A Component-level Analysis of an Academic Search Test Collection

Part I: System and Collection Configurations

Florian Dietz and Vivien Petras

Berlin School of Library and Information Science
Humboldt-Universität zu Berlin
Dorotheenstr. 26, 10117 Berlin, Germany
florian.dietz@alumni.hu-berlin.de | vivien.petras@ibi.hu-berlin.de

This is an author's accepted manuscript version of a conference paper published in the International Conference of the Cross-Language Evaluation Forum for European Languages CLEF 2017: Experimental IR Meets Multilinguality, Multimodality, and Interaction within the Springer Lecture Notes in Computer Science book series (LNCS, volume 10456).

The final publisher's version is available online at:
https://doi.org/10.1007/978-3-319-65813-1_2

A Component-level Analysis of an Academic Search Test Collection Part I: System and Collection Configurations

Florian Dietz and Vivien Petras

Berlin School of Library and Information Science
Humboldt-Universität zu Berlin
Dorotheenstr. 26, 10117 Berlin, DE
florian.dietz@alumni.hu-berlin.de | vivien.petras@ibi.hu-berlin.de

Abstract. This study analyzes search performance in an academic search test collection. In a component-level evaluation setting, 3,276 configurations over 100 topics were tested involving variations in queries, documents and system components resulting in 327,600 data points. Additional analyses of the recall base and the semantic heterogeneity of queries and documents are presented in a parallel paper. The study finds that the structure of the documents and topics as well as IR components significantly impact the general performance, while more content in either documents or topics does not necessarily improve a search. While achieving overall performance improvements, the component-level analysis did not find a component that would identify or improve badly performing queries.

Keywords: academic search; component-level evaluation; GIRT04

1 Introduction

The basis of success for every IR system is the match between a searcher's information need and the system's content. Factors that contribute to the success of such a match have been studied at length: the underlying information need, the searcher's context, the type of query, the query vocabulary and its ambiguity, the type, volume and structure of the searched documents, the content of the documents, the kind of expected relevance of the documents and finally - the primary focus of IR - the IR system components, i.e. the preprocessing steps, ranking algorithm and result presentation. While all these aspects have been shown to impact search performance, it is also common knowledge that a successful configuration of these aspects is highly contextual. There is no one-size-fits-all solution.

Earlier initiatives such as the Reliable Information Workshop [12] and the TREC Robust Track [23] used TREC test collections to study what causes differences in search performance. They showed that search performance depends on the individual searcher, the search task, the search system and the searched documents. This did not only motivate new research on processing difficult queries,

but it also spawned a new research field in query performance prediction [2]. Grid- or component-level evaluation initiatives [7, 11] moved into another direction, focusing on the evaluation of system component configurations to identify optimal retrieval settings.

This paper presents a component-level study in academic search. Academic search presents different challenges from the collections and information needs previously studied: the queries and their vocabulary can be highly technical and domain-specific, and, often, the searched documents just contain the bibliographic metadata of scientific publications. The study utilizes component-level evaluation aspects to find the causes for differences in search performance using the whole pipeline of the search process including the query, the documents and the system components. In a parallel paper [6], we present an analysis of indicators used in query performance prediction to delve even deeper into the causes for successful or unsuccessful search performance based on the queries, particularly trying to identify badly performing queries.

The goal of the research was not to find the best configuration (some state-of-the-art ranking algorithms were not even considered) but to find the most predictive factors for performance differences for this test collection. Future work should then be able to use this approach to extrapolate from the analysis of one collection to compare it with other collections in this domain.

The paper is structured as follows: section 2 describes the area of academic search and discusses relevant research on component-level evaluation. Section 3 describes the test collection GIRT4 (used in the CLEF domain-specific track from 2004-2008) and the experimental set-up including the test configurations used. Section 4 describes the components that were analyzed for their predictive power in determining search performance. Section 5 concludes with an outlook on future work.

