A Point of Interest Intelligent Search Method based on Browsing History

Wei Sun[1], Yang Cong[1], Xiaoli Liu[2] and Chengming Li[2]
[1]*College of Geomatics, Shandong University of Science and Technology, Qingdao, China*
[2]*Chinese Academy of Surveying and Mapping, Beijing, China*

Keywords: Point of Interest Search Method, Spatial Heat and Scoring Algorithm.

Abstract: Point of interest search is one of the core search functions in GIS. The conventional methods may not be able to personalize search results because the user's personal interests are not taken into account. In order to address this problem, this article proposes a point of interest intelligent search method based on the user's browsing history of map tiles. First, analyze the user's map tile browsing history data on a smart city platform and visualize the user's focal hotspots through a heat map to derive the spatial heat. Second, add the spatial heat influence factor to the attribute query to influence the search results of point of interest, so that the results are more consistent with the user's search intention and corresponding services are offered to different users based on their hotspots. Finally, an experiment with point of interest data from the City of Tengzhou verifies the effectiveness and advantage of the method proposed by this article.

1 INTRODUCTION

Point of interest search is one of the core search functions in Geographic Information System (GIS), and the accuracy of the search results is directly related to the practicability of the system. Many scholars have done in-depth research on point of interest search. Zhong (2007) proposed a rule-based point of interest search algorithm, but it can lead to errors if there is an irregular full name in the database. Mao (2008) proposed a point of interest Chinese abbreviation retrieval algorithm based on statistical model. This algorithm does not have the full names based on segmentation information and does not discern the user's characteristics either, so the accuracy of matching is low. Cen (2010) proposed a method of reordering the search results, which uses log information to construct a collection of the user's high frequency words. When the user submits a query request, it combines the content of the collection and search results, and presents reordered result web pages to the user. Miao (2012) proposed a text filtering system based on vector space model (VSM). In the field of GIS, point of interest data accumulate over time, and a large amount of data will lead to increasingly reduced search efficiency. Therefore, a search method is needed to extract the best matching results from the massive amount of data sources and push them to the users based on their interests to improve the search efficiency and make it more intelligent. However, these algorithms do not achieve the effect of intelligent search of point of interest. Therefore, this paper proposes an intelligent search method of point of interest applicable for GIS based on the user's map tile browsing history.

2 GENERAL SEARCH METHODS

In the existing research, the typical point of interest search method is a string-based word segmentation matching method. This process is shown in Figure 1.

First, the search keyword for the address is segmented, as the acquisition of point of interest requires search keyword segmentation and semantic normalization. Second, based on a specific strategy, the segmented string and the entries in the machine dictionary are matched, and the relevance score of the search results are calculated according to the degree of matching. Finally, the search results are obtained based on the order of the relevance score, and the position point coordinates are matched based on the results.

Sun, W., Cong, Y., Liu, X. and Li, C.
A Point of Interest Intelligent Search Method based on Browsing History.
DOI: 10.5220/0007712301890192
In *Proceedings of the 5th International Conference on Geographical Information Systems Theory, Applications and Management (GISTAM 2019)*, pages 189-192
ISBN: 978-989-758-371-1

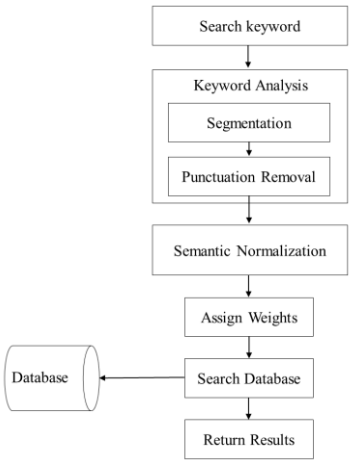

Figure 1: String-based word segmentation matching method.

String-based word segmentation matching method does not take user's interests into account. It only matches the word segmentation results to the database and simply returns the search results. This process leads to the same search results for different user groups if they are using the same keywords. Therefore, understanding and managing user's needs and an active push require a higher standard for the search method.

3 POINT OF INTEREST INTELLIGENT SEARCH METHOD BASED ON USER'S MAP TILE BROWSING HISTORY

3.1 Spatial Heat Calculation

Spatial heat refers to the degree of user's attention to a region. A high spatial heat in an area indicates the area is likely a focal area for the user. In this article, spatial heat is calculated using the user's map tile browsing history.

