ATLANTIS PRESS

Advances in Intelligent Systems Research (AISR), volume 145

2017 International Conference on Electronic Industry and Automation (EIA 2017)

Research on Multi-level Gazetteer Services Based on Object Relational Database

Jieyu DONG
National University of Defense Technology
Changsha, Hunan Province, China
e-mail: dongjieyu11@163.com

Luo CHEN
National University of Defense Technology
Changsha, Hunan Province, China
e-mail: luochen@nudt.edu.cn

Mengyu MA
National University of Defense Technology
Changsha, Hunan Province, China
e-mail: mamengyu10@nudt.edu.cn

Ning JING
National University of Defense Technology
Changsha, Hunan Province, China
e-mail: ningjing@nudt.edu.cn

***Abstract*—Gazetteer service is one of the basic functions of Web map services, and the search efficiency and accuracy is very important to the user experience. Traditional approach is based on full-text indexing technology. The ranking of search results is mainly based on the relevance of the search keywords. However, in practical applications, the search results are often classified and sorted. In this paper, we propose a new method which uses gin index of PostgreSQL to achieve the rapid retrieval of phrases containing geo-location information. Search accuracy is optimized by modification of a word dictionary and graded and classified information. A prototype system is designed and implemented. Experiments on real dataset showed that the proposed method can provide users with a better retrieval experience under the environment of big data.**

Keywords-gazetteer services; index optimization, word frequency dictionary, grade and classification

I. INTRODUCTION

As a basic function of web map application, the Gazetteer service plays an important role in smart city domain and geographic information services [1]. In geographic systems, the matching of geographic names and addresses is one of the most important ways to associate spatial data with non-spatial data, which directly related to the quality of the geographic service platform [2]. However, current service models have little consideration in multi-level of geographic retrieval processing, which needs to improve the accuracy and efficiency of retrieval using natural language [3]. Based on PostgreSQL database, in order to bring a better digital city service experience for the urban citizens, we study methods to make better use of word segmentation dictionary technology to improve the search efficiency and accuracy.

II. KEY TECHNIQUES

Chinese word segmentation and the matching of geographic names and addresses are two key techniques used in multi-level geographic search service [4]. Using SCWS to segment the search keywords and data sets, we can improve the accuracy of word segmentation by modifying the word segmentation dictionary. We established gin index in PostgreSQL to improve the retrieval efficiency, and added grade and classification of information to limit the search range and improve the retrieval accuracy.

A. Chinese Word Segmentation

In Gazetteer service, by analyzing of geographical semantics, we can get some meaningful words from the documents to describe geographic name and addresses and the text that user input [5]. The accuracy of the word segmentation is directly related to whether the search service is able to correctly identify the user's search intent or not, and finally may affect the user's search experience.

Compared with the general text, the phrases containing geographic information have the following characteristics[6]: (1) Generally, the keywords used to retrieve, are relatively short, and there are not enough context information provided for word segmentation; (2) There are more specific and unique words, which may bring uncertainty to word segmentation; (3) Some obvious word identifiers can be used as a hint for word segmentation, such as "road", "town", "supermarket" and so on[7]. Because of the particularity of the phrases containing geographic information, there is a special requirement for the segmentation technique. So far, most of the segmentation techniques are composed of hybrid algorithms, such as SCWS, which is an algorithm and also a system which is using word frequency dictionary.

Using Standard C language, the SCWS has the following advantages in geographic names and addresses system[8]: (1) Without depending on any third-party library function, SCWS provides the C interface and PHP extension. As one of the most convenient open source Chinese word segmentation software for free use, SCWS is suitable for the research of geographic service, which can easily implant any existing software system; (2) SCWS support GBK, UTF-8, BIG5 and other Chinese characters coding. It has a high efficiency in word segmentation, and can improve the efficiency of the retrieval; (3) Using a self-collected word frequency dictionary, supplemented by a certain degree of proper nouns, person names, place names, specific ages and other rules set, SCWS provides a common Internet information thesaurus by default,

and provides the import and export tools written by PHP. Users can customize the text dictionary, and define the weight of words based on some rules. So users can artificially update the dictionary to improve the accuracy of the retrieval; (4) In order to meet the requirements of the full-text index, SCWS specifically provides a compound word segmentation created by itself, which can divide the longer words into shorter ones and recombine scattered words. It is suitable for the application of geographic names and addresses retrieval service, because it can find the most appropriate segment of the phrases containing geographic information.

