PROCEEDINGS OF SCIENCE

Research on An Electronic Map Retrieval Algorithm Based on Big Data

Jian Li[1]
College of Computer Science and Technology，Jilin Agricultural University
Changchun, 130118,China
E-mail: liemperor@163.com

Rujing Yao[a], **Shiji Zhu**[b]
College of Computer Science and Technology，Jilin Agricultural University
Changchun, 130118,China
E-mail: [a]1557652756@qq.com; [b]742435081@qq.com

Today, big data plays an important role. Electronic map has great research value as the core part of the application of big data. Electronic map retrieval is a hot spot in the research field of electronic map. This paper makes some improvement to the defect in traditional ZZL algorithm of electronic map retrieval algorithm that the same tail character string is initialized every time when we retrieve something so that the algorithm performance is reduced. We introduce clustering analysis in the traditional ZZL algorithm and propose an improved ZZL algorithm. The experimental results show that the improved ZZL algorithm has improved the retrieval efficiency by 37.1% on average compared with the traditional ZZL algorithm in the retrieval of geographic name in electronic map retrieval. The improved ZZL algorithm has an exploring significance to the development of the retrieval of electronic map.

CENet2017
22-23 July 2017
Shanghai, China

[1]This work is supported by National Natural Science Foundation Project of China(41601454), Agricultural Science and Technology Achievement Transformation Project of Ministry of Science and Technology of China(2014GB2B100021), China Spark Program Project (2015GA660008), Science and Technology Development Plan Project of Jilin Province(20150204058NY、20160519014JH、20140204045NY), Science and Technology Research Project of "The 13th Five-Year Plan" of The Education Department of Jilin Province(JJKH20170296KJ), The Eighth Batch of Engineering Research Center Project of Jilin Province, University Innovation Team Project of The Education Department of Jilin Province and Campus electronic map Dbank ClickUp of Jilin Agricultural University(No.201510193036) .

PoS(CENet2017)055

1. Introduction

In the era of big data[1-2], data mining[3-4]plays a crucial role when we facing massive data resources. Large databases are numerous, noisy, fuzzy and random. Data mining is the process that finding the potentially useful information and knowledge in large databases. Data mining is also a process of decision support. The common used methods[5]of big data mining are classification[6], regression analysis[7-8], clustering[9], association rules[10], etc. These methods mine data from different perspectives.

Electronic map[11]has great research value as the core of the application of big data. Its appearance makes up for the shortcomings of traditional paper map and makes the application of the map transform fundamentally. However, electronic map retrieval is a hot spot in the research field of electronic map at present and people attach great importance to it. Now, the method of electronic map retrieval mainly includes string matching algorithm[12-13], SQL query algorithm[14]and full text retrieval algorithm[15]. Among them, the most commonly used algorithm is string matching algorithm.

The improved ZZL algorithm proposed in this paper is a combination of classification and association rules in data mining algorithms. When retrieving data, geographic data are classified according to generic term of geographical name and geographical depth level under the aegis of massive geographic data. It can achieve the purpose that using geographic entity database to add spatial database of geographical names. The data at different geographic depth levels are related to the spatial relationship so that the function of depth and reverse query can be realized. At the same time, the retrieval efficiency of the algorithm is greatly improved because of the correlation among the data.

2. Traditional Zzl Algorithm

ZL algorithm is proposed by Ji Fuquan et al and it is a kind of string matching algorithm which can be used as a special purpose[16]. The existing string matching algorithm is a direct comparison, whether it is in accordance with the pattern string from left to right or from right to left. But the core idea of ZZL algorithm is that finding the first letter of the pattern string T in the main string S at first. The positions of each letter which is found is stored and then extract these positions in turn and continue to match the pattern string T from these positions. For the main string which is to match and pattern string which is to match of frequent use, the matching speed is very fast because it has saved all the storage positions of pattern string in the main string in advance.

2.1 Traditional ZZL algorithm processing

The ZZL algorithm is divided into two stages: preprocessing and matching.

1): preprocessing

Preprocessing mainly finishes finding all positions where the first character of the pattern string appears in the main string and it store them in an array.

2): matching

On the basis of preprocessing, string matching algorithm can start from the position where the founded pattern string appears in the main string. Then BF algorithm is used to match the rest after the first letter of the pattern string.

2.2 Performance analysis of traditional ZZL algorithm

The characteristics of ZZL algorithm are simple and easy to implement. In a main string whose length is N, pattern string length is M. We suppose that K is the number of appearance that the first letter of the pattern string appears in the main string. If the preprocessing process of the algorithm are not taken into account, the number of comparisons of ZZL algorithm is k*(M-1) at worst and k*(M-1)<k*M. If the preprocessing process of the algorithm are taken into account, the total number of comparisons are required to add N times.That is to say the number of comparisons is k*M+N.

