SJTU at Trec2004: Web Track Experiments

Yiming Lu, Jian Hu, Fanyuan Ma
(Department of Computer Science & Engineering , S hanghai Jiaotong University , S hanghai 200030*)*
{luyiniao , hujian , ma-fy}@sjtu.edu.cn

Abstract:

This is the first year our lab to participate in Trec. We participate in Mixed-Query task for the Web track. All the runs we submitted are based on the modified Okapi weighting scheme. Besides, we used several heuristics as the re-rank method: site-merging, minimal span weight, and etc. Also, the PageRank of a document is combined with the similarity of the document with the query to obtain an overall ranking of documents.

Especially for the mixed-query task, we try a simple classification method to estimate whether the query is topic distillation or entry-page finding.

Introduction

As we know, there are a lot of rank strategies to estimate relevance of documents and queries, for example, Okapi BM25[2], PDLN (Pivoted Document Length Normalization)[3], Lnu-ltc[5] and Language Models[4]. In our experiment, we use BM25 as our main strategy for estimating the content relevance.

According to the survey of results of the past anticipants in web track, we find that using document structure to provide various document representations was shown to be an effective and stable way to improve the ranking effectiveness. So through html parser we build a lot of indexes, like title, meta, inlink anchor text, outlink anchor text, section headers, etc. To combine these document representations is also a big problem for us, there are several ways to combine the multiple document representations to achieve improvements in retrieval effectiveness, Fox and Shaw[6] examined several combination algorithms, Lee[7] conducted extensive experiments with the Fox and Shaw combination rules. Lee also introduced the normalization method for different rank results and analysis the combination of multiple document representations. In our experiment we tried a lot of results normalization methods and combination algorithms they have proposed.

Recently, the research of web retrieval has focused on link-based ranking methods; however the results of web track history showed that using link-base methods only does not work. Thus, we combine the PageRank[1] result and content-based retrieval result in our experiment.

To improve the precision of retrieval, we also use several heuristics for re-ranking, such as site-merging and minimally matching span for each document title text.

In our experiment, we find that topic distillation and entry-page finding has their own most suitable. Thus, a simple query classification is also included in our experiment.

Data Processing

Pre-processing

Our data pre-processing includes data cleaning and information extraction.

We first clean the documents with on content, like the files with postfix of “jpg” or “gif”, and the redirect html documents.

We then used an HTML parser to extract the following information:
Text Information:

- Title: words in <title>…</title>;
- Meta: words in <meta description = "…"/> or <meta keywords = "…"/>
- Head: words in H1 to H6 tags and words with font type "bold", "underline",etc.
- Anchor text: words in anchor texts extracted from the pages which pointing to this page.
- Image text: words in <alt>…</alt> of image files.
- Plain text: all the rest content words.
- All of the above information are extracted and put into individual index.

Link Information:

- All the hyperlinks and their corresponding anchor text are extracted and store in their indexes, each URL is assigned a unique ID and a URL dictionary is built which is used to construct Web link graph and facilitate the anchor text index building.

Indexing

For each web page, before indexing, we first performed stemming for each word using the traditional porter stemming algorithm, Stop words then remove, we used a stop word list[2]. The term weighting is BM25.

$$\sum \log\left(\frac{(N-n)+.5}{(n+.5)}\right) * \left(\frac{(k1+1)*tf}{(K+tf)} * \frac{(k3+1)*qtf}{(k3+qtf)}\right)$$

$$K = k1*((1-b)+b*dl/avdl)$$

Equation 1: Okapi BM25

Where:

- tf = frequency of occurrences of the term in the document
- qtf = frequency of occurrences of the term in the query
- dl = document length
- avdl = average document length
- N = is the number of documents in the collection
- n = is the number of documents containing the word
- k1 = 1.2
- b = 0.75 or 0.25 (we use .75 for full text and .25 for shorter representations like title, meta, inlink anchor)
- k3 = 7, set to 7 or 1000, controls the effect of the query term frequency on the weight.

Using Document Structure and Data Fusion

From the paragraph above, you can see we build a lot of indexes by extract the text different logic field of html document, such as title, meta, head, plain, inlink anchor and outlink anchor, the first five fields are widely used in information retrieval, we always consider the anchor text is not a part of the document which the anchor text belongs to, instead we think it is description of the documents it pointing to, it's true in some aspect, but as we know navigation is also an important task of a document, the anchor text is also a part of a document. From table below we can see the outlink anchor is much better than the plain text of the document in topic distillation task.