2 Component-level Evaluation in Academic Search

Academic search is defined as the domain of IR, which concerns itself with searching scientific data, mostly research output in the form of publications [17]. It is one of the oldest search applications in IR. Not only were bibliographic information systems one of the first automated information retrieval systems (e.g. Medline [19]), but the first systematic IR evaluations, the Cranfield retrieval experiments, were also performed with an academic search collection [4]. The late 1990s and 2000s saw a renewed interest in these collections when digital libraries became a prominent research topic [3]. Academic search differs from previously tested search environments - mostly newspaper or web documents with general information needs [2] - in significant aspects. Academic search output is still comparatively small: between 1.5 [25] and 2.5 million [24] new articles were reported for 2015 globally. Most academic search collections are focused on one or a small number of disciplines and are therefore significantly smaller.

Documents in an academic search collection have a particular organization - either just the bibliographic metadata or the structure of a scientific publication with further references. Bibliographic metadata could be enriched with technical

vocabulary (such as the MeSH keywords in PubMed), which support searching in the technical language of the documents [21]. When searching the full-text of publications, the references can be a major source for search success [18].

Information needs and their query representations academic search are different as well. While queries were found to be the same in length or longer than in standard web search engines [10], the content differs more dramatically. Particularly, queries contain technical terms or search for particular document components (such as author, title, or keywords) that are specific to the type of documents searched [13]. It appears logical that with highly specific information needs and small document collections, the number of relevant documents for any information need in this domain is also low.

Finally, these different documents and queries also demand different processing [22] and different ranking algorithms [1]. The CLEF domain-specific track, first established as a track in TREC8, provided a test collection to evaluate IR systems in this domain [14]. This study uses the CLEF domain-specific test collection GIRT4, which was released in 2004.

As a possible solution to the problem of finding the root causes for IR challenges of this type of test collection, component-based evaluation might be a suitable approach. For analyzing their impact on search performance, it is important to understand the impact of individual IR system components and parameters which may make a difference in retrieval. To measure the effect of those factors independent from others, a sufficiently dimensioned amount of data is necessary. Component-based evaluation takes a parameter and averages the measured effects of all other parameters while keeping the respective factor in focus [11].

The amount of generated ground truth data is important. Scholer and Garcia [20] report that a diversity of factors is needed to make evaluation effective. They criticize the evaluation methods for query performance prediction, because usually only one retrieval system is used. The concentration on just one system distorts the results, as the effect of a different system can make significant changes in terms of prediction quality. This is also true when searching for root causes for query failure. However, testing in a large dimensional space of IR system components requires a large-scale effort [7, 8]. Kürsten [16] used GIRT4 as a test collection for a component-level anaylsis.

Component-level evaluation has not been studied extensively in academic search. De Loupy and Bellot [5] analyzed query vocabulary and its impact on search performance in the Amaryllis test collection, a French collection of bibliographic documents, which was also used in CLEF. Kürsten's component-level evaluation of the GIRT4 collection [16] found that utilizing the document structure (different document fields) did not impact retrieval performance. Other aspects were not considered in detail. In this paper, query and document structure, IR system preprocessing filters and ranking algorithms will be analyzed to determine which factors contribute most to search success.

3 Study Configurations

3.1 Test Collection & Test IR System

Following the Cranfield approach [4], the GIRT4 test collection consists of documents (metadata about social science publications), matching topics representing information needs and relevance assessments [15]. The GIRT4 English collection contains 151,319 English-language documents, consisting of structured bibliographic data: author, title, year, language code, country code, controlled keyword, controlled method term, classification text and abstract. Most fields have very little searchable text. Only very few documents contain an abstract (less than 13% of the collection).

GIRT4 queries are TREC-style topics, prepared by expert users of the documents. The topics contain three fields, which can be utilized for search: title, description and narrative. The binary relevance assessments are based on pooled retrieval runs from the CLEF domain-specific track. As topic creators and relevance assessors were not the same, certain information about the searcher (e.g. their context) remains unknown and can therefore not be measured or evaluated. Altogether, 100 topics and their relevance assessments from the years 2005-2008 were used.

