Asymmetrical Query Recommendation Method Based on Bipartite Network Resource Allocation

Zhiyuan Liu
Department of Computer Sci. & Tech.
State Key Lab on Intelligent Tech. & Sys.
Tsinghua University, Beijing 100084, China
liuliudong@gmail.com

Maosong Sun
State Key Lab on Intelligent Tech. & Sys.
National Lab for Information Sci. & Tech.
Tsinghua University, Beijing 100084, China
sms@tsinghua.edu.cn

ABSTRACT

This paper presents a new query recommendation method that generates recommended query list by mining large-scale user logs. Starting from the user logs of click-through data, we construct a bipartite network where the nodes on one side correspond to unique queries, on the other side to unique URLs. Inspired by the bipartite network based resource allocation method, we try to extract the hidden information from the Query-URL bipartite network. The recommended queries generated by the method are asymmetrical which means two related queries may have different strength to recommend each other. To evaluate the method, we use one week user logs from Chinese search engine Sogou. The method is not only 'content ignorant', but also can be easily implemented in a paralleled manner, which is feasible for commercial search engines to handle large scale user logs.

Categories and Subject Descriptors: H.3.3[Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation.

Keywords: Asymmetrical query recommendation, user log analysis, network resource allocation, bipartite network.

1. INTRODUCTION

With the development and popularity of WWW, billions of web pages are retrievable via search engines like Google. Despite it is not a perfect method to find what we want, most search engines still use keywords in documents and queries to calculate the relevance. As the only interface for users accessing tremendous web pages, queries are one of the most important factors that affects the performance of search engines. However web pages returned from search engines are not always relevant to user search intentions. An independent survey of 40, 000 web users found that after a failed search, 76% of them will try to rephrase their queries on the same search engine instead of resorting to a different one [3]. Thus an effectively improved method for commercial search engines is to recommend related queries. Query recommendations not only enhance the hit rate of search engines, but also help users to find the target information more quickly and conveniently.

There have been many related works on query recommendation which were usually carried out on bipartite networks constructed from user logs with one node set contains only

queries and the other contains only URLs. In Ref. [3], a 'content ignorant' method using bipartite-network-based iterative clustering was proposed to cluster both queries and URLs and then generate a list of related query formulations for selected queries. Ref. [5] proposed a more well designed solution for query clustering by combining content based clustering and link-based clustering together. And in Ref. [1], click-through data were used to get recommended queries. The closest work to our method is by Ref. [2] where query relations were extracted from the Query-URL bipartite network based on the set relations among URLs that connect two query nodes. In fact query recommendation is an important application of Recommendation System [4].

The existing methods mostly recommend queries symmetrically, which means two related queries recommend each other with the same strength. However in most instances two related queries should be assigned different strength to recommend each other. For example, the recommending strength for the query '2008' with '奥运'(Olympics) may be stronger than that for '奥运' with '2008'. With the query '2008' Chinese users most likely want to get the information on the 2008 Olympic Games, therefore we may recommend '奥运' strongly. While with the query '奥运', users may have more complicated intensions and we may not strongly recommend '2008'. It's obvious that this feature may significantly affect the recommendation performance. In Ref. [2] asymmetrical relations can be extracted only if one URL set clicked by a query completely covers another URL set. In this paper we propose to use an asymmetrical query recommendation method based on bipartite network resource allocation dynamics [6] using user logs which can extract more general asymmetrical relations. The method is originally applied as a personal recommendation algorithm. It is reported that, in spite of its simplicity, the method performs much better than the most commonly used global ranking method as well as the collaborative filtering method [6], and it has three prominent feature, i.e. asymmetrical, parallelable and 'content ignorant'.

To implement the method, we need to construct a weighted Query-URL bipartite network from query log data. The click frequency in a user log dataset from one query to one URL is usually different, which suggests the matching degree between search intensions and semantics behind the URLs. Thus it is essential to assign weights to the links between queries and URLs based on their click frequencies. To recommend queries for the query Q based on the network, we initially assign Q with resource r indicating the user search intensions behind it. Then the resource-allocation

Table 1: Examples of recommended queries based on our method.

Query	Recommendations
2008	奥运, 奥运会, 2008奥运会, 北京奥运会, 北京奥组委, 北京奥运, 北京奥组委官方网站, 2008北京奥运会, 2008奥运
奥运	奥运会, 2008奥运会, 北京奥运会, 北京奥运, 北京奥组委, 北京奥组委官方网站, 2004奥运会, 2008北京奥运会, 2008

Table 2: Comparisons of query recommendation results to '小说'(novel) based on our method(AQRM) with Baidu and Google.

