

Predicting the Effectiveness of Queries and Retrieval Systems

CLAUDIA HAUFF

Chapter 1

Introduction

The ability to make accurate predictions of the outcome of an event or a process is highly desirable in several contexts of human activity. For instance, when considering financial benefit, a very desirable prediction would be that of the successful anticipation of numbers appearing in an upcoming lottery. A successful attempt, in this context, can only be labeled a lucky guess, as previous lottery results or other factors such as the number of lottery tickets sold have no impact on the outcome (assuming a fair draw). Similarly, in terms of financial gain, a prediction on the stock market's behavior would also be very desirable. However, in this case, as opposed to lottery draws, the outcome can be predicted to some extent based on the available historical data and current economical and political events [28, 86, 172]. Notably, in both previous instances a rational agent may be highly motivated by the prospect of financial gain to make a successful guess on a future outcome but only in the latter are predictions to some extent possible.

In this work, the investigation theme is that of predictions and the factors that allow a measurable and consistent degree of success in anticipating a certain outcome. Specifically, two types of predictions in the context of information retrieval are set in focus. First, we consider users' attempts to express their information needs through queries, or search requests and try to predict whether those requests will be of high or low quality. Intuitively, the query's quality is determined by the outcome of the query, that is, whether the results meet the user's expectations. Depending on the predicted outcome, action can be taken by the search system in view of improving overall user satisfaction. The second type of predictions under investigation are those which attempt to predict the quality of search systems themselves. So, given a number of search systems to consider, these predictive methods attempt to estimate how well or how poorly they will perform in comparison to each other.

1.1 Motivation

Predicting the quality of a query is a worthwhile and important research endeavor, as evidenced by the significant amount of related research activity in recent years. Notably, if a technique allows for a quality estimate of queries in advance of, or

during the retrieval stage, specific measures can be taken to improve the overall performance of the system. For instance, if the performance of a query is considered to be poor, remedial action by the system can ensure that the users' information needs are satisfied by alerting them to the unsuitability of the query and asking for refinement or by providing a number of different term expansion possibilities.

An intuition of the above may be provided with the following simple example. Consider the query “jaguar”, which, given a general corpus like the World Wide Web, is substantially ambiguous. If a user issues this query to a search engine such as A9¹, Yahoo!² or Google³, it is not possible for the search engine to determine the user's information need without knowledge of the user's search history or profile. So only a random guess may be attempted on whether the user expects search results on the Jaguar car, the animal, the Atari video console, the guitar, the football team or even Apple's Mac OS X 10.2 (also referred to as Jaguar). When submitting the query “jaguar” to the Yahoo! search engine on August 31, 2009, of the ten top ranked returned results, seven were about Jaguar cars and three about the animal. In fact, most results that followed up to rank 500 also dealt with Jaguar cars; the first result concerning Mac OS X 10.2 could be found at rank 409. So a user needing information on the operating system would likely have been unable to acquire it. It is important to note that an algorithm predicting the extent of this ambiguity could have pointed out the unsuitability of the query and suggested additional terms for the user to choose from as some search engines do.

A query predicted to perform poorly, such as the one above, may not necessarily be ambiguous but may just not be covered in the corpus to which it is submitted [31, 167]. Also, identifying difficult queries related to a particular topic can be a valuable asset for collection keepers who can determine what kind of documents are expected by users and missing in the collection. Another important factor for collection keepers is the findability of documents, that is how easy is it for searchers to retrieve documents of interest [10, 30].

Predictions are also important in the case of well-performing queries. When deriving search results from different search engines and corpora, the predictions of the query with respect to each corpus can be used to select the best corpus or to merge the results across all corpora with weights according to the predicted query effectiveness score [160, 167]. Also, consider that the cost of searching can be decreased given a multiple partitioned corpus, as is common practice for very large corpora. If the documents are partitioned by, for instance, language or by topic, predicting to which partition to send the query saves time and bandwidth, as not all partitions need to be searched [12, 51]. Moreover, should the performance of a query appear to be sufficiently good, the query can be improved by some affirmative action such as automatic query expansion with pseudo-relevance feedback. In pseudo-relevance feedback it is assumed that the top K retrieved documents are relevant and so for a query with low effectiveness most or all of the top K documents would be irrelevant. Notably, expanding a poorly performing query leads to

¹<http://www.a9.com/>

²<http://www.yahoo.com/>

³<http://www.google.com/>

query drift and possibly to an even lower effectiveness while expanding queries with a reasonable performance and thus a number of relevant documents among the top K retrieved documents is more likely to lead to a gain in effectiveness. Another recently proposed application of prediction methods is to shorten long queries by filtering out predicted extraneous terms [94], in view of improving their effectiveness.

