Web Search Based on Micro Information Units

Xiaoli Li, Bing Liu, Tong-Heng Phang, and Minqing Hu
School of Computing
National University of Singapore
3 Science Drive 2, Singapore 117543
{lixl, liub, phangth, humq}@comp.nus.edu.sg

ABSTRACT
Internet search is one of the most important applications of the Web. One shortcoming of existing search techniques is that they do not give due consideration to the micro-structures of a Web page. A Web page is often populated with a number of small information units, which we call *micro information units* (MIU). Each unit focuses on a specific topic and occupies a specific area of the page. During the search, if all the keywords in the user query occur in a single MIU of a page, the top ranking results returned by a search engine are generally relevant and useful. However, if the query words scatter at different MIUs in a page, the pages returned can be quite irrelevant. The reason for this is that although a page has information on individual MIUs, it may not have information on their intersections. In this paper, we propose a technique to solve this problem. At the off-line pre-processing stage, we segment each page to identify the MIUs in the page, and index the keywords of the page according to the MIUs in which they occur. In searching, our retrieval and ranking algorithm utilizes this additional information to return those most relevant pages. Experimental results show that this method is able to dramatically improve the search precision.

Keywords
Web search, Micro information Units, Web page segmentation.

1. INTRODUCTION
Web search engines allow the user to specify keywords to retrieve those Web pages that contain the keywords. The key issue in Web search is how to efficiently retrieve relevant Web pages with high precision for its top ranking results. A major shortcoming of the current techniques is that they do not consider different topic areas of a page. Typically, the contents of a Web page encompass several related or even unrelated topics. For example, a bookstore Web page selling books may include other diverse information like stock market quotations and weather forecasting. A personal homepage may contain information on different interests of its owner. Each topic usually occupies a separate area in the page. We call each topic area a *Micro Information Unit* (MIU).

MIU is a coherent topic area according to its content, and it is usually also a visual block from the display point of view. If the user's query terms (or keywords) occur in a single MIU of a Web page, the pages returned by a search engine are generally relevant and useful. However, if the keywords scatter at different MIUs, it can cause low precision of the returned search results.

In this paper, we propose a technique to deal with the problem. The key idea is to segment each Web page to identify different micro information units or topic areas according to its HTML tags and contents. In searching, if the keywords of a query occur in the same MIU, the Web page will be given a higher ranking score. Otherwise, it will be given a lower ranking score. In the proposed technique, page segmentation and indexing according to MIUs in a Web page is done in off-line pre-processing. We show that the additional information on MIUs can be naturally integrated with inverted lists indexing commonly used by Web search engines. In on-line search, our retrieval and ranking algorithm makes use of this MIU information to sort the relevant pages. Due to seamless integration of MIUs with inverted lists, additional computation required in searching is minimum.

The proposed technique is intended to be used as an advanced search option or technique for a search engine (which we also call the *base search engine*). That is, when the precision of the results returned by the base search engine is low, we can employ the proposed technique to re-rank the results. To evaluate the proposed technique, we use Google as the base search engine. Experimental results show that our method is able to improve Google's search precision dramatically.

2. SEGMENTATION AND RANKING
Our proposed technique first builds a *HTML tag tree* for each Web page using the nested structure of HTML source codes. A *node* in the tag tree contains a tag name, content text and its display attributes (color, font, size, etc). Using the tag tree, we segment the Web page into various MIUs. However the tag tree is often too refined and is solely based on presentation features of the page. Hence, we merge some nodes in the tree to form a coherent topic or information units. Merging of nodes is done by: (1) merging each heading and its immediate content paragraph; (2) merging two adjacent text paragraphs.

As in normal search, we also use inverted lists to store the information of the Web pages. Thus, the search technique we adopted is similar to those in a normal search engine [1]. The main difference is we need to index and retrieve MIUs of each page. We simply add an extra data structure to each inverted list node to indicate in which MIUs each word appears.

Our ranking algorithm computes two scores for each page, a *primary score* and a *secondary score*. The primary score is the maximum number of query terms that occur in a MIU of a page. If the primary score of the page is less than the number of query terms (i.e., not all query terms are covered), we compute the secondary score, which takes into account of the neighboring MIU on the right of each MIU in the same sub-tree. During ranking, pages with higher primary scores are ranked at the top, followed by pages with higher secondary scores.

Note that a normal search engine typically considers many factors in its ranking algorithm, e.g., *hyperlink information*, *word count-weight*, *type-weight* (title, anchor, URL, font size, etc), and *type-prox-weight* (how close multi-words occur in every type) [1]. In our ranking algorithm, we only focus on whether the query terms occur in a single MIU (or 2 neighboring MIUs within the same sub-tree) of a page. Since the proposed technique should be used as an *advanced search* method for a *base search engine*, we make use of our MIU-based information and also the ranking information from the base search engine in our final ranking

process. That is, when the primary/secondary scores are the same for some pages, we follow the ranking of the base search engine. Hence, we do not need to consider other factors except our MIU-based factor in our ranking algorithm. If we have access to a search engine system, all factors should be integrated in a more sophisticated manner. Section 3 shows that even this simple approach is already able to produce remarkably good results.

