Using a Medical Thesaurus to Predict Query Difficulty

Florian Boudin[1], Jian-Yun Nie[2], and Martin Dawes[3]

[1] Université de Nantes, florian.boudin@univ-nantes.fr
[2] Université de Montréal, nie@iro.umontreal.ca
[3] University of British Columbia, martin.dawes@ubc.ca

Abstract. Estimating query performance is the task of predicting the quality of results returned by a search engine in response to a query. In this paper, we focus on pre-retrieval prediction methods for the medical domain. We propose a novel predictor that exploits a thesaurus to ascertain how difficult queries are. In our experiments, we show that our predictor outperforms the state-of-the-art methods that do not use a thesaurus.

1 Introduction

The task of predicting query difficulty is to quantify, for a given query, the expected quality of a list of documents returned by an Information Retrieval (IR) system. Even for systems that perform very well on average, the quality of results is known to be poor for some of the queries [6]. Ideally, a system that can predict difficult queries can adapt its search parameters to suit the query. It can also give feedback to the user, for example by reporting confidence scores or showing hints to reformulate a more precise query.

Methods for predicting query performance fall into two groups: pre-retrieval and post-retrieval approaches. Pre-retrieval methods analyze the query expression and try to determine its performance without performing an actual retrieval. However, the query alone is often too short for a reliable prediction. On the other hand, post-retrieval approaches analyze the retrieval results to predict query difficulty. Post-retrieval methods are known to perform better than pre-retrieval methods, but they also require an additional retrieval process, which is often too costly to do in practice.

In this paper, we study the task of predicting query difficulty in the medical domain. Different from most proposed approaches, we take advantage of the existing domain-specific resources to analyze the query expression and ascertain how difficult this query is from a retrieval system point of view.

2 Related Work

Query performance prediction have received much attention recently and many different approaches have been proposed. In this paper, we concentrate on pre-retrieval predictors. This section presents the previous work relevant to this type of difficulty estimation.

R. Baeza-Yates et al. (Eds.): ECIR 2012, LNCS 7224, pp. 480–484, 2012.

Several proposed predictors on query difficulty are based on the Inverse Document Frequency (IDF) measure. The average IDF of query terms was used by Cronen-Townsend et al. [3]. According to this predictor, the more discriminative the query terms are on average, the better the query will perform. Later, Scholer et al. [5] reported that using the maximum IDF value of any term in a query gives the best correlation on the TREC web data. These results were confirmed and extended to other TREC collections [7].

He and Ounis [4] proposed and evaluated a variety of pre-retrieval predictors. Their experiments show that predictors measuring the divergence between a query model and a collection model have the strongest correlation with query performance. If the models are disparate, then the query identifies a subset of documents which is likely to be relevant to the query.

More recently, Zhao et al. [7] presented two families of pre-retrieval predictors. The first is based on the similarity between a query and the overall document collection, the second focuses on the variability in how query terms are distributed across documents. Results show that these predictors give more consistent performance across a variety of data types and search tasks. However, in most previous studies, one usually assumed that no general ontology or thesaurus of sufficient coverage is available. Few studies have examined the use and impact of a thesaurus on the prediction of query difficulty. Indeed, even if a good thesaurus of wide coverage does not exist for general domain, in specialized domains such as medicine, we do have high quality ontologies and thesauri. It is thus possible to examine the impact of such a thesaurus for the prediction of query difficulty in a specialized domain.

3 Predicting Query Difficulty

In this work, we use the Medical Subject Headings (MeSH)[1] thesaurus. MeSH is a comprehensive controlled vocabulary for which the main purpose is for indexing journal articles and books in the life sciences. The 2011 version contains a total of 26,142 subject headings (also known as descriptors) arranged in a hierarchy. Most of these are accompanied by a list of entry terms (i.e. orthographic variants; "*Vitamin C*", for example, is an entry term to "*Ascorbic Acid*").

Our predictor tries to exploit two different aspects of query difficulty: term variability and term generality/specificity. Our general assumption is that if a query contains concepts that are highly variable in their expressions and very general, then the query tends to be more difficult. The medical terminology is known to be precise and well defined. However, the increasing production of literature in the medical domain has contributed to raise the number of terms that may refer to the same concept. The most-used term referring to each query concept should then be used to maximize the retrieval performance. Here, we use MeSH to estimate the level of usage of a term in relation to its concept.

