A Framework for Designing Retrieval Effectiveness Studies of Library Information Systems Using Human Relevance Assessments

Christiane Behnert, Dirk Lewandowski
Hamburg University of Applied Sciences

christiane.behnert/dirk.lewandowski@haw-hamburg.de

This is a preprint of a paper to be published in The Journal of Documentation.
Please cite as:
Christiane Behnert, Dirk Lewandowski, (2017) "A framework for designing retrieval effectiveness studies of library information systems using human relevance assessments", Journal of Documentation, Vol. 73 Issue: 3, pp.-, doi: 10.1108/JD-08-2016-0099

Abstract:

Purpose
- This paper demonstrates how to apply traditional information retrieval evaluation methods based on standards from the Text REtrieval Conference (TREC) and web search evaluation to all types of modern library information systems including online public access catalogues, discovery systems, and digital libraries that provide web search features to gather information from heterogeneous sources.

Design/methodology/approach
- We apply conventional procedures from information retrieval evaluation to the library information system context considering the specific characteristics of modern library materials.

Findings
- We introduce a framework consisting of five parts: (1) search queries, (2) search results, (3) assessors, (4) testing, and (5) data analysis. We show how to deal with comparability problems resulting from diverse document types, e.g., electronic articles vs. printed monographs and what issues need to be considered for retrieval tests in the library context.

Practical implications
- The framework can be used as a guideline for conducting retrieval effectiveness studies in the library context.

Originality/value
- Although a considerable amount of research has been done on information retrieval evaluation, and standards for conducting retrieval effectiveness studies do exist, to our knowledge this is the first attempt to provide a systematic framework for evaluating the retrieval effectiveness of twenty-first-century library information systems. We demonstrate which issues must be considered and what decisions must be made by researchers prior to a retrieval test.

Introduction

When it comes to information seeking, users expect modern library information systems (LIS) "to look and function more like search engines" (Connaway & Dickey, 2010, p. 5). Typical user information behaviour is characterized by a strong preference for the first results in the ranked search results list, short queries of only two to three terms, and users relying on default search settings. This can be observed in both web search (Barry & Lardner, 2011; Jansen & Spink, 2006; Pan et al., 2007; Y. Zhang, Jansen, & Spink, 2009) and library search (Antelman, Lynema, & Pace, 2006; Asher, Duke, & Wilson, 2013; Hennies & Dressler, 2006).

When libraries began integrating search engine technology into their catalogues, they moved away from the Boolean searching paradigm ("exact match" approach) towards ranking search results by relevance ("best match" approach) and a combined search interface allowing immediate access to multiple, heterogeneous information sources (Lewandowski, 2010). These so-called discovery systems are designed with the user in mind. They also seek to provide Google-like advanced search and retrieval functionality that goes beyond metasearch systems that were introduced as a first step to enable unified search on multiple sources, but "do not cover the whole spectrum of library resources" (Sadeh, 2015, p. 214). For example, discovery systems provide simple keyword search and ranked results lists based on relevance criteria as well as user-friendly search features like auto-completion and spell-checking for search queries (Chickering & Yang, 2014). However, whether these features contribute successfully to satisfying the information needs of users in a library context has not yet been extensively researched. Most of the studies that investigate discovery systems focus on the process of evaluating and selecting discovery software, determining whether certain criteria are fulfilled, or which features are available in the evaluated tool (Moore & Greene, 2012). This common procedure was also noted by Hofmann & Yang (2012) who studied 260 academic library catalogues in the United States and Canada regarding the implementation of discovery layers or features. They found that the number of libraries offering discovery tools had increased within two years from 16% to 29% (Hofmann & Yang, 2012).

Despite the fact that Discovery Systems are being implemented by libraries, one must ask how well they perform in terms of search and retrieval. These issues have been evaluated to a lesser extent. Studies on the retrieval effectiveness of a particular information system typically involve assessing the relevance of documents for a search query or information need. Relevance assessments are commonly used for information retrieval (IR) evaluation, particularly within TREC (*Text REtrieval Conference*), whereas relevance as a "fundamental concern" for OPAC evaluation had been ascertained decades ago (O'Brien, 1990). Although there are some studies on testing the retrieval effectiveness of OPACs or discovery systems, there are no consistent criteria for evaluation, as they are met within TREC.

In this paper, we introduce a framework for designing retrieval effectiveness studies of modern library information systems[1] that is based on the standard methods for IR evaluation. The framework

[1] In this paper, we use the term *library information system* for all kinds of library catalogues, discovery systems or digital libraries, as there is no clear and unequivocal internationally standardized definition of the term "digital library." Although it is often referred to as a tool to discover digitized cultural heritage, at the same time many library websites that include search entry fields refer to themselves as digital libraries. Since these systems all have implemented search engine technology and provide access to library materials, the term library information system can be used as a general term.

specifically focuses on the different data types from multiple, heterogeneous sources which need to be integrated into a single results list, ranging from short surrogates to full texts of monographs.

The rest of this paper is structured as follows. First, we provide an overview of conventional procedures and their application to web search followed by evaluation studies of library information systems. After describing our methods, we present the five-part framework and discuss decisions that need to be made with regard to (1) search queries, (2) search results, (3) assessors, (4) testing procedures and (5) data analysis. We conclude with practical implications and acknowledge the limitations of the framework.

Literature review

In this section, we review the literature on standard methods of IR evaluation and their application to web search evaluation; on approaches taken to evaluate library information systems; and on empirical findings from studies concerning these systems.

1. TREC and other standards

In the late 1950s, Cleverdon and his colleagues laid the foundation for systematic IR evaluation based on a formal framework (C.W. Cleverdon, 1960). They conducted a pioneering study on "factors determining the performance of indexing systems" in an experimental environment (Cyril W. Cleverdon & Keen, 1966; Cyril W. Cleverdon, Mills, & Keen, 1966a, 1966b). In what became known as Cranfield experiments, an inverse relationship was observed between recall and precision, which remain the standard concepts behind performance measurement for IR evaluation today. The retrieval tests within the Text Retrieval Conference (TREC)[2] started in 1992 and followed the Cranfield paradigm, i.e., using test collections including a set of documents and search queries to evaluate different ranking methods by comparing the results produced with (expert) relevance assessments (Baeza-Yates & Ribeiro-Neto, 2011, p. 134).

