



encourage research on Web information retrieval. Homepage finding (HP) and named page finding (NP) is to model two types of navigational queries. The difference is that a homepage finding query is the name of a site while a named page finding query is the name of a non-homepage that the user wishes to reach. Topic distillation (TD), on the other hand, is to model informational queries. It was first proposed by Bharat and Henzinger [3] to refer to the process of finding quality document on a query topic. They argued that it is more practical to return quality documents related to the topic than to exactly satisfy the users' information need since most short queries do not express the need unambiguously. In TRECs, a topic distillation query describes a general topic and requires retrieval systems to return homepages of relevant sites. Until now, these three types of queries are acknowledged and TRECs cumulated valuable data through years for related research.

URL, as a Uniform Resource Locator [19] for each Web page, usually contains meaningful information for measuring the relevance of the Web page to a query. Related works can be roughly grouped into 3 categories: one is to use the length or depth of a URL as query-independent evidence in ranking [9][21][12][6]; another is to use URL-based sitemap to enhance topic distillation [20][18]; the other addresses the issue of word break in URLs [5][12].

Kraaij et al [9] found that the probability of being an entry page, i.e. homepage, seems to inversely related to the depth of the path in the corresponding URL. They classified URLs into four types in terms of the depth, estimated prior relevance probability for each type, and integrated the priors in the language model. Their experimental results verified that the depth is a strong indicator for homepages. By doing some extension, Ogilvie and Callan [12] reported improvements on mixed homepage/named-page finding task. However, by closely observing the URL priors in [12], we found that the priors for homepage finding queries are quite different from those for named page finding queries (see Section 2 for details). Thus the priors may hurt named-page finding while improving homepage finding.

In this paper, we aim to find a kind of stable priors to enhance retrieval performance for various kinds of queries. We observe that the occurrence location of the query terms in a URL is an effective indicator of the quality and relevance of a Web page. Especially, a URL with some query term appearing near to its tail promises to be a relevant domain, directory or file. Our statistics on queries of past TREC experiments verify this observation. Therefore, we treat the occurrence location of the query terms in a URL as a good prior for the relevance of a page. We call this kind of priors the URL hit priors as a hit refers to a query term occurrence.

The effectiveness of URL hit priors relies on the capability of detecting the hits of query terms in URLs. To increase the hit rates of query terms, we explore three successive methods to recognize terms in URLs. First, a simple rule is used to recognize most of acronyms in URLs. Second, the recognition of concatenations is formulated as a search problem with constraints. Third, prefix matching is used to recognize other fuzzily matched words. With this 3-step approach, it is shown on the TREC data that the recall of URL hits is doubled from 33% to 66% while the precision is close to 99%.

We integrate the URL hit priors into the probabilistic model. Experimental results, on seven TREC Web Track datasets, indicate that, with the URL hit priors and URL

hit recognition methods, the performance is consistently improved across various types of queries.

The rest of the paper is organized as follows. Section 2 introduces the related work. In section 3, we give the details of URL hit priors, URL hit recognition methods, and how to combine URL hit priors into the probability model. We conduct experiments to verify the proposed methods in Section 4. Conclusion and future work are given in Section 5.

## 2 Related Work

As mentioned in the introduction, several URL-related approaches have been proposed to enhance Web search or recognize more query terms. In this section, we will briefly review four latest and representative works.

Kraaij et al found that the URL depth is a good predictor for entry page search [9]. Four types of URLs are defined in their work [21] as follows:

*“ROOT*: a domain name, optionally followed by 'index.html'.

*SUBROOT*: a domain name, followed by a single directory, optionally followed by 'index.html'.

*PATH*: a domain name, followed by a path with arbitrarily deep, but not ending with a file name other than 'index.html'.

*FILE*: any other URL ending with a filename other than 'index.html'.”

The priori probability of being an entry page is elegantly integrated in the language model. As a result, the performance is improved by over 100%. About 70% of entry pages are ranked at No.1. The TREC2001 evaluation confirmed some successful exploitation of URL depth in entry page search [7][13].

Ogilvie and Callan extends the usage of URLs in TREC2003 [12]. A character-based trigram generative probability is computed for each URL. A shortened word or a concatenation of words is handled by treating a URL and a query term as a character sequence. Another extension is that they include named page in the estimation of URL depth priors.