The second aspect we wanted to capture is the hypernymy/hyponymy level of query terms. Queries containing broader terms are more difficult to answer

[1] `http://www.nlm.nih.gov/mesh/`

because the size of the expected result set is much larger. We use the hierarchical structure of MeSH to differentiate between narrower and broader terms. A narrower term is defined as close to a leaf of the concept tree. Given a query $Q = (t_1, t_2, \cdots t_n)$, our query difficulty predictor is computed as:

$$\textit{MeSH-QD}(Q, \mathcal{T}) = \sum_{t \in Q} \overbrace{\frac{df(t)}{\sum\limits_{t' \in V(t)} df(t')} \cdot ln\Big(1 + \frac{N}{df(t)}\Big)}^{\text{term variability}} \cdot \overbrace{\frac{\text{depth}(t)}{\text{length}(t)}}^{\text{term generality}} \tag{1}$$

with $df(t)$ the number of documents containing the term t, $V(t)$ the set of alternative expressions of the concept t in the thesaurus, N the size of the collection, depth(t) the depth of t in the concept tree, and length(t) the maximum depth of the branch containing t. The higher *MeSH-QD*, the less the query difficulty.

4 Experimental Settings

In this study, we evaluate the above new predictor using the CLIREC test collection [1], made of 155 clinical queries, 2596 relevance judgments and 1.5 million documents extracted from PubMed[2], one of the most searched medical resources.

In order to use our predictor, query terms have to be mapped to MeSH. However, mapping terms to an existing resource is a difficult task. Spelling problems, synonyms, or term ambiguity are some of the difficulties that can introduce errors. To estimate the impact of the mapping quality on the performance of our prediction method, we performed this process both manually and automatically. In the manual mapping, two annotators were asked to map query terms to MeSH. 78 queries were fully mapped at the phrase level. All the experiments in this study are conducted on this subset of queries. We used Metamap[3] to perform the automatic mapping. In comparison to the manual mapping, Metamap achieves a recall of 83.2% and a precision of 85.4%.

We evaluate prediction quality by measuring the correlation between the actual performance of queries (as determined by using relevance judgments) and the difficulty scores assigned by the predictors. In previous work, two evaluation methodologies were used, comparing prediction scores with individual retrieval models (e.g. [7]) or with the average performance of several models (e.g. [2]). In this study, we use the latter. Retrieval tasks are performed using the Lemur toolkit[4]. We experiment with three retrieval models: tf.idf, Okapi BM25 and a language modeling approach with Dirichlet prior smoothing ($\mu = 2000$). Retrieval accuracy is evaluated in terms of Mean Average Precision (MAP).

Three correlation coefficients are commonly used in the query difficulty estimation literature: Pearson product-moment correlation, Spearman's rank order correlation and Kendall's tau. As there is currently no consensus on which correlation measure is the most appropriate, all the three measures are reported.

[2] `http://www.pubmed.com`
[3] `http://metamap.nlm.nih.gov`
[4] `http://www.lemurproject.org`

Table 1. Correlation tests between query performance (MAP scores) and each prediction method († and ‡ indicate significance at the 0.01 and 0.001 levels respectively)

Predictor	**Kendall (τ)**	**Pearson (ρ_1)**	**Spearman (ρ_2)**
MaxIDF	0.114	0.010	0.164
MaxSCQ	0.292‡	0.215	0.433‡
MeSH-QD mapping auto	0.321‡	0.402‡	0.454‡
MeSH-QD mapping man	0.368‡	0.294†	0.532‡

5 Results

We compared the prediction quality of our predictor with that of two other pre-retrieval predictors: the maximum IDF value of the query terms (*MaxIDF*) [3] and the Maximum of Collection Query Similarity (*MaxSCQ*) [7]. They were shown to be very effective on the TREC data. Results are reported in Table 1. Overall, results show that the proposed method (*MeSH-QD*) can provide useful information for predicting how well a query will perform. *MeSH-QD*, either with manual or automatic mapping, outperforms the two baselines. We observe that the errors committed by the automatic query mapping have a strong impact on two of the three correlation measures. However, the performance of Metamap, although not perfect, may be sufficient for a query difficulty prediction task.

6 Conclusions

We have introduced a new domain-specific pre-retrieval predictor that uses MeSH to ascertain how difficult queries are. The proposed prediction method is based on estimating query variability and query specificity. Experiments show that the proposed method outperforms existing predictors.

The performance of the proposed predictor depends on many factors such as the coverage of the thesaurus or the query mapping quality. In future work, we intend to study the impact of these factors on the proposed predictor. We also want to extend this work to other domains where resources are available.

References

1. Boudin, F., Nie, J.-Y., Dawes, M.: Positional Language Models for Clinical Information Retrieval. In: Proceedings of the EMNLP Conference (2010)
2. Carmel, D., Yom-Tov, E., Darlow, A., Pelleg, D.: What makes a query difficult? In: Proceedings of the SIGIR Conference (2006)
3. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the SIGIR Conference (2002)
4. He, B., Ounis, I.: Query performance prediction. Information Systems (2006)

5. Scholer, F., Williams, H.E., Turpin, A.: Query association surrogates for web search. JASIST (2004)
6. Voorhees, E.M.: Overview of the TREC 2004 robust retrieval track. In: Proceedings of the TREC Conference (2004)
7. Zhao, Y., Scholer, F., Tsegay, Y.: Effective Pre-retrieval Query Performance Prediction using Similarity and Variability Evidence. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 52–64. Springer, Heidelberg (2008)