An alternative to Cranfield's and TREC's standard system-oriented approach in IR evaluation is the user-oriented approach, which regards the users themselves as best able to judge the relevance of results to their own search query or information need. With respect to other evaluation methods, a variety of TREC Tracks has been introduced in order to investigate new aspects, for example, the Web Track, Question Answering Track, and the Interactive Track (Dumais & Belkin, 2005)[3]. Although the Interactive Tracks applied methods to accommodate user interaction, they were within the general framework and experimental design of TREC. An attempt to overcome the methodological limitations was made by Borlund (2003b) who proposed a framework that combines system-oriented with user-oriented elements: The framework addresses realistic search behaviour (1) by confronting the user with a simulated work task situation, and (2) by allowing the user to judge relevance on a non-binary scale, whereas variables are being kept under control through an experimental setting.

With regard to the IR evaluation process, Tague-Sutcliffe (1992) follows a holistic approach to providing a guide to conduct IR experiments while emphasizing the importance of skills and knowledge in research design, statistical analysis, project management, and human factors in

[2] http://trec.nist.gov/overview.html

[3] A useful overview on the TREC evaluations and other evaluation campaigns such as the Cross-Language Evaluation Forum (CLEF) or INitiative for the Evaluation of XML retrieval (INEX) is provided by D. Harman (2011).

research (Tague-Sutcliffe, 1992, p. 468). She describes ten decisions that need to be made by the investigator based on the classic “Cranfield paradigm” as summarized in the following list (Tague-Sutcliffe, 1992, pp. 487–489):

1. *Need for testing.*
2. *Type of test.*
3. *Definition of variables.*
4. *Database development.*
5. *Finding queries.*
6. *Retrieval software.*
7. *Experimental design.*
8. *Data collection.*
9. *Data analysis.*
10. *Present results.*

These guidelines can be seen as a checklist that provides a useful aid for designing and conducting retrieval effectiveness tests.

2. *Application to web search engines*

Since the traditional Cranfield paradigm cannot keep pace with the high volume of information accessible via web search engines and the preferences of the modern user regarding results presentation, standard procedures for retrieval effectiveness studies reached their limits some time ago.

An early example in the context of web search engine evaluation focusing on user interaction is the Spink (2002) study. Spink followed a user-centric approach, letting participants complete pre- and post-questionnaires to "capture the state of each user in a number of areas before and after their [web search engine] interaction" (Spink, 2002, p. 407) in addition to relevance judgments. The disadvantage of this alternative to the Cranfield paradigm, however, is the controlled situation which, although it allows us to understand user interactions during the information seeking process, also impedes replicating the study in the sense of statistical analysis (Carterette, Kanoulas, & Yilmaz, 2012). It is important to emphasize that despite the use of standard evaluation methods, standard measures of precision and recall are not applicable when it comes to user-based evaluation. For example, with respect to a search session involving reformulating the initial search query and therefore interacting with the system, session-based measures need to be applied.

Regarding the fact that in web search, full recall is not feasible, a large-scale set of queries and documents should be used in web search engine retrieval tests. Thus, in the TREC Web Tracks (see, for example, Collins-Thompson, Macdonald, Bennett, Diaz, & Voorhees, 2015), methods other than the standard TREC model have been applied for a variety of reasons including document types, interlinking between single web pages, the enormous number of search queries submitted, the predominantly short query lengths, and the types of searches conducted on the web (Hawking & Craswell, 2005, p. 200).

In other retrieval effectiveness studies besides TREC, some alterations to the standard methods have been proposed with respect to the web search context: Gordon & Pathak (1999) proposed seven

features for evaluation to ensure highly accurate and informative results as follows (Gordon & Pathak, 1999, pp. 146–147):

1. Tasks should be based on genuine information needs.
2. The description of the information need should contain as much context information as possible.
3. The number of search tasks being conducted should be sufficiently large.
4. The systems to be tested should include most major search engines.
5. When comparing different search engines, their special features should be exploited.
6. Relevance judgments should be made by the actual user, not by the investigators.
7. The experiment should be conducted properly by following an appropriate design and using statistical analysis with accepted IR measurements.

However, when they reviewed search engine retrieval effectiveness studies published between 1995 and 1997, they found that none of them met these requirements. Studies on the retrieval effectiveness of web search engines are generally based on standard methods following the Cranfield paradigm or TREC (Tague-Sutcliffe, 1992) in combination with adjusted methods with respect to web search (see, for example, Gordon & Pathak, 1999; Griesbaum, 2004; Hawking et al., 2001). Lewandowski (2012) introduced a framework for evaluating the retrieval effectiveness of search engines which is also based on the above methods. It includes the following five parts: “queries selection, results collection, results weighting, results judgment and data analysis” (Lewandowski, 2012, p. 463). With respect to this framework, an overview of eighteen search engine retrieval effectiveness studies is provided by Lewandowski (2015) showing, however, that web search engine evaluations are usually conducted as small-scale studies, i.e., the number of queries is relatively low. In his study, by contrast, 1,000 queries were representatively selected from transaction logs of a German search portal in order to compare Google and Bing using human relevance judgments. Thanks to the crowdsourcing approach used, all 1,000 search tasks had been judged within seven hours (Lewandowski, 2015, p. 1769). Thus, the above-mentioned issue of large-scale results in web search is addressed by the large number of queries used for relevance judgment, and the efficient tools used to conduct the study.

3. Evaluation of library information systems

Since OPACs were first employed in libraries in the mid-1970s, they have been studied with one major objective: to allow the systems to be more effective. Large & Beheshti (1997) reviewed OPAC studies published between 1990 and mid-1996 concerning different methods used for OPAC research including interviews and observations of users conducting search tasks, think-aloud techniques, and transaction log analysis. They noted that relevance judgments present “a particularly ubiquitous problem in IR research of any kind” (Large & Beheshti, 1997, p. 114) and describe the fundamental problems of judging relevance, e.g., whether relevance can be assessed by the user based on the surrogate at all or whether it can only be assessed after examining the complete document. Further, they point out that some variables can cause comparability problems when they are not considered carefully, e.g., users (age, gender, knowledge, etc.), search tasks (different search types demand different methods and metrics), the OPAC or system (features, interfaces), the kind of library, and the type of data collection (researchers collect their data in different ways) (Large & Beheshti, 1997, pp. 115–116).

Following a methodological framework involving the application of standards can help to prevent such comparability problems. However, to date, there is no evaluation framework for retrieval studies of library information systems, i.e. OPACs, their advanced catalogue successors discovery systems, or digital libraries based on the evaluation standards described above. Although there are a few frameworks for evaluating library systems in general, they mainly focus on interfaces and user levels regarding digital libraries (Y. Zhang, 2010), or on the user demand regarding usability and the availability of certain features of discovery systems (Moore & Greene, 2012).