Cost considerations are also the prevalent driving force behind the research in predicting the ranking of retrieval systems according to their retrieval effectiveness without relying on manually derived relevance judgments. The creation of test collections, coupled with more and larger collections becoming available, can be very expensive. Consider that in a typical benchmark setting, the number of documents to judge depends on the number of retrieval systems participating. In the data sets used throughout this thesis, for example, the number of documents judged varies between a minimum of 31984 documents and a maximum of 86830 documents. If we assume that a document can be judged for its relevance within 30 seconds [150], this means that between 267 and 724 assessor hours are necessary to create the relevance judgments of one data set – a substantial amount.

Moreover, in a dynamic environment such as the World Wide Web, where the collection and user search behavior change over time, regular evaluation of search engines with manual assessments is not feasible [133]. If it were possible, however, to determine the relative effectiveness of a set of retrieval systems, reliably and accurately, without the need for relevance judgments, the cost of evaluation would be greatly reduced.

Correctly identifying the ranking of retrieval systems can also be advantageous in a more practical setting when relying on different retrieval approaches (such as Okapi [125] and Language Modeling [121]) and a single corpus. Intuitively, different types of queries benefit from different retrieval approaches. If it is possible to predict which of the available retrieval approaches will perform well for a particular query, the best predicted retrieval strategy can then be selected. Overall, this would lead to an improvement in effectiveness.

The motivation for this work is to improve user satisfaction in retrieval, by enabling the automatic identification of well performing retrieval systems as well as allowing retrieval systems to identify queries as either performing well or poorly and reacting accordingly. This thesis includes a thorough evaluation of existing prediction methods in the literature and proposes an honest appraisal of their effectiveness. We carefully enumerate the limitations of contemporary work in this field, propose enhancements to existing proposals and clearly outline their scope of use. Ultimately, there is considerable scope for improvement in existing retrieval systems if predictive methods are evaluated in a consistent and objective manner; this work, we believe, contributes substantially in accomplishing this goal.

1.2 Prediction Aspects

Evaluation of new ideas, such as new retrieval approaches, improved algorithms of pseudo-relevance feedback and others, is of great importance in information retrieval research. The development of a new model is not useful if the model does not substantially reflect reality and does not lead to improved results in a practical setting. For this reason, there are a number of yearly benchmarking events, where different retrieval tasks are used to compare retrieval models and approaches on common data sets. In settings such as TREC⁴, TRECVID⁵, FIRE⁶, INEX⁷, CLEF⁸, and NTCIR⁹, a set of topics t_1 to t_m is released for different tasks, and the participating research groups submit runs, that are ranked lists of results for each topic t_i produced by their retrieval systems s_1 to s_n . The performance of each system is determined by so-called relevance judgments, that is manually created judgments of results that determine a result's relevance or irrelevance to the topic. The retrieval tasks and corpora are manifold - they include the classic adhoc task [62], the entry page finding task [89], question answering [147], entity ranking [47] and others on corpora of text documents, images and videos.

The results returned thus depend on the task and corpus - a valid result might be a text document, a passage or paragraph of text, an image or a short video sequence. In this thesis, we restrict ourselves to collections of text documents and mostly the classical adhoc task.

For each pairing (t_i, s_j) of topic and system, one can determine a retrieval effectiveness value e_{ij} , which can be a measure such as average precision, precision at 10 documents, reciprocal rank and others [13]. The decision as to which measure to use is task dependent. This setup can be represented by an $m \times n$ matrix as shown in Figure 1.1. When relevance judgments are *not* available, it is evident from Figure 1.1 that the performances of four different aspects can be predicted. In previous work, all four aspects have been investigated by various authors and are outlined below. We also include in the list a fifth aspect, which can be considered as an aggregate of evaluation aspects EA1 to EA4.

(EA1) How difficult is a topic in general? Given a set of m topics and a corpus of documents, the goal is to predict the retrieval effectiveness or difficulty ranking of the topics *independent* of a particular retrieval system [7, 30], thus the topics are evaluated for their inherent difficulty with respect to the corpus.