	MAP	R-P	real-ret
Outlink anchor	0.0677	0.0665	300/516
Plain text	0.1117	0.1207	303/516

Table 1: outlink anchor and plain text results for topic distillation queries

Fox and Shaw[6] introduced several combination methods such as CombMax, CombMin, CombSum, CombANZ, CombMNX and CombMed, and they found CombSUM to be the best performing combination method. Lee[7] conducted extensive experiments with Fox and Shaw combination method based on the TREC data, and he found CombMNZ emerges as the best combination rule.Vogt and Cottrel[8] improved the CombSUM and proposed the linear combination method. In our experiment all of the following methods were tried and compared.

CombMNZ = SUM(Individual Similarities) * Number of Nonzero Similarities （1）

CombMax = Max(Individual Similarities) （2）

$sim_{new} = \sum_{i=1}^{n} wi \cdot sim_i$ （3）where w_i is the relative weight of run_i

Similarity score distributions may differ radically across runs, so instead of directly applying the methods to the retrieval status values(RSV), we need to normalize them to a standard value scope. In our experiment we use the max-min norm:

$sim_{neorm} = \frac{sim_{original} - sim_{\min}}{sim_{\max} - sim_{\min}}$ （4）sim_{min} (sim_{max})is the minimal(maximal) RSV score in the run.

We use the trec12 topic distillation queries and home page finding queries separately, and compare the three combination methods. For topic distillation queries we base on the head+plain and outlink anchor indexes. From table2 we find that linear combination is better than CombMNZ and CombMax for topic distillation queries.

	MAP	R-P	real_ret
CombMax	0.0974	0.08	349/516
CombMNZ	0.1424	0.1572	335/516
Linear combination	0.1456	0.1572	366/516

. Table2 : results of CombMax,CombMNZ and linear combination for TD queries

For home page finding queries we base on the title+meta and inlink anchor indexes. From table3 we find that through the MRR of CombMax result is still lower than the others, the S@10 is much higher than them.

	MRR	S@10
CombMNZ	0.60	0.727
CombMax	0.51	0.777
Linear combination	0.6049	0.727

Table3: results of CombMax,CombMNZ and linear combination for home queries

Link Structure

Site Unit

The definition of key resource in topic distillation task implied that only one page can be a key resource among pages from an identical site. As we know the pages from an identical site especially the site use the same template sometimes have the same title and many identical outlink anchor and plain text words, so when they are ranked by content-base retrieval method, they might be ranked

adjacently. Sometimes might be ranked high and thus made a possible key resource from another site lower. So we should try to find as many different websites as possible within the top ten results. In our experiment we allow each site can only have at most 3 pages in top 1000 pages.

Using the trec12 queries of topic distillation task, we compared the retrieval results. From table below we can see the result of the rank added site unit is almost the same as the one not add.

	MAP	R-P	real-ret
Use site unit	0.1894	0.1912	401/516
Not use site unit	0.1894	0.1912	402/516

Table4: results for topic distillation queries

Link Analysis

We use the PageRank score [1] as the measure of the quality of the Web Page.

Minimal Span Weight

Experimental research on the seeking behavior of human searchers using a web search engine[9], has shown that most users only consider the top ten, so web retrieval system should opt for high precision ,and proximity-base retrieval seems to be a natural way to accomplish this. The minimal span weight [10] algorithm is a kind of proximity-base retrieval method widely used in current question answering systems. It depends on three factors.

1.document similarity: The document similarity is computed for the whole document. Similarity scores are normalized with respect to the maximal similarity score for a query.

2. span size ratio: The span size ratio is the number of unique matching terms in the span over the total number of tokens in the span.

3. matching term ratio: The matching term ratio is the number of unique matching terms over the number of unique terms in the query, after stop word removal.

$$RSV(q,d) = \lambda RSV_n(q,d) + (1-\lambda)\left(\frac{|q \cap d|}{1+\max(\text{mms})-\min(\text{mms})}\right)^{\alpha}\left(\frac{|q \cap d|}{|q|}\right)^{\beta}$$

Equation2: minimal Span weight

where $RSV_n(q,d)$ is the normalized document similarity, $\left(\frac{|q\cap d|}{1+\max(\text{mms})-\min(\text{mms})}\right)$ is the span size ratio, $\left(\frac{|q\cap d|}{|q|}\right)$ is the matching term ratio, $\alpha = 1/8$, $\beta = 1$.

After analysis of web track 2002, 2003 results, we found that title text is very important for topic distillation and known-Item finding task, the title text of the answers for a certain query is of high quality, it includes nearly all of the query words. So we use minimal span weight algorithm in the title index.

