Evaluation, Best Practice & Collaboration for Multilingual Information Access

BEST PRACTICES FOR TEST COLLECTION CREATION AND INFORMATION RETRIEVAL SYSTEM EVALUATION

Mark Sanderson, University of Sheffield, UK

Martin Braschler, Zürcher Hochschule für Angewandte Wissenschaften, Switzerland

Edited by:
TrebleCLEF Consortium

November 2009

www.trebleclef.eu

BEST PRACTICES FOR TEST COLLECTION CREATION AND INFORMATION RETRIEVAL SYSTEM EVALUATION

Mark Sanderson, University of Sheffield, United Kingdom

Martin Braschler, Zürcher Hochschule für Angewandte Wissenschaften, Winterthur, Switzerland

Edited by:
TrebleCLEF Consortium

November 2009

TrebleCLEF: Evaluation, Best Practices & Collaboration for Multilingual Information Access
Coordination Action: 215281 Kickoff: 1 January 2008 Duration: 24 months

Abstract

There is a widely held perception that the evaluation of searching systems is a difficult, time consuming, expensive process to get right. A great deal of that perception is due to the impression given by a number of annual academic evaluation campaigns that appear to need large quantities of money and a substantial community effort in order to conduct evaluation to a necessary level of accuracy. However a large quantity of research has been conducted showing that evaluation can be conducted far quicker than is generally thought. The broad conclusions of this research has not yet been collated into a single publication.

In addition, almost all publications on evaluation of information retrieval systems are geared towards academic research. The needs of this community are not the same as the needs of the users, administrators and designers of actual searching systems. While the academic community is willing to work with shared testing resources that are perhaps overly abstract representations of a searching situation, practitioners cannot use these resources; they need a way to test on their actual data sets in order to understand exactly which queries are working and which are failing. If they are trying to decide which of two commercial systems they will choose to purchase, it will be critical that they are able to test on their own datasets.

Surprisingly little has been written for practitioners on how to assess the quality of an operational search system, even less has been written about how to conduct such an evaluation quickly with minimal use of resource. This document attempts to address this gap in publications by providing a guide on how to conduct an evaluation. The choice of collection, how to source topics, how to conduct relevance assessments, and which of the many available evaluation measures to use is described in this document. The issues surrounding use of testing multilingual search are also addressed. The document also describes two case studies, illustrating, in the first, how one of the large research oriented evaluation campaigns constructs test collections; in the second, how a comparison between two searching systems was conducted by an organisation on its own data.

Preface

The popularity of the Internet and the consequent global availability of networked information sources and digital libraries have led to a strong demand for multilingual access and communication technologies. These technologies should support the timely and cost-effective provision of knowledge-intensive services for all members of linguistically and culturally diverse communities. This is particularly true in the multilingual setting of Europe. Despite recent research advances, there are still very few operational systems available, and these are limited to the most widely used languages.

The Treble-CLEF project was launched to build on and extend the results already achieved by the Cross Language Evaluation Forum (CLEF). The challenge that has been faced is how to best transfer the impressive research results to a wider market place. The aim has been both to support the development and consolidation of expertise in the multidisciplinary research area of multilingual information access (MLIA) and also to disseminate this know-how to the application communities through a set of complementary activities, with the following goals:

- To promote high standards of evaluation in MLIA systems using three approaches: test collections; user evaluation; and log file analysis
- To sustain a MLIA evaluation community by organizing annual evaluation campaigns and providing high quality access to past evaluation results
- To disseminate knowhow, tools, resources and best practice guidelines, enabling system developers to make content and knowledge accessible, usable and exploitable over time, over media and over language boundaries.

In addition to the annual CLEF evaluation campaigns, the project has organised a summer school and number of workshops, and is responsible for the production of a large volume of literature and training material on Research and Development in the MLIA domain. The results are now being transferred to the application communities through the promotion of dissemination events, and via the implementation of a project portal to be made public in Autumn 2009.

TrebleCLEF has produced three Best Practice Guidelines:

- Best Practices in Language Resources for Multilingual Information Access
- Best Practices in System and User Oriented Multilingual Information Access
- Best Practices for Test Collection Creation & Evaluation Methodologies

For more information on TrebleCLEF activities and results see http://www.trebleclef.eu/.

Carol Peters, ISTI-CNR, Pisa, Italy
TrebleCLEF Coordinator

Table of Contents

1 Introduction

Evaluation has always been a critical component of Information Retrieval (IR). Of particular interest to users, designers, researchers and administrators of IR systems has been measuring *effectiveness*: determining the *relevance* of items, retrieved by a search engine, relative to a user's query.

To understand why evaluation has always held such importance in IR, it is helpful to consider what one actually does. Van Rijsbergen (1979) contrasted IR systems with databases, which typically search over highly structured, strongly typed data, where care is taken to ensure that objects within a database can be unambiguously located using a unique key. Consequently, the queries executed on such systems are generally assumed to be complete: precisely defining what data is to be retrieved.

Van Rijsbergen pointed out that the query to an IR system is generally incomplete in specifying the user's request: not all aspects of the user's information need are explicitly expressed within it. To illustrate, a user entering the query "nuclear waste dumping" into a research paper search engine is probably looking for multiple documents describing this topic in detail, the user probably prefers to see documents from reputable sources and is willing to examine a largish amount of material. A user querying on a web search engine for "BBC" is probably looking only for the official home page of the corporation and isn't interested in much else. A lawyer searching for prior art that might invalidate a patent application is likely to have the time and patience to examine every potentially relevant document a retrieval system can find.

The IR systems receiving such queries need to fill in the gaps of the users' underspecified query. The fact that the content being searched is typically unstructured and its components (i.e. words) are usually ambiguous, often have synonyms and maybe incorrectly entered merely adds to the challenge of locating relevant items. In contrast to a DB system, whose search outputs are deterministic, the accuracy of an IR system's output cannot be predicted with any confidence prior to the system being used. Therefore, search accuracy needs to be evaluated.

A strong focus of IR research has been on measuring the *effectiveness* of an IR system: determining the *relevance* of retrieved items, relative to a user's information need. One approach to measuring effectiveness is to issue a set of queries to that system and examine the search output, judging the relevance of the retrieved documents and determining if the system was successful for those queries. To achieve this, one needs to conduct the following work:

- the documents to be searched need to be collected and (often) copyright or access issues need to be resolved;
- a set of queries need to be gathered from some source perhaps from logs or from a requirements analysis taken from users;
- the retrieved documents have to be assessed for relevance; and
- if one is interested in understanding what relevant documents were missed, a large number of documents in the whole collection need to be assessed also.

From the example queries described above, one can also see that different searching contexts require different search algorithms. Consequently, when testing an IR system that is intended for use in a particular *operational setting*, it is necessary to ensure that the documents, queries and relevance judgments used in testing reflect those that will be seen in the operational setting. These testing sets are commonly known as *test collections*. The classic components of which are as follows:

- a collection of documents; each document is given a unique identifier, the *docid*;
- a set of topics (also referred to as queries); like the documents, each topic is given an id number, typically called the query id (*qid*); and
- a set of *relevance judgments* (referred to as *qrels*) composed of a list of qid/docid pairs, detailing the relevance of documents to topics.
- In addition, an *evaluation measure* that reflects the searching behaviour of the typical users of an IR system is selected

Together, the collection and chosen evaluation measure provide a simulation of the users of a searching system in a particular operational setting. Using test collections, researchers can assess a retrieval system in isolation helping locate points of failure, but more commonly, a collection is used to compare the effectiveness of multiple retrieval systems. Either rival systems are compared with each other, or different configurations of the same system are contrasted. In such situations, a determination is made if one system is better than another or if the systems are equally effective. Such determinations, by implication, predict how well the retrieval systems will perform relative to each other if or when they were deployed in the operational setting simulated by the test collection.

In the possession of an appropriate test collection and with an evaluation measure selected, users of an IR system can test the system simply by loading the documents into their system and submitting the topics one-by-one. The list of the docids retrieved for each of the topics is concatenated into a set, which is commonly referred to as a *run*. The builder then examines the content of the run to determine which of the documents retrieved were present in the qrels and which were not. The evaluation measure is then used to summarize the effectiveness of that run.

A key innovation in the Information Retrieval academic community was the early recognition of the importance of building and crucially sharing test collections. Through sharing, others benefit from the initial effort put into the creation of a test collection by re-using the collection in other experiments. Groups evaluate their own IR systems on a shared collection and make meaningful comparisons on the effectiveness of their system against others that were tested on the same collection. Others can reproduce past work and by testing on a shared test collection, can measure if their re-production is correct. Shared test collections provide a focus for many international collaborative research exercises. Experiments using them constitute the main methodology for validating new retrieval approaches in the academic community.

Such is the importance of test collections to the IR research field that at the time of writing, there are many conferences and meetings devoted purely to evaluation of research ideas: including three international conferences, TREC, CLEF and NTCIR, which together have run more than thirty times since the early 1990s.

- TREC – in 1990, the US government agency DARPA funded the National Institute of Standards and Technology (NIST) to build a large test collection to be used in the evaluation of a text research project, TIPSTER. In 1991, NIST proposed that this collection (several orders of magnitude larger than previous collections) be made available to the wider research community through a program call TREC – the Text REtrieval Conference. Since then, TREC has created and fostered the creation of a wide range of test collections covering a large number of document domains and query types.
- CLEF[1] – The annual Cross Language Evaluation Forum (funded through a range of sources within the EU's frameworks of funding) focuses strongly on search across European languages; though in recent years it has diversified into other forms of search such as images (Martin Braschler & Carol Peters 2004). CLEF has also re-examined the search of library catalogue data with a collection composed of 3 million multi-lingual records (Agirre et al. 2008).
- NTCIR – The NII Test Collection for IR Systems is a regular evaluation exercise held every 18 months in Japan. Funded by the NII (National Institute for Informatics), NTCIR has focused on cross language search for Asian languages such as Japanese, Chinese and Korean. A particular focus has been on patent search (Kando 2003). The first NTCIR evaluation exercise used a collection of the title and abstracts of several hundred thousand scholarly articles (Kando et al. 1999).