For the experiments, the open source retrieval toolkit Lemur[1] was used. The software offers different retrieval ranking algorithms, which can be adjusted and further specified by several parameters: the Vector Space Model (VSM), Okapi BM25 (BM25) and Language Modeling (LM) with either Dirichlet-Prior-Smoothing, Jelinek-Mercer-Smoothing, or absolute discount smoothing. Lemur also provides the Porter and Krovetz stemmers and stopword list integration for preprocessing of documents and queries. The Lemur toolkit allows easy configuration of system components, which allows a component-level evaluation.

For evaluation, the trec_eval program[2] was used. As most analyses were done on a topic-by-topic basis, average precision (AP) per query was chosen as a metric. All experiments were performed on an Ubuntu operating system. The result sets were automatically structured and evaluated with Python scripts.

3.2 Component-level Configurations

Following the component-level evaluation approach, different configurations of document fields, topic fields and IR system components were compared.

For the document collection, different combinations of the title (DT), abstract (AB) and an aggregated keyword field (CV), which consisted of the controlled keywords, method terms and classification terms were compared. All other fields were discarded due to a lack of relevant content when comparing them to the information needs represented in the topics. All possible document field combinations make a total of seven document configurations. However, the AB-only variant was not analyzed, because too few documents would remain in the collection for retrieval.

[1] https://www.lemurproject.org/lemur.php, last accessed: 04-30-2017
[2] http://trec.nist.gov/trec_eval, , last accessed: 04-30-2017

The three topic fields title (T), description (D) and narrative (N) were also used in every possible configuration, totaling seven topic configurations. Although the narrative was originally not intended for retrieval but to help the relevance assessors determine the most relevant documents per query, it was still used as a query field.

Every preprocessing option provided by the Lemur toolkit was included as well: Porter stemmer, Krovetz stemmer, no stemming, use of a stopword list, and no stopword list. A general stopword list for the English language was used, adjusted with a small number of non-relevant topic words (such as: find, documents) to improve the performance of all topic fields.

To create a reliable amount of results, the inclusion of multiple retrieval models is a critical requirement [20]. All Lemur ranking models were used for the experiments. To analyze the impact of model parameters, the VSM term weights in documents and queries were parameterized (different TF, IDF variants). Overall, 13 different ranking approaches were tested.

Table 1 summarizes the possible configurations that were used for experiments. For each topic, 3,276 configurations were tested, totaling in 327,600 data points for the 100 topics.

Table 1: Component-level configurations

Component	**Configuration variables**	**Configurations**
Document fields	DT, AB, CV	6
Topic fields	T, D, N	7
Stemming	Krovetz, Porter,none	3
Stopwords	List, none	2
Ranking models	VSM, BM25, LM (different smoothing)	13

3.3 Analysis Steps

For every topic, the AP over every configuration was calculated to reach an impression of the topic's performance. The AP differs widely across the test set. Some topics perform very well overall, while others seem to fail in every tested configuration. As a starting point, we separated between good and bad topics similarly to [9]: the median of the APs (per topic over all configurations) represents the threshold between well and badly performing topics.

The analysis was divided in several parts. One part consisted of an overall evaluation of all components of the retrieval process reported in this paper. Another part looked more specifically at various query and collection factors, where a relation to retrieval performance was assumed [6].

Every single aspect (document and topics fields, preprocessing components, ranking algorithms) was looked at while keeping every other component in the tested configuration stable. To measure the impact of a component on the retrieval success, all results with a specific component in the configuration were compared against all results without the tested component. The significance of the impact of a specific configuration compared to others was measured using the Wilcoxon signed-rank test. When a specific configuration component significantly increased the average AP per topic, we concluded that this component had an impact on the search performance.

4 Analyzing Component Performance

This section reports the results for specific IR process aspects under consideration. Sections 4.1 studies the impact of specific document or topic field configurations. Section 4.2 compares the impact of IR system preprocessing components and ranking algorithms.

4.1 Document and Topic Structure

In academic search, where the amount of textual content is limited, the appropriate use of document and topic structures is important. By including different fields in the retrieval process, the searchable content is influenced. This section determines the impact of different field configurations while keeping every other component stable.

