A Component-level Analysis of an Academic Search Test Collection

Part II: Query Analysis

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 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_3

A Component-level Analysis of an Academic Search Test Collection Part II: Query Analysis

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 causes for successful and unsuccessful search performance in an academic search test collection. Based on a component-level evaluation setting presented in a parallel paper, analyses of the recall base and the semantic heterogeneity of queries and documents were used for performance prediction. The study finds that neither the recall base, query specificity nor ambiguity can predict the overall search performance or identify badly performing queries. A detailed query analysis finds patterns for negative effects (e.g. non-content-bearing terms in topics), but none are overly significant.

Keywords: academic search; query analysis; query prediction; GIRT04

1 Introduction

Several IR research areas study the aspects that contribute to a successful search performance of an IR system. Grid- or component-level evaluation initiatives [5, 6] focus on the large-scale and fine-grained evaluation of system component configurations for optimal retrieval. Query performance prediction studies [2] zoom in on individual pre- and post-retrieval indicators to predict the level of difficulty for queries in a particular system configuration to perform well [8, 9].

This paper reports on a search performance study combining both component-level evaluation approaches with query performance prediction on a particular academic search test collection. The study's objective is 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. The research question is stated as: "Which aspects in the search process best predict performance differences in an academic search collection?"

Academic search is defined as the domain of IR, which concerns itself with searching scientific data, mostly research output in the form of publications [12]. Both information needs and document collections present unique challenges in academic search: document collections are smaller than web search or newspaper collections; document and query terminology can be highly technical and domain-specific; documents may contain only sparse text (just metadata); and

information needs can be very specific with subsequent fewer relevant results in smaller collections [4].

In a parallel paper [4], we present the system configurations that were tested in a component-level setting, while in this paper, query performance indicators for an academic search collection are studied.

The paper is structured as follows: section 2 discusses relevant previous research on query performance indicators. The next section briefly describes the test collection GIRT4, a test collection representative for academic search, and the experimental set-up for the query performance analysis. Section 4 describes the factors that were analyzed for their predictive power in determining search performance. The paper concludes with some speculations on the nature of academic search and the performance success factors of IR systems in this context.

2 Query Performance Prediction

Query performance is an important topic in IR, as it influences the usefulness of an IR system - no matter in which domain. The field of query performance or query performance prediction addresses query behavior in IR. It is a known fact that some queries nearly always perform well while other examples fail in every circumstance [19]. Query performance prediction tries to distinguish between those good and bad queries without necessarily going through the actual retrieval process. The goal is to help make retrieval more robust so that a user always gets at least somewhat relevant results. In Carmel & Yom-Tov [2], the authors provide a broad overview of state of the art query performance predictors and an evaluation of published work.

The TREC Robust track was established to study systems that could deal robustly even with difficult queries [18, 20]. Unlike this paper, most of the research coming out of the Robust track attempted to predict the performance, but the causes for the performance variations are not focused on.

Query performance research can be divided into two different areas: pre-retrieval prediction and post-retrieval prediction. While pre-retrieval aspects are analyzed before the actual retrieval process and are much easier to compute, post-retrieval indicators are often found to be more precise, but are much more time consuming to calculate.

The inverse document frequency (IDF) of query terms is used quite often for pre-retrieval performance prediction. The average IDF (avgIDF) of a query and its relation to performance [3] takes the relative term weights of all query terms and averages them. The maximum IDF (maxIDF) per query is very similar as it takes the highest IDF per term of a query and matches it with the average precision [17]. Several IDF indicators will be tested in this study.

An example for an effective post-retrieval indicator is the Clarity Score [3], which measures the coherence of the vocabulary of the result list of a query and the vocabulary of the whole collection with regards to the query terms. The assumption is that if the language of the retrieved documents differs from the general language of the collection, the degree of difference is an indicator for performance. A pre-retrieval version of the Clarity Score is the Simplified Clarity

Score (SCS) [9], which compares term statistics of the query and the document collection, showing a strong correlation to performance for TREC collections. The SCS will also be calculated for this study.