Existing frameworks

Blandford et al. (2008) introduced the *PRET A Rapporter* approach to evaluating digital libraries. This framework involves six stages for designing an evaluation study which the authors then used to analyze case studies investigating the usability and user experience of working with digital libraries (Blandford et al., 2008, p. 9):

1. *The purpose of evaluation: what are the goals of the study, or the detailed questions to be answered in the study?*
2. *Resources and constraints: what resources are available for conducting the study and, conversely, what constraints must the study work within?*
3. *Ethics: what ethical considerations need to be addressed?*
4. *Techniques for gathering data must be identified.*
5. *Analysis techniques must be selected.*
6. *Reporting of findings is (usually) the final step.*

They conclude that, when compared with the evaluation methods by Tague-Sutcliffe (1992) and Borlund (2003), their framework presents questions at an abstract level in order to take into account a “broader work setting” regarding digital libraries. Although this framework involves useful and general steps, concrete details for the purpose of conducting retrieval effectiveness studies are missing.

A recent work by Deodato (2015) includes a step-by-step guide for evaluating and selecting web-scale discovery services, i.e., in contrast to federated search tools that search multiple indexes, web-scale means performing the search in a single, combined index. One of the test trials to evaluate how a certain system works with the library’s own data included a test on relevance ranking as an important feature: 18 subject specialists constructed a total of 26 search tasks related to their area of expertise, performed the searches and assessed the relevance of each of the top ten results themselves. “Results were recorded in the exact order retrieved and ranked on a scale of 0–3 (0 = not relevant, 1 = somewhat relevant, 2 = relevant, 3 = very relevant)” (Deodato, 2015, p. 33). With this procedure, an order effect would not have been eliminated, which questions the validity of the experimental setting. Although the overall evaluation plan in order to select a discovery system considers all the important requirements, it cannot provide the methodological steps and details to design a library information system’s retrieval effectiveness study with respect to the standard procedures. However, some empirical studies on OPACs or discovery systems have been conducted, and an overview is provided below.

Empirical studies

"The most common forms of evaluation of OPACs have been (1) surveys, (2) comparative studies, and (3) transaction log analysis" (O'Brien, 1990, p. 269). One might assume this quotation would now be outdated, but it actually does not appear to be.

Studies conducted to evaluate OPACs or discovery tools often involve *observing users in laboratory situations*. This common method of user research has, for example, been applied by Asher, Duke, & Wilson (2013). They observed students completing research tasks with discovery systems and Google Scholar and interviewed the students' on their individual search processes which provided quantitative and qualitative information on the students' information seeking and selection behaviour. The results showed that students rely heavily on default search settings and the system's ability to present the best, i.e., most relevant, search results within the first results page.

Gathering data for evaluation studies through user observation is not possible without accepting a certain bias: as users are put in an unnatural laboratory situation, the investigators need to rely on the participants' statements to be as genuine as possible. To gain knowledge about actual user behaviour in a non-artificial environment instead, transaction log analysis (or log file analysis) is frequently used.

Log file analysis

Early log file analysis studies with LIS have been reviewed by Peters (1993). Studies involving log files of modern library information systems show that user search behaviour is highly influenced by web search.

Lau & Goh (2006) extracted 641,991 queries from the OPAC of Nanyang Technological University (NTU) in Singapore. The results show that search queries most frequently consist of two terms, which is consistent with other studies. They also observed an increase in search failure correlating with increasing query length. In OPACs, not only query length, but also browsing behaviour is comparable to web search (Hennies & Dressler, 2006).

A transaction log analysis with discovery tools was carried out by Niu, Zhang, & Chen (2014). They studied search activities on two discovery systems (VuFind and Primo) at an academic library with a total of 52,601 log entries of 18,264 search sessions. Results implied that users invest minimum search effort, which is analogous to search behaviour in web search engines like Google or Bing. Since transaction log analysis does not provide information on the user context, they combined this method with a user study in a usability laboratory, i.e., they tested users' search task performance while applying the think-aloud protocol method.

Applying combined methods is not unusual in IR studies. Analyzing the transaction log files of any kind of IR system provides genuine data on user behaviour and gives details on what, when and how a search has been performed. Nonetheless, it cannot answer why some results were selected by the information seeker, and to what degree the selected information was able to satisfy the seeker's information need. While reasons behind the choice of selection may not be fully understood, knowledge about how well a system retrieves relevant search results for a given query can be obtained by conducting retrieval effectiveness studies.

Retrieval effectiveness studies

It seems difficult to believe that the first retrieval effectiveness study using data specific to library catalogues was not published until 1990. Dillon & Wenzel (1990) analyzed the "retrieval effectiveness of enhanced bibliographic records", i.e., the influence of abstracts and tables of contents on the retrieval ability of the system. Unfortunately for the purpose of this literature review, their paper does not cite any references. Although this omission is not in compliance with rules of good academic practice, the method used does include common IR evaluation standards. The researchers allowed one juror, a staff member with expertise in the particular subject area, to judge the relevance of the top 50 results to 20 queries in a random order by using a 3-point relevance scale: 0 (not relevant), 1 (possibly relevant), 2 (highly relevant). The bibliographic records were taken from an Online Database (Wiley Catalog Online file on DIALOG).

The first experimental study involving web OPACs was conducted by Hildreth (2001), following a user-oriented approach. The experimental design is not in compliance with the IR evaluation standards described above because it was not the system's ability to produce relevant results that was researched, but rather users' search performance and level of satisfaction with the system's interface. Results of the experiment indicate that the ease of use of web-based OPACs with graphical user interfaces affects user assessments of search results, i.e., users are often satisfied with search results even if the results are rather poor.

Since academic search engines, i.e., special search engines with indexed resources of academic publications, are often acknowledged as competitors of academic LIS, an interesting research question is to compare discovery systems with Google Scholar. Zhang (2013) compared the search performance of Google Scholar with Primo using an experimental design: twelve participants rated the relevance of results for four search tasks on a scale from 1 – 7, as well as the usability of the systems. It seems, however, that the source of the results was not anonymized, so the participants knew the systems being evaluated which, in turn, may have influenced ratings. Ciccone & Vickery (2015) also evaluated the search performance of Google Scholar in comparison to the discovery systems Summon and EBSCO Discovery Service. They used a sample of 225 search queries extracted from log files, but, again, the sources had not been made anonymous.

