for a normal ad-hoc track. The score per system is not calculated by the arithmetic mean of all topics, but by the geometric mean. The geometric mean reduces the influence of topics which were solved with very good results. The concept of robustness is extended in TREC 2005. Systems need to perform well over different tracks and tasks (Voorhees 2005). For multilingual retrieval, robustness is also an interesting evaluation concept because the performance between queries differs greatly. The issue of stable performance over all topics instead of high average performance has been explored at CLEF 2006 for six languages (Di Nunzio et al. 2007). For the top systems, a high correlation between standard and robust measures was found. However, further analysis revealed that the robust measures lead to very different results with a growing number of topics, especially if the percentage of low performing topics is high. Because this is the case in multi-lingual retrieval settings, robust evaluation is of high importance for multi-lingual technology (Mandl, Womser-Hacker et al. 2008).

For several other tasks, the traditional measures have been considered to be inadequate. For the Web tasks, for example, Web-user-oriented measures were sought. For the navigational tasks, the mean reciprocal rank of the target item was established. For informational tasks, early precision measures were used. Often, the precision at ten documents is used. The recall power of a system can be neglected when taking into consideration an underlying user model with the average Web user who is seeking only a few hits.

One concern about evaluations of large collections is the percentage of judged documents. The effort spent on relevance assessment remains constant. As collections grow, only a small fraction of documents is actually being assessed by humans. Many of the documents retrieved by systems are not judged. These documents are considered as not relevant. Results might be unreliable because most documents in the result lists are not judged. A new measure has been proposed and meanwhile adopted for most experiments with large collections. The binary preference (Bpref) metric takes only documents into account which were judged by a human juror. They are disregarded and the new measure checks how many times a system retrieves a relevant document before a document is judged as not relevant (Buckley & Voorhees 2004).

Still, it remains unclear how evaluation results relate to user satisfaction. For a small experiment with simple search tasks, no correlation between evaluation measures and user satisfaction was found (Turpin & Scholer 2006). The relation between system performance and the perception of the user needs to be the focus of more research.

Many novel retrieval measures have been developed in the past years. Nevertheless, the classic measures are still being widely used. Overall, there is a consensus that these new measures might reveal something important that is not covered by recall and precision. However, it is not yet well understood what this “something” is (Robertson 2006).

8 Summary

As this overview shows, the evaluation of information retrieval systems has greatly diversified in recent years. Research has recognized that evaluation results from one domain and one application cannot be transferred to other domains. Evaluation campaigns need to continuously re-consider their tasks, topics and evaluation measures, in order to make them as similar to real-world tasks and information needs as possible.

In the future, the diversification will continue as further tasks are being explored. The evaluations of multimedia data and of Web resources are likely to converge because more and more multimedia data is available on the Web. Further evaluation initiatives are being established. In 2008, the new Indian evaluation campaign FIRE (Forum for Information Retrieval Evaluation) will run for the first time and provide test environments for the major languages spoken in India .

9 References

[1] Allen, James (2004). HARD Track Overview in TREC 2004 High Accuracy Retrieval from Documents. In: Buckland, Lori; Voorhees, Ellen

(Eds.). *The Thirteenth Text Retrieval Conference (TREC 2004)* NIST Special Publication: SP 500-261. http://trec.nist.gov/pubs/trec13/t13_proceedings.html

[2] Artemenko, Olga; Mandl, Thomas; Shramko, Margaryta; Womser-Hacker, Christa (2006). Evaluation of a Language Identification System for Mono- and Multi-lingual Text Documents. In: *Proceedings of 2006 ACM SAC Symposium on Applied Computing (SAC)* April, 23-27, 2006, Dijon, France. pp. 859-860.

[3] Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (1999). Retrieval Evaluation. In: Baeza-Yates, R.; Ribeiro-Neto, B. (eds.): *Modern Information Retrieval*. Addison-Wesley. pp. 73-97.