Based on the TREC2003 data, we did some statistics about the distributions of URL depth types for different retrieval tasks. The results are shown in Table 1. It is clear that most of the relevant documents for HP queries have the ROOT type URLs, while the majority of NP queries tend to have the FILE type URLs for their relevant documents. For TD queries, more than half of relevant documents' URLs are with the FILE type, whereas the distributions in the other three URL types are quite even. Therefore, the computed priors based on URL depth are unlikely to benefit all query types.

Craswell et al [6] use URL length in characters as query independent evidence and propose a function to transform the original depth for effective combination. Their results show a significant improvement on a mixed query set. And their finding is that the average URL length of relevant pages is shorter than that of the whole collection.

Chi et al [5] reported that over 70% URL words are "compound word", that means multiple words are concatenated to form one word. Such phenomenon is caused by the

special of URLs. Some of the most frequent delimiters, such as spaces, in a document are not allowed to appear in URLs [19]. Consequently, webmasters have to concatenate multiple words when creating a URL. These compound words cannot be found in the ordinary dictionaries. Thus they proposed to exploit maximal matching, a Chinese word segmentation mechanism, to segment a “compound word”. An interesting idea is that title, anchor text and file names and alternated text of embedded objects are used as a reference base to help disambiguate segmentation candidates. Although the authors aim to recover the content hierarchy of Web documents in terms of URLs, the approach is also a good solution for recognizing URL hits. We have not implemented their approach because this paper focuses on the effectiveness of URL hit priors for search and their approach does not handle individual shortened words. In addition, our recognition methods do not use any dictionary but the query only. Another solution worth mention was proposed by Ogilvie and Callan [12]. They treat a URL and a query term as a character sequence and compute a character-based trigram generative probability for each URL.

**Table 1.** Distributions of URL depth types (TREC2003)

URL Depth Type	HP	NP	TD
ROOT	103	1	79
SUBROOT	33	8	65
PATH	13	11	77
FILE	45	138	295

### 3. Our Approach

In this section, we first define a new classification of URL types and the related URL priors called URL hit priors. Then three methods are described to recognize URL hits. Finally, we introduce how to combine the URL hit priors into the probabilistic model and for improving retrieval performance.

#### 3.1 URL Hit Priors

A query term occurrence in a URL is called a URL hit. We assume that the location of a URL hit may be a good hint to distinguish a good answer from other pages. For example, when a user is querying "wireless communication" and 2 URLs below are returned, U2 is more probably to be a better answer because it seems to be a good entry point, neither too general nor too narrow.

*U1: cio.doe.gov/wireless/3g/3g\_index.htm*

*U2: cio.doe.gov/wireless/*

When “ADA Enforcement” is queried, U3 looks like a perfect answer as a URL hit occurs in the file name.

*U3: <http://www.usdoj.gov/crt/ada/enforce.htm>*

Given the query of “NIST CTCMS”, U4 is easy to beat other pages like U5 and again the URL hits appear in a good position.

*U4: <http://www.ctcms.nist.gov/>*

*U5: <http://www.ctcms.nist.gov/people/>*

Given a URL, slashes can easily split the URL into several segments (the last slash will be removed if there is no character followed by it). U2, U3 and U4 are similar for the last URL hit occurs in the last segment. Therefore, we define four kinds of URL hit types:

**Hit-Last:** A URL, in which the last URL hit occurs in the last segment;

**Hit-Second-Last:** A URL, in which the last URL hit occurs in the second last segment;

**Hit-Other:** A URL, in which all the URL hits occur in other segment than the last two;

**Hit-None:** A URL, in which no URL hit is found.

In our examples, U2, U3 and U4 belong to the type of “Hit-Last”, U5 is of the type “Hit-Second-Last”, while U1 is of the type “Hit-Other”.

We perform a statistical analysis base on the TREC2003 data. The distribution of URL hit types is shown in Table 2. There are two important observations from the statistics. First, a large portion of good answers have query term hits in their URLs. Second, the distributions of good answers in different types are quite consistent across different query types. Except for the "Hit-None" type, most of the good answers fall into the URL type "Hit-Last" for all the three query types HP, NP and TD. Also, type "Hit-Second-Last" has more good answers than type "Hit-Other". Thus, we expect to find a stable prior relevance probability for the URL hit types, which can be uniformly used in various tasks.