To sum up, only very few retrieval effectiveness tests of library information systems have been performed and, of this limited number, only a few studies use genuine search queries; a sample size we can assume is large enough; and relevance assessments by human jurors other than the investigators themselves. We note that (1) there is a lack of consistent criteria for conducting empirical retrieval studies in LIS, and (2) there are deficiencies in describing the methods used in the published research. Describing the methods and research design in detail is indispensable for transparency, comparability, and reproducibility. We hope that our framework may provide helpful advice to address these issues.

Methods

We developed a framework for designing and conducting retrieval effectiveness studies with LIS specific data, based on the conventional procedure for retrieval effectiveness tests following the Tague-Sutcliffe (1992) approach. Although this approach is system-oriented, adopting it is a first step towards establishing retrieval effectiveness studies in the LIS context. Our framework is partly overlapping with the comparable framework for web search engine retrieval tests by Lewandowski (2012), as we also consider the issues that derive from web search engines. This is due to the fact

that library information systems nowadays deal with large data sets from multiple, heterogeneous sources (so-called web-scale discovery systems) or have implemented search engine technology, and moreover, the similar user behaviour in search.

A framework for designing retrieval effectiveness studies of library information systems

When evaluating library information systems, specific characteristics of library collections should be taken into account. Library materials nowadays not only contain data from heterogeneous sources including surrogates, they also link to full texts available via open access or via subscription content as well. Since discovery systems are based on search engine technology and library materials also involve typical "web content" to a large extent, e.g., electronic articles, videos, and audio files, the logical consequence would be to apply standard IR evaluation methods that have already been used in the context of web search engines. IR evaluation methods must nonetheless also take into account traditional library holdings such as printed materials. Metadata (surrogates) and full-texts, both print and digital, provide different levels of information as the basis for assessing the relevance of a document to a search query and/or information need. These different types of data generate some comparability issues that need to be considered and decisions that have to be made before the beginning of the evaluation. These decisions involve five issues that are reflected by the five parts of the framework: At first, we need *search queries* to produce *search results* that can be assessed by *assessors* during the *testing* phase in order to perform the *data analysis*. Within these parts or steps, there are several issues to be taken into account and questions to be answered, such as "*Who* should be assessing *what* in *which way*?" In the next sections, we answer these questions and show examples in the context of library materials, both electronic and printed. Since individual questions arise across all five parts, the subtopics of the framework have not been presented in a strict chronological order. Figure 1 depicts the overall framework.

Figure 1: A framework for designing retrieval effectiveness studies of library information systems

1. *Search queries*

With search queries, there are four main issues: How should they be presented to the assessor? What kind of queries should be used? Where do they come from? How many queries should be used?

Query types

The different types of web search queries (Broder, 2002) can be equated to the library context (Lewandowski, 2010) and serve different kinds of underlying information needs (Frants, Shapiro, & Voiskunskii, 1997), as shown in Table 1. Informational or topic search queries usually require a set of results to satisfy the user's information need, whereas navigational or known-item queries only demand the one correct result at the top position of the search results list. With respect to the particular research question, the choice of whether to test with topic or known-item search queries influences the choice of metrics as well as the number of search results (see section 2) and the number and selection of assessors (see section 3). This means that one should already decide which metrics to use at this early stage in the research design, even though data analysis is the final part of the experiment.

Table 1: Query types in web search and in the library context

Query presentation

Judging the relevance of a document with respect to a particular search query is not without difficulty because a query represents the underlying information need in a limited way, i.e., without context information or the user's background knowledge. It has been known for years that particularly with regard to topic searches, people do face problems when articulating their information needs, and they have difficulties formulating precise queries. Keeping this in mind, it is plausible to provide additional information on the (topic search) query that produced the results to be assessed. A description of the underlying (problem-oriented) information need is also part of the query representation in TREC, where a narrative section that states "what makes a document relevant" (D. K. Harman, 2005, p. 23) is added as well. Since relevance is a concept intuitively understood by everyone (Saracevic, 1996), including a narrative section seems rather optional. It is assumed the assessors will understand the concept of the desired document based on the query and description, for example:

> Query: "corporate happiness"
> Description: "What does corporate happiness mean and how can it be achieved? Is there a measure for corporate happiness?"

Sources

Genuine queries for testing can be obtained using different approaches including: (1) directly entered by users themselves, i.e., queries extracted from transaction logs or by observing users as they interact with the search system, (2) memorized by users, i.e., reported in interviews or questionnaires, or (3) selected by field experts. Although it may seem an obvious choice to give log file analysis preference due to the large amount of queries available, asking users in person has a major advantage: they can explain their individual information need directly, which is the most realistic description possible. However, in practice, extracting queries from transaction logs would be less time-consuming and furthermore, more effective, considering the number of queries needed. Drawbacks to using queries from log files are, however, that (1) log files are only available for live systems, and (2) researchers must find ways to enrich the queries with information needs descriptions.

With the decision to use log files, information needs must be described by a person other than the actual user. One preferred method accomplishes this by letting several participants formulate the possible underlying information need. This approach has, for example, been followed by Huffman & Hochster (2007), who asked multiple jurors for information need statements and selected the one statement described by the majority of jurors.

Number of queries

In previous studies on web search engine retrieval effectiveness, the number of queries used for relevance assessment tasks usually ranges from 5 to 100 queries (Lewandowski, 2015, p. 1765), with a median of 45.5. In early TREC tracks, 50 topics were generated to formulate search queries (D. K. Harman, 2005, p. 39), while a similar number was applied to web tasks, whereas the track involving

finding a name or homepage (navigational or known-item queries, respectively) used up to 150 topics (Hawking & Craswell, 2005, p. 203). However, Voorhees & Buckley (2002) found that error rates based on the number of queries (topics) used in several TREC evaluation runs and the observed scores were larger than expected. They conclude that researchers should do a deeper analysis on the compared retrieval systems' differences before drawing valid conclusions solely based on retrieval scores. They state that "retrieval effectiveness depends on which question is asked, which retrieval mechanism is used, and how the retrieval mechanism deals with the question type" (Voorhees & Buckley, 2002, p. 316).In practice, the number of queries is often a question of feasibility because that choice influences the number of assessors needed as well as the number of results to be judged per assessor and within one task. The possibility to give the assessor a form of incentive, e.g., a voucher or cash, can be decisive for recruitment. Since the most likely participants in this kind of experiment are students, we can assume that the greater the financial incentives available to offer, the more tasks can be processed and therefore, the more queries can be considered.

2. Search results

In order to assess search results, one must consider which elements to display, e.g., in a screenshot, because not every detail displayed in the original record is necessary or useful to the assessors. What elements can be assessed and how can they be presented to the subject? What is an appropriate number of results for relevance assessment and what decisions need to be made regarding the actual assessment tasks?