[4] Bollmann, Peter (1984). Two Axioms for Evaluation Measures in Information Retrieval. In: *Proceedings of the Third Joint BCS/ACM Symposium on Research and Development in Information Retrieval (SIGIR 1984)* Cambridge, 2-6 July 1984. pp. 233-245

[5] Braschler, Martin; Peters, Carol (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements. In: *Information Retrieval* no. 7. pp. 7-31.

[6] Broder, Andrei (2002). A taxonomy of web search. In: *ACM SIGIR Forum* vol. 36(2) pp. 3–10.

[7] Buckley, Chris; Voorhees, Ellen (2005): Retrieval System Evaluation. In: *TREC: Experiment and Evaluation in Information Retrieval*. Cambridge & London: MIT Press. pp. 53-75.

[8] Clarke, Charles; Craswell, Nick; Soboroff, Ian (2004). Overview of the TREC 2004 Terabyte Track. In: Buckland, Lori; Voorhees, Ellen (Eds.). *The Thirteenth Text Retrieval Conference (TREC 2004)* NIST Special Publication: SP 500-261. http://trec.nist.gov/pubs/trec13/t13_proceedings.html

[9] Cleverdon, Cyril (1997). The Cranfield Tests on Index Language Devices. In: Sparck-Jones, Karen; Willett, Peter (Eds.): *Readings in Information Retrieval*. Morgan Kaufman. pp. 47-59.

[10] Clough, Paul; Müller, Henning; Deselaers, Thomas; Grubinger, Michael; Lehmann, Thomas; Jensen, Jeffery; Hersh, William (2005). The CLEF 2005 Cross-Language Image Retrieval Track. In: *Working Notes of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF*. Sep. 2005, Vienna, Austria. http://www.clef-campaign.org/

[11] Cormack, Gordon (2006). TREC 2006 Spam Track Overview. In: Voorhees & Buckland (2006)

[12] Craswell, Nick; Hawking, David; Wilkinson, Ross; Wu, Mingfang (2004). Overview of the TREC 2003 Web Track. In: *Proceedings Text Retrieval Conference (TREC)*. http://trec.nist.gov/pubs/trec12/t12_proceedings.html

[13] Craswell, Nick; Hawking, David (2004). Overview of the TREC-2004 Web Track. In: Voorhees & Buckland 2004.

[14] Downie, Stephan (2003). Toward the Scientific Evaluation of Music Information Retrieval Systems. In: *Intl Symposium on Music Information Retrieval (ISMIR)* Washington, D.C., & Baltimore, USA. http://ismir2003.ismir.net/papers/Downie.PDF

[15] Downie, Stephen; West, Kris; Ehmann, Andreas; Vincent, Emmanuel (2005). The 2005 Music Information retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview. In: *6th International Conference on Music Information Retrieval (ISMIR)* London, UK, 11-15 Sept. pp. 320-323.

[16] Eguchi, Koji; Oyama, Keizo; Aizawa, Akiko; Ishikawa, Haruko (2004). Overview of the Informational Retrieval Task at NTCIR-4 WEB. In: *NTCIR Workshop 4 Meeting Working Notes.* http://research.nii.ac.jp/ntcir-ws4/NTCIR4-WN/index.html

[17] Fuhr, Norbert (2003). Initiative for the Evaluation of XML Retrieval (INEX): *INEX 2003 Workshop Proceedings,* Dagstuhl, Germany, December 15-17. http://purl.oclc.org/NET/duett-07012004-093151

[18] Gey, Fredric; Larson, Ray; Sanderson, Mark; Bischoff, Kerstin; Mandl, Thomas; Womser-Hacker, Christa; Santos, Diana; Rocha, Paulo; Di Nunzio, Giorgio; Ferro, Nicola (2007). GeoCLEF 2006: the CLEF 2006 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, Carol et al. (Eds.). *Evaluation of Multilingual and Multi-modal Information Retrieval. 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, Revised Selected Papers. Berlin et al.: Springer [LNCS 4730] pp. 852-876.

