Finding "Abstract Fields" of Web Pages and Query Specific Retrieval -- THUIR at TREC 2004 web track*

Yiqun Liu, Canhui Wang, Min Zhang, Shaoping Ma
State Key Lab of Intelligent Tech. & Sys., CST Dept, Tsinghua University, Beijing 100084, China
Liuyiqun03@mails.tsinghua.edu.cn, wangcanhui@tsinghua.org.cn, {z-m, msp}@tsinghua.edu.cn

1 Introduction

In this year's TREC Web Track research, THUIR participated in the Mixed-query Task. This task involves a single query set comprising 3 kinds of queries (Homepage Finding, Named Page Finding and Topic distillation) which are mixed and unlabelled.

Efforts have been made on two directions: to find a strong and robust unified approach which works well for all kinds of queries, and to build a query-specific retrieval strategy that classifies queries by types and perform specific approaches. The using of non-content information has been studied in both approaches.

With topic distillation and navigational search tasks in the last year, we are able to build a training set with 150 topics and corresponding relevant qrels. This training set is used to evaluate effectiveness of different methods in mixed query search. Experiments in section 2, 3 and 4 are all based on this set.

2 Non Query Specific Retrieval

Our efforts on non query specific retrieval are focus on the following aspects:

[1] How is effectiveness of using in-link anchor text in full text retrieval?

[2] Should we give a larger weight to several important fields in a HTML document in content retrieval? Such fields include TITLE, first appearing BOLD text, first appearing HEAD text and the first N words of straight text.

[3] Is word pair method useful in content retrieval? It means if query word pair appears in a web document, we should give this document a higher rank in the result list.

2.1 In-link anchor text retrieval

It is known that in-link anchor text is a useful source of information for navigational search (Home page finding and named page finding). In last year's web track experiments, we also found that anchor text based retrieval outperformed full text retrieval. So in-link anchor text is supposed to be useful for mixed-query task.

Data set	MAP	S@5
.Gov full text	0.3336	51.33%
.Gov anchor only	0.3872	60.00%
.GOV anchor combined with full text	0.5003	71.33%

From experiment results shown above we can see that in-link anchor text retrieval gets higher rank than full text retrieval for .GOV corpus. Meanwhile the retrieval on the combination of anchor text and full text gets more surprising result: 49.97% improvement in MAP and 38.96% in S@5 comparing with full text only. It validates the hypothesis that anchor texts provide a credible description to the

* Supported by by Chinese National Key Foundation Research & Development Plan (973) (Grant No.2004CB318108) and Chinese Natural Science Foundation (Grants No. 60223004, 60321002, 60303005).

page it links to and this description is another point of view as for the page. Although one destination page may not include words in one query, its in-link anchor does; and this page can be found by IR systems.

2.2 "Abstract fields" term weighting

We find out the phenomena that named page finding type query words often appears in several particular fields of its corresponding result page. These fields include TITLE, first appearing BOLD text, first appearing HEAD text and the first N words of straight text. In a sample of 15 NP type queries and its result pages, only 2 pages don't have query words in the fields proposed above. We call these fields "abstract fields" because they provide a brief summarization of the web page and people always refer to this page with content of these fields.

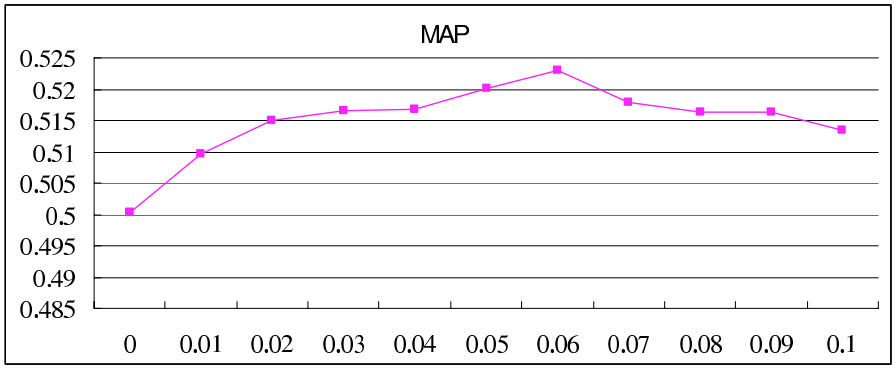
With this analysis, known item search results would improve if we gave larger term weights to abstract field words.

	Original MAP	MAP with abstract field
.Gov full text	0.3336	0.3409
.Gov anchor	0.3872	0.3887
.Gov combination	0.5003	0.5003

Experiments show that this method isn't as effective as we expected for the mixed-query task. It may help improve NP finding task performance (see section 3), but overall results don't benefit a lot.

2.3 Word pair method

Word pair search proves effective in TREC 2003 robust track. It aims at enhancing system performance in terms of precision and thus useful for web track tasks because of the size explosion of web documents. The basic idea of word pair is that if adjacent words in the query are also adjacent in the document, then the document would be more likely to be relevant. More implemental details are available in [4].



Experiment results show that MAP improves with word pair method and the improvement is stable with different parameters.

3 Query Specific Retrieval

In query specific retrieval, queries are classified and retrieved with different ranking methods. Many features were involved in our experiment to find the difference between 3 kinds of queries but only a few of them are useful. With these features we were able to separate most TD type queries from HP/NP queries. How to achieve better result with a partly-classified query set is studied in our research.

3.1 Query classification technologies

The features studied in query classification research are divided into two classes: features of the query

itself, and features involving retrieval process with the query.