Documents For the document fields, there are six different configurations to compare. Intuitively, the more text is available, the better the performance should be. Table 2 compares the term counts for the respective document fields. The title field contains a higher number of unique terms than the controlled term field and should thus yield more available terms for retrieval. The abstract field contains the highest number of unique terms. However, because only 13% of all collection documents contain abstracts, its impact may not be as high. A retrieval run just on the abstracts was not attempted as too few documents would have been available to search.

Table 2: Number of Terms per Document Field

	Title	Controlled term fields	Abstract
terms	2,157,680	4,841,399	3,776,509
terms w/o stopwords	1,445,065	4,380,116	1,871,727
unique terms	38,919	4,326	69,526
unique terms w/o stopwords	38,359	4,196	68,923

Table 3 shows the MAP over all topics for the different document field configurations. The MAP is averaged over every possible retrieval component configuration with the respective document field. The combination of all fields contains the most searchable text, but does not achieve the best search performance.

Table 3: Document Field Configurations and MAP

DT	CV	DT + CV + AB	DT + AB	CV + AB	DT + CV
0.1205	0.1242	0.1776	0.1153	0.1175	0.1961

While the title and controlled term fields perform similarly (according to a Wilcoxon signed-rank test, the difference is not significant, DT vs. CV: Z=-1.2103, p=0.226), adding the abstract text tends to slightly deteriorate the performance although the difference is not significant for either combination (Wilcoxon signed-rank test for DT vs. DT+ABS: Z=-0.2785, p=0.779; CV vs. CV+AB: Z=-1.0177, p= 0.308). The best configuration for search performance

seems to be a combination of the title and controlled term fields. It performs significantly better than the combination of all fields (DT+AB vs. DT+CV+AB: Z=-3.8612, p=0.000) showing a negative impact of the abstract field after all. The better performance of the combined title and keyword fields shows that the controlled terms are not only different from the title terms, but add relevant content to the documents.

About a third of the relevant documents contain an abstract (compared to only 13% of the documents in the whole collection), so the finding that the abstract field deteriorates the search performance is even more puzzling. Abstracts may have a negative effect by containing misleading terms for many queries, but a high number of abstracts in the relevant documents suggests that for a small number of queries, abstract terms provide relevant matching input. It is important to note that the number of abstracts in relevant documents could concentrate on the relevant documents for a small number of queries - the differences between the number of relevant documents per query are surprisingly high. One possible explanation is also the fact that the academic search collection and topics seem to rely on a combination of highly specific words and more general method terms.

Topics The topic fields suffer from similar problems. The three fields - title, description and narrative - are different from each other. A first point to observe is their different lengths (table 4). As expected from TREC-style topics, the title field is shorter than the description, which is shorter than the narrative. Observing the average length after stopword removal shows that the title field consists of mostly content-bearing terms, while the description and narrative fields are reduced by half.

Table 4: Average Number of Terms per Topic Field

	Title	Description	Narrative
terms	3.79	12.67	32.28
terms w/o stopwords	2.83	6.04	14.64

Table 5 lists the MAP of all topic field configurations while keeping the other factors stable. A similar image to the document field analysis emerges. The shortest field (title) achieves the best results when compared on an individual field basis, while the longest field (narrative) performs significantly worse. The combination of title and description appears to achieve even better results than the title field alone, but the difference is not significant (T vs. TD: Z=-1.1828, p=0.238). The addition of the longer narrative field to the title and description field configuration seems to deteriorate the performance, but the difference is also not significant (TD vs. TDN: Z=-1.8326, p=0.067).

Table 5: Topic Field Configurations and MAP

T	D	N	TD	TN	DN	TDN
0.1704	0.1256	0.0849	0.1786	0.1380	0.1308	0.1657

A possible explanation for these results could be that the description of many topics just repeats the title terms, which may improve the term weights, but does not add content-bearing terms. Although the narrative contains the highest number of terms, it has a mostly negative effect on retrieval performance, probably because the field also contains instructions for the relevance assessors and may add terms that divert from its topical intent.