⁴Text REtrieval Conference (TREC),

<http://trec.nist.gov/>

⁵TREC Video Retrieval Evaluation (TRECVID),

<http://trecvid.nist.gov/>

⁶Forum for Information Retrieval Evaluation (FIRE),

<http://www.isical.ac.in/~fire/>

⁷INitiative for the Evaluation of XML Retrieval (INEX),

<http://www.inex.otago.ac.nz/>

⁸Cross Language Evaluation Forum (CLEF),

<http://clef.iei.pi.cnr.it/>

⁹NII Test Collection for Information Retrieval Systems (NTCIR),

<http://research.nii.ac.jp/ntcir/>

- (EA2) How difficult is a topic for a particular system?** Given a set of m topics, a retrieval system s_j and a corpus of documents, the aim is to estimate the effectiveness of the topics given s_j . A topic with low effectiveness is considered to be difficult for the system. This is the most common evaluation strategy which has been investigated for instance in [45, 71, 167, 175].
- (EA3) How well does a system perform for a particular topic?** Given a topic t_i , n retrieval systems and a corpus of documents, the systems are ranked according to their performance on t_i . This approach is somewhat similar to aspect EA4, although here, the evaluation is performed on a per topic basis rather than across a set of topics [50].
- (EA4) How well does a system perform in general?** Given a set of n retrieval systems and a corpus of documents, the aim is to estimate a performance ranking of systems independent of a particular topic [9, 114, 133, 135, 161].
- (EA5) How hard is this benchmark for all systems participating?** This evaluation aspect can be considered as an aggregate of the evaluation aspects EA1 to EA4.

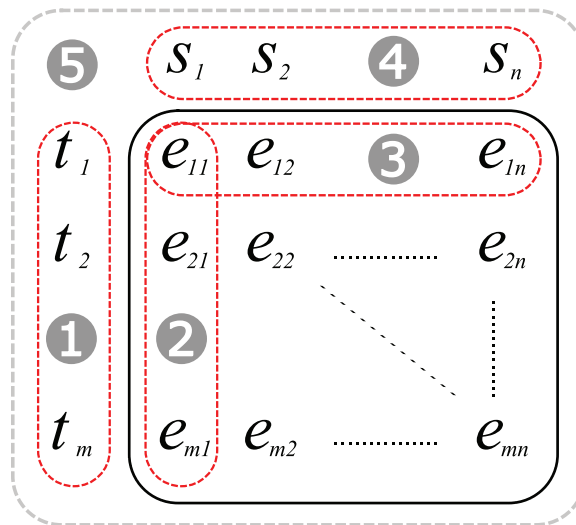


Figure 1.1: A matrix of retrieval effectiveness values; e_{ij} is the retrieval effectiveness, system s_j achieves for topic t_i on a particular corpus. The numbered labels refer to the different aspects (label 1 corresponds to EA1, etcetera).

A topic is an expression of an information need – in the benchmark setting of TREC it usually consists of a title, a description and a narrative. A query is a formulation of the topic that is used in a particular retrieval system. Often, only the title part of the topic is used and a formulation is derived by, for instance, stemming and stopword removal or the combination of query terms by Boolean operators. As the input to a retrieval system is the formulation of an information need, that is a query, this concept is often expressed as query performance or query effectiveness prediction.

Search engines, even if they perform well on average, suffer from a great variance in retrieval effectiveness [63, 151, 152], that is, not all queries can be answered with equal accuracy. Predicting whether a query will lead to low quality results is a challenging task, even for information retrieval experts. In an experiment described by Voorhees and Harman [148], a number of researchers were asked to classify a set of queries as either easy, medium or difficult for a corpus of newswire articles they were familiar with. The researchers were not given ranked lists of results, just the queries themselves. It was found that the experts were unable to predict the query types correctly and, somewhat surprisingly, they could not even agree among themselves how to classify the queries¹⁰. Inspired by the aforementioned experiment, we performed a similar one with Web queries and a Web corpus; specifically, we relied on the newly released ClueWeb09 corpus and the 50 queries of the TREC 2009 Web adhoc task. Nowadays (as opposed to the late 90’s time frame of [148]), Web search engines are used daily, and instead of information retrieval experts, we relied on members of the Database and the Human Media Interaction group of the University of Twente, who could be considered expert users. Thirty-three people were recruited and asked to judge each of the provided queries for their expected result quality. The users were asked to choose for each query one of four levels of quality: *low*, *medium* and *high* quality as well as *unknown*. The quality score of each query is derived by summing up the scores across all users that did not choose the option *unknown* where the scores 1, 2 and 3 are assigned to *low*, *medium* and *high* quality respectively. Note that a higher quality score denotes a greater expectation by the users that the query will perform well on a Web search engine. Named entities such as “volvo” and “orange county convention center” as well as seemingly concrete search request such as “wedding budget calculator” received the highest scores. The lowest scores were given to unspecific queries such as “map” and “the current”. The correlation between the averaged quality scores of the users and the retrieval effectiveness scores of the queries evaluated to $r = 0.46$. This moderate correlation indicates, that users can, to some extent, predict the quality of search results, though not with a very high accuracy, which denotes the difficulty of the task¹¹.