We stored geographic names and addresses information in PostgreSQL database, and use a segmentation tool named zhparser based on SCWS for word retrieval. We established the word frequency dictionary according to word frequency of the object, the procedure is shown in Fig. 1. Then we split words according to the dictionary, which is the base of names and addresses full-text search.

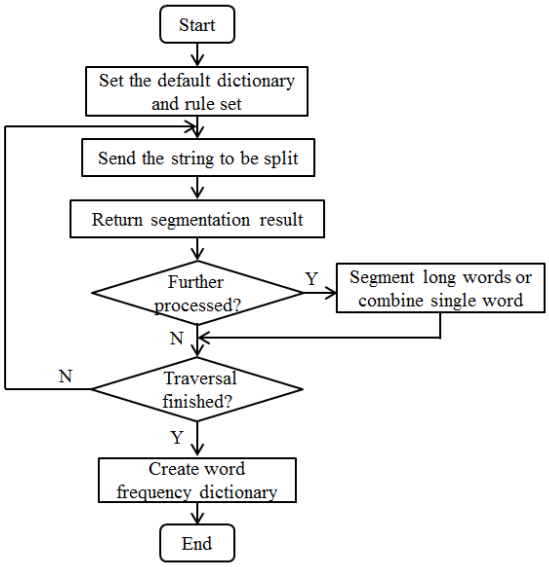

Figure 1. The establishment of word frequency dictionary

B. *Word Segmentation Dictionary Modification*

In Gazetteer service, the word frequency dictionary established by SCWS includes many proper nouns, which may affect the accuracy of the retrieval system. So we provide the correction service in the system, which allows users who discover some errors when they use the service to submit application, and then we modify the dictionary according to opinions of users.

We can use the import and export tool written by PHP to import the dictionary into the text file. According to this way we can add new words to the dictionary, change the weight of words, and re-edit the dictionary. The dictionary can be manually upgraded, so accuracy of geographic names and addresses service can be improved. The process of correcting dictionary service is shown in Fig. 2.

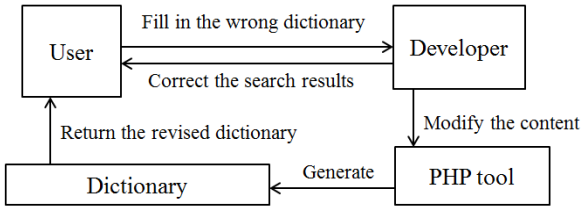

Figure 2. The service flow of correcting the word frequency dictionary

C. *Suitable Index Establishment*

The technique of matching geographic names and addresses plays a very important role in the service. We need to compare user's input containing the geographic information and the data in geocoding library according to a certain matching strategy, find the matching object, and then get the correct spatial location [9]. We can greatly improve the efficiency of geographic names and addresses search service if we index the search field.

During the retrieval, B-tree index, which is commonly used, is not the best one because of more word segmentation and fuzzy queries. So we need to establish some other more effective index, such as gin index, which is suitable for multi-field and fuzzy search, and is suitable for geographic retrieval services. So gin index is set for the keywords field to optimize the retrieval efficiency.

In the service, we should firstly get the keywords which need to be indexed, and then segment the words, and set up index for words after segmentation, finally get the index file. The keywords which users input when they retrieve will be segmented according to the same word segmentation dictionary firstly, and then the searcher will retrieve the index file, and finally return the search results, just as is shown in Fig. 3.

D. *Classification Retrieval*

In the case of large amounts of data, there will be lots of data having the same names, which will greatly affect the retrieval accuracy. For example, users may use "hot pot" as keywords when they want to enjoy hot pot, but there will be some shops in which hot pot spices are sold in the search result, which need to be classified. Give another example, if users use "Industrial and Commercial Bank (bus station)" as keywords, they will get 31 bus stations, as is shown in Table 1, which located in many different provinces and cities. Information about bus stations which are not in the province or city we want is the redundant information. Complex and redundant information will interfere with the user's judgment, and then affect the search service. So we need to divide geographic names and addresses data into different levels and different types. It will improve the accuracy of the search, and give users a better search experience.

