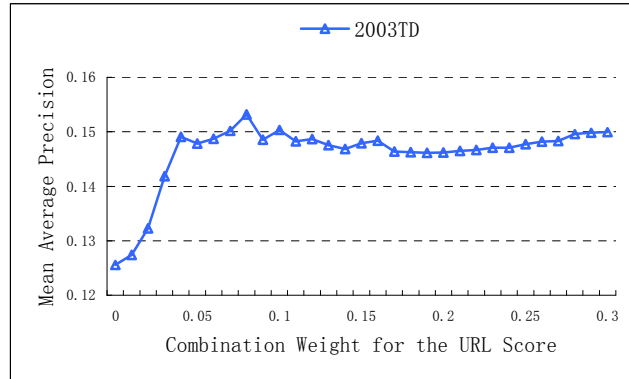


(a)



(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.

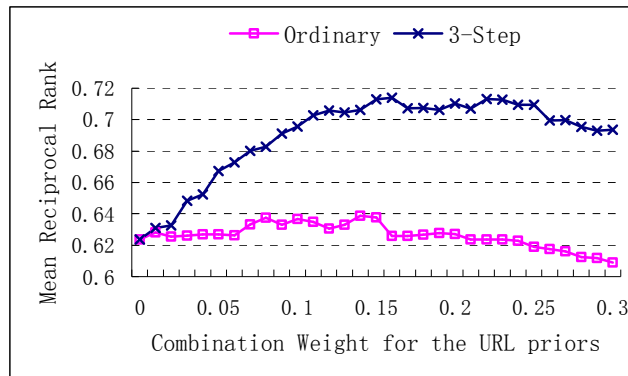
#### 4.4 Experiments on Using 3-Step Recognition Method vs. Not Using

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