2.3 Defects of the traditional ZZL algorithm

In the traditional ZZL algorithm, if the first letter of the pattern string appears a lot of times in all pattern string set, the preprocessing is carried every time in the process of matching pattern string. Then the BF algorithm is used to try to match it from each match point. However, in all matches, it is not only to match one pattern string. Even if the tail character pattern string is same, initialization is carried every time. As a result, the performance of the algorithm is seriously affected.

3. Improved Zzl Algorithm

In order to reduce the influence of the performance of the algorithm caused by repeated preprocessing of the same tail character pattern string set, this paper improves the algorithm from two aspects. On the one hand, the repeated preprocessing of the same tail character pattern string set is analyzed. The same tail character pattern string set is pre-processed uniformity by introducing clustering analysis to improve the performance of the algorithm. On the other hand, the last character often represents a type of thing in most retrieval environments. We change the situation that all positions of appearance where the first characters of the pattern string appears in the main string is stored in an array into the situation that all positions of appearance where the tail characters of the pattern string appears in the main string is stored in an array.

3.1 the improved ZZL algorithm based on clustering analysis

We use the algorithm to cluster analysis on the pattern string set, and then use the algorithm to deal with the same tail character pattern string together with a class cluster, at the same time, and will find the pattern string first characters in the main string in all of the location of the stored in an array to find the pattern string to find the end of the string in the main string of all the location is stored in an array. In matching the BF algorithm is used to match the pattern string and the position of the pattern string from the top to the position in the main string to the position of the pattern string in the main string. Algorithm is described as follows:

1. Get the pattern string set of pre clustering analysis;
2. Enter the string s, the point will be assigned to 0;
3. To pretreatment s, find the tail of the pattern string collection of characters in the main string in the location of the index, and save it in a dictionary A;
4. The pointer P points to the first item in the dictionary A;
5. Read the indexed set which the one that pointer P points to the dictionary A corresponds to and assign it to array m;
6. The pointer Q points to the first item in the array m;

7. Read the pointer Q points to a m and assigned to k;
8. The string s in the index for the point to K part of the match, if the match is successful, jump to step 11, otherwise the next step;
9. The K assigned to the point;
10. Move the pointer Q to the next item in the array m;
11. If the pointer Q points to the address is not empty, then return to step 7, otherwise the next step;
12. Move the pointer to the next item in the dictionary A p;
13. The point value is 0;
14. If the pointer P points to the address is empty, then the algorithm ends, otherwise the return step 5.

In order to achieve the improved ZZL algorithm, we add the following four constraints:

Constraint condition 1: The matching pattern string set contains at least one pattern string and each pattern string in the collection cannot be empty and can not contain only one character.

Constraint 2: The length of the string s is at least greater than any one of the string set of the matching pattern and must contain at least one pattern string.

Constraint 3: the item number of array A which is obtained after the preprocessing of string s must be greater than one and the first item cannot be empty.

Constraint 4: The array A that is obtained after the preprocessing of the string s. In addition to the first item, the latter must not contain the string contained in the last item.

3.2 Performance analysis of improved ZZL algorithm

The improved ZZL algorithm is able to deal with a class of clusters in the pre process, so that it can avoid the repeated pretreatment of the same class clusters. In a length of N in the main string, there is a P average length of M of the fixed pattern string and the search has been pre cluster analysis, the pattern string collection can be divided into Q clusters, if the mode string tail letters appear in the number of the number of the main string is k, the total number of P mode string is N*Q+K*M*P, the average number of each mode string is K*M+N*Q/P, that is, when P>Q, the improved ZZL algorithm has more advantages than the traditional ZZL performance. In the worst case, each pattern string is a class cluster, at this time P=Q, compared with the number of K*M+N, the number of comparisons with the traditional algorithm.

4. The Improved Zzl Algorithm In The Realization Of Electronic Map Retrieval

Place name retrieval is one of the most important functions of electronic map, and the string matching algorithm is the core of the search. The improved ZZL algorithm is proposed in this paper to advance the pattern of clustering analysis, and can be used to cluster the geographical names names of the standard, is conducive to the geographical name by cluster analysis, then, using improved ZZL algorithm. The name changes of long period and high clustering and improved ZZL algorithm can greatly exert its advantages.