In recent years many different kinds of predictions in information retrieval have been investigated. This includes, for instance, the prediction of a Web summary’s quality [83] and of a Q&A pair’s answer quality [22, 81], as well as the prediction of the usefulness of involving the user or user profile in query expansion [92, 93, 137]. Further examples in this vain include predicting the effect of labeling images [85], predictions on the amount of external feedback [53] and predicting clicks on Web advertisements [18] and news results [88].

In this thesis we focus specifically on predicting the effectiveness of informational queries and retrieval systems, as we believe that these two aspects will bring about the most tangible benefits in retrieval effectiveness and in improvement of user satisfaction, considering that around 80% of the queries submitted to the Web are

¹⁰The highest linear correlation coefficient between an expert’s predictions and the ground truth was $r = 0.26$, the highest correlation between any two experts’ predictions was $r = 0.39$.

¹¹The user study is described in more detail in Appendix A.

informational in nature [78]. Furthermore, the works cited above depend partially on large-scale samples of query logs or interaction logs [92, 137] which cannot be assumed to be available to all search systems.

1.3 Definition of Terms

As of yet there is no widely accepted standard terminology in this research area. Depending on the publication forum and the particular author different phrases are used to refer to specific evaluation aspects. As can be expected the same term can also have different meanings in works by different authors. In this section, we explicitly state our interpretation of ambiguous terms in the literature and we use them consistently throughout.

Firstly, while generally - and also in this thesis - *to predict* and *to estimate* a query’s quality are used interchangeably, in a number of works in the literature a distinction is made between the two. *Predicting* the quality of a query is used when the algorithms do *not* rely on the ranked lists of results, while the quality of a query is *estimated* if the calculations are based on the ranked list of results. In this work the meaning becomes clear in the context of each topic under investigation, it is always explicitly stated whether a ranked list of results used.

Throughout, the term *query quality* means the *retrieval effectiveness* that a query achieves with respect to a particular retrieval system, which is also referred to as *query performance*. When we investigate *query difficulty* we are indirectly also interested in predicting the retrieval effectiveness of a query, however we are only interested whether the effectiveness will be low or high. We thus expect a binary outcome - the query is either classified as easy or it is classified as difficult. In contrast, when investigating *query performance prediction* we are interested in the predicted effectiveness score and thus expect a non-binary outcome such as an estimate of average precision.

EA1	collection query hardness [7], topic difficulty [30]
EA2	query difficulty [167], topic difficulty [30], query performance prediction [175], precision prediction [52], system query hardness [7], search result quality estimation [40], search effectiveness estimation [145]
EA3	performance prediction of “retrievals” [50]
EA4	automatic evaluation of retrieval systems [114], ranking retrieval systems without relevance judgments [133], retrieval system ranking estimation
EA5	-

Table 1.1: Overview of commonly used terminology of the evaluation aspects of Figure 1.1.

In Table 1.1 we have summarized the expressions for each evaluation aspect as they occur in the literature. Evaluation aspect EA2 has the most diverse set of labels, as it is the most widely evaluated aspect. Most commonly, it is referred to as *query*

performance prediction, *query difficulty* as well as *query effectiveness estimation*. Evaluation aspect EA3 on the other hand has only been considered in one publication so far [50], where the pair (t_i, s_j) of topic and system is referred to as “retrieval”.

Aspect EA4 of Figure 1.1 was originally referred to as *ranking retrieval systems without relevance judgments*, but has also come to be known as *automatic evaluation of retrieval systems*. We refer to it as *retrieval system ranking estimation* as in this setup we attempt to estimate a ranking of retrieval systems.

1.4 Research Themes

The analysis presented in this thesis aims to provide a comprehensive picture of research efforts in the prediction of query and retrieval system effectiveness. By organizing previous works according to evaluation aspects we methodically clarify and categorise the different dimensions of this research area. The thesis focuses on two evaluation aspects (enumerated in full in Section 1.2), in particular, EA2 and EA4, as their analysis has value in practical settings as well as for evaluation purposes.