Other approaches zoom in on the specific linguistic features of queries. Mothe & Tanguy [14] analyzed 16 different linguistic features of a query to predict the search performance. Among those were the number of words per query (i.e. query length), the average word length per query, the syntactic structure of a query or the average polysemy value. The latter averages the number of meanings all of the query terms can adopt. This indicator is based on the synsets from the linguistic database WordNet, which was used to get the number of synsets for every query term. The same procedure was also tested in this study.

This paper utilizes query prediction indicators not for performance prediction, but to identify which factors cause a good or bad performance. As Buckley [1] motivates, it is more important to try to identify root causes for query failure first than to come up with new approaches to improve the retrieval techniques for such queries - especially in domains where the failure causes are still relatively unsolved as is the case for academic search environments studied here.

While component-level evaluation as presented in another paper [4] focuses on the IR system components, but neglects to delve more deeply into query or document analyses, this paper delves deeper into the query and document analysis.

3 Study Configurations

3.1 Test Collection & Test IR System

For the analysis, the GIRT4 test collection, containing bibliographic metadata records on social science publications such as journal or conference articles, was used. The GIRT4 test collection consists of 151,319 documents with author, title, different types of keywords and abstract information and a 100 matching topics and relevance assessments, which were used in the CLEF domain-specific tracks from 2005-2008 [11]. The GIRT4 queries are TREC-style topics, consisting of three fields with different amounts of content (title, description and narrative). The relevance assessments were created with pooled retrieval runs from the CLEF domain-specific track. All three query fields were used for retrieval in all possible combinations. Three document fields were used: document title, controlled vocabulary (consisting of controlled keywords, controlled method terms and classification text) and the abstracts. Most of the document fields contain very little searchable content. Some documents contain more content than others, making them more likely to be retrieved. The test collection and component-level experimental setup is described in more detail in [4].

For the experiments, the open source retrieval kit Lemur[1] was used. For the component-level evaluation [4], nearly every possible option of Lemur in combination with the document and topic fields was used for retrieval, resulting in a total of around 327,600 data points. The retrieval experiments conducted

[1] https://www.lemurproject.org/lemur.php, last accessed: 04-30-2017

with Lemur were evaluated with the trec_eval program[2]. All experiments were performed on an Ubuntu operating system and automatically structured with Python scripts.

3.2 Analysis Steps

For evaluating the query performance on a topic-by-topic basis, the component-level approach from [4] proved to be the optimal basis. Analog to the first part, the average AP over every configuration was calculated to classify the topics by their respective performance. While some topics perform very well, others seem to have vast problems independent of the configuration that was used.

One aspect of the analysis addresses the problem of an insufficient recall base, i.e. the number of relevant documents per topic. The retrieval performance was compared to the number of relevant documents per topic to verify if the number of relevant documents impacts the search performance.

The linguistic specificity of the collection documents and queries and the query ambiguity were also examined. Specificity determines the uniqueness of a term for a collection [7]. Different values for query term specificity were calculated. The query length was also associated with its query's respective retrieval performance. As the narrative was not meant for retrieval, containing a lot of descriptive terms for the relevance assessors, only the lengths of title (short queries) and the combination of title and description (longer queries) were used for evaluation.

The next part delves deeper, using IDF measures. The IDF has been shown to have a relation to search performance, thus the measure was used in different variations to see if there are any connections to performance. Another measure already tested in other articles, the Simplified Clarity Score (SCS) was calculated for every query and compared to its AP [9].

The last part of the query analysis addressed the linguistic ambiguity of the query terms. For this, the WordNet database's synsets were utilized, as was suggested by Mothe & Tanguy [14]. Synsets are not synonyms, but the number of senses a word can have when used in natural language. As an example, a word like "put" has a lot of synsets and is thus considered more ambiguous, while a word like "soccer" has a small number of synsets, making it more specific. The ambiguity of a query is then measured by extracting the number of synsets for each query term. The results of these analysis parts were checked for significance by computing the Spearman rank correlation coefficient. The Simplified Clarity Score was also tested with the Pearson correlation efficient to compare the results to previous research. When one of the factors showed a significant correlation with the averaged AP per topic (over all configurations), we concluded that this aspect impacts the search performance.