4.1 The Relations between Generic Term of Geographical Name and Geographical Entity

In order to promote the sharing of geographic information and improve the level of geographic information services, In 2011, the national mapping Geographic Information Bureau issued a “geographic information public service platform geographic entities and geographic

name address data specification".The specification gives a clear definition of geographic information applications, geographical entities and place names and their means and methods. Geographical entities are entities in the geographic database, which can not be divided into the same phenomenon in the real world, a place name is the name of a natural or human geographical entity that is given to a particular space.

In the information retrieval of geographical name, adding generic term of geographical name can quickly locate the geographical entity category. Especially in the local map, generic term of geographical name can accurately be defined. As in the electronic map, you can define generic term of geographical name "province", "city" and "area", they all belong to the category of building, but added more humane factors, more detailed and specific, which in its area of clear, specific categories clearly understand its representative geographical entities. Thus, fusing generic term of geographical name into the map especially local map geographic information retrieval, better retrieval to the user desired geographic entity.

Geographical names and the names of geographical entity information retrieval significance mainly includes the following two points:

(1)Expands the number of retrieved objects and types， enriched the content of the search. (2)The geometric types of retrieval objects are more abundant, so that the spatial relationship between geographical elements is accurate and reasonable.

For example, Search for "Jilin province Changchun city", "Jilin province" as a complex geographical entity, can very accurately express and "Changchun city" is a space that contains the "province" clear "Jilin province" category. For the lower geographical entity, Changchun City, in the Jilin Province, this part of the map, the city refers to its category. So just keep in "Jilin province" belongs to the "city" refers to the lower category of geographical entity search "Changchun city", rather than text matching way to retrieve the full name of "Jilin Changchun city", which greatly improves the efficiency of geographic information retrieval on the map, reduce the consumption of server and ensure the retrieval accuracy, and a single point data is not very accurate method can express the retrieval conditions, only to text matching. Therefore, it can be seen that the combination of geographic information and geographic entities to carry out geographical information retrieval can bring people to the reality of the understanding of the transfer to retrieve, improve the efficiency and quality of retrieval.

4.2 Implementation of improved ZZL Algorithm in Electronic Map Retrieval

In the electronic map, the retrieval of geographic name is very important. By using the improved ZZL algorithm, we proceed cluster analysis to related attributes of geographic entity, spatial information and generic term of geographical name in the retrieval system. Geographic entities are defined as the base objects of the information retrieval of geographical name, and the characteristics and spatial relations of geographic entities constitute the feature items of the retrieval. By using this, we design the algorithm and apply it to the electronic map. Experiments show that this method can realize the information retrieval of geographic names and this retrieval method is based on the characteristics and spatial relations. What's more, the recall and precision are basically maintained at more than 90%. The retrieval quality is good. By a series of contrasts, we found that the precision of the retrieval method and retrieval method of comparative experiments are almost the same.But in the recall, there are obvious advantages. In the search for "Changchun city of Jilin province", the reference volume of the retrieval request "Jilin province" and the "Changchun city" of the target body have a topological relationship,

"Jilin" is the map of the father of geographic entity, "city" is the "Changchun city" names, refer to the type, retrieval method for mining overall and part of the hierarchical relationship between entities, referring to the body and the object unified into the retrieval model, not only improves the accuracy of search results, but also enriches the forms of the results of reference and target were presented and expressed in different forms, so users can easily understand retrieval the meaning of the results.

The program will building map corresponding to the alias, and referred to as the general geographical name into the database and the clustering analysis of geographical names, with obvious characteristics of clustering principle, each geographic name has the obvious "cluster", has the same geographical name of the name can be grouped into a cluster. We use the improved ZZL algorithm to achieve the electronic map retrieval function.

5. Application Of Improved Zzl Algorithm

The purpose of the improved ZZL algorithm is to minimize the influence of the repeated preprocessing of the same tail character string set on the performance of the algorithm and therefore the runtime is be reduced. However, The introduction of the clustering analysis makes the pretreatment process more complex. The efficiency of clustering analysis is the key to improve the efficiency of the algorithm. Therefore, the advantages and disadvantages of the improved ZZL algorithm are reflected in the different factors affecting the efficiency of cluster analysis, such as the number of pattern strings and nature of cluster.

