Deriving Implicit User Feedback from Partial URLs for Effective Web Page Retrieval

Rongmei Li
Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente
Enschede, the Netherlands
lir@cs.utwente.nl

Theo van der Weide
Institute for Computing and Information Sciences, Radboud University
Nijmegen, the Netherlands
tvdw@cs.ru.nl

ABSTRACT

User click-throughs provide a search context for understanding the user need of complex information. This paper re-examines the effectiveness of this approach when based on partial clicked data using the language modeling framework. We expand the original query by topical terms derived from clicked Web pages and enhance early precision via a more compact document representation. Since our URLs of Web pages are stripped, we first reconstruct them at different levels based on different collections. Our experimental results on the GOV2 test collection and AOL query log show improvement by 31.7% and 28.3% significantly in *statMAP* for two sources of reconstruction and 153 ad-hoc queries. Our model also outperforms pseudo relevance feedback.

Categories and Subject Descriptors

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

Keywords

Parsimonious Models, Query Log, Query Expansion

1. INTRODUCTION

Information needs in Web page search has both a topical interest and preference that varies by the context. For instance, the query "black jaguar" is about "animal" for some users and about "car" for the others. Understanding complex needs expressed as a short keyword query thus is very important for focused retrieval. A query log provides us with an economical and unobtrusive way for deriving user's search context and their language usage for a topic domain. The recorded search history usually includes query terms, retrieved documents and ranks, clicked documents, date and time of the search action, and user IDs. This first-hand user data has been used for enhancing retrieval performance [2, 4, 8]. However, knowledgeable users wish to control the access to their data due to privacy concerns. In this work, we use partially shared query log, specifically, the stripped click-through data for improving retrieval performance. We re-examine the effectiveness of this approach in the language modeling (LM) framework and focus on the following research questions: Can we extract common topical context from partially available click-through data? Can we represent the topical context effectively? What is the performance of this approach compared to pseudo relevance feedback (the state-of-the-art implicit feedback)?

RIAO'10, 2010, Paris, France.

2. STRIPPED URL RECONSTRUCTION

We derive the context knowledge via the query-click relation in a query log as it suggests the relevance information to the query. Our query log is the stripped AOL log [7]. It has 657K users' search information in three months in 2006 and 10,154,742 unique queries with 19,442,629 click-throughs. We carry out the reconstruction process by matching stripped URLs with other URLs of Web pages in the most similar collection (e.g. GOV2 or DMOZ) using *partial matching at domain level (noted as server level)* and *exact matching at the level of the stripped URL (noted as URL level).*

In our setting, there are many choices of Web page collections for facilitating URL reconstruction. The intuitive option is the test collection itself since the context information can distinguish the clicked pages from others more. Another choice is the open Project (also known as DMOZ), as it is one of the largest and most comprehensive human-edited directories on the Web.

3. PARSIMONIOUS LANGUAGE MODELS

Following the bag-of-words assumption [1], topical terms can be extracted from documents to construct the so-called *parsimonious (topical) language model* [3].

With a similar approach in [9], we use the topical model as query-dependent evidence to update the original query:

- Expand the original query with top-K terms computed from the corresponding restored pages.
- Expand the original query by top-N selected terms from the top 10 pages of the above result.
- Identify restored pages that appear in top-M ranked pages of the initial result and apply query expansion using top-N selected terms from those pages.

Following [3], the document representation can be improved by removing non-informative terms. As a result, documents with more precise information are ranked higher. Instead of using the ML document model $P_{ml}(t|\theta_D)$, we estimate the parsimonious model $P(t|\theta'_D)$ by the EM algorithm.

4. EXPERIMENTS

We extended the standard language model generating search engine Indri to obtain topical (or ML) models. The evaluation tool for the Million Query (MQ) Track (TREC 2007) is used to measure the performance. A two-tailed paired T-test is computed at the 0.05 (•) level of significance.

Retrieval model: Inspired by the relevance model [3, 5, 6], we rank documents based on cross-entropy scores for the (topic-based) query model and the smoothed document model: $\sum_{i=1}^{l}\left[P(t_i|\theta'_Q)\log(\lambda P(t_i|\theta'_D)+(1-\lambda)P_{ml}(t_i|\theta_C))\right]$, where l is query length, $P(t_i|\theta'_D)$ a ML or parsimonious document model, and $P_{ml}(t_i|\theta_C)$ the ML background model.

