Improving Retrieval Accuracy by Weighting Document Types with Clickthrough Data

Peter C. K. Yeung, Charles L. A. Clarke, Stefan Büttcher
University of Waterloo
Waterloo, Canada
{p2yeung, claclark, sbuettch}@plg.uwaterloo.ca

ABSTRACT

For enterprise search, there exists a relationship between work task and document type that can be used to refine search results [3]. In this poster, we adapt the popular Okapi BM25 scoring function to weight term frequency based on the relevance of a document type to a work task. Also, we use click frequency for each task-type pair to estimate a realistic weight. Using the W3C collection from the TREC Enterprise track for evaluations, our approach leads to significant improvements on search precision.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Search process

General Terms

Performance, Human Factors

Keywords

Enterprise Search, Clickthrough Data

1. INTRODUCTION

Recently, Robertson et al. [4] introduced a modified version of Okapi BM25 to incorporate *weights* into different fields of a structured document. The intuition is to consider structured documents and rank them according to the importance of each structure. Although the modified Okapi BM25 was intended for weighting fields of a structured document, it can be used to weight another useful piece of information of a document: *document type*. This poster also introduces an approach to use *clickthrough data* to estimate a realistic weight for each *work task-document type* pair.

Previous research [3] has shown that there exists a relationship between work task and document type (or genre) in an enterprise search environment. A *document genre* is a class of documents, grouped together based on similar subject, form, and content. For our purpose, we would consider *document type*. A document type defines the source of a document (i.e., WWW pages, emails, discussion threads, etc). If a user's work task is known to a retrieval system, retrieval accuracy can be improved by returning documents from those relevant types and ranking them higher in the result list. Therefore, document type is an important factor to consider in the retrieval process.

Clickthrough data is a history about user-submitted queries and user-selected documents on the corresponding search result page. Although clickthrough data does not provide direct indication on document relevance, it provides useful hints for determining which document (or type of documents) is relevant to a user's need. Many different approaches of utilizing clickthrough data to improve retrieval performance have previously been proposed (e.g. [1]). In this poster, we take a simpler approach of utilizing clickthrough data in the retrieval process.

In our approach, clickthrough data are grouped together based on different task-type pairs. To determine the weight for each task-type pair, we consider the click frequency of the document type when the work task was given. For example, given a work task, if type A is clicked more frequently than type B, then type A's weight would be larger than type B's. Depending on the document's type and on the given work task, we apply the corresponding weight to the modified BM25 to compute the relevance score.

2. WEIGHTING DOCUMENT TYPES

The extended version of Okapi BM25 outputs a relevance score for each document by computing a linear combination of term frequencies and field weights. For query terms $Q_1, Q_2, ..., Q_n$, the weighted BM25 relevance score of a document D is

$$S^{(D)}_{BM25F} = \sum_{i=1}^{n} w_{Q_i} * \frac{(k_1+1) * f'_{D,Q_i}}{f'_{D,Q_i} + k_1 * ((1-b) + b * \frac{|D|}{avgdl})} \quad (1)$$

where $|D|$ is the length of D, and $avgdl$ is the average document length. k_1 $(= 1.2)$ and b $(= 0.75)$ are free parameters. w_{Q_i} is the *inverse document frequency* weight. f'_{D,Q_i} is the weighted term frequency of Q_i. It is a combination of its unweighted frequency f_{D,Q_i} and the corresponding weight w_j. Suppose that there are N different fields to be weighted,

$$f'_{D,Q_i} = \sum_{j=1}^{N} w_j * f_{D,Q_i} \quad (2)$$

For our purpose, *document type* is the only field that would be weighted. If work task is known, a search system can use the corresponding set of weights for document types to calculate relevance scores.

To determine a realistic estimate of the weight for each

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

document type, we consider click frequency for each document type and work task. Each weight should have these properties:

- the weight is one if click frequency is zero;
- the weight increases monotonically with click frequency; and
- the weight increases to an asymptotic maximum.