As valuable as these collections are to the research community, to those who wish to deploy a searching system in an operational setting or who wish to research a field of IR in which no test collection exists, it is necessary to build a test collection "from scratch".

[1] Pronounced "clay", from the French word for key.

There exist a number of historical and more recent perspectives on test collection based evaluation, including the review of past work in chapter seven of Van Rijsbergen's book (1979); Spärck Jones's edited articles on Information Retrieval experiments (1981); and Salton's evaluation chapter (1968, chap.5). There are also notable publications addressing particular aspects of evaluation: Voorhees and Harman's book amongst other topics details the history of the TREC evaluation exercise and outlines evaluation methods used (2005); a special issue of Information Processing and Management reflected the state of IR evaluation in 1992 (Donna Harman 1992); another special issue on evaluation in the Journal of the American Society for Information Science provided a slightly later perspective from 1996; more recently, Robertson published his personal view on the history of IR evaluation (2008).

However, none of these publications described in detail the means by which test collections can be built quickly and efficiently, despite a wealth of research published on this matter in the past few years. It is this topic that this document attempts to address.

In this handbook, the best practices for constructing a test collection are described along with two case studies that detail how different searching situations require different approaches to building a test collection. The aim is to provide commercial system developers as well as users and administrators of systems with guidelines that will enable them to build their own test collection for system testing and tuning. The study starts with a description of each of the components of a test collection, which is followed by an explanation of the workings of a number of common evaluation measures. This is followed by a tutorial on use of significance when comparing two runs. Issues related to multi-lingual test collections are finally described before, finally case studies are detailed.

Test collections aren't everything

It is important to recognise that evaluation of search based simply on a count of the number of relevant documents retrieved has long been recognised as a limited approach: in 1966, Cleverdon et al (1966, p.4) listed a range of factors that affect the success of a searching system amongst which are the following

1. *"The interval between the demand being made and the answer being given (i.e. time)*
2. *The physical form of the output (i.e. presentation)*
3. *The effort, intellectual or physical, demanded of the user (i.e. effort)."*

To this list one could add:

- the ability of the user at specifying their need;
- the usability of the searching system;
- the context in which the user's query was issued;
- the eventual use for the information being sought.

Many of the later factors relate to usability studies or to wider questions of the evaluation of information systems. All of these approaches to evaluation are important and none are addressed by the test collection methodology. However, they are beyond the scope of this document, readers interested in such testing are directed to the works of Ingwersen & Järvelin (2005), Saracevic (1995) and Borlund & Ingwersen (1997), Borlund (2003).

2 Designing a Test Collection based Evaluation

When creating a test collection for evaluating a searching system, one needs to consider the following questions that will influence the design and will impact on the resources needs to conduct the evaluation.

What is the purpose of the evaluation?

There are three main types of evaluation of a searching system: comparing two competing systems against each other; using evaluation to optimise a searching system; and long term monitoring of a searching system. In all these forms of evaluation, collections are gathered, topics are formed, documents are judged for relevance and evaluation measures are selected; however the differing forms of evaluation will require different designs.

When comparing two different systems (e.g. determining which to purchase), the evaluation will likely be run once. When optimising a system, the evaluation will likely be run multiple times as the parameters of the search system are adjusted to produce the best retrieval results. However, the same topics will be run each time the system is evaluated, with a set of topics held out for final test once the optimization is complete. In the case of monitoring, evaluation will be run at regular intervals. As over time the types of queries issued to an IR system may alter, the topics tested each time could potentially be different.

What resources are available to conduct the evaluation?

Considering what resources will be available when constructing an evaluation is an important early consideration to make so as to scope the exact nature of the evaluation.

Human and financial resources will be needed to generate each of the components of a test collection. In some cases this can represent a substantial quantity of resource. Some of the test collections built by the well known US government funded evaluation exercise, TREC, require several person months of effort from TREC organizers and (when salary costs are taken into account) thousands of dollars are spent, as well as many extra months of effort and cost from across participating research groups. Some evaluation exercises and research groups have also developed specialist software to enable components of test collections be built quicker, the creation and maintenance of which also requires resource. However, because of the scale of the involvement by organizations like TREC, many assume that conducting their own evaluation will be beyond their means.

For many evaluations, substantially smaller quantities of human, software, and financial commitment are required. Creating test collections with the size, detail, or accuracy of TREC is not often needed.

What sort of searching is typically conducted on the search engine(s) under test?

Another critical question to ask when designing the evaluation, is what sort of searching is conducted on the IR system(s) under test? As described in the introduction, different types of search engines are built to work well on different types of query. Some users wish to find a small number of documents, the relevance of which a wide range of people could reliably make a judgment on. In other situations, users may wish to locate substantial numbers of documents, or the documents they seek may require expert knowledge in order to understand their relevance.

As with considering resources, understanding the nature of the searching being undertaken must be conducted as part of the evaluation design.

What do you know about the IR system(s) being tested?

The final aspect to consider is what are the qualities of the IR system(s) being tested? If the retrieval effectiveness of two different IR systems is being compared, for example, do the systems output retrieval results in the same way? Are search features of the engines the same? If there are differences,

might those differences make comparison difficult? If the evaluation task relates to optimization of a searching system, the types of queries that might be helped by the optimization should be considered when designing the test collection.

Once these factors have been considered, it then becomes possible to choose how the four components of the test collection can be constructed. In the next four Sections the issues involved in obtaining, selecting or generating these components are considered.

3 Collections

It may seem to many that considering what collection should be searched is an odd question. In many operational settings, the collection is simply the set of documents that a searching system indexes. Before testing, however, it is worth considering if the searching system is indexing all the information held by an organisation. An evaluation might determine that a system is producing poor retrieval results, but one of the common reasons for this is the IR system not having access to all the documents that could be searched.

If comparing two searching systems, one should consider if the two are searching over the same collection. Several research studies have shown that when evaluating using certain measures, such as precision measured at rank 10 (see below), if one system in the comparison is searching over a larger collection than the other, better retrieval results will be produced even if the internal search algorithms of the two systems are identical.

If for some reason, it is not possible to access the collection that the search engine under test will eventually be deployed to, it is then necessary to locate a collection that provides the best possible simulation of the un-reachable collection. Similarly, if one is a researcher wishing to test a new search algorithm which could potentially be applied to a hypothesised setting, a collection needs to be sought. Despite many years of studying the evaluation of IR systems, there is barely any research from the academic community on how to characterise and generate collections that are accurate simulations of the intended setting of the searching system. The level at which such characterization work has been conducted is identifying the broad genre of a collection (e.g. newspaper articles, web pages, email, etc) and locating matching materials. A number of such collections can be located from evaluation research groups, such as TREC, CLEF, and NTCIR.

4 Queries

When considering the topics of a test collection, a number of questions arise: where can a set be obtained; are all topics the same; how many topics are needed to produce accurate measures; and what information should be recorded with each topic? Each of these issues is now discussed in turn.

Obtaining topics

The intention of the test collection methodology is to create a simulation of an operational setting. It is important that the topics used in the collection are representative of that setting. Almost all early test collections were composed of a set of topics that were written by researchers without a great deal of checking that the topics were representative of the types of queries submitted to search engines. In part, this approach to test collection creation was used as it was hard to obtain logs of user interactions with search systems. There is also evidence of a belief in the research community that topics to search engines were all very similar to each other; hand generated would be little different from actual.

Evidence for this can be found in the way that early web based test collections were created at TREC, where topics were created by hand with little examination of query logs and user behaviour on web search engines. However, researchers using the collections started to report unexpected results: Hawking et al (2000) and Hawking (2001) reported that link based methods in web search, such as

PageRank, appeared to provide little or no value in improving search effectiveness. It was eventually realised that the link based methodologies weren't at fault; the problem was with the test collection topics that were not representative of the types of queries submitted to actual web search engines.

A common source for obtaining a representative sample of topics is from a log of queries submitted to an operational search engine. A simple analysis of logs can locate queries that are popular. Examination of queries where no retrieved documents were examined by the user, or locating queries where a user appears to quickly reformulate their query can used as a way to find failing queries where a search engine is not returning good quality results.

Using this source is not without its challenges, logs will at most record the text of the query and the location of any retrieved documents that were clicked by the user. Relevance assessors who will identify relevant documents for the test collection will require clear instruction on how to assess relevance. Therefore, the broader underlying information need that lies behind a particular query has to be inferred from the recorded data. As surveyed by Jansen and Spink (2006), there is a large body of research conducted on understanding user behaviour from query logs, to which the reader is referred.

Are topics the same?

Understanding the source of the problem in the early web collections at TREC was in part informed by Broder's analysis of web search engine logs (2002) which revealed that user queries could be grouped into three basic categories, each of which required a different approach to searching.

1. So called *informational queries* were submitted by user requiring access to documents describing relevant information.
2. *Navigational queries*, were submitted by users seeking a home page, typically that of an organization or person.
3. *Transactional queries* were queries where the user was seeking to find a web-based service that they could conduct a transaction (e.g. buying tickets for travel, purchasing goods, etc).

Broder's work sparked a wide range of research on classifying user queries and consequently studying how search systems can best respond to each query type: see for example the analysis conducted by Rose (2004). When gathering the topics for a test collection, it is important to consider if there might be different types of topics that the searching system should be tested against. It is not unusual for example for there to be informational and navigational queries submitted to the intranet search engine of an enterprise (e.g. find me a page describing this topic, find me the home page of this department). An understanding of the range of query types that might be important in an evaluation is still being addressed by the research community.

What information should be recorded with each topic?

A lesson learned from the construction of early test collections was the importance of not simply recording the query text. Armed only with this limited information, assessors judging the relevance of documents are unlikely to understand what the broader information need of the user was and so may not be accurate in judging relevance. Consequently, the topics of most modern test collections are composed of the query text and a description of what the user's information need was. Sometimes the same query text can potentially address multiple information needs and recent test collections have started to produce multiple descriptions with each topic, one description for each distinct need.