Query Expansion: In general all runs outperform the baseline in all metrics. The best runs (Table 1) show 31.7%, 33.4%, 39.5% for GOV2 and 28.3%, 32.1%, 34% for DMOZ significant improvements. Though reconstruction at the *server level* might have larger deviation from real clicks than that at *URL*, the performance of this level is better. Roughly, EM estimate generates better expansion terms.

Pseudo Relevance Feedback: All runs are improved from the PRF baseline. The best runs (Table 1) show 17.9%, 21.7%, 27.1% increase for GOV2 and 17.5%, 20.9%, 27.3% for DMOZ.

On the whole, 1.6∼3.9 times of MQ07 queries are boosted rather than hurt by QE and PRF for both GOV2 and DMOZ. One of the beneficial queries is topic 1439 (*signs of getting your period*) with 3 clicks that focus on health topic. But queries like topic 5695 (*map of the united states of america*) are hurt due to clicks on diverse topic of interest. Besides, reconstruction from GOV2 is not always superior to DMOZ.

Selected Relevance Feedback: When the target collection is used for restoring the stripped URLs, the intial result can be used to select the click-throughs. Our result shows consistent improvements for all numbers of initial results (Figure 1). For reference pages of top 1,000, the retrieval is more accurate as the baseline run returns more relevant pages at early precision.

Parsimony of Document Models: With parsimonious document representation, retrieval performance gains moderate improvement compared to ML document counterparts.

5. CONCLUSION

This paper aims at improving effectiveness of ad-hoc Web IR with the help of topical context from a degraded query log (AOL). We focus on the aggregated query-click relation and the content of clicked Web pages because they provide topical information. To identify those pages, we first reconstruct URLs from their domain names at the *URL* or *server* level. We represent the topical context by weighted topical terms derived from a target Web corpus (the GOV2 collection) and an external corpus (DMOZ). We illustrate strategies to integrate the query, the topical context, and the parsimonious document representation with the LM framework. One is QE while another is PRF for re-ranking the results of QE. To remove restored Web pages that might not be clicked in reality we use the retrieval result of our baseline to select restored pages that are ranked higher. We further apply QE on the selected relevance feedback. Our extensive experiments prove that all strategies improve the effectiveness of retrieval significantly compared to the standard query likelihood model. Strategy 2 shows 17.9% and 17.5% *statMAP* improvement for GOV2 and DMOZ at the 0.15 level of significance when compared to the stronger baseline of PRF.

Table 1: Results of Topic-based query models

models/level.terms	performance metrics		
	statMAP	statMRP	statMPC_10
QE Results on GOV2 (upper) and DMOZ (lower)			
baseline (JM)	0.1843	0.1777	0.1794
server.EM.100	**0.2700** •	**0.2668** •	**0.2967** •
server.EM.100	**0.2569** •	**0.2618** •	**0.2719** •
PRF Results on GOV2 (upper) and DMOZ (lower)			
baseline (EM.100)	0.2400	0.2349	0.2476
server.EM.100-50	**0.2924**	**0.2999**	**0.3396**
server.EM.100-50	**0.2910**	**0.2969**	**0.3407**

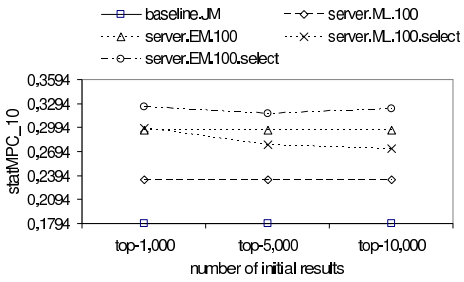

Figure 1: Improvement with reference pages

Table 2: Results of Parsimony on GOV2

models/level.terms	performance metrics	
	statMPC_10	increase over ML
server.EM.100	0.3147	+5.72%
server.EM.100-50	0.3541	+4.09%
server.EM.100.select	0.3140	-3.35%

6. REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] H. Cui, J. Wen, J. Nie, and W. Ma. Probabilistic query expansion using query logs. In *Proceedings of WWW*, pages 325–332, 2002.

[3] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings of SIGIR*, pages 178–185, 2004.

[4] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of SIGKDD*, pages 133–142, 2002.

[5] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR*, pages 111–119, 2001.

[6] V. Lavrenko and W. Croft. Relevance models in information retrieval. In *Language Modeling for Information Retrieval*, pages 11–56, 2003.

[7] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proceedings of InfoScale*, page 1, 2006.

[8] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accracy, 2006.

[9] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM*, pages 403–410, 2001.