Results presentation

When it comes to decisions on how to present the results, one must weigh comparability of results against a real-life situational test environment: what is the user supposed to assess – full-texts or surrogates, and which metadata are useful to include?

In web search, we can distinguish between the result descriptions (snippets) and the linked web page presented in the search results list, and design retrieval effectiveness studies accordingly (Lewandowski, 2008). This is a major difference to LIS because library holdings still include printed materials. Thus, due to the nature of (meta)data in LIS, the results being assessed are technically the descriptions (surrogates) of the actual documents, e.g., surrogates of full texts of articles or e-books. If the study participant is supposed to assess actual full texts, with respect to the comparability of results, printed materials would have to be made available for judgment, as well. In contrary, links to full texts could be removed or made non-clickable to ensure each assessor is making their judgments on the basis of the same prerequisites. Providing printed materials would, however,require (a) inviting the participants to the same location or (b) preparing analog content as digitized files.

Although there are many factors that may influence the relevance judgment of information seekers, the selection of metadata should primarily be based on the research question or study objective. To present surrogates as realistically as possible, all elements of the records in the library catalogue should be included. When testing and comparing two or more systems, document sources should be anonymized to avoid influencing the test persons in any way, unless this would incompatible with the study's objective. Such data might, for example, identify the library system, holdings location, or library branch (including specific icons). However, we recommend that the information on the (electronic) availability of a document should not be removed, as it may have a strong influence on the assessors' decisions.

Further, the original position of the search results should not be visible to the assessors, to avoid order effects. Thus, the order of the results for each task should be randomized, while the data need to be analyzed according to the original results position produced by the system. To ensure this, all identifiers, e.g., record ID or URL, need to be stored in the background.

Number of results

In order to get meaningful results, a certain amount of assessment data is required. With respect to the web-scale character of discovery systems, the number of results to be assessed should be as large as possible. However, due to (a) the cut-off value, (b) the number of systems to be evaluated, and (c) the number of duplicate results, the number of results can grow exponentially, making a feasible limit necessary (see also section 4 on *pool building*). For example, a cut-off value of 10 would reflect the typical user behaviour of only considering the results on the first results page. A pretest can provide a baseline for the length of time required by one assessor to judge a given number of results.

Results selection

With respect to comparability issues, a decision must be made about whether (a) all types of documents should be included in the pool of results to be judged, i.e., all books, articles, working papers, even audio or video files, (b) whether documents in a certain language should be filtered out, and (c) whether only documents accessible by all test persons should be considered. In a real-life search session, a user, for example, might focus on research articles available via Open Access and written in Spanish, or seek information on an artist within a Digital Library of Cultural Heritage Collections expecting images of the artist's works. To avoid discrimination against documents that do not fulfill certain prerequisites, it would be helpful to exclude in advance documents with type, language, or access issues that are not suitable for judgment. Of course, these decisions must be made with respect to the research questions and goals pursued, taking characteristics of the target group(s) into account.

Assessment

The relevance of a document (surrogate) can be judged with binary and scale assessments. Binary assessment allows for only two conditions, i.e., a document is either relevant or it is not. With user-based models for retrieval tests, binary assessment is insufficient (Carterette et al., 2012). For a more differentiated assessment, scale assessment such as a 5-point Likert scale or a slider with a scale from 0 – 100 points should be used to observe graded relevance. We recommend a combination of both binary and scale assessment, because, even if a document is judged as non-relevant when there are only two possible options, it can receive some points on a scale or slider indicating it is relevant to a certain degree.

Regarding specifics of the target user group or user models, a description of the usage situation provides information about the information seeking context. At the same time, it allows constraints on the user group, time or location such as: "You are at home preparing a seminar paper when you realize you need information on topic XY. Given your deadline, you are only interested on documents you have immediate access to."

3. Assessors

Although the relevance of a retrieved set of search results should "be judged subjectively by the individual searcher rather than objectively in some way by the researcher" (Large & Beheshti, 1997, p. 114), there is a problem with human relevance assessments: inconsistency, which means that the overlap of judgments between different test groups is not large (Buckley & Voorhees, 2005; Saracevic, 2008). Nonetheless, this only shows again the human notion of relevance and "despite observed relevance problems from the human side, IR systems improved from the systems side" (Saracevic, 2008, p. 780). If time and budget allow, a researcher can also decide to have each document judged by more than one assessor.

Selection of assessors or assessor groups

Assessors should represent certain user groups, as we cannot assume that users are more or less the same. For example, users that are on-site at the library building may not mind browsing printed library collections, while users performing searches off-site or at home presumably prefer electronic documents.

An important criterion for selecting assessors is prior knowledge. Tasks involving topic search queries should be assessed by users with a certain amount of background in the respective subject, e.g., researchers, professors or students. Otherwise, a lot of results may be judged irrelevant because the assessor simply did not understand certain terms. In contrast, when it comes to known-item searches, subject knowledge is not necessary for assessment, because these items only demand one correct answer and it is an easy task to verify if this is at the top position of the results list. Thus, there is no need to recruit assessors; instead, investigators themselves can perform such tasks.

Individual assessors or assessor groups should also be selected with consideration for demographic issues because a diverse group of assessors helps to ensure transferability of results to other user groups or parts of the population. For example, assessor groups could be of mixed gender, age, or level of education.

Location

As described above, selecting search results involves decisions regarding comparability. The location of the assessors undertaking the tasks influences the accessibility of documents: a) on-campus or at home via virtual private network (VPN) access, the assessor could profit from licensed institutional access whereas b) off-campus without VPN access, the assessor may not be able to examine every document without exception. If the assessors are allowed to choose the location where they will complete the tasks, the lack of uniform conditions means comparability is not certain.

Number of assessors

The total number of assessors needed depends on the number of search tasks and the amount of time needed for judging. In general, one assessor would be allowed to work on several tasks, whereas one task would only be completed by one assessor. It can be assumed that, at some point, for instance, once a certain number of results have been judged, the concentration will decline. Thus, a task should not consist of too many results, and one assessor would only be able to judge so many tasks in a row. The estimated duration for assessing one task can be determined by measuring start and end times as part of a pretest.

Recruitment

Being confident with the decisions made with regard to assessor selection does not necessarily lead to the desired sample size. Most frequently, students participate in experiments because they are quite easy to reach, e.g., via institutional mailing lists or postings on notice boards. Nonetheless, incentives for participation within the limits of available financial resources would be more professional and increase the certainty of acquiring the necessary assessors. However, it is useful to contact and recruit more participants than actually needed to compensate for any problems that occur such as participants who cancel last minute or when plausibility checks suggest an assessor has provided implausible or dishonest assessments (see also section 4 on *organization*).