**Table 2.** Distribution of URL hit types (TREC2003)

URL Hit Type	HP	NP	TD
Hit-Last	136	86	129
Hit-Second-Last	21	17	21
Hit-Other	8	12	6
Hit-None	29	43	360

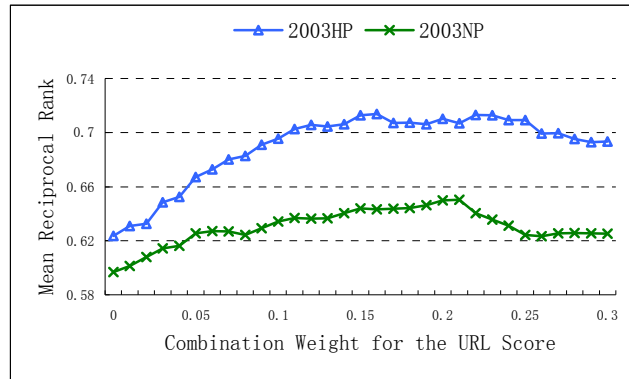
Based on the above observations, we target to assign each URL a prior relevance probability. Given a hit type  $t$ , this prior is consistently used for HP, NP and TD queries. Given a query  $q$  and a page with URL  $u$ , we denote  $P(t)$  as the probability of URL  $u$  having hit type  $t$  for the query. We denote  $P(R)$  as the probability of  $u$  being relevant to query  $q$ . And  $P(TD)$ ,  $P(HP)$ , and  $P(NP)$  are denoted respectively as the



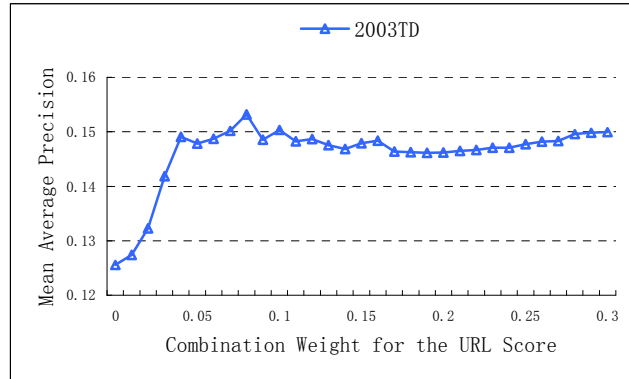








(a)



(b)

**Fig. 1.** Tuning the combination weight on TREC2003 data. (a) shows the results for HP and NP task in terms of MRR and (b) shows the result for TD in terms of MAP

On the test set, the URL hit priors improve MRR by about 4% for named page finding queries and by about 11% for homepage finding queries. And it also improves MAP by about 10% for topic distillation queries (See Table 4). Therefore, it is safe to conclude that the improvement with the usage of URL hit priors is stable for different types of queries. In addition, the improvement for NP tasks are less than those for the HP and TD tasks, which may be caused by the relatively rare occurrences of query terms in file names.

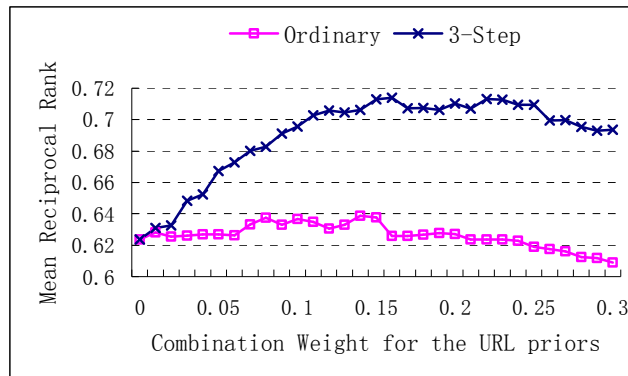
#### 4.4 Experiments on Using 3-Step Recognition Method vs. Not Using

It is necessary to evaluate how the URL hit recognition affects URL hit priors and the retrieval performance. Therefore, we use the ordinary word break method to recognize URL hits and apply the same approach to estimate the URL hit priors. And we redo the retrieval experiments of combining the priors with the basic content score. Figure 2 shows the results on HP task of TREC2003. There is a big gap between priors based

on different recognition methods. The same gaps are also found for other query sets and data sets. We omit the figures due to space limitation. In summary, the URL hits recognition methods are essential for fully taking advantage of the URL hits priors. If not sufficient URL hits are detected, the URL hit priors are less useful for improving retrieval performance.