The analyses of the impacts of document and topic structures show that the content of either components can have a decisive impact on the search performance. More terms do not automatically lead to a better performance - this is true for both documents and topics, although the specific structure of the test collection needs to be taken into account. For all retrieval scenarios, the impact of topic and document terms needs to be looked at in combination, because both factors are connected. Another paper [6] focuses on such combinatorial analyses.

4.2 IR System Components

After analyzing the document and topic structure, this section takes a closer look at the preprocessing steps and the influence of the ranking algorithms.

Stopwords. A simple factor to analyze is the influence of the stopword list as there are only two datasets to compare - all experiments with or without applying the stopword list. For this test collection, the removal of stopwords improved the AP by 30% on average, a significant difference.

The stopword list helps retrieval in two ways. One, it helps to reduce the amount of terms that need to be searched. After stopword removal, ideally, only content-bearing terms should remain. It also helps in optimizing term weights, because the important keywords are more exposed. Especially longer queries benefit from this effect. The positive impact seems to affect both documents and topics.

An example for the positive effect of stopword removal is topic 128-DS[3], which receives a boost of 70% in AP (before stopword removal, averaged AP=0.1883, after, AP=0.3196). The number of topic terms is reduced from 52 to 23, removing terms such as "their" and also explanatory phrases like "relevant documents". Left over are stronger verbs like "discussed", which can help to identify the scientific methods applied in the academic publication.

The narrative field length changes most with stopword removal (Table 4). Table 6 shows that it also benefits most from it in retrieval performance when compared to the other fields. The lowest effect is observable for the title field, because stopword removal does not change the query as much.

Removing stopwords can also have negative effects. A small number of topics suffered from the removal of supposedly unimportant terms. The problem are terms, which might be unimportant in a different context but are content-bearing

[3] T: Life Satisfaction; D: Find documents which analyze people's level of satisfaction with their lives.; N: Relevant documents report on people's living conditions with respect to their subjective feeling of satisfaction with their personal life. Documents are also relevant in which only single areas of everyday life are discussed with respect to satisfaction.

Table 6: MAP for Topic Fields with and without Stopwords

	T	D	N
all terms	0.1682	0.1088	0.0669
w/o stopwords	0.1725	0.1424	0.1028

in these topics. There a two very figurative examples, which show this particular problem. The title field of topic 177-DS has the following terms: "unemployed youths without vocational training". After stopping, the word "without" is removed, reversing the information intent. Consequently, the AP suffers a drop of over 10%. Another example is topic 151-DS, searching for "right-wing parties". After applying the stopword list, the term "right" is removed.

These examples might occur more often in academic search with queries in a specific or highly technical language, which may be adversely effected by a conventional stopword list. The usage of the stopword list is dependent on the relationship between information intent, query terms and the document collection. For topic 151-DS, the AP stays relatively stable although the actual information need is not represented anymore. This is because the collection does not contain a lot of documents distinguishing right-wing and left-wing parties, which means the same documents are retrieved. In larger document collections, these distinctions may have a much bigger impact.

Stemming. Stemmers have been shown to increase the search performance, because different word forms are reduced (plural and conjugated forms are stemmed to their respective stems), unifying the vocabulary and thus making it easier to match queries and documents. In the test collection, stemming has positive effects. While the Porter stemmer leads to an average improvement of 52% in AP over all configurations, the improvement of the Krovetz stemmer is around 30%.

Also stemmers can have negative effects on the search performance, deriving from the same cause as the positive effects. In unifying the vocabulary to stems, errors occur when the stemming is too strict or too soft. Looking at the topic terms and comparing the original forms to the Porter and Krovetz stems, one can see a variety of over- and understemming occurring. The examples are relatively rare and do not have the biggest influence on performance, but in some cases are measurable. A figurative example is topic 210-DS (table 7).

Table 7: Topic 210-DS: Stemming of Narrative Terms. Original Narrative: The activities of establishing business in the new German federal states are relevant. What characterized the boom in start-up businesses following reunification?

