Computers in Human Behavior 30 (2014) 378–388

Contents lists available at SciVerse ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Exploiting location information for Web search

Jie Zhao [a], Peiquan Jin [b,*], Qingqing Zhang [b], Run Wen [a]

[a] School of Business, Anhui University, 230039 Hefei, China
[b] School of Computer Science and Technology, University of Science and Technology of China, 230027 Hefei, China

ARTICLE INFO

Article history:
Available online 4 May 2013

Keywords:
Ranking algorithm
Spatial textual query
Web search
Location extraction

ABSTRACT

Most Web pages contain location information, which are usually neglected by traditional search engines. Queries combining location and textual terms are called as spatial textual Web queries. Based on the fact that traditional search engines pay little attention in the location information in Web pages, in this paper we study a framework to utilize location information for Web search. The proposed framework consists of an offline stage to extract focused locations for crawled Web pages, as well as an online ranking stage to perform location-aware ranking for search results. The focused locations of a Web page refer to the most appropriate locations associated with the Web page. In the offline stage, we extract the focused locations and keywords from Web pages and map each keyword with specific focused locations, which forms a set of <keyword, location> pairs. In the second online query processing stage, we extract keywords from the query, and computer the ranking scores based on location relevance and the location-constrained scores for each querying keyword. The experiments on various real datasets crawled from nj.gov, BBC and New York Time show that the performance of our algorithm on focused location extraction is superior to previous methods and the proposed ranking algorithm has the best performance w.r.t different spatial textual queries.

1. Introduction

Nowadays, Web search engines such as Google and Bing have become necessary tools in people's daily life. Generally, the effectiveness of those search engines is mostly determined by ranking algorithms, e.g., the *Pagerank* algorithm (Brin & Page, 1998; Haveliwala, 2002). Unfortunately, traditional ranking algorithms are based on link analysis and textual relevance (Bordin, Roberts, et al., 2005), and are hard to satisfy different querying needs. For example, such queries like "Nike sales information in Massachus" are very difficult to be evaluated in Google and other similar search engines, because they all treat the location name "Massachus" as a textual keyword. Therefore, the Web pages reporting Nike sales information in Peak Stone, which is a town in Massachus, will not be returned or ranked correctly, because the search engines are not aware of the spatial containment relationship between Massachus and Peak Stone. Another problem is that the textual-relevance-based ranking approach does not consider the relationship between textual keywords and location names in a query. For instance, for the query specifying "Nike sales information in Massachus", if a Web page first reports the Nike sales information in Boston and then describes a sports meeting in Massachus, it will be returned by traditional search engines and with a high ranking score, because all the querying keywords appear in this Web page. The main reason is that they have not considered much on the relationship between the querying keywords ("Nike sales information" in the above example) and locations ("Massachus" in the example).

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 (Sanderson and Kohler, 2004) reported that among 2500 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 then use them in Web search process has been a hot and critical issue in current research on Web search.

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 relevance 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 introduce the focused locations of Web pages into the ranking process, and present an effective algorithm to extract the focused locations of Web pages.
(2) We propose a new ranking algorithm named *MapRank* for spatial textual Web queries. MapRank is implemented using a two-staged strategy, namely an offline stage extracting and

* Corresponding author. Tel.: +86 13955162813.
E-mail address: jpq@ustc.edu.cn (P. Jin).

0747-5632/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.chb.2013.04.023

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.

(3) We conduct comparison experiments on various real datasets crawled from nj.gov, BBC and New York Time, to measure the performance of focused locations extraction as well as the MapRank algorithm. The experimental results show that the performance of our algorithm on focused location extraction is superior to previous methods and the proposed MapRank algorithm has the best performance with respect to different spatial textual queries.

The remainder of the paper is organized as follows. In Section 2, we survey the related work. In Section 3 we introduce the extraction of the focused locations for Web pages. Section 4 describes the details about the MapRank algorithm. In Section 5, the experiments and the performance evaluation results are discussed. Finally, Section 6 concludes the paper.

2. Related work

2.1. Location extraction for Web pages

There are a lot of related works in locations detection. In (Wang et al., 2005a), the authors proposed an algorithm to extract dominant locations from Web queries, while in this paper we deal with the locations in Web pages. Queries are often short, and they use location keywords, combined with search logs and top search results to help finding locations by most of people who know the answer to the query. Wang et al. (2005b) classified locations of Web resources into three kinds, namely provider location, content location and serving location. And they used hyperlinks, user logs and Web content to detect those three types of locations. However, the locations in Wang et al. (2005b) represent GPS positions, which are not suitable in Web search, because it is not possible for users to input a GPS position as Web query.