**Table 4.** Integrating URL Hit Priors in the Probability Model

Query	$S_D$	$S_{combi}$	Improve
2002NP	0.6294	0.6529	+3.73%
2004NP	0.557	0.5818	+4.45%
2004HP	0.5404	0.6002	+11.07%
2004TD	0.13	0.1436	+10.46%



**Fig. 2.** Comparison of priors based on the ordinary word break method and our 3-step method

## 5. Conclusion and Future Work

Through observation and statistics, we found that the location of a query term appearing in a URL is closely related to whether the corresponding document is a good answer for homepage finding, named-page finding and topic distillation queries. However, shortening and concatenating make it difficult to match a URL word with query terms. We proposed three steps together to recognize URL hits. Such method improves the recall of URL hits from 33% to 66% for relevant URLs of TREC data of three years. Based on recognized URL hits, URL hit priors are estimated and integrated into the probability model. Experiments conducted on the TREC datasets show that the URL hit priors can achieve stable improvement across various types of queries.

In the current implementation, URL hits are detected when a query is submitted to the search engine. This requires additional time in processing the query, which could be an issue when the approach is used in a real large-scale search engine. We will

leave offline URL hit recognition as our future works. Our current experiments are based on TREC dataset which have little spam. As a next step, more experiments can be done for current real Web data to further test the effectiveness of our approach.

## References

1. R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval.*, ACM Press, 1999.
2. J. Berger. *Statistical decision theory and Bayesian analysis.* New York: Springer-Verlag, 1985.
3. K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In 21st Annual International ACM SIGIR Conference, pages 104--111, Melbourne, Australia, August 1998.
4. A. Border. A taxonomy of Web search. *SIGIR Forum*, 36(2), 2002
5. C.-H. Chi, C. Ding and A. Lim. Word segmentation and recognition for web document framework. *CIKM'99*, 1999
6. N. Craswell, S. Robertson, H. Zaragoza and M. Taylor. Relevance weight for query independent evidence. In *Proceedings of ACM SIGIR'05*, Salvador, Brazil, 2005
7. D. Hawking, E. Voorhees, N. Craswell, and P. Bailey. Overview of the TREC-8 web track. In *The Eighth Text Retrieval Conference (TREC8)*, NIST, 2001
8. Y. Hu, G. Xin, R. Song, G. Hu, S. Shi, Y. Cao and H. Li. Title extraction from bodies of HTML documents and its application to Web page retrieval. In *Proceedings of SIGIR'05*, Salvador, Brazil, 2005
9. W. Kraaij, T. Westerveld and D. Hiemstra. The importance of prior probabilities for entry page search. *SIGIR'02*, 2001
10. U. Lee, Z. Liu and J. Cho. Automatic identification of user goals in Web search. In the *Proceedings of the Fourteenth Int'l World Wide Web Conference (WWW2005)*, Chiba, Japan, 2005
11. G. Marchionini. Interfaces for End-User Information Seeking. *Journal of the American Society for Information Science*, 43(2):156-163, 1992.
12. P. Ogilvie and J. Callan. Combining structural information and the use of priors in mixed named-page and homepage finding. *TREC2003*, 2003
13. D.-Y. Ra, E.-K. Park, and J.-S. Jang. Yonsi/etri at TREC-10: Utilizing web document properties. In *The Tenth Text Retrieval Conference (TREC-2001)*, NIST, 2002
14. S. E. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In the *Eighth Text Retrieval Conference (TREC 8)*, 1999, pp. 151-162.
15. S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society of Information Science*, Vol. 27, No. May-June, 1976, pp. 129-146.
16. TREC-2004 Web Track Guidelines. [http://es.csiro.au/TRECWeb/guidelines\\_2004.html](http://es.csiro.au/TRECWeb/guidelines_2004.html)
17. D. E. Rose and D. Levinson. Understanding user goals in Web search. In *Proceedings of the Thirteenth Int'l World Wide Web Conference (WWW2004)*, New York, USA, 2004
18. T. Qin, T.-Y. Liu, X.-D. Zhang, Z. Chen and W.-Y. Ma. A study on relevance propagation for Web search. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, Salvador, Brazil, 2005
19. Universal Resource Identifiers. [http://www.w3.org/Addressing/URL/URI\\_Overview.html](http://www.w3.org/Addressing/URL/URI_Overview.html)
20. J.-R. Wen, R. Song, D. Cai, K. Zhu, S. Yu, S. Ye and W.-Y. Ma, Microsoft Research Asia at the Web Track of TREC 2003. In the *Twelfth Text Retrieval Conference*, 2003
21. T. Westerveld, W. Kraaij, and D. Hiemstra. Retrieving web pages using content, links, URLs and anchors. *TREC2001*, 2001