1) The first step is to obtain and filter the log data in order to get the Web Map Tile Service (WMTS) path and number of calls invoked by the user. Log collection is performed using Filebeat. Elastic.co provides the most comprehensive support for Filebeat and easy access to Elasticsearch, the widely used analysis and retrieval system. Filebeat has a lower resource overhead compared to Logstash. It is characterized by minimal amount of coding and easy optimization. Therefore, Filebeat has better performance compared to the traditional log acquisition methods.

2) There is information redundancy in the data collected by Filebeat, and the search results need to be filtered again in Elasticsearch to eliminate invalid data. Elasticsearch is a distributed storage system based on full-text indexing. It expands and encapsulates the Lucene service engine with a user interface, so the operation commands can be called directly by the user to complete a specified task. In addition, the full-text indexing in Elasticsearch can create segments and shards of multiple sizes. It can also create several shard replicas stored on each data node of the Elasticsearch cluster. Meanwhile, it also provides the ability to process, distribute, and assist these shards and replicas, ensuring organized, balanced, and automatically routed communication among the nodes.

3) Using Elasticsearch to parse the service path can obtain the tile specific location, service resolution, range, and coordinates of starting point. Analyzing the map tile browsing history can yield tile names, paths, zoom levels, row and column numbers, and number of calls. Based on this information, the coordinates of the corresponding tiles are calculated. The region information is obtained by matching the coordinates to the database. User's spatial heat is generated based on the number of calls and the name of the location of the map tiles. The formula for calculating geographical coordinates is as follows:

$$\text{lon} = x * (res * twidth) + XOrigin + \left(\frac{twidth}{2} * res\right) \quad (1)$$

$$\text{lat} = YOrigin - y * (res * theight) + \left(res * \frac{theight}{2}\right) \quad (2)$$

In Formula (1) and (2), lon is the center point longitude, lat is the center point latitude, x is the tile row number, y is the tile column number, res is the resolution level, theight is the tile length, twidth is the tile width, XOrigin is the x axis starting point, and YOrigin is the y axis starting point.

3.2 Scoring Algorithm and Optimization

After matching to a set of documents, they need to be sorted by the level of relevance. Not all documents contain all the query words, and some words are more important than others. The relevance score of a document partially depends on the weight of each query word in the document. The commonly used scoring algorithm employs the Boolean model to find matching documents and calculates the relevance using the practical scoring function. In this article, the

spatial heat influencing factor is added to optimize the common algorithm. The formula is as follows:

$$fscore(q,d) = score(q,d) + B_k * num_k * \lambda_k \quad (3)$$

$$\text{score}(q,d) = coord(q,d) * queryNorm * \sum_{t \ in \ q} (tf(t \ in \ d) * idf(t)^2 * t.getBoost() * norm(t,d)) \quad (4)$$

$$idf(t) = 1 + log(\ numDocs/(docFreq + 1)) \quad (5)$$

$$tf(t \ in \ d) = \sqrt{frequency} \quad (6)$$

$$\lambda_k = \frac{\text{overlap}}{max\ o\ verlap} \quad (7)$$

$$B_k = \frac{T_k}{\sum_{k=1} T_j} \quad (8)$$

Formula (3) is the optimized scoring algorithm, and Formula (4) is the common scoring algorithm.

In Formula(4), coor$d(q,d)$ is the coordination factor, which is based on the number of query items contained in the document. This factor weights documents that contain more query items using a method similar to AND.

$queryNorm$ is the normalized value for each search, which is the square sum of the weights of each query item. The shorter the field, the higher the weight of the field. If a word appears in a field such as the title, it is more relevant than when it appears in a field such as the body of content.

$idf(t)$ is the inverse document frequency (idf), which is the frequency of a term appearing in all documents in the collection. The higher the frequency, the lower the weight. It is used to measure the "uniqueness" of an item. Terms with a higher frequency has a lower idf, and uncommon terms have a higher idf. The inverse document frequency is the logarithm of the number of documents in the index, divided by the number of documents that contain the term, the calculation is as Formula(5).

$tf(t \ in \ d)$ is the frequency of a term appearing in the document. The higher the frequency, the higher the weight. A field that mentions the same term five times is more relevant than a field that mentions it only once. Term frequency is the square root of the number of times the term appears in the document, the calculation is as Formula(6).

$t.getBoost$ is the weight of query item. $norm(t,d)$ is a weighted factor related to length.

In Formula(3), num$_k$is the number of calls.