	Original	**Porter**	**Krovetz**
narr. terms	activities establishing business german federal states characterized boom start businesses reunification	activ establish busi german feder state character boom start busi reunif	active establish busy german federal state characterize boom start businesse reunification
AP	0.0184	0.0619	0.0159

While the AP is low overall, the search performance is greatly improved by using the Porter stemmer (AP +230%), but suffers when stemmed with the Krovetz stemmer (AP -13%). The Porter stemmer reduces the number of unique

terms, while Krovetz does not. The Porter stemmer changes term weights, because the important terms appear more often, Krovetz just changes their form. Even worse is the transformation of the variations for "business", one of the key terms for this topic. While the plural form is stemmed to "businesse", the singular is changed to "busy". This might be the cause for the negative effect of Krovetz: the new stem distorts the informational intent as well as the term weights.

Analyzing the effect of the Krovetz and Porter stemmers over all documents in the collection, this observation remains stable. Porter drastically reduces the amount of unique terms while Krovetz stems much more cautiously. This means that Porter has a more significant effect on retrieval because of its dual impact: reducing the word forms also changes the term weights.

Ranking Algorithms. The influence of the ranking algorithms was analyzed as well. While the study did not aim at finding the best ranking algorithm for the test collection, it tested whether the optimization of parameters significantly impacted the search performance for the test collection.

For one ranking algorithm, the Vector Space Model, all available variations in term weighting parameters were tested. Altogether, Lemur offers three different term weighting options for documents and queries: raw frequency, logarithmic frequency or the weighting function of the Okapi retrieval model. Nine different configurations were evaluated for the study. While different document term weighting parameters show an impact on the search performance, different query term weighting schemes did not significantly change the results, probably because query term frequencies are small. According to the averages over all configurations per ranking model, the best weighting option for the Vector Space Model is the Okapi term weighting for document terms (averaged AP=0.1717). It significantly outperforms both the logarithmic weighting schemes (Z=-8.5064, p=0.000) and the raw frequency weighting for document terms (Z=-8.4858, p=0.000).

When comparing the best Vector Space Model configuration to the other ranking models, the differences are small. Table 8 shows the average AP over all experiments performed with the respective ranking model. There is no significant difference between the different models, although some differences may have been observed if the other ranking algorithms were optimized for the test collection. All of the analyzed ranking algorithms use term weights to determine the relevant documents and thus resemble each other if no other ranking signals are used.

Table 8: MAP for different Retrieval Models (LM-ads=Language modeling with absolute discounting, LM-jms=Language modeling with Jelinek-Mercer smoothing, LM-ds=Language modeling with Dirichlet smoothing)

LM-ads	LM-jms	LM-ds	BM25	VSM
0.1700	0.1666	0.1452	0.1654	0.1717

Looking at the variations in search performance on a topic-by-topic basis, the ranking algorithms do not impact good or bad queries. While some models

manage to improve the performance of the good queries, queries that were categorized as performing badly over all configurations also performed badly over all ranking models. The RIA workshop [12] also reported that badly performing queries do not improve when ranked by a different retrieval model, so different components may be more important here.

5 Conclusion

The study has shown that for this particular academic search collection in a component-level evaluation:

- Document collection fields have an impact on success (but more text does not necessarily lead to better performance);
- Topic fields also have an impact (but longer queries tend to decrease performance);
- Applying a stopword list has a significant positive impact on search success;
- Stemming can have a significant impact depending on the chosen stemmer (Porter better than Krovetz);
- Different ranking algorithms based on term weights did not show a significant impact.

The study confirmed previous research that a variety of factors - and particularly their combined (interfering or compounding) influence - impact the search performance. While changing components did change the retrieval performance overall, it did not improve the performance of bad queries. As a matter of fact, the study could not identify a single aspect or component where bad queries would significantly be improved when changing the component.

At first view, these results are frustrating - no matter what component was looked at, a strong correlation between good or bad queries could not be found.

More work could be invested into analyzing different IR ranking models who could deal with documents with sparse and ambiguous text such as LSI or query enrichment strategies such as blind relevance feedback. However, first we will focus on query performance indicators, which delve deeper into the terminology of queries and documents to see whether we can identify badly performing queries this way [6].

