Ranking Web Pages by Associating Keywords with Locations

Peiquan Jin, Xiaoxiang Zhang, Qingqing Zhang, Sheng Lin, and Lihua Yue

University of Science and Technology of China, 230027, Hefei, China
jpq@ustc.edu.cn

Abstract. Many Web queries contain both textual keywords and location words. When answering such queries, the association between the textual keywords and locations in a Web page should be taken into account. In this paper, we present a new ranking algorithm for location-related Web search, which is called MapRank. Its main idea is to extract the associations between keywords and locations in Web pages and further use them to improve ranking effectiveness. We first determine map each keyword with specific locations and form a set of < keyword, location > pairs. Then, we compute the location-constrained score for each keyword and combine it into the ranking procedure. We conduct comparison experiments on a real dataset and use the metrics including MAP and NDCG to measure the performance of MapRank. The results show that MapRank is superior to previous methods with respect to different symbolic-location-related queries.

Keywords: ranking algorithm, symbolic location, Web search, association.

1 Introduction

Ranking algorithms, e.g., the *Pagerank* algorithm, have been one of the major technologies in search engines. Unfortunately, traditional ranking algorithms are based on link analysis and textual relevance, and are hard to satisfy different querying needs. Besides, the textual-relevance-based ranking approach does not consider the relationship between textual keywords and location names in a query. On the other hand, many Web pages are associated with certain locations, e.g., news report, retailer promotion and so on. The study in the literature [1] reported that among 2,500 queries, 18.6% of them contained a geographic predicates and 14.8% of them included a location name. Therefore, how to extract locations for Web pages and use them in Web search has been a hot and critical issue in current research on Web search [2-4].

In this paper, we present a new location-aware ranking algorithm for Web search, which is called *MapRank*. MapRank aims to improve the ranking performance for spatial textual Web queries that contain both textual keywords and location words. The algorithm considers both textual and location relevance between Web pages and querying terms when returning the results, and can improve the effectiveness of Web search engines. The contributions of the paper can be summarized as follows:

(1) We propose a new ranking algorithm named *MapRank* for spatial textual Web queries. MapRank is implemented using a two-staged strategy, namely an offline

J. Wang et al. (Eds.): WAIM 2013, LNCS 7923, pp. 613–618, 2013.

stage extracting and building <*keyword*, *location, score*> pairs for Web pages and an online stage computing the final ranking score. The new algorithm considers both textual and location relevance in the ranking process, and also takes into account the relationship between keywords and locations in a Web page.

(2) We conduct comparison experiments on various real datasets crawled from New York Time, to measure the performance of the MapRank algorithm. The experimental results show that the proposed MapRank algorithm has the best performance with respect to different spatial textual queries.

2 Related Work

Spatial textual queries are usually represented as a triple <*what*, *relation*, *where*>. However, as the "In" relation is the most appropriate one for Web search, spatial textual queries in Web search engines can be simplified as <*what*, *where*>. Location-related ranking algorithms are specially designed to cope with spatial textual queries in Web search engines. Basically, a location-related ranking algorithm has to consider both textual relevance and location relevance between query and Web pages. The challenging issues are to determine location relevance and combine textual and location relevance during the ranking process.

There are two major methods to compute the location relevance. One of them is to utilize the relation words in queries [5-7]. For example, many spatial queries contain some spatial relation words such as "inside", "overlap", and "nearby". The other type does not use the explicit spatial relation words in queries, but determines spatial relations based on the geographic attributes of the spatial objects in queries [8, 9]. For instance, the spatial object "Paris" in the query can be mapped into a determined geographic extent.

Answering spatial textual queries need to consider both textual and location relevance. The native way is to combine them using a linear-weighted method. Some other works do not use the combination of textual relevance and location relevance. Li et al. [10] introduced a topic model to determine the topic of Web queries as well as Web pages. As a topic usually has a distribution among geographic extent, they proposed to utilize a Gauss formula to simulate the geographic distribution of a topic. The geographic distribution model of topics is then taken to determine the ranking scores of Web pages. Some researchers proposed ranking mechanisms which combines various metrics in textual relevance computation and location relevance determination. For example, Martins et al. [11] proposed an SVM-based optimized MAP method, called SVMmap, and Cai et al. [12] proposed the GeoVSM, which is a geographic optimized VSM model.

