USING WEB MINING TO LOCATE UNREGISTERED TOPONYM IN WEB MAP SERVICES

Yingwei Luo, Haibo Wang, Xiaolin Wang and Xiao Pang

Dept. of Computer Science and Technology, Peking University, Beijing, China, 100871
lyw@pku.edu.cn

1. INTRODUCTION

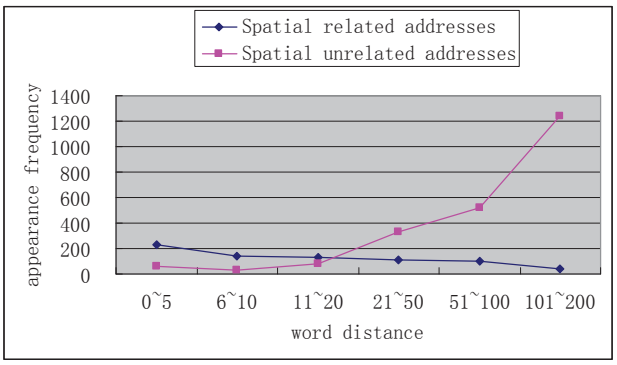
Over the past several years, the Web map service is booming, yet there exist some deficiencies in it. According to the investigation of China Web map service users in 2006 [1], in the statistics of "User's opinion to the Web map service engine", 40% of the users are not satisfied with the map service for lacking of data and updating lag behind, 25% complain the search speed, while 21% claim that the location are not accurate enough. Among all these deficiencies, lacking of data and updating lag behind are mainly to blame.

We search a Beijing toponym "SanHeZhuangSheng Building Co., Ltd." in Baidu Web map service engine and Google Web map service engine, both engines return nothing but some advises on improving search result. Obviously, the toponym doesn't exist in Baidu or Google's spatial database. We define a toponym like this which doesn't exist in spatial database of Web map service engine as Unregistered Toponym. Due to the cost and other factors, it is impossible for the Web map services providers to update their spatial databases in the near future. In this case, how can we return some useful information to the user without update spatial database? As we all know, the Internet is a huge information set. It may contain a great many of address information, while the update rate surpasses any single spatial database. If we can find out the addresses which are included in the spatial database and near to the unregistered toponym, and return them as a result, it may helpful to the user. We have made an investigation about this method, and the statistic result shows that: In the Web page context, the nearer an address to the unregistered toponym, and the more times it appears, it is more possible that the address is closer to the real address of the unregistered toponym in the real world.

Based on this statistic result, we propose a method to locate the unregistered toponym, which does not exist in the spatial database. It takes advantage of Web mining to find out the location of an unregistered toponym according to the fact that there is a lot of related address information on the Web. The principle is: Firstly, it collects Web pages which contain the unregistered toponym and extracts addresses in the pages while also existing in spatial database (those addresses are also called recognized addresses); secondly, it computes the spatial relationship based on the word distance between the unregistered toponym and each recognized address, here we use credibility to express the spatial relationship; thirdly, it tries to improve the credibility by making use of the cluster effect of relative spatial address; lastly, it returns the addresses with highest weight which will help the user to locate the unregistered toponym and shows these addresses on a map. Currently, some people have done some other related work to find addresses in Web pages, so as to provide better address search function in search engine. But our work is quite different. Our goal is to locate unregistered topoym for Web map services.

2. STATISTIC-BASED HYPOTHESIS

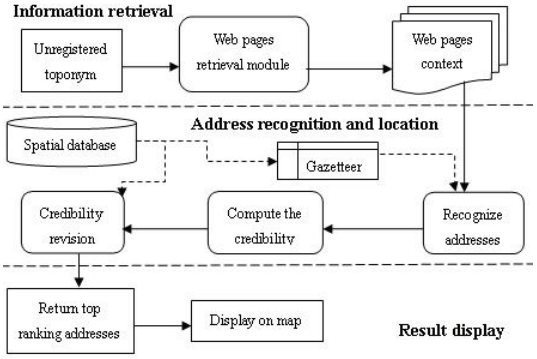
We made an investigation to the unregistered toponym and the distribution of addresses in the context so as to observe their relationship. We randomly selected 42 toponyms as sample, and these toponyms include corporations, restaurants, entertainment places and real estate buildings. Then we collect a lot of Web pages which contain these toponyms through search engine, and treat these pages as sample data set. We artificially statistic and analyze the spatial relationship between the toponyms and the addresses in the context in the sample data set. First, take down the word distance of every address to the nearest toponym; Second, judge artificially whether the address spatially near to the toponym in real world. The right figure shows the distribution (word distance to the toponym) of these addresses. The figure shows that the appearance