[19] Gonzalo, Julio; Clough, Paul; Vallin, Alessandro (2006). Overview of the CLEF 2005 Interactive Track In: Peters, Carol; Gey, Fredric C.; Gonzalo, Julio; Jones, Gareth J.F.; Kluck, Michael; Magnini, Bernardo; Müller, Henning; Rijke, Maarten de (Eds.). *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, Vienna, Austria, Revised Selected Papers. Berlin et al.: Springer [LNCS 4022] pp. 251-262.

[20] Harman, Donna; Buckley, Chris (2004). The NRRC reliable information access (RIA) workshop. In: *Proceedings of the 27th annual international conference on Research and development in information retrieval (SIGIR)*. pp. 528-529.

[21] Hersh, William; Bhuptiraju, Ravi; Ross, Laura; Johnson, Phoebe; Cohen, Aaron; Kraemer, Dale (2004). TREC 2004 Genomics Track Overview. In: Buckland, Lori; Voorhees, Ellen (Eds.). *The Thirteenth Text Retrieval Conference (TREC 2004)* NIST Special Publication: SP 500-261. http://trec.nist.gov/pubs/trec13/t13_proceedings.html

[22] Hersh, William; Cohen, Aaron; Roberts, Phoebe; Rekapalli, Hari Krishna (2006). TREC 2006 Genomics Track Overview. In: Voorhees & Buckland (2006)

[23] Kando, Noriko and Evans, David (2007). *Proceedings of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access*. National Institute of Informatics, Tokyo, Japan. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/index.html

[24] Kluck, Michael (2004). The GIRT Data in the Evaluation of CLIR Systems - from 1997 until 2003. In: *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003,* Trondheim, Norway. Revised Selected Papers. Springer: LNCS 3237. pp. 376-390

[25] Leveling, Johannes (2006). A baseline for NLP in domain-specific information retrieval. In: Peters, Carol et al. (eds): *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, Vienna, Austria, Revised Selected Papers Berlin: Springer [LNCS 4022] pp. 222-225.

[26] Mandl, Thomas; Womser-Hacker, Christa (2005). The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. In: *Proceedings ACM SAC Symposium on Applied Computing (SAC)*. Santa Fe, New Mexico, USA. March 13.-17. pp. 1059-1064.

[27] Mandl, Thomas; Gey, Fredric; Di Nunzio, Giorgio; Ferro, Nicola; Larson, Ray; Sanderson, Mark; Santos, Diana; Womser-Hacker, Christa; Xing, Xie (2008). GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, Carol et al. (Eds.). *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007,* Budapest, Hungary, Revised Selected Papers. Berlin et al.: Springer [LNCS] to appear. Preprint at: http://www.clef-campaign.org

[28] Mandl, Thomas; Womser-Hacker, Christa; Ferro, Nicola; Di Nunzio, Giorgio (2008). How Robust are Multilingual Information Retrieval Systems? In: *Proceedings ACM Symposium on Applied Computing (SAC)* Fortaleza, Brazil. pp. 1132-1136.

[29] McNamee, Paul; Mayfield, James (2004). Character N-Gram Tokenization for European Language Text Retrieval. In: *Information Retrieval*, vol. 7 (1/2). pp. 73-98.

[30] Mittal, Ankush (2006). An Overview of Multimedia Content-Based Retrieval Strategies. In: *Informatica* 30. pp. 347-356.

[31] Di Nunzio, Giorgio; Ferro, Nicola; Mandl, Thomas; Peters, Carol (2007). CLEF 2006: Ad Hoc Track Overview. In: Peters, Carol et al. (Eds.). *Evaluation of Multilingual and Multi-modal Information Retrieval. 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, Alicante, Spain, Revised Selected Papers. Berlin et al.: Springer [LNCS 4730] pp. 21-34.