4 Analyzing Query Performance

This section reports the results for specific query and collection aspects under consideration. Section 4.1 studies the impact of the recall base per topic. Sections

[2] http://trec.nist.gov/trec_eval, last accessed: 04-30-2017

4.2 and 4.3 focus on the linguistic features of queries and the collection by analyzing the impact of query specificity and ambiguity.

4.1 Recall Base

Buckley [1] states that the most important factor for search performance is probably the collection, i.e. the documents that are available. The recall base is the number of relevant documents in a collection available for a particular query. It is the basis for every successful search - without relevant documents, an information need cannot be fulfilled.

As a first step in the study, the available recall base for each topic was checked to see whether it had an influence on the performance. In the GIRT4 collection, some of the topics have very few relevant documents. Topic 218-DS[3], for example, has only three relevant documents within the GIRT4 collection and performs badly in any configuration.

To verify that a small recall base is not the reason for bad search performance, the average AP per topic over all configurations was correlated with the number of relevant documents. Figure 1 shows the distribution of averaged AP for the number of relevant documents per topic. The figure reveals no discernible correlation. The statistical verification also shows this trend, but is inconclusive (r_s=0.17, p<0.1).

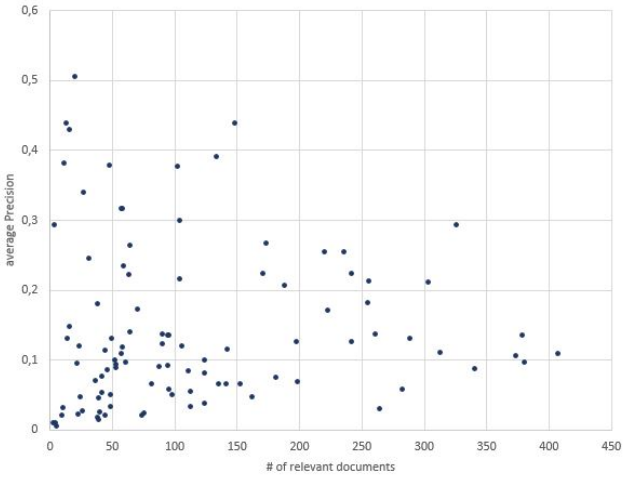

Fig. 1: Recall Base per Topic compared to AP

The explanation for search failure may not be the small number of relevant documents, but the content and structure of those relevant documents. Looking

[3] T: Generational differences on the Internet; D: Find documents describing the significance of the Internet for communication and the differences in how people of different ages utilize it.; N: Relevant documents describe the significance and function of the Internet for communication in different areas of society (private life, work life, politics, etc.) and the generation gap in using various means of communication.

at the stopped topic 218-DS (which has only three relevant documents in the collection), one can find the following terms: "generational differences internet significance internet communication differences people ages utilize significance function internet communication areas society private life work life politics generation gap means communication". The gist of this query concentrates on generational differences in using the internet. The narrative emphasizes internet use in everyday life. Looking at the content of the three relevant documents, they do not contain the term "internet", but use a different wording - "media", for example. The query cannot be matched with the relevant documents because those documents do not contain the terms used in the query. Vice versa, the collection contains about 30 other documents containing both "internet" and "generation", none of which was assessed as relevant. Vocabulary problems (synonymous, but unmatched term use in queries and documents) as well as insufficient relevance assessments might be the cause for the performance problems with this test collection.

4.2 Topic Specificity

Topic specificity refers to the uniqueness of the query terms in relationship to the document collection. It is two-sided: it considers the vocabulary that topics and documents share. Specificity has been found to influence the search performance [7] as it considers the number of matchable terms as well as the discriminatory nature of those terms in the retrieval process. Topic specificity can be measured in many different ways.

Query Length. The easiest way to calculate specificity is the query length. Intuitively, query length makes sense as an indicator for a good query. The more terms one issues to an IR system, the more accurate the results could be. However, this theoretical assumption does not hold true in every environment. For example, synonym expansion may add more noise to a query because of the inherent ambiguity of terms or because relevant documents only use one of the synonymous words.