4. Testing

In order to start the actual test, the decisions on selecting search queries, results and assessors contribute to building the pool of documents to be assessed. We follow the approach used in TREC (D. K. Harman, 2005) and in IR evaluation on web search engines (Lewandowski, 2015).

Pool building

In order to obtain results, the search queries are entered into all the test systems (see Figure 2). Duplicate results, i.e., results produced by more than one system, should only be judged once per task, but re-included for data analysis. Thus, they need to be identified and removed from the pool, e.g., by comparing each result's URL or identifier. For example, if we obtain the top ten results for one query from three different systems, we would get 30 documents. Assuming there are 10 duplicates, we then would have 20 documents left to assess for the corresponding task.

A pretest assesses the research design and provides information on, for example, project-specific adjustments to the test design or evaluation tool and the number of duplicate results produced with a particular cut-off value. Further, it allows an estimate of the amount of time needed to complete one task to be calculated in order to determine the number of tasks in each evaluation run and thus the number of assessors needed. However, a pretest may uncover unforeseen problems, which is very valuable for the study.

Organization

To ensure reliable results, it may be helpful to integrate a plausibility check. For example, if a participant judges several documents in a row within a much shorter time frame than others, it might be that the assessments are not trustworthy because the evaluator did not take the task seriously. Another option can be to include some kind of fake document that would trigger a certain assessment (behaviour). Of course, these plausibility checks should also be implemented prior to the pretest, and trustworthy assessors other than the investigators themselves should be recruited, as well.

Another organizational task involves the preparation of information for the assessors. They need information on how to access the evaluation instrument, some information on the study (who, why, how, contact details) to show professionality and what it is exactly that they have to do in order to receive the compensation.

Figure 2: Example of pool building

5. *Data analysis*

In IR evaluation, there are many metrics to choose from, ranging from the classic *Precision* and *Recall* to user-oriented measures and metrics that combine precision with graded relevance assessments.[4] All have their advantages and disadvantages. The choice of metrics is affected by three decisions: (1) the research question, (2) the query types used to get the results for assessment and (3) the type of assessment applied.

Much as the research question influences the research design, it also informs the choice of metrics used for data analysis. It is essential to select the metrics that will answer the research question. Then, all data need to be collected or obtained within the study according to the chosen metrics. Moreover, an additional analysis based on, for example, demographic data or the time needed per task or result, can provide insights into the target group's behaviour or other issues that had not been focused on at the beginning. For this purpose, it would be useful to gather as many data as possible during the evaluation process.

Considering the type of queries, for *topic search results* we can evaluate *relevance values*, as for *known-item queries* the results can be measured with *success rates*, because the evaluation of known-item searches refers to the research question: is the IR system able to present the one correct search result at the top of the results list? With *Success @n* we can evaluate what proportion of queries produce the correct document until position *n*, e.g., the first, fifth or tenth position (Craswell & Hawking, 2005; Lewandowski, 2011). A standard measure applicable to navigational queries or known-item queries is the Mean Reciprocal Rank (MRR), which is also used in TREC. It was introduced by Kantor & Voorhees (2000) as the average score of all reciprocal ranked positions on which the known item was found while considering only the first correct result of a ranked list.

For analysis of assessment data on *topic search* results, we have to consider the differentiations between binary and scale assessments. For *binary relevance scores*, a Precision Graph of the top n results of all tasks can be created. The Mean Average Precision (MAP) has one disadvantage, as it does not distinguish between retrieved documents that are judged as non-relevant and those that are not judged at all. As a solution, BPREF was introduced by Buckley & Voorhees (2004) as a measure for binary relevance to defining a preference relation, i.e., "any relevant document is preferred over any nonrelevant document for a given [task]" (Buckley & Voorhees, 2004, p. 26). Thus, not the specific rank of a document in a list is considered but the relative ranks between judged relevant and non-relevant documents. This may be useful for comparing information systems with a large number of documents to be judged but it would not be consistent with the user behavioural focus on only a small number of the first ranked search results.

As scale assessments offer *graded relevance*, the choice of metrics must also be suitable for graded relevance assessments. Precision graphs can also be created for graded relevance assessments with Graded Average Precision (GAP), a measure proposed by Robertson, Kanoulas, & Yilmaz (2010). GAP

[4] An overview of IR metrics with calculation examples is provided by Baeza-Yates & Ribeiro-Neto (2011, pp. 134–158).

is built on Average Precision (AP), whereas the "AP of a ranked list is the average of the precisions at each relevant document in that list" (Carterette et al., 2012, p. 113).

Although binary-based metrics were used within TREC, there have been several attempts to introduce graded relevance metrics (Kekäläinen, 2005), for example, cumulated gain-based measures. In statistics, cumulated frequencies are the sum of the frequencies until they reach a certain boundary (e.g., 1.0 or 100%). Thus, cumulated measures consider human behaviour regarding scanning a search results list, i.e., top-down rather than bottom-up. In contrast to precision, cumulated gain-based measures "allow researchers to test different weighting schemes for relevant documents, which reflect different user scenarios" (Kekäläinen, 2005, p. 1022). To analyse the systems' ability to take the presentation of results in descending order, the discounted cumulative gain (DCG) was introduced as a novel measurement by Järvelin & Kekäläinen (2000). A corrected version of DCG is the normalized discounted cumulative gain (NDCG) because it uses normalized figures. NDCG is a measure regularly used in IR evaluation.

Conclusion and limitations

With the development of advanced library information systems as user-oriented, web-scale search systems, the need to evaluate their retrieval effectiveness has also increased. Prior LIS research often neglected to follow the standard procedures for IR evaluation. These criteria and methods must first be adapted to library information systems evaluation, whereas further research on retrieval evaluation methods of LIS needs to adopt user-oriented evaluation features, as pointed out in interactive information retrieval (IIR) evaluation.

As a first step in establishing a framework for designing retrieval effectiveness studies of library information systems, we introduced a framework based on the conventional, system-oriented procedures. It deals with the major aspects from selecting search queries, results, human assessors, the testing phase, and data analysis. In addition, it can be understood as a guide for systematically describing the methods used in LIS retrieval studies, following the five parts of the overall framework.

From our point of view, the major benefit of the proposed framework is that it will help standardizing retrieval effectiveness studies with LIS, and making the results from these studies comparable. The literature review clearly showed that the lack of such standards has led to rather unsystematic evaluations. The proposed framework is a first step towards standardizing evaluations of LIS.