Creating your own topics

Not all organisations log the queries submitted to their search engine, others are un-willing to release such data under any circumstance. In such situations, it is necessary to either locate a source of existing queries or to generate queries in some way. As to the question of how best to obtain such a set, similarly to document collections, the research community has not addressed this question in much detail. In the past, if researchers did not have access to query logs, queries were created in some

manner. We detail past research methodologies here, however, it is worth noting that comparisons between the described methods do not appear to have been conducted.

- **Known Item Queries** – The first publication about a test collection being built was by Thorne (1955), who along with his colleague, Cleverdon, was evaluating the effectiveness of a library catalogue system. They suggested that if the queries of library users were logged, then a sample of those should be used in a test collection. However, no such log was being kept, therefore they created a set of so-called *know item queries*. Each query was based on a document known to be catalogued in the library. They tried to imagine what sort of query would be written by a user who wished to locate the information held in the document. The test was: would such a query retrieve the document it was derived from? Such an approach has been used by many researchers when constructing test collections.

- **Searching for topics** – An alternative approach that was used in a number of evaluation exercises, such as TREC and CLEF was to trial topics on a searching system. As stated in the TREC overview papers (e.g. (D. K. Harman 1993); (E. M. Voorhees & D. K. Harman 1999)), the creators of the topics would create a set of candidates. These were tried out by searching on the documents of the chosen collections to estimate how many relevant documents each topic would return. Topics with too few relevant documents were rejected. In addition, because of concerns about the number of relevant documents that might have to be assessed, topics that appeared to have too many relevant documents were also rejected. This later choice was made so as to ensure that assessors had a manageable number of documents to judge. There was a danger, however, that this choice introduced a bias in the test collection topics that could affect the collections ability to effectively simulate an operational setting.

Increasingly in the research community, these generational approaches are falling out of favour with sampling from query logs being used more often. However, to the best of our knowledge, this isn't being done because of published research showing that these artificial methodologies are poor, there is just an assumption that logs are a better source.

How many topics should go into a test collection?

A key question when constructing a test collection, is how many topics are needed in order to be confident about any results produced? In the IR research community, there is a general consensus that around 50 topics is sufficient and that any number less than 25 is likely to produce largely meaningless results; there is a small body of research confirming these numbers Voorhees (1998), Sanderson and Zobel (2005).

However, it is hard to be confident that such figures will be generally applicable across all searching situations. For example, there are publications from some of the web search engine companies describing test collections with many more topics: White and Morris (2007) mentioned a collection at Microsoft with 10,680 "*query statements*"; Carterette and Jones (2007) described a collection in Yahoo! with 2,021 queries. However, generating test collections of this size requires a great deal of resource and it is unlikely that such detailed collections are needed in most testing situations.

As will be described in the following section on QRELS, it will be seen that the question of the correct number of topics is related to how much assessor effort is devoted to locating the relevant documents for each topic. Given a particular balance between numbers of topics and quantities of assessment per topics (as described in a later Section) there are statistical methods such as significance tests that one can employ to determine if measurements gained from a test collection are reliable or not.

Ultimately, the answer to this question depends on the magnitude of the difference between searching systems that one is trying to measure. As will be seen in one of the case studies, important differences between searching systems can be revealed using just fifty topics based on a limited number documents assessed per topic.

5 Creating Relevance Judgements – QRELS

There is a perception that obtaining the qrels for a test collection is a process that requires a large investment of time and human resources. However, building qrels is only a time consuming process if it is necessary to quantify the number of relevant documents that were not retrieved by the searching system(s) under test. If this information is not needed, then building the qrels is relatively straightforward. In the initial part of this section, questions about how qrels can be gathered, who should assess qrels and what sort of assessments should be made is described. This is followed by a description for those who need to know it, of how at least some of the missing relevant documents can be found.

How to gather qrels

The simplest approach to gathering qrels is to sit an assessor down in front of a computer terminal ask them to systematically type in the queries of the test collection and examine the retrieved documents noting down those that were relevant to the query. Such an approach is simple and quick to achieve and is likely to be workable in many testing situations.

However, there are a series of potential biases that could be introduced by this process. There is relatively little research studying relevance assessors, however, there is some literature on the behaviour of users with search engines. Joachims et al (2005) identified two forms of user bias when examining the logs of a search engine, what they called *trust bias* (in other publications, this was called *presentation bias*) and *quality bias*. Trust bias was given its name due to users' willingness to trust the search engine to find the most relevant item in the top ranks. With the second form of bias, Joachims showed that when the overall quality of search results was poor, users tended to click on less relevant documents. There is a danger that a relevance assessor may be influenced by the same biases seen in Joachims et al's users.

If the evaluation is a comparison of two searching systems that have a different way of presenting search output, there is a danger that the presentation might influence the way that assessors judge the relevance of retrieved documents.

A common approach to tackling these biases is to remove the retrieved documents from the search engine interface and present them to the assessor in a normalized manner, mixing the order in which they are presented. It is relatively straightforward to write an application that does this sort of processing. One of the better known examples is the DIRECT system produced by researchers working within CLEF (Dussin & Ferro 2009). This system not only presents documents in an unbiased form, but also allows assessors to enter lists of keywords that can be highlighted in the documents so as to speed assessors' ability to locate relevant passages.

Unfortunately, these systems tend not to be publically available and for most people conducting some form of search evaluation, it will be necessary to create one of these programs from scratch.

Who should do the assessing?

Early in the history of search evaluation, there was a concern expressed that a test collection methodology wouldn't work because judging if a document was relevant was, potentially, a highly subjective process. Relevance could be influenced by the prior knowledge of the assessor as well as by the documents seen during the assessment process. An early critic of test collections, Fairthorne (1963) argued for relevance judgments made by groups rather than individuals. Later, Katter stated "*A recurring finding from studies involving relevance judgments is that the inter- and intra-judge reliability of relevance judgments is not very high*" (1968). After such criticism, a series of studies were conducted to examine how variations in judgments might affect measuring effectiveness. One example is the work of Lesk and Salton (1968) who perhaps surprisingly found that even with quite high levels of disagreement between assessors the way in which different searching systems were evaluated against each other was relatively unaffected. This experiment was repeated several times by others over the years and has drawn broadly similar results.

These works provide strong evidence that only one assessor per topic needs to be assigned and that assessors can be selected without a great deal of concern about their background. This sort of liberal approach to picking assessors has been used in a great many evaluation campaigns. The one exception to this is the US-funded evaluation campaign TREC, which generally uses retired intelligence analysts as assessors. However, it is likely that these assessors are used more for their ability to read quickly.

There is some evidence from recent research that subject expertise can have an impact on an evaluation. Bailey et al (2008) repeated the Lesk and Salton's experiment drawing on different sets of assessors based on their knowledge about topics. The assessors were classed as gold, silver, and bronze judges. Gold and silver were subject experts with the gold judges having a more intimate knowledge of the data set being searched. Like the previous results, Bailey et al showed the qrels from the gold and silver judges produced similar evaluation results. However, there was a measurable difference between the results produced from gold and bronze judges.

There is some additional empirical evidence that the cultural background of assessors can be important for some types of search, as can local knowledge of a place for queries related to locations. This topic has not been extensively researched; however, it does suggest that if the evaluation is of a searching task or of a data set that would not be easily understood by an average user, it may be necessary to recruit more expert users.

How many assessments to make

Each query will potentially retrieve a large number of documents. One question to be considered is how many documents should be assessed. The answer is inevitably linked to the question of how many topics have been selected and how many documents the relevance assessors can examine in the available time. Researchers have examined the question of what is the right balance between these factors. Carterette et al (2008) described how they were able to test whether – as Sanderson and Zobel described (2005) – a "wide and shallow" approach of minimally examining many queries was better than a "deep and narrow" detailed examination of a few topics. They found that assessing more topics with fewer judgments was the best use of assessor time.

The question of how many assessments to make is also influenced by the choice of evaluation measure that will be used. As will be seen in the Evaluation Section, there are a wide range of measures. For example, a common one often used is the average number of relevant documents in the top 10 (known as precision at 10). If this is the only measure to be used, then only the first ten results from each search system needs to be assessed.

If the evaluation is a comparison between the outputs of two retrieval systems, time can be saved by only assessing the documents in the union of the outputs (i.e. removing duplicates). However, further time savings are possible, Carterette et al (2006) pointed out that if one is only interested in understanding the magnitude of the difference between two systems, it is only necessary to assess the retrieved documents that will make a difference in the evaluation measure calculated on the two outputs. For example, if the evaluation measure is precision at 10; then only the documents in the symmetric difference[2] of the two outputs needs to be examined. This approach is used in some parts of the TREC evaluation exercise; however, this methodology is only workable if the document selection process is supported in an assessment system.

What are the assessors asked to do?

The question of what judgement an assessor makes is also one that has long been discussed. Although not the choice for the first test collections built, the dominant approach is to simply ask assessors to judge if a retrieved document is relevant or not: so called binary relevance. However, multi-level assessments are becoming more common, with a ternary scheme of highly relevant, partially relevant, and not relevant being a frequent choice.

[2] The elements of the output that are unique to one of the search engines.

One of the reasons for this approach being adopted has been a growing realization that degrees of relevance were commonly being used in the internal test collections of web search companies. To illustrate, White and Morris (2007, p.256) mentioned a form of test collection within Microsoft with relevance judgments "assigned on a six-point scale by trained human judges". Carterette and Jones (2007) described a collection within Yahoo! used for advertising retrieval with five levels of relevance ("Perfect, Excellent, Good, Fair, and Bad"); Huffman and Hochster (2007) described work in Google where assessors judged relevance of retrieved documents on a "graduated scale". (Note, although the dates of these publications are more recent, they are the best examples found of the companies revealing some aspects of their internal testing; it is thought very likely that this approach to relevance has been used for sometime in search companies.)