In this study, query length was defined as the number of terms - other definitions calculate the number of characters a query consists of [14]. Query length based on the number of non-stop query terms was already tested on a TREC collection, showing no significant correlation with AP [9].

The average query length varies widely between the different topic fields [4]. Our analysis showed that query length correlates very weakly with AP over all configurations when all query terms are considered (r_s=-0.20, $p<0.005$) and still weakly, but slightly better, when stopwords are removed (r_s=-0.28, $p<0.005$). Stopword removal improves the predictive power of query length as only content-bearing words are considered. The negative correlation seems to support this - the longer the query is, the more general terms can be found. The amount of non-relevant words affects the performance.

Table 1 shows the impact of shorter and longer queries on AP by comparing queries using only the title (short query) or the title and description together (longer query) as was also calculated by [9]. The correlation is stronger once stopwords are removed from the queries, but very similar. The correlations for

stopped queries are somewhat stronger than what [9] found. It is possible that for academic search, query length may be a weak indicator for search performance.

Table 1: Spearman Rank Correlation for Query Length and AP for Shorter and Longer Queries (T=title, TD=title + description)

	all terms		w/o stopwords	
	T	TD	T	TD
r_s	-0.31	-0.19	-0.35	-0.37
p	0.001	0.046	0.000	0.000

Inverse Document Frequency. In contrast to query length, IDF has been shown to predict the search performance [7, 9]. An IDF measure shows how rare query terms might be in the collection. A somewhat rarer query term is more helpful for discriminating between relevant and non-relevant documents. IDF was calculated in its original form [16] with Python scripts, although it should not make any difference when computed in a modified form [15]. These indicators were calculated:

- maxIDF (maximum IDF of all query terms)
- minIDF (minimum IDF of all query terms)
- avgIDF (average of query term IDFs)
- devIDF (standard deviation per query term IDF)
- sumIDF (sum of all query term IDFs)

Table 2 lists the results for the correlation of the IDF indicators with AP after applying a stopword list while searching document title and the controlled vocabulary fields. There are no significant correlations between the various IDF indicators and AP, except avgIDF, which shows a very weak correlation. Although [7] reports maxIDF as a relatively stable indicator, the results here could not really be used to determine a query's search performance. The term weights alone do not provide enough information about the possible performance of a query.

Table 2: Spearman Rank Correlation for IDF Indicators and AP (using DT+CV documents fields, the Porter stemmer & a stopword list)

	maxIDF	minIDF	avgIDF	devIDF	sumIDF
r_s	0.14	0.08	0.24	0.15	-0.17
p	0.160	0.405	0.015	0.143	0.098

Simplified Clarity Score. As another measure of specificity, the Simplified Clarity Score (SCS), introduced by [9], was also tested. This pre-retrieval indicator is based on the Clarity Score [3], but focuses on the whole collection instead of a retrieved result list. It is calculated by the following formula:

$$\mathrm{SCS} = \sum_{Q} \mathrm{P_{ml}(w|Q)} \cdot \log_2\left(\frac{P_{\mathrm{ml}}(w|Q)}{P_{\mathrm{coll}}(w)}\right)$$

$P_{ml}(w|Q)$ is given by Qtf/QL, where Qtf is the frequency of a term w in query Q and QL is the query length. It is the maximum likelihood model of a query term w in query Q. $P_{coll}(w)$ defines the collection model and is given by $tf_{coll}/token_{coll}$, which is the term frequency of the term in the whole collection divided by the overall number of terms in the collection.

The SCS was calculated for eight different configurations. Four topic field combinations were tested against the document title and controlled vocabulary (DT+CV) fields, because this is the best document field combination for optimal retrieval performance [4]. Forms with and without stopword removal for each configuration were evaluated to test the impact of query length[4]. The correlation results for the SCS with the respective AP of all configurations with the tested topic fields are listed in table 3.