The other aspects are not directly considered. Evaluation aspect EA1, which assumes the difficulty of a topic to be inherent to a corpus, is mainly of interest in the creation of benchmarks, so as to, for instance, choose the right set of topics. In an adaptive retrieval system, where the system cannot choose which queries to answer it is less useful. The same argumentation applies intuitively to aspect EA5. As part of the work on evaluation aspect EA4 in Chapter 5 we will briefly discuss EA3.

Four main research themes are covered in this work and will now be explicitly stated. The first three (**RT1**, **RT2** and **RT3**) are concerned with evaluation aspect EA2, while the last one (**RT4**) is concerned with evaluation aspect EA4 (and partly with EA3). The backbone of all results reported and observations made in this work form two large-scale empirical studies. In Chapter 2 and Chapter 3 a total of twenty-eight prediction methods are evaluated on three different test corpora. The second study, discussed in detail in Chapter 5 puts emphasis on the influence of the diversity of data sets: therein five system ranking estimation approaches are evaluated on sixteen highly diverse data sets.

RT1: Quality of pre-retrieval predictors Pre-retrieval query effectiveness prediction methods are so termed because they predict a query’s performance before the retrieval step. They are thus independent of the ranked list of results. Such predictors base their predictions solely on query terms, the collection statistics and possibly external sources such as WordNet [57] or Wikipedia¹². In this work we analyze and evaluate a large subset of the main approaches and answer the following questions: on what heuristics are the prediction algorithms based? Can the algorithms be categorized in a meaningful way? How similar are different approaches to

¹²<http://www.wikipedia.org/>

each other? How sensitive are the algorithms to a change in the retrieval approach? What gain can be achieved by combining different approaches?

RT2: The case of the post-retrieval predictor Clarity Score The class of post-retrieval approaches estimates a query’s effectiveness based on the ranked list of results. The approaches in this class are usually more complex than pre-retrieval predictors, as more information (the list of results) is available to form an estimate of the query’s effectiveness. Focusing on one characteristic approach, namely Clarity Score [45], the questions we explore are: how sensitive is this post-retrieval predictor to the retrieval algorithm? How does the algorithm’s performance change over different test collections? Is it possible to improve upon the prediction accuracy of existing approaches?

RT3: The relationship between correlation and application The quality of query effectiveness prediction methods is commonly evaluated by reporting correlation coefficients, such as Kendall’s Tau [84] and the linear correlation coefficient. These measures denote how well the methods perform at predicting the retrieval performance of a given set of queries. The following essential questions have so far remained unexplored: what is the relationship between the correlation coefficient as an evaluation measure for query performance prediction and the effect of such a method on retrieval effectiveness? At what levels of correlation can we be reasonably sure that a query performance prediction method will be useful in a practical setting?

RT4: System ranking estimation Substantial research work has also been undertaken in estimating the effectiveness of retrieval systems. However, most of the evaluations have been performed on a small number of older corpora. Current work in this area lacks a broad evaluation scope which gives rise to the following questions: is the performance of system ranking estimation approaches as reported in previous studies comparable with their performance on more recent and diverse data sets? What factors influence the accuracy of system ranking estimation? Can the accuracy be improved when selecting a subset of topics to rank retrieval systems?

1.5 Thesis Overview

The organization of the thesis follows the order of the research themes. In Chapter 2, we turn our attention to pre-retrieval prediction algorithms and provide a comprehensive overview of existing methods. We examine their similarities and differences analytically and then verify our findings empirically. A categorization of algorithms is proposed and the change in predictor performance when combining different approaches is investigated. The major results of this chapter have previously been published in [65, 69].

In Chapter 3, post-retrieval approaches are introduced, with a focus on Clarity Score for which an improved variation is proposed and an explanation is offered as to why some test collections are more difficult for query effectiveness estimation than others. Part of this work can also be found in [70].

The connection between a common evaluation measure (Kendall's Tau) in query performance prediction approaches and the performance of retrieval systems relying on those predictions is evaluated in Chapter 4. Insights in the level of correlation required in order to ensure that an application of a predictor in an operational setting is likely to lead to an overall improvement in the retrieval system are reported. Parts of this chapter have been described in [64, 67].

Chapter 5 then focuses on system ranking estimation approaches. A number of algorithms are compared and the hypothesis that subsets of topics lead to a better performance of the approaches is evaluated. The work of this chapter was initially presented in [66, 68].

The thesis concludes with Chapter 6 where a summary of the conclusions is included and suggestions for future research are offered.