[32] Di Nunzio, Giorgio; Ferro, Nicola; Mandl, Thomas; Peters, Carol (2008). CLEF 2007: Ad Hoc Track Overview. In: Peters, Carol et al. (Eds.). *8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007*, Budapest, Hungary, Revised Selected Papers. Berlin et al.: Springer [LNCS] to appear. Preprint: http://www.clef-campaign.org

[33] Robertson, Stephan (2006). On GMAP: and other transformations. In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM)* Arlington, Virginia, USA. pp. 872-877

[34] Sanderson, Mark; Zobel, Justin (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2005)* Salvador, Brazil. ACM Press. pp. 162 - 169.

[35] Sigurbjörnsson, Börkur; Kamps, Jaap; Rijke, Maarten de (2006). Overview of WebCLEF 2005. In: Peters, Carol; Gey, Fredric C.; Gonzalo, Julio; Jones, Gareth J.F.; Kluck, Michael; Magnini, Bernardo; Müller, Henning; Rijke, Maarten de (Eds.). *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*, Vienna, Austria, Revised Selected Papers. Berlin et al.: Springer [LNCS 4022] pp. 810-824.

[36] Sigurbjörnsson, Börkur; Kamps, Jaap; de Rijke, Maarten (2005). Blueprint of a Cross-Lingual Web Retrieval Collection. In: Journal of Digital Information Management, vol. 3 (1) pp. 9-13.

[37] Smeaton, Alan (2005). Large Scale Evaluations of Multimedia Information Retrieval: The TRECVid Experience. In: CIVR 2005 – International Conference on Image and Video Retrieval, Springer: LNCS 3569, pp 11-17.

[38] Turpin, Andrew; Scholer, Falk (2006). User performance versus precision measures for simple search tasks. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*, Seattle, Washington, USA, August 6-11. ACM Press. pp. 11-18.

[39] Oard, Douglas W.; Wang, Jianqiang; Jones, Gareth; White, Ryen; Pecina, Pavel; Soergel, Dagobert; Huang, Xiaoli; Shafran, Izhak (2006). Overview of the CLEF-2006 Cross-Language Speech Retrieval Track. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): *CLEF 2006 Working Notes*. http://www.clef-campaign.org/2006/working_notes

[40] Ounis, Iadh; Rijke, Maarten; Macdonald, Craig; Mishne, Gilad; Soboroff, Ian (2006). Overview of the TREC-2006 Blog Track. In: Voorhees & Buckland (2006)

[41] Oyama, Keizo; Ishida, Emi; Kando, Noriko (2002) (eds.). *NTCIR Workshop 3: Proceedings of the Third NTCIR Workshop on research in Information Retrieval, Automatic Text Summarization and Question Answering* (Sept 2001-Oct 2002) http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/index.html

[42] Vallin, Alessandro; Giampiccolo, Danilo; Aunimo, Lili; Ayache, Christelle; Osenova, Petya; Peñas,

Anselmo; de Rijke, Maarten; Sacaleanu, Bogdan; Santos, Diana; Sutcliffe, Richard (2006). Overview of the CLEF 2005 Multilingual Question Answering Track. In: Peters, Carol; Gey, Fredric C.; Gonzalo, Julio; Jones, Gareth J.F.; Kluck, Michael; Magnini, Bernardo; Müller, Henning; Rijke, Maarten de (Eds.). *Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005,* Vienna, Austria, Revised Selected Papers. Berlin et al.: Springer [LNCS 4022] pp. 307-331.

[43] Voorhees, Ellen; Buckland, Lori (2006) (eds.). *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)* NIST Special Publication SP 500-272. National Institute of Standards and Technology. Gaithersburg, Maryland. Nov. 2006. http://trec.nist.gov/pubs/trec15/

[44] Voorhees, Ellen (2005). The TREC robust retrieval track. In: *ACM SIGIR Forum* 39 (1) pp. 11-20.

[45] Voorhees, Ellen (2006). Overview of TREC 2006. In: Voorhees & Buckland (2006)