Table 3: Simplified Clarity Score per Topic Field and AP (Correlation with Spearman and Pearson)

	all terms				w/o stopwords			
	T	D	TD	TDN	T	D	TD	TDN
r_s	0.1	0.12	0.12	0.13	0.16	0.17	0.17	0.04
p	0.346	0.215	0.226	0.195	0.107	0.927	0.096	0.681
r_p	0.15	0.23	0.22	0.16	0.19	0.26	0.27	0.08
p	0.126	0.017	0.022	0.106	0.055	0.009	0.007	0.404

No matter which topic field combination is tested, the SCS of a query does not correlate with its performance using the Spearman Rank Correlation. We verified the results with the Pearson Correlation (which was used in [9]), although not all of the involved variables may have a linear distribution. These calculations show a weak correlation for stopped topics, but are smaller than those reported by [9] (except for the TDN topic field combination, which is not significant). This could be due to numerous reasons. One cause might be differences in collection size. Also, the average lengths of the topic fields in this study are longer than in [9], possibly weakening the effect of the SCS. Most likely, it is again the nature of the test collection, resulting in variations because of specific terms in this academic search context, which distort the calculations of term discriminativeness.

4.3 Semantic Heterogeneity of Topic Terms

Query ambiguity is another possible indicator for search performance. Both WordNet [14] and Wikipedia [10] were suggested as good resources to study the semantic heterogeneity of topic terms. This study used WordNet for analyzing topic term ambiguity. The count of synsets for every topic term and its respective part of speech (noun, verb or adjective) were downloaded. For each

[4] Note that the number of data points becomes smaller with every fixed factor such as DV+CV documents. Thus, the amount of data per query for testing SCS is smaller than in other calculations in this article. Still, the data is comparable to the amount of data used by [9].

topic, the average synsets (avgSyn) and the sum of synsets (sumSyn) were calculated and further grouped by the three word classes to get an indicator for query ambiguity (each also with and without stopword removal).

WordNet appeared to be the best option for academic queries, as it offers a good mixture of technical terms and normal language. Hauff [7] warns that Wordnet might not be reliable enough to use when predicting performance, because the database has problems with proper names and year dates. In this study, missing terms such as these were assigned the equivalent of one synset uniformly as dates and proper names (in this case mostly authors) would take on only one meaning.

The results for the average number of synsets per topic after stopword removal are shown in figure 2. The Spearman Rank Correlation test showed no significant correlation between AP and the average number of synsets (r_s=-0.09, $p<0.38$ w/o stopword removal; r_s=-0.13, $p<0.19$ with stopword removal).

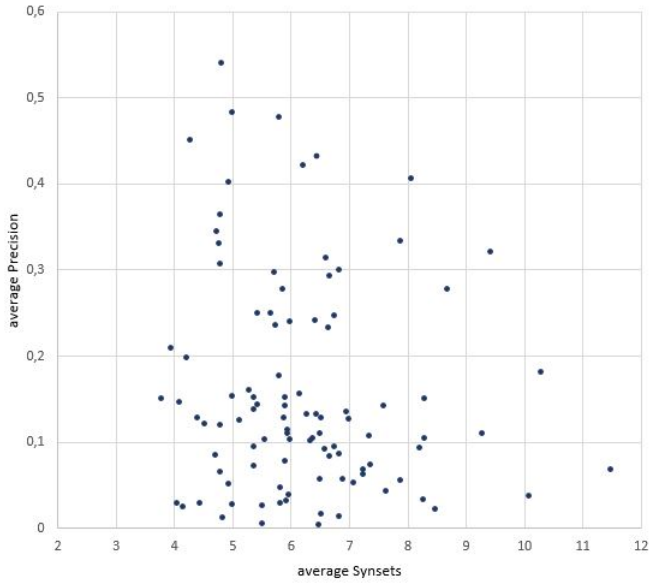

Fig. 2: Average Number of Synsets per Topic after Stopword Removal compared to AP

When comparing the absolute number of synsets over all topic terms, a slightly different image emerges as shown in figure 3. The correlation test revealed a weak correlation, which was slightly stronger after stopword removal (r_s=-0.25, $p<0.011$ w/o stopword removal; r_s=-0.28, $p<0.005$ with stopword removal). Similar observations appear when correlating the indicators of the individual word classes with the AP.