frequency of spatial relative address is reversed to the distance to the corresponding toponym; while the situation of spatial unrelated address is on the opposite. So, we can get two conclusions: 1) the smaller an address's distance to the toponym, the more possible that the address is spatial relative to the toponym; 2) the more frequently an address appears in the context near the toponym, the more possible that the address is spatial relative to the toponym. That is, an address is nearer the toponym and appears more frequently, it is more possible that the address is spatial relative to the toponym.

3. A PROTOTYPE SYSTEM TO LOCATE UNREGISTERED TOPNYM

In order to experiment the strategy's feasibility and accurate, we designed and implemented a prototype system to locate unregistered toponym, which is focused on toponyms in Beijing. The system can return location information of unregistered toponym in Beijing, and show the result in Beijing map. It can be applied to other cities in case of changing the spatial database and gazetteer. The prototype system mainly includes three parts: information retrieval module, address recognition and location module and result display module, as shown in right figure. The figure shows the overall flow of the prototype: When the user enters an unregistered toponym, the information retrieval module retrieves Web pages that contain the toponym, and then extracts the context around the unregistered toponym. Based on a spatial database, it then builds a gazetteer, whose element is as form as "address longitude latitude", one element each line. The address recognition and location module extracts the address information from above extracted context. Each extracted address should be included in the gazetteer, and is called as recognized address. The address recognition and location module then computes the credibility of each recognized address, next it tries to improve the credibility based on addresses' coordinates (longitude latitude), and sort addresses by credibility and return several addresses with the highest credibility. At last, the result display module displays the returned addresses on a map.

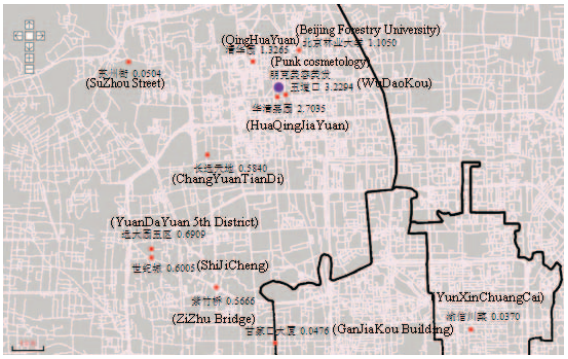


4. COMPUTING CREDIBILITY FOR A RECOGNIZED ADDRESS

We use credibility to indicate the spatial relationship between recognized address and the unregistered toponym. Based on the statistic result, we give a higher weight to the address's credibility if the distance of the address to the unregistered toponym is small, while a lower weight if the distance is large. In this case, the credibility directly expresses the spatial relativity of the address and unregistered toponym. In order to compute the credibility, it firstly define a proper function based on the word distance between the recognized address and the unregistered toponym, with this function it get the primary weight; then multiply the count that a recognized address appears in the context of all retrieved Web pages and finally we get the credibility. Also, we use location based (or geography coordinate based) clustering method to improve credibility for a recognized address. Through clustering, we can categorize the geography close enough recognized addresses, and increase their credibility in direct proportion to the number of addresses in this category.

5. EXPERIMENT AND ASSESSMENT

We take the toponym "Punk cosmetology (朋克美容美发)" as an example. The right figure shows the distribution of the recognized spatial related addresses after the credibility revision. The blue point is the real location of toponym "Punk cosmetology", the red points are the location of each recognized spatial related address on the top of the credibility rankings, with the format "name credibility(revised)".



6. CONCLUSION

This paper proposes a method to locate the unregistered toponym which cannot be located in the Web map service. The method takes advantage of the great many of addresses existing in the Internet to locate the unregistered toponym. The Web map service can use this method to make up for its "lacking of data and updating lag behind" shortcoming.

[References are omitted here]