More recent works on spatial textual query processing are conducted by Gao Cong et al. [2, 3], and a spatial textual search engine called SWORS [2] is designed. The locations in SWORS are with the similar semantics in traditional geographical information systems. Therefore, spatial queries involving geographical relations such as "nearby" and "close to" have to be resolved by some new indexing structures, e.g., IR-Tree [3]. Regarding the ranking techniques, SWORS considered both textual and spatial relevance, which was similar with the previous solutions.

3 The MapRank Algorithm

3.1 The Basic Idea

MapRank is a location-aware ranking algorithm for spatial textual Web search. It considers both textual relevance and location relevance of Web page. The basic idea of MapRank can be described as follows:

(1) MapRank considers the association of keywords and locations when computing the scores of Web pages. In particular, we map each keyword in a Web page with a specific focused location and then calculate the location-constrained score of each keyword. As a result, we construct a set of <*keyword*, *location*, *score*> pairs for each Web page, in which the location represents the most relevant focused location of the given keyword.

(2) We use a two-staged design to implement the MapRank algorithm. The first offline stage is to construct the <*keyword*, *location*, *score*> pairs for each Web page. The second online stage is to compute the final ranking scores for all the Web pages. In the second stage, we combine two factors, namely the relevance between the querying location and the focused locations in Web page, and location-constrained keyword score, to achieve a tradeoff between location relevance and textual relevance.

The main difference between MapRank and other existing ranking algorithms is that it combines the focused locations of Web page into the ranking algorithm. Furthermore, the <*keyword*, *location*, *score*> mapping policy also introduces a reasonable solution to integrate textual and location relevance into the ranking algorithm.

3.2 Constructing <keyword, location, score> Pairs

The focused location refers to the most appropriate location associated with a Web page. The extraction of focused locations is performed by the algorithm discussed in our previous work [13]. It returns a set of focused locations for each Web page.

Generally, a spatial textual query contains several keywords and one location word. Moreover, the keywords and the location word in most spatial textual queries usually imply some relationships which represent users' indeed searching needs. For example, a query "Massachusetts population statistics" actually means that users want to find the population statistics of the state "Massachusetts". Here, the text keywords "population statistics" and the location word "Massachusetts" in the query have an intrinsic relationship. In general, the relationship between a keyword and a location can be represented as a pair <*keyword*, *location*>, which indicates that the given keyword is mostly related with the location.

Current search engines deals with the text keywords and location words individually and ignores the relationship between the keywords and locations. Our MapRank algorithm is designed to present a better solution for this problem. We will consider the relationship between keywords and locations when performing the ranking task. This idea is motivated by the temporal textual ranking approach proposed in [14]. Basically, we first find the most relevant focused location for each keyword, and construct <*keyword*, *location*> pairs. After that, we compute the location-constrained ranking score of each keyword, which finally forms the list of <*keyword*, *location*, *score*>.

We use three constant values to measure the scores for <keyword, location> pairs, namely TITLE_SCORE, SENT_SCORE, and PARA_SCORE. The TITLE_SCORE is used to represent the score of <keyword, location> when the keyword and associated location word are both contained in the title of the Web page. The SENT_SCORE is used when the keyword and associated location word are both contained in the same sentence, but not in the title. The PARA_SCORE is used when the keyword and associated location word are both contained in the same paragraph, but not in the same sentence. Generally, we have the following assumption: TITLE_SCORE > SENT_SCORE > PARA_SCORE.

3.3 Computing the Ranking Scores of Web Pages

When users post a query to the search engine, the final ranking scores of Web pages are computed dynamically. In this paper, the final ranking score of a Web page is calculated based on two factors, namely the relevance between the querying location and the focused locations in Web page, and location-constrained keyword score.