Whichever type of relevance an assessor is asked to mark documents with, they will need instructions on how to assign documents to a particular categories particularly for borderline cases. They may also need training if the documents, queries or assessment criteria are complex.

Finding the missing relevant

For many evaluations, knowing if one system is better than another is all that is needed. However, for some forms of search evaluation, e.g., legal, medical or patent search, knowing how many documents were missed is important. Finding the most efficient way of locating all such documents has been a major preoccupation of the IR evaluation research community.

The very first test collections were composed of a few hundred short documents, consequently, it was possible to check every document for relevance against every topic. As collections grew, this approach required too much assessment resource. What replaced it was a technique called *pooling*. The researchers who gave this approach its name were, Spärck-Jones and Van Rijsbergen who suggested that for a particular topic, assessors judge the documents retrieved by "*independent searches using any available information and device*" (1975, p.13). The aim of pooling was to create a small subset of documents that would hopefully hold the vast majority of relevant documents in the collection. The set would be composed of the union of documents retrieved from variations of a topic and by running the topic through different search engines each using (presumably) different retrieval algorithms. If the subset was built carefully, it would be possible to locate nearly all documents relevant to a particular topic in a manageable amount of time.

Large evaluation campaigns use what is sometimes called *system pooling*, where a large number of research groups are encouraged to submit runs for the topics of a test collection. The hope being that each group will use different searching techniques, which when combined produces a high quality pool. This approach has been tested in a number of empirical experiments and shown to be effective.

The success and high profile of the large campaigns, such as TREC and CLEF, appear to give the impression that drawing in a large number of research groups is the only way to create a pool. However, much of the early work on pooling assumed that a different approach would be used, namely using the same searching system, but varying the wording of queries to locate as many relevant documents as possible, sometimes called *query pooling*. This approach was most recently tested by Cormack et al's who called this technique *Interactive Search and Judge* (ISJ). Here the assessor would search the test collection for relevant documents, issuing multiple queries, noting down relevant documents found and searching until they could find no more relevant for a particular topic. Cormack et al reported that this approach was highly effective. It has been used in a number of evaluation campaigns since then.

6 Measures

The final component of evaluation using a test collection is the evaluation measure. The measure provides a simple simulation of a user's behaviour. A great many such measures have been proposed, here only the most popular are described.

Precision measured at a fixed ranking

A common option for measuring is to decide that a user will choose only to examine a certain number of ranked results and calculate precision at that fixed rank position.

$$P(n) = \frac{r(n)}{n}$$

Where *r(n)* is the number of relevant items retrieved in the top *n* documents. The choice of *n* is often influenced by the manner in which results are displayed. Precision measured at rank 10 is the commonest approach reflecting the current convention of most web search engines displaying 10 search results per web page.

This form is appealing as the number it calculates reflects the experience a user will typically encounter when using a search engine across a set of topics. Note *P(n)* ignores the rank position of relevant documents retrieved above the cut off and ignores all other relevant documents, also if a topic has fewer than *n* relevant documents in the collection being searched, *P(n)* for that topic will always be <1. However, there is little evidence these features of the measures are problematic.

Mean average precision

A commonly used measure is *mean average precision*. The first reference to this measure was in Harman (1993), where the measure was called *non-interpolated average precision*. It is defined as follows

$$AP = \frac{\sum_{rn=1}^{N} (P(rn) \times rel(rn))}{R}$$

Here, *N* is the number of documents retrieved, *rn* is the rank number; *rel(rn)* is the relevance of the document retrieved at *rn*; *P(rn)* is the precision measured at rank *rn* and *R* is the number of relevant documents for this particular topic. Simply, the measure calculates precision at the rank position of each relevant document and takes the average. Note, by summing over the *N* retrieved documents and dividing by the *R* relevant documents, in effect, precision is measured as being zero for any un-retrieved relevant document.

If one calculates AP for each of a set of topics and takes the mean of those average precision values, the resulting calculation is known as *Mean Average Precision* (MAP). Voorhees appears to be the first to describe the measure as mean average precision (1993), though it took several years for MAP to become its universally accepted name. MAP has become one of the primary measures used in many evaluation exercises as well as a large quantity of published IR research.

Graded relevance measures

Measures such as Precision and MAP can really only be used with binary judgments. Assuming that one can transform grades of relevance assessment to numerical values, Järvelin and Kekäläinen (2000) proposed a suite of measures that evaluated the effectiveness of a retrieval system regardless of the number of levels of relevance. Their simplest measure, *cumulative gain* (CG), is the sum of relevance values (*rel*) measured in the top *n* retrieved documents.

$$CG(n) = rel(1) + \sum_{i=2}^{n} rel(i)$$

Examining some example rankings: in the left hand rank in **Figure 1**, CG measured at rank 5, CG(5)=6. However, because CG ignores the rank of documents we see that this is also the value of CG in the poorer rank on the right in **Figure 1**. However, *discounted cumulative gain (DCG)*, where the relevance values are discounted progressively as one moves down the document ranking used a log-based discount function to simulate users valuing highly ranked relevant documents over the lower ranked.

Rank	Rel
1	2
2	1
3	2
4	0
5	1

Rank	Rel
1	1
2	0
3	2
4	1
5	2

Figure 1: Two document ranks

Rank	Rel	Disc	Rel/Disc	DCG
1	2	1.00	2.0	2.0
2	1	1.00	1.0	3.0
3	2	1.58	1.3	4.3
4	0	2.00	0.0	4.3
5	1	2.32	0.4	4.7

Rank	Rel	Disc	Rel/Disc	DCG
1	1	1.00	1.0	1.0
2	0	1.00	0.0	1.0
3	2	1.58	1.3	2.3
4	1	2.00	0.5	2.8
5	2	2.32	0.9	3.6

Figure 2: Document ranks with discount values

$$DCG(n) = rel(1) + \sum_{i=2}^{n} \frac{rel(i)}{\log_b(i)}$$

Järvelin and Kekäläinen suggested setting *b* to 2. The ranks in **Figure 2** show the values of the discount function and the DCG scores for each rank position. We see that DCG(5) of the left hand rank in **Figure 2** is 4.7 and 3.6 in the right hand poorer rank. In a follow up paper (2002) Järvelin and Kekäläinen added a third measure, *normalized DCG (nDCG)*. Here DCG was normalized against an ideal ordering of the relevant documents, IDCG, see **Figure 3**.

Rank	Rel	Disc	Rel/Disc	IDCG
1	2	1.00	2.0	2.0
2	2	1.00	2.0	4.0
3	1	1.58	0.6	4.6
4	1	2.00	0.5	5.1
5	0	2.32	0.0	5.1

Figure 3: Perfect ordering of relevant documents

$$nDCG(n) = \frac{DCG(n)}{IDCG(n)}$$

The value of *nDCG* range between 0 and 1. The *nDCG(5)* of the left and right rankings in **Figure 2** are 4.7/5.1=0.92 and 3.6/5.1=0.71. Al-Maskari et al (2007) pointed out that in certain circumstances *nDCG* can produce unexpected results. To illustrate, for both rankings in Figure 4, there are only three known relevant documents, though the topic on the right has three highly relevant documents, the other has three partially relevant documents. For both topics the rankings are ideal, so the *nDCG* (DCG÷IDCG) in both cases is 1, which is perhaps a counter intuitive result.

Rank	Rel	Disc	Rel/Disc	DCG	IDCG
1	2	1.00	2.0	6.7	6.7
2	2	1.00	2.0	8.7	8.7
3	2	1.58	1.3	10.0	10.0
4	0	2.00	0.0	10.0	10.0
5	0	2.32	0.0	10.0	10.0

Rank	Rel	Disc	Rel/Disc	DCG	IDCG
1	1	1.00	1.0	4.6	4.6
2	1	1.00	1.0	5.6	5.6
3	1	1.58	0.6	6.3	6.3
4	0	2.00	0.0	6.3	6.3
5	0	2.32	0.0	6.3	6.3

Figure 4: Rankings from two different topics that result in the same *nDCG*

It would appear that discounted cumulative gain measures are commonly used by search engine companies, though not all use the Järvelin and Kekäläinen version. Burges et al (2005) described a version of NDCG, for which the DCG component more strongly emphasized the high ranking of the most relevant documents:

$$DCG(n) = \sum_{i=1}^{n} \frac{2^{rel(i)} - 1}{\log(1+i)}$$

7 Beyond the Mean: Comparison and Significance

Whichever evaluation measure one uses, the effectiveness of one run will almost always be compared to the effectiveness of another. In a great deal of published IR research, such comparisons consider the (absolute or relative) difference between the runs averaged across all topics. However, such a straightforward approach can hide important detail, which is illustrated with an examination of three runs: *a*, *b*, *c*. The MAP for the three runs is 0.281, 0.324 and 0.373 respectively. With similar sized gaps between the runs, one might view the comparisons to be revealing similar differences. However, if one graphs comparisons between *a* & *b* and *b* & *c* – plotting the AP scores across each of the 50 topics used in the collection – a more complex picture is revealed; see Figure 5 and Figure 6.

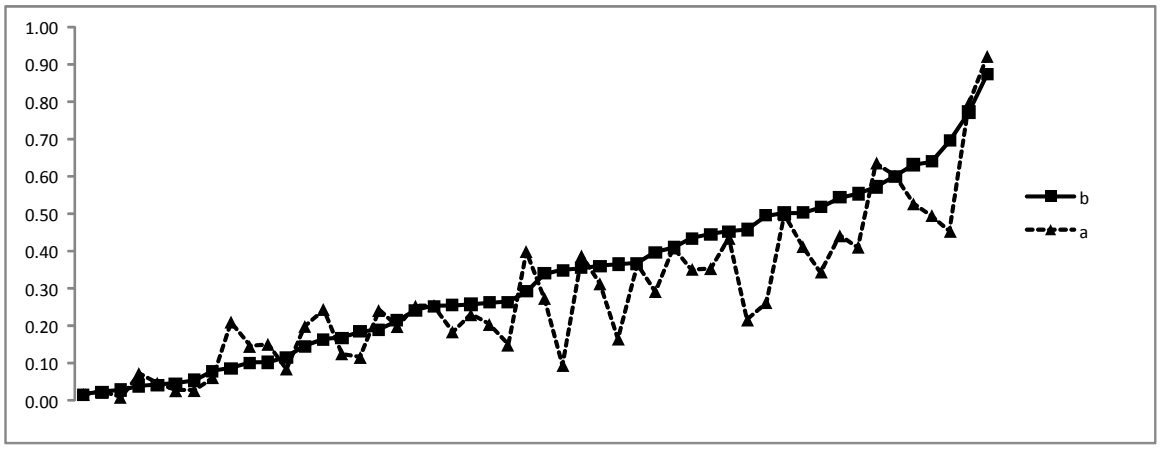

Figure 5: Topic-by-topic comparison of two TREC-8 runs based on average precision scores.