Query length does not significantly change the correlation. It was tested by comparing the results for shorter (T) and longer queries (TD) as in section 4.2. As before, the average synsets indicator does not have a significant correlation with performance. Regardless of query field or query length, the correlation is

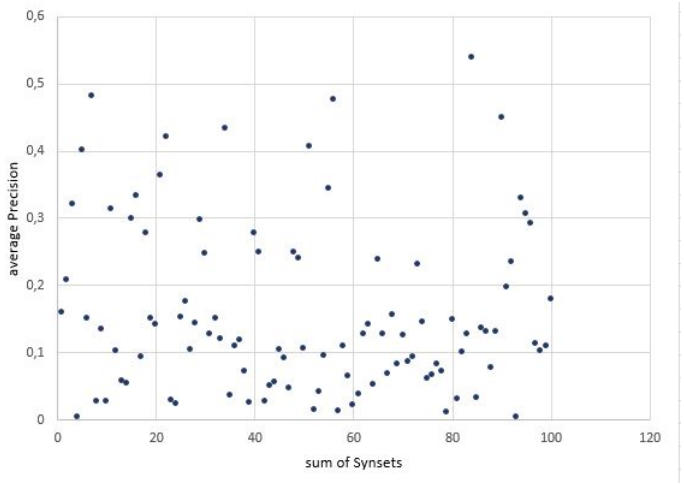

Fig. 3: Sum of Synsets per Topic after Stopword Removal compared to AP

not significant. For the sum of all synsets per query term, the correlation test reveals similar results compared to the sum of the synsets of all query terms - with shorter queries showing just the slightest increase in correlation strength ($T = r_s$=-0.29, $p<0.002$; $D = r_s$=-0.24, $p<0.015$; $TD = r_s$=-0.27, $p<0.006$).

Overall, there is a very weak correlation between the total number of synsets of a query and their respective AP over all configurations. This could be due to numerous causes. A reason might be that there is indeed no strong connection between the synsets of a query and the possible performance. A logical next step would be to test this with other sources or on different document fields. An assumption is that the controlled term fields contain more technical language and are therefore less ambiguous than the title or abstract fields.

5 Conclusion

The query analysis has shown that for this particular academic test collection:

- The recall base (number of relevant documents per topic) does not impact the search performance or distinguishes between good or bad topics;
- Query length could be a weak indicator for retrieval success as opposed to other studies using non-academic search test collections;
- Inverse document frequency in contrast to other studies does not predict search performance in any of its variations;
- The Simplified Clarity Score in contrast to other studies also does not predict search performance;
- The semantic heterogeneity (as expressed by the WordNet synsets) also does not predict the performance of the query.

For GIRT4 at least, the conventional query performance predictors could not distinguish between good and bad topics. Summarizing the results of the component-level analysis [4] and the query performance analysis presented here,

we could not identify a query aspect, test collection or IR system component which would help in distinguishing successful from unsuccessful queries, which would also explain search performance.

The analyses did show that the performance is dependent more on the query - collection combination than on any other IR process aspect, for example the ranking algorithm. The test collection has some aspects that differed from previously researched collections, showing the importance of domain-specific evaluation. In particular, document structure and the mixture of technical language with general language appeared an interesting research challenge when individual topic examples were studied. However, to validate this observation, more academic search test collections need to be analyzed with the same indicators to extrapolate to the whole domain.

GIRT4 is an example for a typical academic search collection with bibliographic metadata only, but even for academic search, it contains few documents, making it a particularly challenging example. Most academic search collections today contain at least the abstracts of the documents, GIRT4 only for 13% of the documents. More documents would certainly increase the recall base, but also introduce more language problems. In contrast to bibliographic documents, full-text academic search could be more related to other query performance research and could show different results. The recall base analyses of queries and their requisite relevant documents showed that without synonym expansion, queries might not match relevant documents especially because text is so sparse. The same analysis also showed that the test collection may also contain other relevant documents that were not assessed. A larger test collection (such as iSearch, which also contains full-text documents [13]) could alleviate some of these problems.