Unfortunately, in practice, there is still no technical support for IR evaluation in the library context as it exists for retrieval tests of web search engines. For example, in past studies on web search engines, the *Relevance Assessment Tool* (Lewandowski & Sünkler, 2013) has been used to scrape the search results, provide the web-based environment for relevance assessments following a crowdsourcing approach that includes a reward system, and export all data into Excel tables. A similar web-based evaluation tool for conducting retrieval tests with library systems and library-specific data would contribute to IR research. The library community would greatly benefit from software supporting retrieval tests, as such tests need to be conducted on a regular basis to monitor LIS performance, as more content is added to the system and user needs change.

Acknowledgements

The framework presented in this article has been developed as part of the research project *LibRank – New Approaches to Relevance Ranking in Library Information Systems* funded by the German Research Foundation (DFG – Deutsche Forschungsgemeinschaft) from March 2014 until February 2016.

References

Antelman, K., Lynema, E., & Pace, A. K. (2006). Toward a twenty-first century library catalogue. *Information Technology & Libraries*, *25*(3), 128–139.

Asher, A. D., Duke, L. M., & Wilson, S. (2013). Paths of discovery: Comparing the search effectiveness of EBSCO Discovery Service, Summon, Google Scholar, and conventional library resources. *College & Research Libraries*, *74*(5), 464–488.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval : the concepts and technology behind search* (2. ed.). Harlow [u.a.]: Addison-Wesley/Pearson.

Barry, C., & Lardner, M. (2011). A Study of First Click Behaviour and User Interaction on the Google SERP. In J. Pokorny, V. Repa, K. Richta, W. Wojtkowski, H. Linger, C. Barry, & M. Lang (Eds.), *Information Systems Development* (pp. 89–99). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-9790-6_7

Blandford, A., Adams, A., Attfield, S., Buchanan, G., Gow, J., Makri, S., … Warwick, C. (2008). The PRET A Rapporter framework: Evaluating digital libraries from the perspective of information work. *Information Processing & Management*, *44*(1), 4–21. https://doi.org/10.1016/j.ipm.2007.01.021

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3). Retrieved from http://www.informationr.net/ir/8-3/paper152.html

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum*, *36*(2), 3–10. https://doi.org/10.1145/792550.792552

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04* (pp. 25–32). New York, New York, USA: ACM Press. https://doi.org/10.1145/1008992.1009000

Buckley, C., & Voorhees, E. M. (2005). Retrieval System Evaluation. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: Experiment and Evaluation in Information Retrieval* (pp. 53–75). Cambridge, MA ; London, UK: MIT Press.

Carterette, B., Kanoulas, E., & Yilmaz, E. (2012). Evaluating web retrieval effectiveness. In D. Lewandowski (Ed.), *Web search engine research* (pp. 105–137). Emerald Group Publishing.

Chickering, F. W., & Yang, S. Q. (2014). Evaluation and Comparison of Discovery Tools: An Update. *Information Technology and Libraries*, *33*(2), 5–30. https://doi.org/10.6017/ital.v33i2.3471

Ciccone, K., & Vickery, J. (2015). Summon, EBSCO Discovery Service, and Google Scholar: A Comparison of Search Performance Using User Queries. *Evidence Based Library and Information Practice*, *10*(1), 34. https://doi.org/10.18438/B86G6Q

Cleverdon, C. W. (1960). The Aslib Cranfield research project on the comparative efficiency of

indexing systems. *Aslib Proceedings*, *12*(12), 421–431. https://doi.org/10.1108/eb049778

Cleverdon, C. W., & Keen, M. (1966). Aslib Cranfield research project - Factors determining the performance of indexing systems; Volume 2, Test results. Retrieved from http://hdl.handle.net/1826/863

Cleverdon, C. W., Mills, J., & Keen, M. (1966a). Aslib Cranfield research project - Factors determining the performance of indexing systems; Volume 1, Design; Part 1, Text. Retrieved from http://hdl.handle.net/1826/861

Cleverdon, C. W., Mills, J., & Keen, M. (1966b). Aslib Cranfield research project - Factors determining the performance of indexing systems; Volume 1, Design; Part 2, Appendices. Retrieved from http://hdl.handle.net/1826/862

Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., & Voorhees, E. M. (2015). TREC 2014 Web Track Overview. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA618628

Connaway, L. S., & Dickey, T. J. (2010). The digital information seeker: Report of findings from selected OCLC, RIN and JISC user behaviour projects. Retrieved from http://www.jisc.ac.uk/media/documents/publications/reports/2010/digitalinformationseekerreport.pdf

Craswell, N., & Hawking, D. (2005). Overview of the TREC-2004 Web Track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*. National Institute of Standards and Technology. Retrieved from http://trec.nist.gov/pubs/trec13/papers/WEB.OVERVIEW.pdf

Deodato, J. (2015). Evaluating Web-Scale Discovery Services: A Step-by-Step Guide. *Information Technology and Libraries*, *34*(2), 19–75. https://doi.org/10.6017/ital.v34i2.5745

Dillon, M., & Wenzel, P. (1990). Retrieval effectiveness of enhanced bibliographic records. *Library Hi Tech*, *8*(3), 43–46. https://doi.org/10.1108/eb047797

Dumais, S. T., & Belkin, N. J. (2005). The TREC Interactive Tracks: Putting the User into Search. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: experiment and evaluation in information retrieval* (pp. 123–152). Cambridge, MA ; London, UK: MIT Press.

Frants, V., Shapiro, J., & Voiskunskii, V. (1997). *Automated information retrieval: theory and methods*. San Diego: Academic Press.

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing & Management*, *35*(2), 141–180. https://doi.org/10.1016/S0306-4573(98)00041-7

Griesbaum, J. (2004). Evaluation of three German search engines: Altavista.de, Google.de and Lycos.de. *Information Research*, *9*(4). Retrieved from http://www.informationr.net/ir/9-4/paper189.html

Harman, D. (2011). *Information Retrieval Evaluation*. *Synthesis Lectures on Information Concepts, Retrieval, and Services; 19*. Morgan & Claypool. https://doi.org/10.2200/S00368ED1V01Y201105ICR019

Harman, D. K. (2005). The TREC Test Collections. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: Experiment and Evaluation in Information Retrieval* (pp. 21–52). Cambridge, MA ; London, UK: MIT Press. https://doi.org/10.1162/coli.2006.32.4.563

Hawking, D., & Craswell, N. (2005). The Very Large Collection and Web Tracks. In E. M. Voorhees & D.