TABLE I. PARTIAL SEARCH RESULTS OF "ICBC (BUS STATION)"

Name	Province Name	City Name	District Name
Industrial and Commercial Bank (bus station)	Hebei Province	Xingtai City	Lincheng County
Industrial and Commercial Bank (bus station)	Hebei Province	Xingtai City	Neiqiu County
Industrial and Commercial Bank (bus station)	Hunan Province	Changde City	Wuling District
Industrial and Commercial Bank (bus station)	Hunan Province	Loudi City	Shuangfeng County
Industrial and Commercial Bank (bus station)	Shandong Province	Texas city	De City District
Industrial and Commercial Bank (bus station)	Shandong Province	Dongying City	Dongying District
Industrial and Commercial Bank (bus station)	Shandong Province	Dongying City	Kenli County
Industrial and Commercial Bank (bus station)	...	...	...

On the one hand, we divide POI (Points of Interest) into three levels — province, city and district[10]. By this means, users can select the level information, and then limit the scope of the search results to improve the retrieval accuracy.

On the other hand, we divide POI data into many different types, which include three levels. In the search service, users can select the type information, and the limit the scope of the search results to improve the retrieval accuracy. Part of the classification criteria is shown in Fig. 4.

III. EXPERIMENTAL RESULTS

We performed some experiments to test the quality and efficiency of our Gazetteer service search. In order to ensure the accuracy of the data, we decided to use 6 million data containing geographic information collected from AMAP.

TABLE II. EXPERIMENT ENVIRONMENT

CPU	Intel i5-6400
	2.70GHz * 4
Memory	8GB
Hard Disk	800GB
OS	Ubuntu 14.04
Database	PostgreSQL 9.5.4
Language	Python 2.7

When users use a same keyword, they will get different results if they use graded and classified information or not. For example, if users want to eat hot pot in Changsha, they will get different results under different retrieval conditions, as is shown in Fig. 5. The histogram in Fig. 5 (a) shows all the results returned when the "hot pot" is retrieved, among which only the hot pot restaurants in the type of "food" service in Changsha are what the user actually needs. Different retrieval conditions lead to different results. It can be seen from Fig. 5(a) that there are many useless results if users do not use graded and classified information during retrieval. The pie chart in Fig. 5 (b) shows the results returned when users search for the "hot pot" in Changsha, where the blue part is the result which meets the user's need and the red part is redundant information. The pie chart in Fig. 5 (c) shows the return result when users search the "hot food " in the "food" type, where the blue part is the result that the user needs, and the red part is redundant information. As can be seen from fig. 5, if the grade and classified information is not clear there will be a lot of redundant results, which will cause interference to the users.

Based on the research and optimization of the method of geographic names and addresses search, we designed and implemented the retrieval service, and integrated into the "higis" system in our laboratory. It lays the foundation for the realization of the path planning function. The location can be geographically retrieved and displayed on the map in our service. Our service is integrated into the path planning function and can complete the keywords automatically when user's input information is incomplete.

In the environment of large data, the change of retrieval time with the increase of the amount of data is particularly important. We tested the performance of our geographic names and addresses search service, and result of retrieval time under different conditions is shown in Fig. 6.The curves show the change of retrieval time with the growing of the amount of data when users just use keywords or they add graded and classified information. It can be seen from the figure that the search time just has little change with the increase of the amount of data, which indicates that our service has high operating efficiency in the case of large amount of data and apply to the system of large geographic information.