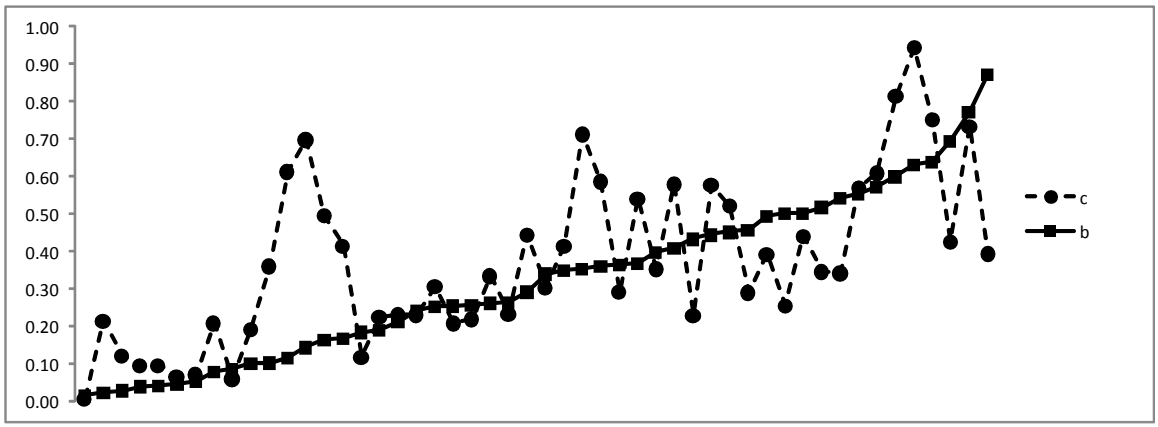

Figure 6: Topic-by-topic comparison of two TREC-8 runs based on average precision scores – taken from Harman's (2002).

The order of topics in the graphs is sorted by the topic effectiveness scores of the *b* run. It can be seen that there is great variation in AP ranging from 0.01 to 0.87. In Figure 5, the topic-by-topic effectiveness of *a* follows a similar pattern and has a similar range of scores. Buckley and Harman (2004) reported on a detailed study of run comparisons and stated that for most runs, the relatively similar performance on topics shown here is typical. Having the worse run, in this case *a*, being the same or a little better for some topics (16 of the 50 topics here) is also common. The absolute difference between *a* and *b* is 4.3% and relative difference is 15.4%. Both the difference in scores and an examination of the graph in Figure 5, would lead most to agree that in this case *b* is better.

Harman (2002) highlighted the comparison shown in Figure 6 where it is arguably harder to determine the better run. Again the topics in the graph are ordered by the AP scores of run *b*. The per-topic scores of *c*, are different. The absolute and relative differences between *b* and *c* are 4.9% and 15.1% respectively, similar to the two runs above, however an examination of the graph might not lead to as unanimous a view on which run is better. While Buckley and Harman's work (2004) showed that most run comparisons are like the case in Figure 5, situations such as those found in Figure 6 are not exceptional. Therefore, it is necessary to examine more than the single value effectiveness measure calculated for a particular run.

Significance tests

A common approach to more fully understanding the comparison of two runs is to use one or more significance tests. These tests estimate the probability *p* of observing the data being examined given that a so-called *null hypothesis* (H_0) is true. In the context of IR experiments, H_0 states that the systems producing the two runs under examination have effectively the same retrieval characteristics and that any difference between the runs occurred by random chance. The convention when using such tests is that if it is found that *p* is below a certain threshold – typically either 0.05 or 0.01 – it is concluded that H_0 is unlikely and consequently should be rejected. Although it is not universally agreed upon, the common interpretation of rejecting H_0 is to conclude that an alternate hypothesis, H_1 is true. This hypothesis states that the two IR systems have different retrieval characteristics, leading the experimenter to conclude that a significant difference has been observed. The exact nature of H_1 depends on whether a *one* or a *two-tailed* test is chosen. This topic is discussed below in Section 0.

The tests are not infallible and can make errors, which have been classified into Type I and Type II errors. Type I errors are false positives: leading the experimenter to incorrectly reject H_0 and conclude that H_1 is true; Type II errors are false negatives: leading the experimenter to incorrectly conclude that they have insufficient evidence to reject the null hypothesis. In IR parlance, Type I measures the precision of the test, Type II measures its recall. Different significance tests tend to produce a different balance between these two errors. For example, the sign test is known for its high number of Type II errors whereas the t-test is known for producing Type I.

The creators of the tests make assumptions on the underlying data being examined and it is important for experimenters to be aware of a test's assumptions before applying it. Tests that have fewer assumptions tend to generate more Type II errors and are said to have less *power*. So called *non-parametric tests*, fit these profile, the best known in IR research are the *Wilcoxon signed ranks test* and the *sign test*. More powerful tests generate fewer Type II errors but make more assumptions about the data being tested. If such tests, known as *parametric tests* are applied to data in violation of such assumptions, Type I errors can result. The best known parametric test used in IR is the *t-test*.

Assuming each run has been conducted on the same topics in a test collection, the significance test is usually operated as a paired test. If the runs being compared use different sets of topics (an uncommon situation in test collection based experimentation) an independent or two-sample version of the tests can be used. Although there is a wide range of tests available to the IR researcher, the three mentioned so far are the most often used. In one of his early books (1968, p.310), Salton discussed use of the sign and the t-test in IR experiments. In the same year, Ide described using the Wilcoxon signed-rank test along with the t-test to examine the significance of retrieval results (1967).

Sanderson and Zobel (2005) compared the properties of the t-test, Wilcoxon and sign test in retrieval experiments. Their conclusions were that use of the t- and the Wilcoxon tests allowed for more accurate prediction over the sign test of which run was better. However, the differences between the t-test and Wilcoxon were small. They also pointed out that even if one observed a significant difference between two runs based on a small number of topics (≤25), one should not be confident that the same observed ordering of runs will be seen in the operational setting.

More recently developed significance tests have also been proposed. Savoy (1997) proposed use of the *bootstrap* test[3] and Smucker et al (2007) explored use of the *randomization* or *permutation* test. They found the t-test, bootstrap and randomization produced similar results.

One or two-tail tests?

So far H_0 has been described, but H_1 has not. There are two types of hypothesis that can be chosen for H_1, which correspond to different types of test: a one- and a two-*tailed* test (also known as a one or two *sided* test). In a two-tailed test, H_1 states that the two systems under examination are not equal, e.g. from the runs in Figure 5, H_1 would state that system *a* does not have the same retrieval characteristics as system *b*. Comparing *a* and *b*, a two tailed t-test of H_0 returns p=0.002; the Wilcoxon signed rank test returns p=0.004; and the sign test p=0.015. Assuming a 5% threshold, regardless of which test was used, the experimenter would reject the null hypothesis and consider the difference between *a* and *b* to be significant.

Since IR experiments are often concerned with determining if a new type of IR system is better than an existing baseline, experimenters sometimes use a form of significance test that focuses only on the question of difference in one direction between two runs: this is the one-tailed test. Here the experimenter predicts before conducting the experiment that one of the systems will be better than the other and sets H_1 to reflect that prediction. Taking this time the comparison from Figure 6, if system *b* is a baseline and system *c* is a new system under test, the experimenter sets H_1 to predict that $c>b$. A one-tailed t-test returns p=0.036; the Wilcoxon p=0.026; and the sign test, p=0.102[4]. Despite the lack of significance in the sign test, most experimenters would consider the improvement of *c* over *b* as significant. The one-tail test is recommended for use in IR experiments by Van Rijsbergen (1979, chap.7) and more recently by Croft et al (Croft et al. 2009), however, it is worth noting that in some areas of experimental science, such as medicine, the one-tailed test is viewed as almost always inappropriate (Altman 1990, p.171).

The one-tail version of a significance test has a p value that is half that of the two-tail version, which makes it a tempting choice for experimenters as its use doubles the chance of finding significance. Note for example that all of the two-tailed tests comparing *b* and *c* would have failed to reject the H_0. However, if using the one-tail, it is important to understand what its use entails. If from Figure 5 an experimenter had incorrectly predicted that system *a* > baseline *b* and chose to use a one-tailed test; upon discovering that a < b, the experimenter would *have* to conclude that they had failed to reject H_0. In other words, the experimenter would be obliged to report that *a* and *b* had the same retrieval characteristics, no matter how much worse *a* was compared to *b*; to many a strange conclusion to draw. The experimenter could of course conduct the one-tailed test in the opposite direction, but this second test could *only* be conducted on a new data set.

Recalling the example in Figure 6, another abuse of significance tests would arise if an experimenter decided to use a two-tailed test, when comparing *c* and *b*, found no significance and so switched to a one tail test in the favourable direction in order to search out significance.

The choice of a one or two-tailed test needs to be made before analyzing the data of an experiment and *not* after. If you are not sure of the direction of difference you wish to test for when comparing two systems, a two tailed test is the appropriate choice. If you are certain that you only wish to test for a

[3] Savoy, cites Efron & Tibshirani (1986) and (1993); Léger et al., (1992) as the originators of the test.