Simple Query Classification

This year's task is Mixed-Query task, considering that one rank strategy does not fit all queries, we developed a simple query classification method according to the web track 2003 results, thus we

can tune the ranking parameters according to the query type. The algorithm can be shown as following

1) If the query include words like “home”, ”home page”, ”administration”, ”agency” ,it is a home page finding query.
2) Else if the query’s length <= 2, it is a topic distillation query
3) Else the query is a mixed query.

To the mixed query we mix the retrieval results: we use the top 10 of the results tuned for name page finding queries as the final top 10 results, the other 990 we use the top 990 results tuned for topic distillation queries.

Experiment

Runs

We totally submitted the following five official runs for the mixed query task:

- SJTUMIX1 – Simple task classification divide the queries into topic distillation queries and known-item queries. Linear combination of top 1000 of BM25 on the 4 word-based stemmed indexes: head+plain, outlink anchor, inlink anchor, title+meta. Minimal span weighting on title index is used to post-process the ranking results.
- SJTUMIX2 – Simple task classification, linear combination of top 1000 of BM25 like SJTUMIX1.
- SJTUMIX3 – only use Linear combination of top 1000 of BM25 on the 4 word-based stemmed indexes: head+plain, outlink anchor, inlink anchor, title+meta. Then we mixed all the results according to the method for the mixed query.
- SJTUMIX4 – SJTUMIX2 is first used to get a result list. Then the PageRank value is used to re-rank the result mentioned in link analysis paragraph.
- SJTUMIX5 –SJTUMIX2 is first used to get a result list. Then the site unit is used to post-process the ranking results and only allow at most 4 pages from identical site in the top 1000 result.

Results

The results of the official runs for the mixed-query task are shown bellow; we divided the results in 3 tables according to the query type.

Run identifier	AveP	R-P	P@10
SJTUMIX1	0.1228	0.1413	0.1747
SJTUMIX2	0.1248	0.1405	0.1640
SJTUMIX3	0.1253	0.1391	0.1640
SJTUMIX4	0.1271	0.1491	0.1733
SJTUMIX5	0.1294	0.1556	0.1893

Table5: Results for topic distillation queries

Run identifier	AveP	recip_rank	Suc@10
SJTUMIX1	0.5142	0.5154	0.7867
SJTUMIX2	0.5402	0.5426	0.7867
SJTUMIX3	0.5376	0.5398	0.7867
SJTUMIX4	0.5079	0.5103	0.7467
SJTUMIX5	0.5016	0.5040	0.7467

Table6: Results for named page finding queries

Run identifier	AveP	recip_rank	Suc@10
SJTUMIX1	0.4566	0.4775	0.7067
SJTUMIX2	0.4698	0.4873	0.6667
SJTUMIX3	0.4720	0.4891	0.6667
SJTUMIX4	0.4421	0.4584	0.6400
SJTUMIX5	0.4400	0.4569	0.6267

Table7: Results for home page finding queries

According to the table 5,6,7 we can see that the performance of simple task classification isn't significant, and the minimal span weighting gives a negative effect, however the site unit improve the effectiveness for topic distillation queries by 13%.

Reference

[1]L. Page, S. Brin, R. Motwani, and T. Winograd (1998). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, Stanford, CA.

[2]S. Robertson, et al., "Okapi at TREC-4", Proceedings of the 4th annual Text Retrieval Conference (TREC-4), NIST, November 1995.

[3] A. Singhal, et al., "Pivoted document length normalization", ACM-SIGIR, 1996

[4]J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In 21st ACM Conference on Research and Development in Information Retrieval (SIGIR'98) 275-281, 1998.

[5] C. Buckley, A. Singhal, and M. Mitra. New retrieval approaches using SMART: TREC 4. In D. Harman, editor, The Fourth Text REtrieval Conference (TREC-4), pages 25 - 48. National Institute for Standards and Technology. NIST Special Publication500-236, 1996

[6] E. A. Fox and J. A. Shaw, "Combination of Multiple Searches," Proceedings of the 2nd Text Retrieval Conference (TREC-2), NIST Special Publication 500-215, pp. 243-252, 1994

[7] J.H. Lee, "Combining Multiple Evidence from Different Properties of Weighting Schemes," Proceedings of the 18th Annual ACM-SIGIR, pp. 180-188, 1995

[8]C. C. Vogt and G. W. Cottrell. Predicting the performance of linearly combined IR systems. In SIGIR' 98, pp. 190 - 196, 1998

[9] B. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management, 36(2):207 - 227, 2000

[10] C. Monz. From Document Retrieval to Question Answering. ILLC dissertation series 2003-04, University of Amsterdam, 2003.