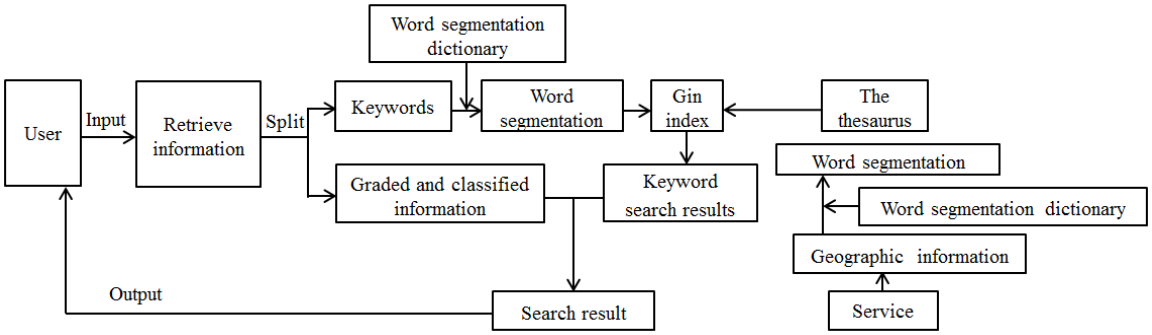

Figure 3. The complete process of retrieve service system

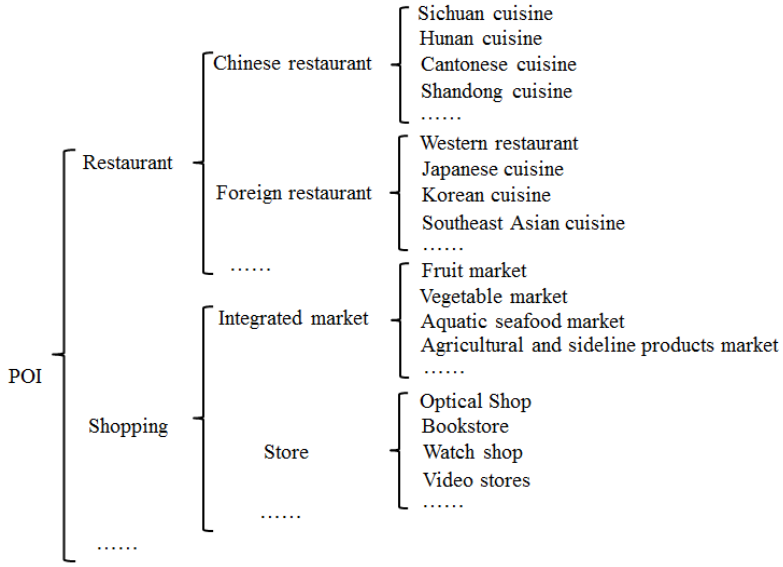

Figure 4. Partial classification criteria

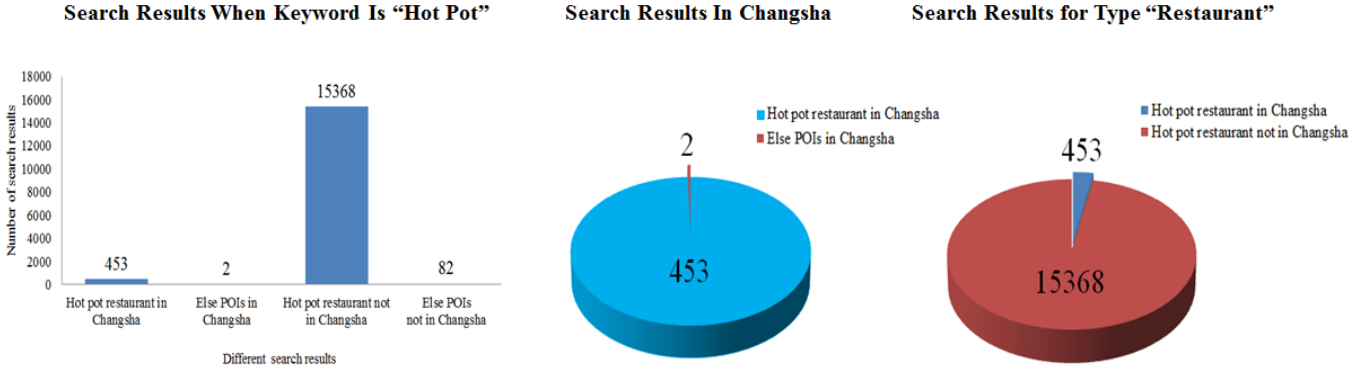

Fig. 5(a) Search results when keyword is "hot pot" Fig. 5(b) Search result in Changsha Fig. 5(c) Search result for type "restaurant"