[4] It is also worth noting in the *c* and *b* comparison how the Wilcoxon and t-tests produced p values below the 0.05 threshold but the sign test did not. The former tests are more influenced by the substantial improvements of *c* over *b* in some topics. The sign test ignores the size of a difference; considering only the sign of the difference.

difference in one pre-selected direction, the one tail test should be used. It is important that the experimenter always states which "tailed version" they used when describing their work.

Consider the data in more detail

It is also always worth remembering that although the use of significance tests can help better understand the differences between two runs, the tests are not oracles, they are merely a generic statistical tool constructed for the purpose of estimating the probability p of observing the experimental results if the null hypothesis, H_0, is true. The value of p is calculated on the sample of topics, documents and relevance judgments in the test collection. If that test set is a representative sample of the broader population of queries, documents and judgments made by users in an operational setting, then the conclusions on whether H_0 can be rejected or not should apply to the population. If, however, as is often the case, the sample is not representative, then conclusions drawn may be unreliable. Sanderson and Zobel (2005) showed a number of examples where a range of significance tests produced p values ≤ 0.05 for a pair of runs on one sample test collection, but for the pair using a different sample collection produced p values >0.05.

Even if the sampling of the components of a test collection are accurate, the experimenter compares two runs using the collection and finds $p \leq 0.05$, there are other questions to be asked. Comparing the runs in Figure 6, if *b* was a baseline system already installed at an organization, even though c is significantly better (according to a one-tailed test), the manager of the existing baseline system might argue it is questionable if users would welcome the new system *c* given that it is notably worse than *b* on 10 of the 50 topics (20%). In a different setting, a manager might conclude that *c* is worth installing because it appears to improve substantially on topics that *b* performed very poorly on, but only reduces somewhat *b*'s top performing topics and those reductions would be acceptable to his/her users. Such issues, which could be critical in deciding the value of one system over another are not addressed by significance tests and can only be answered by a more detailed understanding of the uses and users of an IR system.

It is also important to consider the magnitude of difference between two systems: if a significant difference is not substantial enough, it might not be worth bothering about. In an operational setting, as Cooper pointed out (1973), a better system might require more compute resource than the installed baseline system to produce its improved ranking and the benefits of the new system might not outweigh the disadvantages of the additional resource needed. The opposite could be true: experimentation might show a new system to be better than the baseline, though the improvement is *not* statistically significant; however, if the new system uses fewer resources than the baseline, because the experiments showed that it *may* be better than the baseline and it will run on a cheaper computer, the new system could be the better choice.

The question of what constitutes a sufficiently large improvement in retrieval effectiveness continues to be examined in some detail by the IR community, particularly in the context of measuring the impact of improved rankings on user behaviour. Currently, the published results are contradictory, some papers suggest that very small significant differences in retrieval effectiveness have a measurable material effect on users; others report results showing that very large significant differences in effectiveness have no measurable effect on users or on the success of the tasks they use searching to help with. The research is still on going, for now it is an open question of what significant difference is *practically significant*.

While significance tests are without doubt the most popular of statistical data analysis methods, it is worth remembering that many statisticians feel the tests are over used, abused, and of more concern that their use discourages researchers from examining their data in more detail (Gigerenzer 2004). A significance test provides a binary decision on whether there is something of note in the data or not. On finding significance, researchers may not feel it is necessary to examine their data further. Perhaps more of concern are researchers, who failing to find significance, may not look further at their data, which may prevent them from learning what had gone wrong and/or how to fix any problem.

As pointed out by Gigerenzer, a wealth of other statistical analysis methods exists to allow different forms of analysis to be conducted. One popular alternative is the *confidence interval* (CI) which can be used to compute an interval around a value produced by an evaluation measure. This might commonly be displayed in graphs using an error bar. If when comparing two values, the error bars of the values don't overlap, a researcher can state that the difference between the values is of note. In some scientific fields, confidence intervals have replaced significance tests to become the default method for analyzing experimental data. It would appear they have been chosen because their use encourages more analysis of the properties of the data, than significance testing does. Confidence intervals are sometimes used in IR literature; in describing *statMAP*, Carterette et al defined how to compute CIs over that measure (2008). Cormack et al (2006) described how to calculate an interval on MAP.

8 Multilingual Issues

A number of new issues arise when building test collections in a multilingual context. The research field of "Multilingual Information Access" (MLIA) deals with the construction of systems that access information in a multilingual context. Of immediate interest for MLIA researchers and practitioners is the area of "Cross-Language Information Retrieval" (CLIR). CLIR systems allow users to query across language barriers. In the simplest case, bilingual retrieval, the CLIR system returns a monolingual result list in a language different to the source language of the query. In more complex scenarios, CLIR systems handling any number of different languages are proposed, that return multilingual result lists in response to user requests.

This "crossing of a language barrier" in MLIA/CLIR impacts most aspects of test collection use as introduced earlier. Issues arise in collecting the documents, in topic creation, pooling and relevance assessment. These issues have been tackled in the context of the CLEF (Cross-Language Evaluation Forum) campaign.

Collections

A number of tacit assumptions often underlie the motivations to move towards Cross-Language Information Retrieval: there seems to be an (unspoken) consensus that searching across languages is primarily interesting in cases where the user has to expect poor search results when staying within his or her language of choice. Examples include the search for additional information on a foreign news event that is sparingly reported on in the user's home press, or search for information in the context of travel planning. This type of information need was addressed from the beginning in CLEF by using corpora consisting of articles from news papers and news wires, aligned to cover the same time period across the different languages (Martin Braschler & Carol Peters 2004). In their combination, these corpora form a "comparable multilingual corpus", covering topics unevenly across the different languages.

Queries

Topic creation in such a multilingual setting of unevenly distributed information raises its own issues. Topics addressing events mostly covered in a single language often use vocabulary that is difficult to translate directly or invokes subtle cultural differences. Womser-Hacker (2002) gives some examples of such topics used in the early CLEF campaigns: consider the Dutch single word term "Muisarm" (literally "mouse arm"), which addresses a form of repetitive strain injury (RSI) linked to the use of computer mice. The French translation of the topic containing this term used at CLEF reads "ordinateur: souris et tensions musculaires" (literally "computer: mice and muscle strain"), a much more complex paraphrasing of the term in the Dutch original, that may well have implications on how evaluation results are skewed. The experiences made at CLEF suggest that topic creation needs to carefully address these issues. To this end, CLEF has attempted to provide a mix of open-ended and focused topics, covering a range from global, continental, and local news events. As translation issues strongly influence the results of CLIR systems, there is a potential that different linguistic phenomena

may skew evaluation by being "over-represented". CLEF has strived not to tailor topics to specific linguistic phenomena, with the exception of an attempt to limit the number of topics that can be answered mainly based on a search for proper names to roughly 20% of the overall topic set size. Mandl and Womser-Hacker (2003) report evidence that the scoring of retrieval systems at CLEF for exercises using such comparable news documents was not influenced by the linguistic properties present in the topics. However, clearly, this issue needs to be taken into account further when adapting topic creation to new, specific multilingual scenarios.

The topic creation process adopted at CLEF uses the following basic steps:

1. creation of distinct sets of candidate topics by native speakers in each of the languages to be covered.
2. “pilot searching” in the multilingual corpus to obtain a rough estimate of the distribution of relevant items across languages.
3. a “round-table” meeting by the topic creators to select the topic set.
4. direct translation by the topic creators into all other languages wherever possible, if this is not possible, translation via English as an intermediate language.
5. Cross-checks by professional translators.
6. Final check by the creators of the original topics.

Relevance judgements

As with TREC, the process of forming relevance judgments involved a system pooling process, which is similar to the monolingual case. In CLEF, both bilingual CLIR and multilingual CLIR were offered from the beginning, which helped ensure that there was a sufficient number of documents in the individual languages to build a pool. Retrieval quality of the CLIR systems was initially a concern, as retrieval effectiveness of early CLIR systems in TREC was generally only at around 50% of the monolingual case (M. Braschler 2004). By offering monolingual retrieval exercises in CLEF for all target languages, it was ensured that the pool reaches a sufficient quality. Generally speaking, the additional translation step involved in CLIR should help to ensure a good diversity of approaches contributing to the pool, by providing an additional layer of complexity in which systems can diverge. The experiences of CLEF show that pooling works as expected in the multilingual case: Braschler & Peters (2004) give an overview of the pool quality measured for the first three CLEF ad-hoc tracks, where participants had to search document collections containing up to eight different languages. Testing of pool quality followed a method proposed by Zobel (1998): iteratively, every different single participant's contributions to the pool are removed. The scores are then recomputed and compared to the scores obtained by using the full pool. For the multilingual pools used in the three CLEF campaigns that were studied, no differences larger than 5.99% were found, with the mean differences between 0.48% and 1.02%, depending on the year. Such differences are too small to be likely to obscure real differences that would otherwise be statistically significant.

In CLEF, the pool of relevant documents was judged by the same groups that were responsible for topic creation. The more languages that are covered, the more inevitable it is that multiple assessors are used for a single topic. There is consequently a danger for uneven judgments for the same topic, a problem that has not been extensively analyzed yet. The danger can be minimized by supplying the assessors with comprehensive instructions on what constitutes a relevant item for a particular topic. In case of uncertainties, CLEF has ensured that assessors resolve any questions by direct communications.

Measures

When evaluating multilingual or cross-lingual retrieval runs, essentially the same measures can be used as in the monolingual case. However, work by Mandl (2009) indicates that some measures may exhibit different behaviour for multilingual settings when compared to using them for monolingual runs. Specifically, Mandl showed that in the multilingual case, MAP calculated over a set of topics tend to be dominated by the performance that systems obtain on the "easiest" queries. This same

behaviour was not observed in related monolingual experiments. Mandl showed that using the geometric mean (GMAP) mitigates this effect – topic difficulty has much less influence on overall system rankings in this case. These observations can be taken as strong indications on the benefits of using at least both measures when "robustness", i.e. solid retrieval effectiveness even for hard topics, is a concern. In any case, it may be beneficial to consider both measures (MAP and GMAP) when many languages are used for multilingual retrieval: in such cases, the correlation between the two measures is very low (Mandl et al. 2008), in contrast to usually high correlations for monolingual and bilingual experiments.