The most related works are Web-a-where (Amitay, Har'El, Sivan, & Soffer, 2004) and the evidence-based method proposed in Wang, Zhang, Chen, and Lin (2010). Web-a-where is a four-step heuristics algorithm to determine focused locations for Web pages, in which all names were assigned a location with a confidence score. Based on those confidence scores, as well as other information such as frequency, and location relationships, the focused locations of a Web page are extracted. However, Web-a-where adopts fixed parameters and thresholds, which are not suitable for different kinds of Web pages. In this paper, we use variable parameters and thresholds and get better performance (will be discussed in Section 5). The evidence-based method proposed in Wang et al. (2010) is an effective algorithm for geo-candidates disambiguation, which makes use of metric relation, topological relation and typological relation between an ambiguous geo-candidate and other co-occurring geo-candidates in the context. Those co-occurring candidates are regarded as the evidences of a geo-candidate, which are fused by the Dempster–Shafer (D–S) theory. However, both of Amitay et al. (2004) and Wang et al., 2010 did not consider the changing confidence that a geo-candidate impacts on other ones, which will lead to bad performance of disambiguation. As shown in our experimental results, the evidence-based method has a comparable performance with Web-a-where in resolving place names ambiguity.

As a Web page usually contains two or more location words, it is necessary to find the focused locations of the Web page. The focused locations represent the most appropriate locations associated with contents of a Web page. Generally, we assume that each Web page has several focused locations. The most difficult issue in determine focused locations is that there are GEO/GEO and GEO/NON-GEO ambiguities existing in Web pages. The GEO/GEO ambiguity refers that many locations can share a single place name. For example, Washington can be 41 cities and communities in the United States and 11 locations outside (Washington, 2012). The GEO/NON-GEO ambiguity refers that a location name can be used as other types of names, such as person names. For example, Washington can be regarded as a person name as George Washington and as a location name as Washington, D.C. Sanderson's work (2000) shows that 20–30% extent of error rate in location names disambiguation was enough to worsen the performance of the information retrieval methods. Due to those ambiguities in Web pages, previous research failed to reach a satisfied performance in focused locations extraction.

On the other side, it is hard to resolve the GEO/GEO and GEO/NON-GEO ambiguities as well as to determine the focused locations of Web pages through the widely-studied named entity recognition (NER) approaches. Current NER tools in Web area aim at annotating named entities including place names from Web pages. However, although some of the GEO/NON-GEO ambiguities can be removed by NER tools, the GEO/GEO disambiguation is still a problem. Furthermore, NER tools have no consideration on the extraction of the focused locations of Web pages. Basically, the NER tools are able to extract place names from Web pages, which can be further processed to resolve the GEO/GEO ambiguities as well as the GEO/NON-GEO ones. Thus, in this paper we will not concentrate on the NER approaches but on the following disambiguation and focused locations determination. Those works differ a lot from traditional NER approaches.

2.2. Traditional ranking algorithms

Ranking is one of the core technologies of Web search engines. A lot of ranking algorithms have been proposed so far, which can be classified into three categories.

2.2.1. Link analysis based ranking algorithms

The first one is the ranking algorithms based on link analysis. The most famous algorithms of this kind are *Pagerank* (Brin & Page, 1998; Haveliwala, 2002) and HITS (Brin & Page, 1998). *Pagerank* determines the ranking order of the Web pages according to the number of Web pages that are linked by other pages in the whole Web. The more the linked number of a Web page, the higher its value is. *Pagerank* is an offline algorithm which does not calculate the ranking scores of Web pages during query processing but before this stage. So it is helpful to reduce the response time of query. However, it will lead to a bad sorting result because it ignores the topic relevance between Web pages and user queries. For example, new Web pages will possibly have low ranking scores and will not return to user even if they are mostly topic-related. The HITS algorithm was proposed by Kleinberg at the end of 1990s (Brin & Page, 1998). It assesses the quality of a Web page by two numerical factors, which are content authority (*Authority*) and link authority (*Hub*). The *Authority* of a Web page is related with its referential count in other Web pages (or in other words, its in-link count). A high *Authority* generally means the Web page is frequently referenced in other pages. Similarly, the Hub of a Web page is related with the quality of its hyperlink (out-link). A high Hub means the Web page references many high-quality Web pages. HITS has to compute the *Authority* and *Hub* based on the link relationships between the resulting Web pages and other pages. This makes it difficult to use HITS in a practical application environment.