Features got without retrieval process are non-content attributes of one query, such as query length, existence of abbreviations and existence of named entities. We found that a TD type query tends to have a small term number and not to contain named entity or abbreviation; while known-item search queries (NP & HP type) usually have more terms and a number of named entities.

Features involving retrieval process are based on the following idea: [1] one type of queries may perform well with a ranking strategy, but fails with another. So retrieval results with two ranking strategy help locate those topics who vary most and these topics are likely to be this type of queries. [2] IR system tent to return a home page with the 1st rank in the result list for HP type queries, so those whose result list starts with a home page are likely to be HP type queries.

The features based on retrieval process include: top 1 result URL-type, top 1 result PageRank, top 1 RSV variance between 2 ranking methods, top 10 average-RSV variance between 2 ranking methods etc. However, the effectiveness of these features highly depends on parameter tuning. They show different performance when training set changes and eventually are not used for classification.

	TD	HP/NP
Precision	0.854	0.862
Recall	0.700	0.940

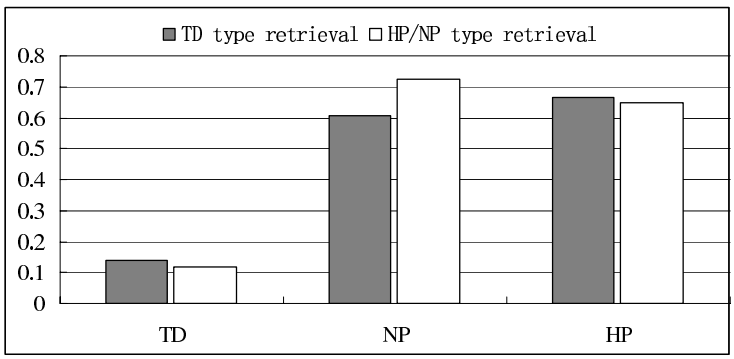
We finally chose query term number and named entity existence to separate TD from others. These two features divide a query set into 2 subsets: TD and HP/NP. Experiment results for training set are shown above in the table.

3.2 Retrieval with different query types

TD type queries and navigational queries are separated with both high precision and recall so different ranking methods may be applied to achieve a better result.

For TD type query, the key idea is to separate key resource pages from ordinary pages with non-content feature analysis before retrieval process [2]. Only key resource pages should be returned as TD task results, so it helps reduce pages which shouldn't be returned by the retrieval system. With features including URL type, document length, site's self link number and in-degree, we obtained a page set which covers about 20% .GOV pages and over 70% TREC 2003 TD task relevant qrels. Retrieval on this page set, which we call "key resource set", got much better ranking than the whole page set of .GOV for TD type queries.

For NP and HP type queries, in-link anchor text is combined with full text for retrieval. Page content analysis is also involved in our study. A larger weight is given to one page if query terms appear in abstract fields (see section 2.2) for NP/HP type retrieval; this helps locate known items more accurately.



TD type retrieval (retrieval on key resource page set) gets better result in TD and HP type queries,

while HP/NP type retrieval (.GOV full text retrieval with emphasis on in-link anchor and abstract field) obtains a higher ranking for NP type queries. Query classification makes it possible to combine these two methods and fit queries for a suitable ranking strategy.

4 Runs submitted and Training Set Evaluation Results

Description:

Runs	Description
THUIRmix041	Content retrieval in full text and in-link anchor, with a larger weight in abstract fields.
THUIRmix042	Content retrieval in full text and in-link anchor of Key resource pages which are selected with non-content features.
THUIRmix043	THUIRmix041 + primary space model weighting (see [1]) in in-link anchor text and abstract field text.
THUIRmix044	Query classification with query length and named entity information. TD topics are assigned to THUIRmix042, while the others are retrieved on THUIR041.
THUIRmix045	THUIRmix041 with a larger weight if word pair appears in page content.

MAP evaluation:

Runs	Mixed	TD	NP	HP
THUIRmix041	0.5108	0.1172	0.7406	0.6547
THUIRmix042	0.4817	0.1393	0.6152	0.6709
THUIRmix043	0.5245	0.1168	0.7565	0.6804
THUIRmix044	0.5111	0.1330	0.7181	0.6625
THUIRmix045	0.5165	0.1131	0.7499	0.6668

S@5 evaluation:

Runs	Mixed	TD	NP	HP
THUIRmix041	73.33%	48.00%	90.00%	82.00%
THUIRmix042	70.00%	50.00%	78.00%	82.00%
THUIRmix043	73.33%	48.00%	88.00%	84.00%
THUIRmix044	74.00%	52.00%	84.00%	86.00%
THUIRmix045	72.67%	46.00%	88.00%	84.00%

Reference

[1] Min Zhang, Ruihua Song, and Shaoping Ma, DF or IDF? On the Use of HTML Primary Feature Fields for Web IR, the 12th World Wide Web conference, (www2003), poster, Hungarian, 2003.

[2] Yiqun Liu, Min Zhang and Shaoping Ma, *Effective Topic Distillation with Key Resource Pre-selection*, Proceedings of the first Asia Information Retrieval Symposium (AIRS2004), Beijing, Oct. 2004.

[3] S. E. Robertson, S. Walker, M.M. Hancock-Beaulieu, and M. Gatford. *Okapi at TREC-3*. In NIST Special Publication 500-225: Overview of the Third Text Retrieval Conference.

[4] Min Zhang et al, *THUIR at TREC 2003: Novelty, Robust and Web*, in NIST Special Publication 500-255:The Twelfth Text REtrieval Conference (TREC 2003).