9 Case studies

This handbook is completed with a description of two case studies of test collection formation. The first describes the methods used to build a classic test collection, the *ad hoc collections* of TREC. The second describes a lightweight evaluation conducted to compare two operational searching systems.

Case study – TREC

As described by Harman (2005), the ad hoc collections were built with the searching task of an information analyst in mind. The analyst was thought of as a person who was given topics to search for on behalf of someone else. The topic given to them was well described and the analyst was expected to locate as much relevant material as possible. The topics for the ad hoc track were created by members of the TREC document assessment team at a rate of fifty per year. The exact procedures, numbers and thresholds for forming the topics varied over the eight years that TREC ad hoc ran, as detailed by Voorhees and Harman (2005, p.28). However certain aspects of the methodology remained constant. As stated in the TREC overview papers (e.g. (D. K. Harman 1993); (E. M. Voorhees & D. K. Harman 1999)), the creators of the topics would create a set of candidate topics. These were then tested by searching on the ad hoc collections to estimate how many relevant documents each topic would return. Topics with too many or too few relevant documents were rejected.

Drawing on lessons learned from earlier test collections, which often specified topics in a single sentence, TREC topics were structured to provide a more detailed statement of the information need that lay behind the query. This was written to help future experimenters understand the topic. The topics were formatted into an XML-like scheme, the format of which was supported by many commonly used IR test beds (e.g. SMART, Lemur, Terrier, Zettair, etc). The structure varied over the years, but its main components were:

- a topic *id*;
- a short *title*, which could be viewed as the type of query that might be submitted to a search engine;
- a *description* of the information need written in no more than one sentence; and
- a *narrative* that provided a more complete description of what documents the searcher would consider as relevant.

The topics, particularly those developed in the early years of TREC were reminiscent of information requests submitted to search intermediaries or librarians.

Obtaining large quantities of text to build a collection involved persuading the copyright owners of a large corpus of material to allow their content to be used. Through connections with news publishers, TREC organizers obtained US and UK newspaper and magazine articles, as well as US government documents. TREC standardized the gathered documents in a similar XML scheme as used in the topics.

Once the documents and topics of a TREC collection were collected and created, the data was sent out to participating groups who were given a limited time to generate and return a series of runs. The qrels were formed from pools, which were the union of the returned runs. Each run contained up to the 1,000 top ranked documents retrieved for each of the TREC topics. The top *n* documents (most often

n=100, or more recently 50) from each run were merged into the pool to be judged. In order to make pool judgment tractable, TREC organizers sometimes had to limit the number of runs that contributed to the pool. In such situations, participating groups nominated a subset of submitted runs they wished to be assessed.

TREC defined two forms of run:

- *automatic* runs, defined as runs where no manual intervention took place between the submission of topics to a group's retrieval system and the outputting of the run.
- *manual* runs, where any amount of human intervention in the generation of search output was allowed. For some manual runs, the list of documents submitted was a concatenation of the best results from multiple queries with documents judged as relevant by the searchers being placed at the top of the run ranking. For details of how individual manual runs were created, see Voorhees and Harman's overview of one of the years of TREC (e.g. 2000). Although such runs appeared to have limited scientific value, TREC organizers encouraged the submission of manual runs as they were found to be rich sources of relevant documents for the pool. Kuriyama et al (2002), showed the importance of manual searching in effective pool formation, when running a TREC-like evaluation exercise, NTCIR.

In order to be seen to be fair to all participants, TREC assessors viewed all top *n* documents in the pool; documents were sorted by docid so that the rank ordering of documents did not impact on the assessment. TREC tried to ensure that the creator of the topic was also the assessor of its qrels. Unlike a number of earlier test collections, which had degrees of relevance, in TREC, assessors made a binary relevance judgment. They were instructed that "*a document is judged relevant if any piece of it is relevant (regardless of how small the piece is in relation to the rest of the document)*"[5], which resulted in a liberal view of what documents were viewed as relevant.

Case study – The National Archives

Introduction – context

The National Archives (TNA) is the UK government's physical and digital repository for all government documents. The archive holds approximately 65 million records; an increasing number of which are digital documents. The archive is spread across sixteen databases. From TNA's web site, there is a so called global search, which allows users to access, through a single search, the records and documents held on nine of the main databases. TNA also made extensive use of sitemap technology[6] to allow large public search engines, such as Google, to index the content of these databases as well.

The TNA's content was therefore searchable by both its internal system (bought from a large well known enterprise search company) and by Google. Employees of TNA were starting to feedback to those in charge of the internal searching system that external search engines, such as Google were producing better quality search results. However, the evidence was at best anecdotal, TNA wished to conduct a study of search quality of the two available approaches: comparing the Internal Search System (ISS) with an external Web Search Engine (WSE).

Staff in the Archive approached researchers at the University of Sheffield to discuss how best to conduct such an evaluation; accuracy, speed, and cost were priorities. Conducting the evaluation as part of a Master student's dissertation was discussed. However, in the end an alternative approach was decided on: making the comparative evaluation of the two search engines part of the coursework for an information retrieval module that was running at the University.

The planning, design, execution, and results of the evaluation are described here.

[5] http://trec.nist.gov/data/reljudge_eng.html (accessed September 2009)
[6] www.sitemaps.org

Methodology

As described above, the three main elements of setting up a retrieval evaluation were considered: collection, queries and qrels.

Collection

In this evaluation, the collection was simply the set of documents and catalogue entries indexed by the two search engines. Both engines indexed the same collection.

Queries

In order to determine the types of queries to be used in the evaluation, a discussion with someone from TNA's search quality team was conducted, during which an initial examination of a series of typical queries was made. From this, it was established that there were two broad classes of queries submitted to the TNA search engines: navigational queries and informational queries. These two query types require different retrieval methodologies in order to locate the most relevant item. Navigational queries are best served by web pages that act as a starting point for the topic expressed in the query. Information queries tend to be better served by documents that describe the query topic in great detail. It was, therefore, decided that a set of queries would be sampled from the logs of the TNA ensuring that an equal number of informational and navigational queries were identified.

A member of the TNA search quality team located in logs of the in-house search engine, the most popular queries as determined by the number of times the query was entered. From the list, a total of 48 queries were created with a 50-50 split between navigational and informational. As with most queries, the searches issued were short. Therefore the TNA team member examined the logs of the retrieved documents clicked on by searchers to try to understand the information need underlying each query. A short description of the information need defining what was considered relevant and what was not was written for each query.

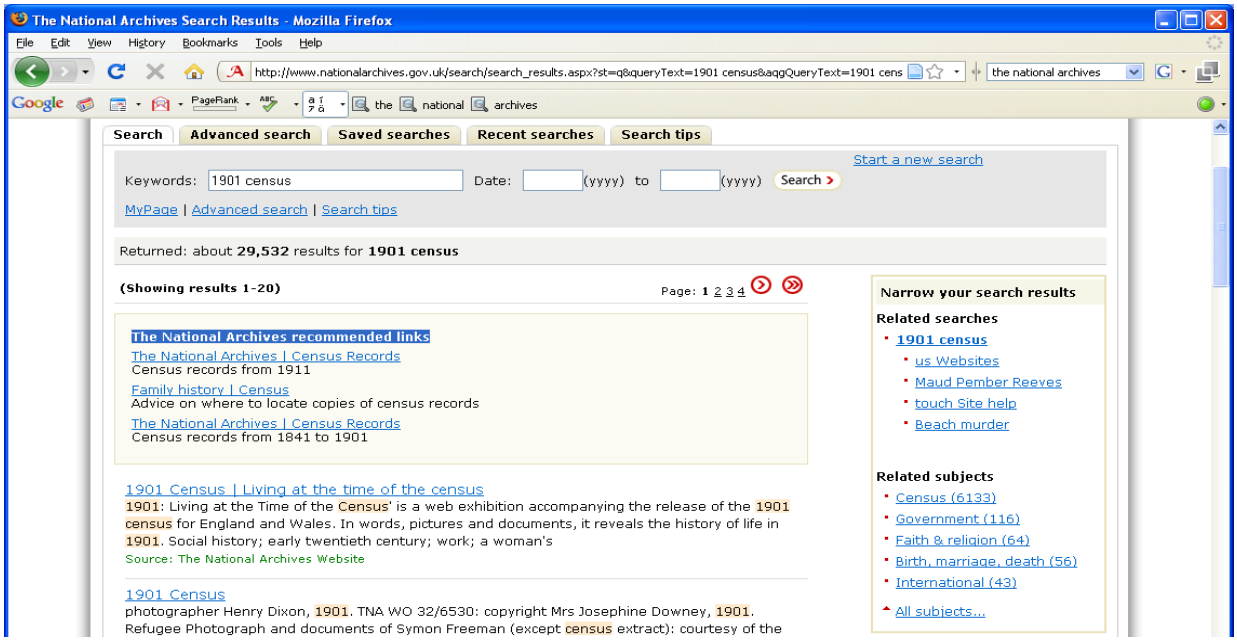

Figure 7: Screenshot of the TNA's in-house search system interface.

Qrels

In the context of this evaluation, two retrieval systems were being compared. The type of searching being conducted by users was judged to be predominantly precision oriented. There was, therefore, no need to compute recall. Simply determining if there was a difference between the two systems in terms of the relevance of top ranked documents would be sufficient.

Because of the relatively general nature of the topics and documents, no special expertise was judged necessary from relevance assessors. In this particular case, a group of 24 University students studying a course on information retrieval were used to assess retrieved documents for each topic. Each student was assigned 4 topics to process. In total 48 topics were assessed, each topic was assessed by two students. Owing to minimal resources being available to run the experiment, no specialised software was used to allow the students to assess documents. Instead, students searched on the two systems for each of their 4 assigned topics.