3. EXPERIMENTAL RESULTS

Evaluation of the ranking effectiveness is hard in the context of Web search because of the difficulties in (i) *choosing queries* and (ii) *evaluating the relevance* of search results. We select queries from two independent sources. First we choose 40 queries from Metaspy of MetaCrawler [5] (which allows users to view others' queries being submitted to the system). We keep 25 queries from 40 queries after we excluded two kinds of queries:

1) Single term queries: The proposed method does not improve search if the user's search query consists of only a singular keyword as page segmentation is irrelevant in this case. However, in reality the average number of terms used in a search query is 2.21 [3] because the desired search results usually cannot be easily captured by a singular keyword.
2) Ambiguous queries: The intent of the query (with a singular meaning) has to be agreed upon by a panel of 3 judges.

We also randomly selected 45 queries from TREC [6] (15 queries from TREC1, TREC 6 and TREC7 each). TREC is a benchmark for text retrieval and provides a standard narrative for each query. The Web pages produced should satisfy the conditions pre-defined by our judges or correspond to the standard narratives provided by TREC. For example, TREC Query 354: Journalist Risks, the narratives stated are "*any document identifying an instance where a journalist has been killed, arrested or taken hostage in the performance of his work is relevant.*" Our judges evaluate the relevance of the search results with such narratives to obtain a consensus on the search precision.

The choice of using Google as a basis for re-ranking (*base search engine*) is because of its state-of-the-art search mechanism. In general, Google performs very well as a general-purpose search engine. However, there exist many query phrases that it fails to perform satisfactorily. Since our purpose is to provide advanced re-rankings, we only consider those queries whose Google's precisions are low. For each query, we re-rank the first 200 search results from Google.

In the context of Web search, many researchers believe that high precision of the top-ranking results returned by a search engine is more important even at the expense of recall [1]. In our experiments, we only use precision of top 20 results to evaluate the performance (shown in table 1). Note that those queries that our method does not make significant improvements are not included. We also include the search results from AltaVista for comparison. The first column gives the query phrases. The second, third and fourth columns list the precisions of the top 20 results from our method, Google and AltaVista respectively.

From Table 1, we observe that the precision after our re-ranking is substantially higher. On average over the 20 search queries, the absolute gain in precision by our system over that of Google is 28% and that of AltaVista is 39%. Figure 1 gives the graphical comparison for average precision of every 5 pages from the 20 results. It shows that our technique improves the precision of other search engines significantly. Most notably, the average precision for the top 5 results increases considerably from 0.34 (Google) to 0.65 (our system). High precisions for the top 5 results are essential and critical in practice.

	Search Query	New	Google	Alta
1	alternative music origins	0.70	0.40	0.05
2	Christmas Island tour	0.60	0.40	0.35
3	decorative candlestick sale	0.70	0.60	0.30
4	free download music	0.60	0.20	0.30
5	html tag tree	0.55	0.50	0.10
6	information history tomatoes	0.70	0.40	0.30
7	literary films list	0.60	0.20	0.10
8	red ladies t-shirt	0.50	0.40	0.40
9	Singapore programming jobs	0.70	0.20	0.05
10	supermodel success stories	0.70	0.50	0.05
11	airbus subsidies	0.70	0.55	0.10
12	British Chunnel impact	0.50	0.25	0.40
13	computer aided crime	0.50	0.20	0.20
14	dismantling Europe's arsenal	0.60	0.20	0.25
15	encryption equipment export	1.00	0.70	0.50
16	journalist risks	0.60	0.05	0.20
17	leveraged buyouts	0.45	0.10	0.05
18	most dangerous vehicles	0.55	0.35	0.40
19	new hydroelectric projects	0.60	0.20	0.40
20	transportation tunnel disasters	0.40	0.30	0.10
	Average	0.62	0.34	0.23

Table 1: Experiment results (first 10 queries from Metaspy, and next 10 queries from TREC)

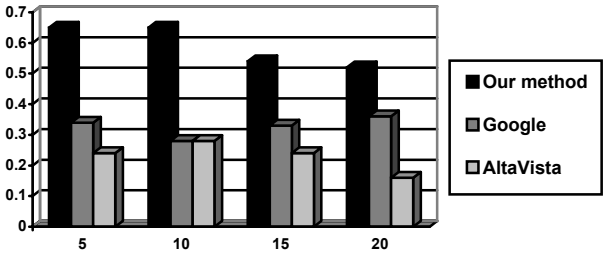

Figure 1: Average precision comparison per 5 pages

4. CONCLUSION

We have presented a technique to improve the precision of Web search, by segmenting each Web page into different MIUs (topic areas) according to its contents and HTML tags. Only the terms in a single MIU or at most two neighboring MIUs are used to match a user's search query. This is different from existing search engines' techniques, which typically employ all terms in the entire page to match the query terms. From experimental results, we observe higher precision of the ranking produced by our method.

5. REFERENCES

[1]. S. Brin, L. Page. The anatomy of a large-scale hypertexture Web search engine. *Computer Networks*, 1998.

[2]. F. Y. Choi. Advances in domain independent linear text segmentation. *NAACL'00*, Seattle, USA, 2000.

[3]. B. J. Jansen, The effect of query complexity on Web searching results. *Information Research*, 6(1), 2000.

[4]. W. S Lee, K. S. Candan, V. Quoc and D. Agrawal. Retrieval and organizing Web pages by Information Unit. *WWW10*, Hongkong, 2001.

[5]. MetaCrawler Search Engine www.metacrawler.com, Metaspy www.metaspy.com.

[6]. Text REtrieval Conference (TREC) Data - English Test Questions (Topics) File List. http://trec.nist.gov/data/topics_eng/index.html