K. Harman (Eds.), *TREC: Experiment and Evaluation in Information Retrieval* (pp. 199–231). Cambridge, MA ; London, UK: MIT Press.

Hawking, D., Craswell, N., Bailey, P., & Griffihs, K. (2001). Measuring Search Engine Quality. *Information Retrieval*, *4*(1), 33–59. https://doi.org/10.1023/A:1011468107287

Hennies, M., & Dressler, J. (2006). Clients information seeking behaviour: An OPAC transaction log analysis. In *click 06, ALIA Biennal Conference*. Perth, AU.

Hildreth, C. R. (2001). Accounting for users' inflated assessments of on-line catalogue search performance and usefulness: an experimental study. *Information Research*, *6*(2). Retrieved from http://www.informationr.net/ir/6-2/paper101.html

Hofmann, M. a., & Yang, S. Q. (2012). 'Discovering' what's changed: a revisit of the OPACs of 260 academic libraries. *Library Hi Tech*, *30*(2), 253–274. https://doi.org/10.1108/07378831211239942

Huffman, S. B., & Hochster, M. (2007). How well does result relevance predict session satisfaction? *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '07*, 567. https://doi.org/10.1145/1277741.1277839

Jansen, B. J., & Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing & Management*, *42*(1), 248–263. https://doi.org/10.1016/j.ipm.2004.10.007

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 - SIGIR '00* (pp. 41–48). New York, New York, USA: ACM Press. https://doi.org/10.1145/345508.345545

Kantor, P. B., & Voorhees, E. M. (2000). The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Information Retrieval*, *2*(2), 165–176. https://doi.org/10.1023/A:1009902609570

Kekäläinen, J. (2005). Binary and graded relevance in IR evaluations—Comparison of the effects on ranking of IR systems. *Information Processing & Management*, *41*(5), 1019–1033. https://doi.org/10.1016/j.ipm.2005.01.004

Large, A., & Beheshti, J. (1997). OPACs: A research review. *Library & Information Science Research*, *19*(2), 111–133. https://doi.org/10.1016/S0740-8188(97)90039-6

Lau, E. P., & Goh, D. H.-L. (2006). In search of query patterns: A case study of a university OPAC. *Information Processing & Management*, *42*(5), 1316–1329. https://doi.org/10.1016/j.ipm.2006.02.003

Lewandowski, D. (2008). The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, *64*(6), 915–937. https://doi.org/10.1108/00220410810912451

Lewandowski, D. (2010). Using search engine technology to improve library catalogs. In *Advances in Librarianship* (Vol. 32, pp. 35–54). https://doi.org/10.1108/S0065-2830(2010)0000032005

Lewandowski, D. (2011). The retrieval effectiveness of search engines on navigational queries. *Aslib Proceedings*, *63*(4), 354–363. https://doi.org/10.1108/00012531111148949

Lewandowski, D. (2012). A framework for evaluating the retrieval effectiveness of search engines. In C. Jouis, I. Biskri, J.-G. Ganascia, & M. Roux (Eds.), *Next Generation Search Engines* (pp. 456–479). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-4666-0330-1.ch020

Lewandowski, D. (2015). Evaluating the retrieval effectiveness of web search engines using a representative query sample. *Journal of the Association for Information Science and Technology*, *66*(9), 1763–1775. https://doi.org/10.1002/asi.23304

Lewandowski, D., & Sünkler, S. (2013). Designing search engine retrieval effectiveness tests with RAT. *Information Services and Use*, *33*(1), 53–59. https://doi.org/10.3233/ISU-130691

Moore, K. B., & Greene, C. (2012). Choosing discovery: A literature Review on the selection and evaluation of discovery layers. *Journal of Web Librarianship*, *6*(3), 145–163. https://doi.org/10.1080/19322909.2012.689602

Niu, X., Zhang, T., & Chen, H. (2014). Study of user search activities with two discovery tools at an academic library. *International Journal of Human-Computer Interaction*, *30*(5), 422–433. https://doi.org/10.1080/10447318.2013.873281

O'Brien, A. (1990). Relevance as an aid to evaluation in OPACs. *Journal of Information Science*, *16*(4), 265–271. https://doi.org/10.1177/016555159001600407

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google We Trust: Users' Decisions on Rank, Position, and Relevance. *Journal of Computer-Mediated Communication*, *12*(3), 801–823. https://doi.org/10.1111/j.1083-6101.2007.00351.x

Peters, T. A. (1993). The history and development of transaction log analysis. *Library Hi Tech*, *11*(2), 41–66. https://doi.org/10.1108/eb047884

Robertson, S. E., Kanoulas, E., & Yilmaz, E. (2010). Extending average precision to graded relevance judgments. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10* (pp. 603–610). New York, New York, USA: ACM Press. https://doi.org/10.1145/1835449.1835550

Sadeh, T. (2015). From Search to Discovery. *Bibliothek Forschung Und Praxis*, *39*(2), 1–17. https://doi.org/10.1515/bfp-2015-0028

Saracevic, T. (1996). Relevance reconsidered. In P. Ingwersen & N. O. Pors (Eds.), *Information science: Integration in perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science (CoLIS 2)* (pp. 201–218). Copenhagen: Royal School of Librarianship.

Saracevic, T. (2008). Effects of inconsistent relevance judgments on Information Retrieval test results: A historical perspective. *Library Trends*, *56*(4), 763–783. https://doi.org/10.1353/lib.0.0000

Spink, A. (2002). A user-centered approach to evaluating human interaction with Web search engines: an exploratory study. *Information Processing & Management*, *38*(3), 401–426. https://doi.org/10.1016/S0306-4573(01)00036-X

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing & Management*, *28*(4), 467–490. https://doi.org/10.1016/0306-4573(92)90005-K

Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '02* (p. 316). New York, New York, USA: ACM Press. https://doi.org/10.1145/564376.564432

Zhang, T. (2013). User-Centered Evaluation of a Discovery Layer System with Google Scholar. In A. Marcus (Ed.), *Design, User Experience, and Usability. Web, Mobile, and Product Design SE - 34* (Vol. 8015, pp. 313–322). Berlin ; Heidelberg: Springer. https://doi.org/10.1007/978-3-642-39253-5_34

Zhang, Y. (2010). Developing a holistic model for digital library evaluation. *Journal of the American Society for Information Science and Technology*, *61*(1), 88–110. https://doi.org/10.1002/asi.21220

Zhang, Y., Jansen, B. J., & Spink, A. (2009). Time series analysis of a Web search engine transaction log. *Information Processing & Management*, *45*(2), 230–245. https://doi.org/10.1016/j.ipm.2008.07.003