The instructions given were to type in the query text specified in the topic and assess the relevance of results of the two searching systems, calculating precision measured at rank 10. The assessors were told to use different judgement criteria for navigational and information topics. For this particular collection, some documents can only be accessed after a payment has been made. Based on feedback from the TNA's search quality team, the assessors were told that if search results included links to relevant but "payment only" documents, they should be regarded as relevant.

As the interfaces and number of results returned was different for the two searching systems, students were instructed on which results to assess and which to ignore. For Google, adverts were not assessed for relevance. The in-house system returned 20 results and showed a combination of query specific recommended links along with the output of the main searching system (see Figure 7). Assessors were instructed to assess only the top 10 returned results starting from the recommended links.

The students were given a period of two weeks to computer the precision at 10 for all four topics across both systems. The results were checked by an overseer who removed a number of results from assessors who appeared to have mis-understood the topic.

Results/Discussion

Despite the relatively small scale of the experiment and the low cost distributed approach used to generating assessments, the results were valuable. An overall comparison of the P(10) scores measured across all 48 topics was that Google was significantly better than the in-house system; the significance test used was a 2-tailed paired t-test ($p<0.01$). What was particularly striking, however was when the two topic sets of navigational and informational queries was examined, it was clear that the difference between the two systems was entirely due to differences in the navigational queries. No significant difference was found in the 24 informational queries, however a strong significant difference was found in the navigational queries ($p>0.001$). Examining the 24 navigational topics, Google was better for 20 of the topics, the in-house system was better on 3 and for 1 topic no difference was observed.

	In house	**Google**	
Overall P(10)	0.39	0.51	**
Navigational	0.34	0.54	***
Informational	0.44	0.49	

As a comparison, the experiment was re-run with a second set of students, with some small modifications to the methodology. The same result occurred: Google produced better retrieval than the in-house system and with almost all the difference being measured in the navigational queries.

The impact of the results of this experiment has been left to TNA to consider.

Conclusion

This simple case study illustrates how with relatively minimal effort it is possible to produce an effective evaluation.

10 References

Agirre, E. et al., 2008. CLEF 2008: Ad Hoc Track Overview. In *Working Notes for the CLEF 2008 Workshop*. Aarhus, Denmark.

Al-Maskari, A., Sanderson, M. & Clough, P., 2007. The relationship between IR effectiveness measures and user satisfaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press New York, NY, USA, pp. 773-774.

Altman, D.G., 1990. *Practical Statistics for Medical Research* 1st ed., Chapman & Hall/CRC.

Bailey, P. et al., 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, pp. 667-674.

Borlund, P., 2003. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10), 913-925.

Borlund, P. & Ingwersen, P., 1997. The development of a method for the evaluation of interactive information retrieval systems. *JOURNAL OF DOCUMENTATION*, 53, 225-250.

Braschler, M., 2004. *Robust Multilingual Information Retrieval*. Ph. D. Dissertation. University of Neuchatel, Switzerland.

Braschler, M. & Peters, C., 2004. Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval*, 7(1-2), 7-31.

Broder, A., 2002. A taxonomy of web search. *SIGIR Forum*, 36(2), 3-10.

Burges, C. et al., 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning,* Bonn, Germany, pp. 89-96.

Carterette, B. & Jones, R., 2007. Evaluating Search Engines by Modeling the Relationship Between Relevance and Clicks. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems (NIPS)*.

Carterette, B. et al., 2008. Evaluation over thousands of queries. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, pp. 651-658.

Cleverdon, C. & Keen, M., 1966. Factors Affecting the Performance of Indexing Systems, Vol 2. *ASLIB, Cranfield Research Project. Bedford, UK: C. Cleverdon*, 37-59.

Cooper, W.S., 1973. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2).

Cormack, G.V. & Lynam, T.R., 2006. Statistical precision of information retrieval evaluation. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, pp. 533-540.

Croft, B., Metzler, D. & Strohman, T., 2009. *Search Engines: Information Retrieval in Practice* 1st ed., Addison Wesley.

Dussin, M. & Ferro, N., 2009. Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns. In *Research and Advanced Technology for Digital Libraries: 13th European Conference. ECDL 2009, Corfu, Greece, September 27-October 2, 2009, Proceedings*. Springer, p. 63.

Efron, B. & Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1), 54-77.

Efron, B. & Tibshirani, R.J., 1993. An Introduction to the Bootstrap. *Monographs on Statistics and Applied Probability*, 57, 1–177.

Fairthorne, R.A., 1963. Implications of test procedures. In *Information Retrieval in Action*. Cleveland, Ohio, USA: Western Reserve UP, pp. 109-113.

Gigerenzer, G., 2004. Mindless statistics. *Journal of Socio-Economics*, 33(5), 587-606.

Harman, D.K., 1993. Overview of the Second Text Retrieval Conference (TREC-2). In NIST Special Publication. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology.

Harman, D.K., 2002. Some Interesting Unsolved Problems in Information Retrieval. Available at: http://www.clsp.jhu.edu/ws02/preworkshop/lecture_harman.shtml [Accessed November 2, 2008].

Harman, D.K. & Buckley, C., 2004. The NRRC reliable information access (RIA) workshop. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, pp. 528-529.

Harman, D., 1992. Evaluation issues in information retrieval. *Information Processing & Management*, 28(4), 439-440.

Harman, D.K., 2005. the TREC Ad Hoc Experiments. In *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*. MIT Press, pp. 79-98.

Hawking, D., 2001. Overview of the TREC-9 Web Track. In NIST Special Publication. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, pp. 87-102.

Hawking, D., Bailey, P. & Craswell, N., 2000. ACSys TREC-8 Experiments. In NIST Special Publication. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, pp. 307-316.

Ide, E., 1967. *Evaluation Parameters*,

Ingwersen, P. & Järvelin, K., 2005. *The Turn: Integration of Information Seeking and Retrieval in Context*, Springer.

Jansen, B.J. & Spink, A., 2006. How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1), 248-263.

Järvelin, K. & Kekäläinen, J., 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4), 422-446.

Järvelin, K. & Kekäläinen, J., 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, pp. 41-48.

Joachims, T. et al., 2005. Accurately interpreting click through data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, pp. 154-161.

Kando, N., 2003. Evaluation of Information Access Technologies at the NTCIR Workshop. In *Comparative Evaluation of Multilingual Information Access Systems. 4th Workshop of the Cross-Language Evaluation Forum, CLEF*. Springer, pp. 29-43.

Kando, N. et al., 1999. Overview of IR tasks at the first NTCIR workshop. In *Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition*. pp. 11–44.

Katter, R.V., 1968. The influence of scale form on relevance judgments. *Information Storage and Retrieval*, 4(1), 1-11.

Kuriyama, K. et al., 2002. Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the First NTCIR Workshop. *Information Retrieval*, 5(1), 41-59.

Léger, C., Romano, J.P. & Politis, D.N., 1992. Bootstrap technology and applications. *Technometrics*, 34(4), 378-398.

Lesk, M.E. & Salton, G., 1968. Relevance assessments and retrieval system evaluation*1. *Information Storage and Retrieval*, 4(4), 343-359.

Mandl, T., 2009. Easy Tasks Dominate Information Retrieval Evaluation Results. In pp. 107-116.

Mandl, T. & Womser-Hacker, C., 2003. Linguistic and statistical analysis of the CLEF topics. *Lecture notes in computer science*, 505-511.

Mandl, T. et al., 2008. How robust are multilingual information retrieval systems? In *Proceedings of the 2008 ACM symposium on Applied computing*. ACM New York, NY, USA, pp. 1132-1136.

van Rijsbergen, C.J., 1979. *Information Retrieval* 2nd ed., Butterworth-Heinemann Ltd.

Robertson, S.E., 2008. On the history of evaluation in IR. *Journal of Information Science*, 34(4), 439-456.

Rose, D.E. & Levinson, D., 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*. ACM Press New York, NY, USA, pp. 13-19.

Salton, G., 1968. *Automatic Information Organization and Retrieval.*, McGraw Hill Text.

Sanderson, M. & Zobel, J., 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*. ACM New York, NY, USA, pp. 162-169.

Saracevic, T., 1995. Evaluation of evaluation in information retrieval. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA, pp. 138-146.

Savoy, J., 1997. Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management*, 33(4), 495-512.

Smucker, M.D., Allan, J. & Carterette, B., 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM New York, NY, USA, pp. 623-632.

Spärck Jones, K., 1981. *Information Retrieval Experiment*, Butterworth-Heinemann Ltd.

Thorne, R., 1955. The efficiency of subject catalogues and the cost of information searches. *Journal of documentation*, 11, 130-148.

Voorhees, E.M., 1993. On expanding query vectors with lexically related words. In NIST Special Publication 500-215. Department of Commerce, National Institute of Standards and Technology, pp. 223-231.

Voorhees, E.M., 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press New York, NY, USA, pp. 315-323.

Voorhees, E.M. & Harman, D.K., 1999. Overview of the Eighth Text REtrieval Conference (TREC-8). In NIST Special Publication. Gaithersburg, MD, USA: Department of Commerce, National Institute of Standards and Technology, pp. 1-24.

Voorhees, E.M. & Harman, D.K., 2005. *TREC: Experiment and Evaluation in Information Retrieval* illustrated edition., The MIT Press.

White, R.W. & Morris, D., 2007. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press New York, NY, USA, pp. 255-262.

Womser-Hacker, C., 2002. Multilingual topic generation within the CLEF 2001 experiments. *Lecture notes in computer science*, 389-393.

Zobel, J., 1998. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press New York, NY, USA, pp. 307-314.

TrebleCLEF Consortium

Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy

Università degli Studi di Padova, Italy

The University of Sheffield, UK

UNED, Madrid, Spain

Zürcher Hochschule für Angewandte Wissenschaften, Winterthur, Switzerland

Center for the Evaluation of Language and Communication Technologies, Trento, Italy

Evaluations and Language resources Distribution Agency, Paris, France