Figure 5. Comparison of the results of different search methods

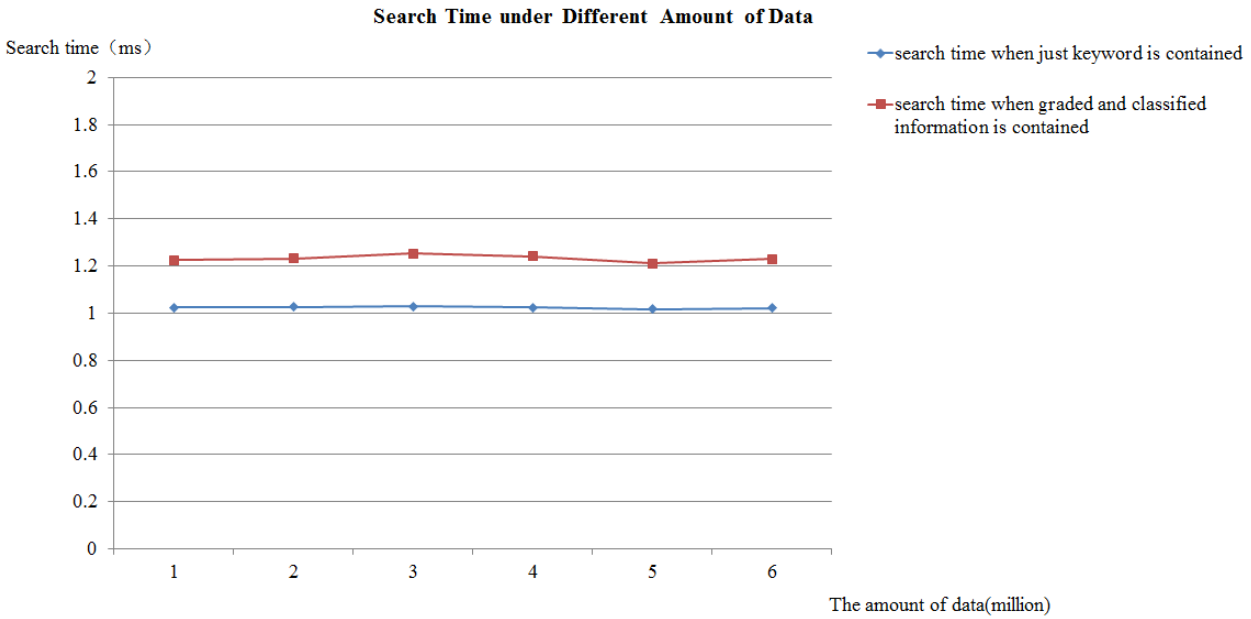

Figure 6. The change of the retrieval time with the different amount of data

IV. CONCLUSIONS

Because of many shortcomings in the practical application of traditional geographic names and addresses search service based on full-text index, we investigate a multi-level approach and an indexing technique based on PostgreSQL to improve retrieval performance of current Gazetteer services. The full-text search technology and optimization strategy for name and addresses retrieval are discussed. An optimized ranking method is proposed by modifying words segmentation dictionary and using graded and classified information. We designed and implemented a prototype system to test our approach, and the evaluation results show that the proposed methods are suitable for large data environment.

ACKNOWLEDGEMENT

This work is supported by the HTRDC(863) under grant No. 2015AA123901, and the Natural Science Foundation of China (41471321)

REFERENCES

[1] HONG Ying. Research on the Matching Method of Urban Geographic Names and Addresses (in Chinese with English abstract) [D]. Liaoning Technical University, 2008.
[2] ZHANG Hong-wen. Research on Methods of Geomorphic Names and Addresses Matching Model Construction (in Chinese) [J]. Science Advisory, 2016 (27): 41-42.
[3] LIU Juan, ZHI Sheng-cui. A Research of the Construction of Address and Geographic Name Database of Municipal Node of World Map (in Chinese with English abstract) [J] .Information and spatial geography information, 2012, 35 (9): 109-110.
[4] CHEN De-quan. Design and Implementation of Key Technologies for GIS Place Search System (in Chinese with English abstract) [J] .Journal of Surveying and Mapping Geography, 2013, 36 (8): 58-60.
[5] SUN Tie-li, LIU Yan-ji. State of The Art and Difficulties in Chinese Word Segmentation Technology (in Chinese with English abstract) [J]. Information Technology, 2009 (7): 187-189.
[6] Information on https://en.wikipedia.org/wiki/Gazetteer
[7] ZHANG Lin-man, WU Sheng. Research on Gazetteer Word Segmentation Algorithm in Geocoding System (in Chinese) [J]. Surveying Science, 2010, 35(2):46-48.
[8] Information on http://www.xunsearch.com/scws/
[9] CHEN Jin-wei. Analysis and Design of City Names and Addresses Management System (in Chinese with English abstract) [D]. Yunnan University, 2013.
[10] Zhang Yan, XU Zhan-hua, XIANG Yu. Design and Implementation of Geographical Address Census System Based on Mobile GIS Technology (in Chinese) [J]. City Survey, 2014(5):59-62.