So what is the logical next step? For this type of search context - sparsity of text and technical language - the term matching aspect in the IR process is crucial. There are only few relevant documents available for any information need and they may contain other terminology than the queries. Future research could concentrate on the complex language problem of considering the collection and topics in tandem. For the IR process, ranking algorithms such as LSI and query feedback approaches could alleviate the synonym problem somewhat. For query performance indicators, future research might concentrate on semantic parsing and understanding of queries and documents that goes beyond the discrimination value of a term (as represented by IDF or SCS) or its language ambiguity (as represented by WordNet synsets) in the collection. Another idea would be to combine factors and look at the possible effects on performance, but those (average) effects shown here would probably still appear. A formal aspect that could also be changed was the distinction of good and bad queries: just taking the median of the averaged AP for all topics to distinguish between good and bad performing topics might be too coarse as a threshold.

The research presented here and in [4] may not yet be able to make representative statements about the nature of the academic search problem or list particular requirements for an IR system to improve performance, but it designed a comprehensive process for analyzing a domain-specific search problem, combining both component-level evaluation of system and collection configura-

tions and query performance analyses. This is a step towards a structured and holistic search process analysis for a specific domain, taking the searchers and their information needs, the collection and the IR system into account.

References

1. Buckley, C.: Why current IR engines fail. Inf. Retr. 12(6), 652–665 (2009)
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. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: SIGIR '02. pp. 299–306. ACM (2002)
4. Dietz, F., Petras, V.: A Component-level Analysis of an Academic Search Test Collection. Part I: System and Collection Configurations. In: CLEF 2017 (2017)
5. Ferro, N., Harman, D.: Grid@clef pilot track overview. In: CLEF 2009. pp. 552–565. Springer (2010)
6. Hanbury, A., Müller, H.: Automated component–level evaluation: Present and future. In: CLEF 2010. pp. 124–135. Springer (2010)
7. Hauff, C.: Predicting the Effectiveness of Queries and Retrieval Systems. Ph.D. thesis, University of Twente, Netherlands (2010)
8. Hauff, C., Azzopardi, L., Hiemstra, D.: The combination and evaluation of query performance prediction methods. In: ECIR 2009. pp. 301–312. Springer (2009)
9. He, B., Ounis, I.: Query performance prediction. Inf. Syst. 31(7), 585 – 594 (2006)
10. Katz, G., Shtock, A., Kurland, O., Shapira, B., Rokach, L.: Wikipedia-based query performance prediction. In: SIGIR '14. pp. 1235–1238. ACM (2014)
11. 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)
12. 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), http://doi.org/10.1016/j.ipm.2017.01.005
13. Lykke, M., Larsen, B., Lund, H., Ingwersen, P.: Developing a test collection for the evaluation of integrated search. In: ECIR'2010. pp. 627–630. Springer-Verlag (2010)
14. Mothe, J., Tanguy, L.: Linguistic features to predict query difficulty. In: Predicting query difficulty workshop. SIGIR '05. pp. 7–10 (2005)
15. Robertson, S.E.: Understanding inverse document frequency: on theoretical arguments for IDF. Journal of Documentation 60, 503–520 (2004)
16. Robertson, S.E., Sparck Jones, K.: Relevance weighting of search terms. In: Willett, P. (ed.) Document Retrieval Systems, pp. 143–160. Taylor Graham (1988)
17. Scholer, F., Williams, H.E., Turpin, A.: Query association surrogates for web search: Research articles. J. Amer. Soc. Inf. Sci. Techn. 55(7), 637–650 (2004)
18. Voorhees, E.M.: Overview of the TREC 2003 Robust Retrieval Track. In: TREC. pp. 69–77 (2003)
19. Voorhees, E.M.: The TREC robust retrieval track. In: ACM SIGIR Forum. vol. 39, pp. 11–20 (2005)
20. Yom-Tov, E., Fine, S., Carmel, D., Darlow, A.: Learning to estimate query difficulty: Including applications to missing content detection and distributed information retrieval. In: SIGIR '05. pp. 512–519. ACM (2005)