λ_k is the level of matching between the keywords and the spatial heat, the calculation is as Formula(7). In Formula(7), overlap is the number of matched terms in the query, maxoverlap is the total number of terms in the query.

$B_k$is the weight of the spatial heat, the calculation is as Formula(8). In Formula(8), $T_k$is the number of calls of the map tile k, $T_j$is the total number of calls for all tiles.

4 EXPERIMENT AND ANALYSIS

4.1 Environment and Data

In this article, the experimental data was the point of interest data from Tengzhou, Zaozhuang, Shandong Province. The index database was constructed using Elasticsearch. The search framework was constructed using front-end web technology. The experimental data was analyzed to empirically verify the method proposed by this article. The steps are as follows:

Step 1: Use Elasticsearch to retrieve the users' map tile call log data.

Step 2: Filter the search results: Retain user records with a Tilematrix value less than 13.

Step 3: Obtain the Tilematrix, Tilecol, TileRow of the called map tile and calculate the center point coordinates and range of the tile.

Step 4: Visualize data to form images, calculate and , and add them to the score sorting of point of interest search results.

Step 5: Compare the optimized point of interest search method to the original method. Carry out experiments with 6 different search terms, and each term is searched 50 times. The returned results are compared, and the corresponding precision ratios are calculated.

4.2 Results and Analysis

Table 1 shows the comparison of experimental data between the improved and the original algorithms.After comparing the results between the two methods, the improved method had a similar precision ratio with the original algorithm when the log data was empty. However, along with the accumulation of the user's map tile browsing history data, the outcome of the optimized search method was significantly improved, whereas the accuracy of the original method was basically unchanged. It demonstrated that the method proposed in this article can return the maximum relevant results based on the user's focal hotspot areas, which makes the search more personalized and intelligent.

Table 1: Comparison of experimental data between the improved and the original algorithms.

Number	Search Term	The 1st Search (Precision of the Improved Algorithm)	The 50th Search (Precision of the Improved Algorithm)	The 1st Search (Precision of the Original Algorithm)	The 50th Search (Precision of the Original Algorithm)
1	Gas Station	0.782	0.832	0.782	0.782
2	Sinopec Gas Station	0.860	0.942	0.857	0.857
3	Police Station	0.754	0.864	0.755	0.755
4	Kindergarten	0.673	0.713	0.672	0.672
5	First Hospital	0.812	0.957	0.812	0.812
6	Dongshahe Hospital	0.946	0.987	0.946	0.946

5 CONCLUSIONS

This article proposes a point of interest intelligent search method based on the user's map tile browsing history, which calculates the user’s focal spatial heat through obtaining and analyzing the user's map tile browsing history and uses the spatial heat as an influence factor to optimize the scoring algorithm. Through the comparison with the original algorithms, the following conclusions can be drawn:

(1) The user’s log data contains information about the interests of the web user. The research based on log data can help us understand and master the real needs of web users. The map tile search records can profile the characteristics of the user's interests.

(2) In the field of GIS, using the map tile search records can achieve intelligent search of point of interest. This article studies the user’s map tile browsing history and proposes an optimized algorithm to add spatial heat to the search method. Through the comparison and analysis of simulation experiments, this improved method can effectively improve the precision ratio of search, and become more intelligent with the increase of data volume.

This article modifies the search results based on the user's spatial heat. However, the in-depth research on the search mechanism and user’s interest model is still lacking. The next steps will involve studying the mechanism and structure of search algorithms and formation and application of user models, in order to improve the quality of point of interest search and efficiency of personalized information services..

ACKNOWLEDGEMENTS

This research was funded by a project supported by the National Key Research and Development Program of China (Grant No. 2016YFF0201305) and National Natural Science Foundation of China(Grant No. 41871375).

REFERENCES

Ceng R, Liu Y, Zhang M. 2010. Search Engine User Behavior Analysis Based on Log Mining. *Journal of Chinese Information Processing*. 24(3): 49-54.

Youdong M, Xipeng Qiu, Xuanjing H. 2012. A FAST ALGORITHM FOR LARGE SCALE WEB PAGE CLASSIFICATION. *Computer Applications & Software*, 29(7):260-263.

Zhujie M, Xuehu Z. 2008. Study of Chinese Points of Interest Search Based on Name Abbreviations. *Journal of Taiyuan University of Technology*, 39:52-55.

Zhong L, Zheng F. 2007. Study on Approaches to Retrieval of Chinese Organization Name Based on its Abbreviated Name. *Journal of Chinese Information Processing*. 21(1): 38-42.