Since this component-level evaluation was performed on one academic search collection, comparison with other test collections is necessary to extrapolate from the results achieved here to the domain in general. This will also be designated as future work.

References

1. Behnert, C., Lewandowski, D.: Ranking search results in library information systems - Considering ranking approaches adapted from web search engines. The Journal of Academic Librarianship 41(6), 725 – 735 (2015)
2. Carmel, D., Yom-Tov, E.: Estimating the query difficulty for information retrieval. Synth. Lect. on Inf. Concepts, Retrieval, and Services 2(1), 1–89 (2010)

3. Chowdhury, G.: Introduction to Modern Information Retrieval. Facet (2010)
4. Cleverdon, C.: The Cranfield tests on index language devices. In: Aslib proceedings. vol. 19, pp. 173–194. MCB UP Ltd (1967)
5. De Loupy, C., Bellot, P.: Evaluation of document retrieval systems and query difficulty. In: LREC 2000, Athens, Greece. pp. 32–39 (2000)
6. Dietz, F., Petras, V.: A Component-level Analysis of an Academic Search Test Collection. Part II: Query Analysis. In: CLEF 2017 (2017)
7. Ferro, N., Harman, D.: Grid@clef pilot track overview. In: CLEF 2009. pp. 552–565. Springer (2010)
8. Ferro, N., Silvello, G.: A general linear mixed models approach to study system component effects. In: SIGIR '16. pp. 25–34. ACM (2016)
9. Grivolla, J., Jourlin, P., de Mori, R.: Automatic classification of queries by expected retrieval performance. In: Predicting query difficulty workshop. SIGIR '05 (2005)
10. Han, H., Jeong, W., Wolfram, D.: Log analysis of an academic digital library: User query patterns. In: iConference 2014. iSchools (2014)
11. Hanbury, A., Müller, H.: Automated component–level evaluation: Present and future. In: CLEF 2010. pp. 124–135. Springer (2010)
12. Harman, D., Buckley, C.: Overview of the reliable information access workshop. Information Retrieval 12(6), 615–641 (2009)
13. Khabsa, M., Wu, Z., Giles, C.L.: Towards better understanding of academic search. In: JCDL '16. pp. 111–114. ACM (2016)
14. Kluck, M., Gey, F.C.: The domain-specific task of CLEF - Specific evaluation strategies in cross-language information retrieval. In: CLEF 2000. pp. 48–56. Springer (2001)
15. Kluck, M., Stempfhuber, M.: Domain-specific track CLEF 2005: Overview of results and approaches, remarks on the assessment analysis. In: CLEF 2005. pp. 212–221. Springer (2006)
16. Kürsten, J.: A Generic Approach to Component-Level Evaluation in Information Retrieval. Ph.D. thesis, Technical University Chemnitz, Germany (2012)
17. Li, X., Schijvenaars, B.J., de Rijke, M.: Investigating queries and search failures in academic search. Information Processing and Management 53(3), 666 – 683 (2017)
18. Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., Mutschke, P.: Bibliometric-enhanced information retrieval. In: ECIR 2014. pp. 798–801. Springer (2014)
19. McCarn, D.B., Leiter, J.: On-line services in medicine and beyond. Science 181(4097), 318–324 (1973)
20. Scholer, F., Garcia, S.: A case for improved evaluation of query difficulty prediction. In: SIGIR '09. pp. 640–641. ACM (2009)
21. Vanopstal, K., Buysschaert, J., Laureys, G., Stichele, R.V.: Lost in PubMed. Factors influencing the success of medical information retrieval. Expert Systems with Applications 40(10), 4106 – 4114 (2013)
22. Verberne, S., Sappelli, M., Kraaij, W.: Query term suggestion in academic search. In: ECIR 2014. pp. 560–566. Springer (2014)
23. Voorhees, E.M.: The TREC robust retrieval track. In: ACM SIGIR Forum. vol. 39, pp. 11–20 (2005)
24. Ware, M., Mabe, M.: The stm report: An overview of scientific and scholarly journal publishing (2015), http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf
25. Web of Science: Journal Citation Report. Thomson Reuters (2015)