Given a work task, assume cf_T represents click frequency of a document type T. A rough model for estimating each weight can be formulated as

$$w_T = |T| * \frac{cf_T + S}{|C| + |T|S} + 1 \tag{3}$$

where $|T|$ is the number of types, $|C|$ is the total number of clicks, and S $(= 1.5)$ is a smoothing parameter.

First, if cf_T is zero, then $w_T \approx 1$ (assume S is relatively small). Second, equation 3 is linear, thus, w_T increases monotonically as click frequency increases. Finally, if a particular document type dominates the clicks, cf_T would equal to $|C|$, which means w_T would have a value close to $|T| + 1$. Hence, the weight increases to an asymtotic maximum. Equation 3 satisfies all properties listed above.

Given the weight of each document type for a specific work task, the weighted term frequency is

$$f'_{D,Q_i} = w_T * f_{D,Q_i}. \tag{4}$$

3. EXPERIMENTAL RESULTS

For our experiments, we employ the W3C collection used in the TREC 2006 Enterprise track [2]. The W3C collection contains 331,037 documents with a total uncompressed size of 5.7 gigabytes. These documents are categorized into six different types: mailing lists, public CVS repository, public Web pages, wiki pages, personal pages for the W3C team, and other pages.

The evaluation is limited by the nature of the TREC Enterprise track. Since the queries were used by the *expert search* task in TREC Enterprise track, they were created with the objective of finding an expert for a particular topic. Thus, there is only one work task—*expert search* task—corresponding to these queries. Our objective is to find relevant document type(s) for the expert search task and rank documents from this type higher to improve search precisions.

We utilized the clickthrough data that were used during the creation of the evaluation topics. The problem is that some clicked-on documents were also listed in the qrels file. It is inappropriate to train our models using these clickthrough data and then evaluate our models using the topics and the qrels file. Therefore, we removed them from our experiments and used only the ones where the clicked-on document is not identified in the qrels file.

Table 1 shows that BM25+CF increases search precision at 5 documents from 0.6286 to 0.7469, a 19% improvement. Our model is statistically significant over BM25 for precision at 5 and 10 documents.

Table 2 shows the number of documents retrieved in each document type for all query topics. The BM25 model mainly retrieved documents from public Web pages and mailing lists, along with a small amount of documents from the other 4 types. However, BM25+CF retrieved the majority of its documents from mailing lists. Therefore, for the expert search task, *mailing lists* is a more relevance document type and retrieval performance can be improved by placing more weights on its documents.

Model	P@5	P@10	P@15	P@20
BM25	0.6286	0.6143	0.6082	0.6031
BM25+CF	**0.7469**	**0.6939**	0.6653	0.6459

Table 1: Precisions for BM25 and BM25+CF.

Type	BM25	BM25+CF
Web pages	21129	444
mailing lists	22043	51930
CVS repositories	4927	0
wiki pages	4078	178
people	4	0
other	353	0

Table 2: Number of Documents Retrieved for Each Document Type.

4. CONCLUSIONS

We have proposed a fundamental approach for weighting document types and estimating the appropriate weight for each document type using click frequency. Click frequency is an indication of users' judgments on each type for the work task. Thus, it is a helpful source for estimating the weights. Our model incorporates these weights to determine a weighted term frequency, which is then used to compute relevance scores. In our experiments, the model improves P@5 by 19%, compared to a BM25 baseline. The improvement is statistically significant according to a paired t-test (confidence level: 95%).

5. REFERENCES

[1] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA, 2006. ACM Press.

[2] N. Craswell, I. Soboroff, and A. de Vries. Overview of the trec-2006 enterprise track. In *Proceedings of the 15th Text REtrieval Conference.* ACM Forum, November 2006.

[3] L. Freund, E. Toms, and C. L. A. Clarke. Modeling task-genre relationships for ir in the workplace. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 441–448, New York, NY, USA, 2005. ACM Press.

[4] S. Robertson, H. Zaragoza, and M. Taylor. Simple bm25 extension to multiple weighted fields. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49, New York, NY, USA, 2004. ACM Press.