5.1 The Comparison of Different Numbers of Retrieved Pattern Strings

Although the increase of the number of pattern strings will extend the time of cluster analysis, it greatly reduces the time of the repeated processing of the same tail character pattern string set.Because the improved ZZL algorithm carries out cluster analysis on the pattern string set in advance, time is obviously lower than the traditional ZZL algorithm and the performance is better than the traditional ZZL algorithm when matching which has different numbers of pattern strings but has same the nature of cluster are disposed.The tests are carried out as follows:

All generic term of geographical names in electronic map data are coincident when the natures of cluster are same (We take 36 clusters as an example). We use the traditional ZZL algorithm and the improved ZZL algorithm directly to deal with the electronic map data of different data size . The results are shown in the following tables:

Retrieval Conditions	the Improved ZZL Algorithm		the Traditional ZZL Algorithm	
	Accuracy(%)	Time(s)	Accuracy(%)	Time(s)
Jilin Province	98.4	37.4	97.8	56.1
Hunan Province	97.5	38.2	97.2	48.7
Hebei Province	97.9	35.7	92.5	52
Changchun,JilinProvince	92.3	52.2	92.1	67.7
Changsha,Hunan Province	92.1	54.7	91.9	64.3
Shijiazhuang,Hebei Province	91.7	49.6	91.8	70.2

Table 1:comparison when the number of pattern strings is 2189

Retrieval Conditions	the Improved ZZL Algorithm		the Traditional ZZL Algorithm	
	Accuracy(%)	Time(s)	Accuracy(%)	Time(s)
Jilin Province	91.2	53.6	91.1	88.6
Hunan Province	92.4	52.7	92.7	87.2
Hebei Province	89.7	54.8	89.5	89.3
Changchun, Jilin Province	75.2	80.2	75.6	134.3
Changsha, Hunan Province	75.9	82.4	75.2	132.7
Shijiazhuang, Hebei Province	74.1	84.7	74.4	135.7

Table 2:comparison when the number of pattern strings is 82940

The comparison tables show that the accurate rate between the improved ZZL algorithm and the traditional ZZL algorithm is similar whether the number of pattern strings is 2189 or 82940 because the number of clusters is 36 taking generic term of geographical name as the clustering standard. However, time of the improved ZZL algorithm is obviously lower than the traditional ZZL algorithm. It shows that the performance of the improved ZZL algorithm is better than the traditional ZZL algorithm.

5.2 the comparison of different nature of cluster of the pattern string

The improved ZZL algorithm is based on cluster analysis,so the performance of algorithm is seriously affected by the nature of cluster of the pattern string. Although the improved ZZL algorithm carries out clustering analysis on the pattern string in advance, the increase of clusters will increase the times of preprocessing. And the ratio between Q and P will become larger. As a result, the performance of the improved ZZL algorithm will drop. The tests are carried out as follows:We use the traditional ZZL algorithm and the improved ZZL algorithm directly to deal with the electronic map data of clustering of different data when the data size is same. The results are shown in the following tables:

Retrieval Conditions		the Improved ZZL Algorithm		the Traditional ZZL Algorithm	
		Accuracy(%)	Time(s)	Accuracy(%)	Time(s)
2189 pattern string	Jilin Province	98.4	37.4	97.8	56.1
	Hunan Province	97.5	38.2	97.2	58.7
	Hebei Province	97.9	35.7	92.5	52.1
	Changchun, Jilin Province	92.3	52.2	92.1	67.7
	Changsha, Hunan Province	92.1	54.7	91.9	64.3
	Shijiazhuang, Hebei Province	91.7	49.6	91.8	70.2
82940 pattern string	Jilin Province	91.2	53.6	91.1	88.6
	Hunan Province	92.4	52.7	92.7	87.2
	Hebei Province	89.7	54.8	89.5	89.3
	Changchun, Jilin Province	75.2	80.2	75.6	134.3
	Changsha, Hunan Province	75.9	82.4	75.2	132.7
	Shijiazhuang, Hebei Province	74.1	84.7	74.4	135.7

Table 3:comparison when the number of clusters is 36

Retrieval Conditions		The Improved ZZL Algorithm		The Traditional ZZL Algorithm	
		Accuracy(%)	Time(s)	Accuracy(%)	Time(s)
2189 pattern string	Jilin Province	94.7	47.4	94.8	57.1
	Hunan Province	94.5	48.7	94.2	58.3
	Hebei Province	95.1	46.3	95.5	56.2
	Changchun, Jilin Province	91.2	61.2	90.8	68.2
	Changsha, Hunan Province	90.6	60.7	91.1	68.3
	Shijiazhuang, Hebei Province	90.7	60.6	91.8	69.6
82940 pattern string	Jilin Province	88.3	63.6	88.1	87.9
	Hunan Province	88.2	64.7	89.1	87.8
	Hebei Province	88.7	64.3	89.5	88.3
	Changchun, Jilin Province	70.2	110.2	70.6	133.7
	Changsha, Hunan Province	70.9	112.7	69.2	134.7
	Shijiazhuang, Hebei Province	70.6	114.4	69.4	134.2

Table 4:comparison when the number of $clusters$ is 127

The contrast tables obviously show that the number of clusters has a little influence on the traditional ZZL algorithm. The number of clusters has a little influence on the accuracy of the improved ZZL algorithm, but time is greatly increased. However, the performance of the improved ZZL algorithm is still better than the traditional ZZL algorithm.