Source	Recommendations
AQRM	玄幻小说, 起点, 小说阅读网, 言情小说, 潇湘书院, 起点中文网, 小说网, 幻剑书盟, 起点中文
Baidu	言情小说, 小说阅读网, 玄幻小说, 小说网, 网游小说, 小说下载, 武侠小说, 网络小说, 小说520
Google	言情小说, 玄幻小说, 小说阅读网, 武侠小说, 小说下载, 小说网, 网游小说, 玄幻小说网, txt小说下载, 言情小说网

is processed which consists of two steps. Firstly, the resource in query nodes (Initially only Q has the resource.) is distributed proportionally according to the link weights to their neighbor URL nodes. Analogously, the resource in URL nodes is proportionally distributed back to their neighbor query nodes. The process can be run recursively. The final resource located in the query nodes denotes the distribution of search intensions behind the original query and thus indicates the recommending strength of each query for Q.

2. EXPERIMENTS AND ANALYSIS

Our study was carried out on the user logs of the first week in March 2007 which contains 1.3 million unique queries and 4.1 million unique URLs, obtained from Chinese commercial search engine Sogou (`www.sogou.com`). For each query, we assigned resource $r = 100$, ran the resource-allocation process with only one loop and recorded a maximum of nine recommendations. Table 1 shows some examples. The recommended queries are listed by recommending strength in reverse order. For the query '2008', '奥运' is positioned at the first place with strength 20.83. While for '奥运', '2008' is positioned at the last place with strength 1.22.

We also compare our method with some commercial search engines, i.e. Baidu (`www.baidu.com`) and Google (`www.google.cn`). As shown in Table 2, we compare the recommendation results for query '小说'(novel). The two search engines only recommend those queries containing the original query as substring. On the contrary our method is based on link analysis and 'content ignorant', therefore we can recommend related queries with no common substrings, which extends the recommendation range widely.

Users' perception is always subjective but still indicates the performance of search engines to some extent. Therefore we use editors' ratings to evaluate our query recommendation method. We randomly select about 180 recommended queries and ask editors to rank the recommended query results from 5 to 0, where 5 means very good and 0 means totally unrelated. For precision, the editors are required to review whether the recommended queries are really related to the original query. For coverage, the editors should de-
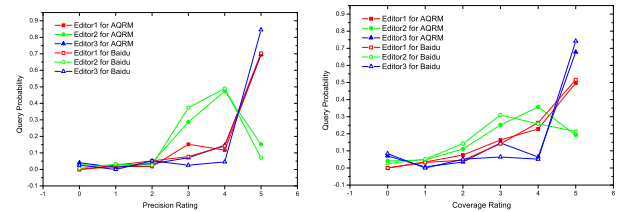

Figure 1: Precision(left panel) and coverage(right panel) ratings for 180 recommended queries.

cide whether they would click the recommended queries in a real search scenario. To compare our method with Baidu, Fig. 1 shows how many recommended queries are ranked in different specific ranks [1]. For our method the average precision ratings are 4.46, 3.61 and 4.36; the average coverage ratings are 4.08, 3.42 and 4.16. While for Baidu the average precision ratings are 4.45, 3.52 and 4.62; the average coverage ratings are 4.19, 3.36 and 4.24. It can be seen that the performance of our method is comparable with Baidu.

3. CONCLUSION AND FUTURE WORK

In this paper we proposed an asymmetrical query recommendation method based on network resource allocation using user logs which is simple to implement with low computational cost. The method is not only 'content ignorant', but also can be easily implemented in a paralleled manner which is convenient for commercial search engines to mine the large-scale user logs.

The approach here is still preliminary and needs further exploration. Future work includes: (1) the content based method, such as the common substring method used by Baidu and Google, is expected to be combined with link analysis for mutual complementation; and (2) more rigorous evaluation will be designed by monitoring the real users choices, and our method needs to be further compared with the related research work given in Ref. [1, 3].

4. ACKNOWLEDGEMENTS

5. REFERENCES

[1] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query recommendation using query logs in search engines. *EDBT 2004 Workshop on Current Trends in Database Tech.*, 2004.

[2] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *Proc. of KDD*, 2007.

[3] D. Beeferman and A. Berger. Agglomerative clustering of a search engine query log. In *Proc. of KDD*, 2000.

[4] M. Deshpande and G. Karypis. Item-based top-n recommendation algorithms. *ACM TOIS*, 22(1), 2004.

[5] J.-R. Wen, N. Jian-Yun, and Z. Hong-Jiang. Query clustering using user logs. *ACM TOIS*, 20(1), 2002.

[6] T. Zhou, J. Ren, M. Medo, and Y. C. Zhang. Bipartite network projection and personal recommendation. *Physical Review E*, 76(4), 2007.

---

[1] All the rating data by three editors can be accessed through `http://nlp.csai.tsinghua.edu.cn/~lzy/qr/qr.zip`.