Given a Web page *D*, a set of querying keywords, say $< w_1, w_2,..w_n >$, and a querying location *g*, the computation of the final ranking score is based on the following algorithm (as shown in Fig.1).

```
Algorithm Location_Ranking(D, Q)

Input: (1) a Web page D, which has the list of <keyword, location, score> pairs K =
           {<k_1, l_1, sc_1>, < k_2, l_2, sc_2>, …,< k_m, l_m, sc_m>}.
       (2) a query Q including a keywords set $W =< w_1, w_2,..w_n >$ and a location g
Output: GS, the ranking score of D according to Q
Preliminary: n is the count of focused locations.

/* Removing the keywords unrelated with Q   from K */
for each <k, l, sc>$\in$ K do
    if k$\notin$ W then
        K = K–<k, l, sc>; // remove unrelated keywords
/* Computing ranking score */
for each <k, l, sc>$\in$ K do
    $GS = GS + \frac{common(l, g)}{max(l, g)} \cdot sc$
return GS;
```

Fig. 1. Computing the ranking score for Web page

In Fig.1, we emphasize the location relevance between the user query and Web pages. This is done by introducing location relevance into the computing procedure of ranking scores, as defined by the formula 3.1.

$$location \quad relevance = \frac{common(l, g)}{max(l, g)} \tag{3.1}$$

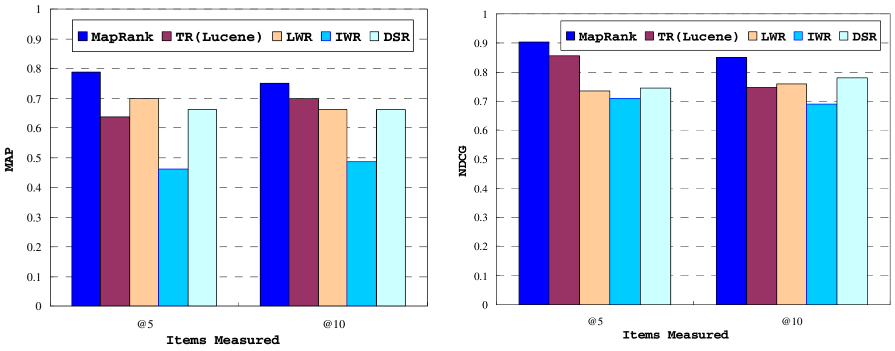
Given two locations *l* and *g*, *common*(*l*, *g*) in Formula 3.1 is defined as the length of the common prefix of *l* and *g* in the location tree generated from Gazetteer, and *max*(*l*, *g*) refers to the maximum length of *l* and *g*. For example, suppose that

l = “USA/Massachusetts/Peak Stone”, g = “USA/Arizona”, *common*(l, g) is 1 and *max*(l, g) is 3.

4 Experimental Results

To evaluate the performance of MapRank, we conduct an experiment on a dataset crawled from New York Time, which contains 311,187 Web pages ranging from 2006 to 2011. For every Web page in the dataset, we cut the page into words using HTMLParser (http://htmlparser.sourceforge.net/) and use using the stop-words database provided by SMART (http://www.lextek.com/manuals/onix/stopwords2.html) to filter stop words. Then we stem the keywords using the Porter Stemmer tool (http://tartarus.org/~martin/PorterStemmer/). After that, we extract the focused locations for each Web page and construct <*keyword*, *location*, *score*> pairs. The extracted keywords are maintained in Lucene 3.5, which will be used when executing comparison algorithms.

In the experiments, we run 16 spatial textual queries and use two metrics to measure the performance of each algorithm, i.e., MAP and NDCG. We implement the algorithms using Java under the developing environment ObjectWebLomboz. The test machine has an Intel Dual Core Processor, 2GB of main memory, and is running Windows XP Professional.