6. Conclusion and Prospect

This paper makes some improvement to the defect in traditional ZZL algorithm of electronic map retrieval algorithm that the same tail character string is initialized every time when we retrieve something so that the algorithm performance is reduced. We introduce clustering analysis in the traditional ZZL algorithm and propose an improved ZZL algorithm. The experimental results show that the improved ZZL algorithm has improved the retrieval efficiency by 37.1% on average compared with the traditional ZZL algorithm in the retrieval of geographic name in electronic map retrieval. The improved ZZL algorithm provides a theoretical basis and a simple application for the development of electronic map retrieval and has a stimulative significance to the development of the retrieval of electronic map. As the same time, in the big data, part of the data with high cluster is valuable data frequently, so the improved ZZL algorithm can also be applied to big data retrieval. In the future, we will further discuss the algorithm and optimize the performance of the algorithm so that the efficiency of retrieval can be improved constantly.

References

[1] Yuanzhuo Wang, Xiaolong Jin, Xueqi Cheng, "*Network big data: present and future,*"[J]. Chinese Journal of Computers, vol.36,n.6,2013.6,pp.1125-1138.(in Chinese)

[2] Katharine Armstrong, "*Big data: a revolution that will transform how we live, work, and think. Information,*"[J]. Communication & Society, vol.17, n.10,2014.10,pp.1300-1302.

[3] Yongchun Zhu, Min Wan, "*A brief analysis on DM technique,*"[J]. Computer Knowledge and Technology, vol.6,n.2,2010.1,pp. 266-267. (in Chinese)

[4] Smith Andrew, Gerstein Mark, "Data mining on the web,"[J]. Science,vol.314,n.5806, 2006.12,pp.1682.

[5] Mingbang Liu, Xiongfei Li, Tao Sun, Xiaoqing Xu, "*Survey of data mining technology standards,*"[J]. Computer Science, vol.35,n.6,2008.8,pp.5-14.(in Chinese)

[6] Gang Wang, Lihua Huang, Chenghong Zhang, Jie Xia, "*Review of classification algorithms in data mining*,"[J]. Science & Technology Review, vol.24,n.12, 2006.12,pp.73-76.(in Chinese)

[7] Liang Ao, Huimin Fu, "*Regression analysis method for interval censored data,*"[J]. Journal of Aerospace Power, vol.22,n.6,2007.6 pp.1013-1017. (in Chinese)

[8] Weiping Huang, Yu Xu, Jie Wang, "*Data association method based on regression analysis*,"[J]. JOURNAL OF XI'AN JIAOTONG UNIVERSITY, vol.45,n.8, 2011.7,pp.92-107.(in Chinese)

[9] Faxin Zhao, Guoye Wang, "*Research of clustering analysis algorithm in data mining*,"[J]. JOURNAL OF TONGHUA TEACHERS COLLEGE, vol.26, n.2, 2005.3, pp.11-13. (in Chinese)

[10] Yuxing Mao, Tongbing Chen, Baile Shi, "*Efficient method for mining multiple-level and generalized association rules*,"[J]. Journal of Software, vol.22,n.12,2011.12,pp.2965-2980. (in Chinese)

[11] Wenhong Cui, "*The application and development trends of electronic map*,"[J].GEOMATICS&SPATIAL INFORMATION TECHNOLOGY, vol.31,n.3, 2008.6,pp.87-89. (in Chinese)

[12] Cheng Wang, Jingang Liu, "*An improved string matching algorithm,*"[J].Computer Engineering, vol.32, n.2, 2005.1, pp.62-64.(in Chinese)

[13] Junjie Zhao, "*A calculate way of rapid string precision used for keyword index matches,*"[J]. Computer Systems & Applications, vol.19, n.2, 2010.2, pp.189-191.(in Chinese)

[14] Aiping Xu, Fuling Bian, *"Change from middle language to SQL in Chinese query system of GIS,"*[J]. Computer Engineering, vol.32,n.22, 2006.11,pp.49-50.(in Chinese)

[15] Aibing Qian, *"Algorithm design of the full text retrieval and summarization of the full text retrieval system,"*[J]. New Technology of Library and Information Service, n.2, 2003.3, pp.42-44.(in Chinese)

[16] Fuquan Ji, Zhanli Zhu, *"A string matching algorithm for special use,"*[J].Computer&InformationTechnology,n.8,2006.8,pp.85-86+89. (in Chinese)