(a) MAP (b) NDCG

Fig. 2. MapRank vs. other four comparison algorithms

Figure 2(a) shows the MAP@5 and MAP@10 scores of MapRank and other four competitor algorithms, namely Textual Ranking (TR), Linear Weighted Ranking (LWR), Improved Weighted Ranking (IWR), and DS Ranking (DSR). The LWR approach uses the linear weighted sum of textual relevance and location relevance as the ranking score of Web page. The IWR approach is similar with the linear weighted ranking one, except that it uses an improved way to determine the parameter ω. In particular, it uses the method in the literature [15] and determines the ω value on the basis of the description of textual words and location words. The DSR approach is based on the Dempster-Shafer (DS) theory of evidence, and uses different types of evidences to determine the possibility of an event [15]. The textual relevance and location relevance can be regarded as two individual evidences for spatial textual ranking. In our experiment, we use the weight of importance for each textual word and location word to determine the uncertainty of those two evidences, and rank Web pages according to the uncertainty. Figure 2(b) shows the different NDCG scores. As shown in Fig.2, MapRank gets the best MAP and NDCG scores in all cases.

5 Conclusion

In this paper, we introduce the MapRank algorithm which is based on the association between the focused locations of Web page and keywords. It presents an appropriate tradeoff between textual relevance and location relevance. The experimental results show that the MapRank algorithm has better performance for spatial textual queries than its competitors. Next we will integrate our algorithm with temporal information in Web pages.

Acknowledgement. This paper is supported by the National Science Foundation of China (No. 60776801 and No. 71273010), the National Science Foundation of Anhui Province (no. 1208085MG117), and the USTC Youth Innovation Foundation.

References

1. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: Proc. of GIR (2004)
2. Cao, X., Cong, G., Jensen, C.S., et al.: SWORS: A System for the Efficient Retrieval of Relevant Spatial Web Objects. PVLDB 5(12), 1914–1917 (2012)
3. Cong, G., et al.: Efficient Retrieval of the Top-k Most Relevant Spatial Web Objects. In: Proc. of VLDB (2009)
4. Lu, J., Lu, Y., Cong, G.: Reverse Spatial and Textual K Nearest Neighbor Search. In: Proc. of SIGMOD, pp. 349–360 (2011)
5. Zhou, Y., Xie, X., Wang, C., et al.: Hybrid Index Structures for Location-based Web Search. In: Proc. of CIKM, pp. 155–162. ACM, New York (2005)
6. Martin, B., Silva, M., et al.: Indexing and Ranking in Geo-IR Systems. In: GIR 2005 (2005)
7. Andrade, L., et al.: Relevance ranking for geographic information retrieval. In: GIR 2006 (2006)
8. Jones, C.B., Alani, H., Tudhope, D.: Geographical Information Retrieval with Ontologies of Place. In: Montello, D.R. (ed.) COSIT 2001. LNCS, vol. 2205, pp. 322–335. Springer, Heidelberg (2001)
9. Larson, R.: Ranking approaches for GIR. SIGSPATIAL Special 3(2) (2011)
10. Li, H., Li, Z., Lee, W.-C., et al.: A Probabilistic Topic-Based Ranking Framework for location-sensitive domain information retrieval. In: Proc. of SIGIR, pp. 331–338 (2009)
11. Martins, B., Calado, P.: Learning to Rank for Geographic Information Retrieval. In: Proc. of GIR (2010)
12. Cai, G.: GeoVSM: An Integrated Retrieval Model for Geographical Information. In: Proc. of GIS, pp. 65–79 (2002)
13. Zhang, Q., Jin, P., Lin, S., Yue, L.: Extracting Focused Locations for Web Pages. In: Wang, L., Jiang, J., Lu, J., Hong, L., Liu, B. (eds.) WAIM 2011 Workshops. LNCS, vol. 7142, pp. 76–89. Springer, Heidelberg (2012)
14. Jin, P., Li, X., Chen, H., Yue, L.: CT-Rank: A Time-aware Ranking Algorithm for Web Search. Journal of Convergence Information Technology 5(6), 99–111 (2010)
15. Yu, B., Cai, G.: A Query-Aware Document Ranking Method for Geographic Information Retrieval. In: Proc. of GIR, pp. 49–54. ACM, New York (2007)