Exploring Traffic Accident Locations from Natural Language Based on Spatial Information Retrieval

Shanshan Wang[1,2,3], Honghui Dong*[,1,2,3], Yue Zhou[1,2,3], Limin Jia[1,2,3], Yong Qin[1,2,3]

1. Beijing Jiaotong University, Beijing 100044
E-mail: wangss@bjtu.edu.cn

2. State Key Lab of Rail Traffic Control and Safety, Beijing 100044
E-mail: hhdong@bjtu.edu.cn

3. Beijing Engineering Research Center of Urban Traffic Information Intelligent Sensing and Service Technologies, Beijing 100044
E-mail: hhdong@bjtu.edu.cn

Abstract: Exploring traffic accident locations is essential for making prevention strategy in order to improve traffic safety proactively. Quite a number of methods have been developed to acquire accident spatial data, by various detectors in roads. However, plenty of locations information is contained in textual alarm information transformed from telephone calls of 122 Alarm Reception. This paper, aims to obtain spatial data, longitude and latitude, of traffic accident locations based on textual alarm information. Natural language processing, database system and Geographic Information System (GIS) are used to get traffic accident spatial data. The proposed method involves three processes, location name identification, spatial information retrieval and information matching respectively. Validation results show that the method is of reasonably high accuracy for exploring traffic accident locations.

Key Words: Traffic Accident Locations, Spatial Data, Location Name Identification, Spatial Information Retrieval, Information Matching

1 INTRODUCTION

In recent years, frequent road traffic accidents not only burden urban road transportation, but also are responsible for heavy societal cost. Moreover, plenty of traffic accidents make managers have consistently been interested in analysis methods to prevent accidents. In particular, historical information of accident locations plays an important role in traffic accidents prevention work because it can reveal detailed spatial distribution features, which can give better effect of various changes on the road structure or environment on the behavior of drivers [1]. Plenty of historical accidents data, however, have long been dominated by traditional textual alarm information which is transformed from telephone calls of 122 Alarm Reception and over the years have yielded invaluable insight, because of natural language form and unstructured characteristic. Therefore, transforming textual alarm information into spatial data, latitudes and longitudes, becomes a primary task of traffic accident prevention.

In the field of acquiring traffic accident locations, several researches about it have pointed out already. For instance, M. J. Corby and F. F. Saccomanno presented that speed of vehicle was a good indicator for disruption in flow, so the speed detectors would be used for analyzing traffic accident locations. Although speed flow criterion was better than occupancy and volume, they found that it would appear to some deviations due to the delay of precise time [2]. Mario Miler and Filip Todić developed a model for validating traffic accident locations based on two key concepts. One is Jaro-Winkler string matching technique and the other one is Inverse Distance Weighting method [3]. Yiming Gu, Zhen (Sean) Qian detected real-time traffic incident using data from twitter of social media [4]. However, this method mainly depends on the twitter that people pushed and cannot cover all accidents. In addition, there are fewer researches about identifying traffic accident locations using historical textual alarm information.

This study aims to present a new method for acquiring traffic accident locations based on natural language processing, database system and GIS. The research results will help transportation professionals propose some specific measures to reduce the occurrence of the traffic accident. This paper is organized as follows. Section 2 presents the methodology of the proposed method, details of the location name identification, spatial information retrieval, and information matching. Section 3 describes data processing which is the core part of this paper. Section 4 presents results and discusses the accuracy based on the experimental results. Finally, conclusions and recommendations for future work are provided in Section 5.

2 METHODOLOGY

In this section, the concept of main methodologies used in this paper is introduced. The essential work of traffic accident locations exploration is acquiring crash locations data of longitudes and latitudes accurately. This study applies processes of location name identification, spatial information retrieval and information matching, into exploring the traffic accident locations. In addition, all of three parts are running in Java programming environment.

This work is supported by the National Key Technology Support Program of China (Grant No.2014BAG01B04).

*Corresponding author

Figure 1 is presenting a framework of methodology in this study. Each part is described as follows in detail.

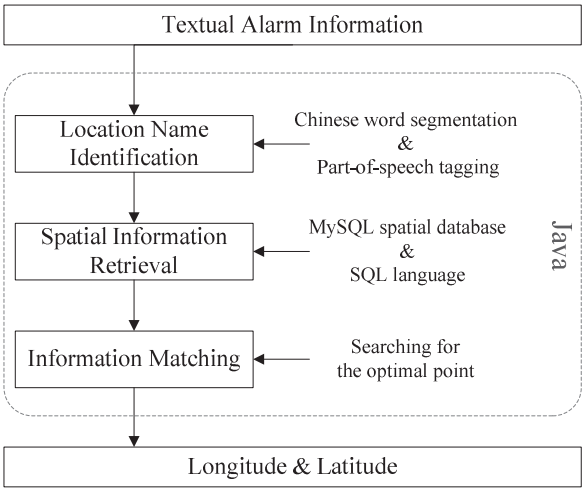


Fig 1. Framework of Methodology

Part 1 is location name identification. After this procedure, precise location name could be captured from large amounts of textual traffic accident alarm information. This method contains two steps. One is Chinese word segmentation which means segmenting a sentence into words sequence, because words are the basic units to comprehend and process [5]. Chinese is unlike English, there is no space between a word and another word. Therefore, Chinese word segmentation is a foundation of this study. The other one is part-of-speech tagging which can distinguish words of location name from words of other part-of-speech. In order to complete two processes of natural language processing mentioned above, Ansj segmentation tool is applied to accomplish location name identification. As an open source word segmentation tool, Ansj has a high accuracy about 98.49% when tested with People's Daily corpus [6].

Part 2 is spatial information retrieval which is the most important methodology in this study. After that, traffic accident location name could be transformed into a series of related points' longitudes and latitudes. This process involves two techniques. One is the MySQL database. It uses to store data of electronic map and provide an environment for retrieving. Compared to some large databases such as SQL Server and Oracle, MySQL is on a small scale, however, it convenient to use and provide enough functions for this study. The other one is Structured Query Language (SQL). It uses to search and retrieve longitude and latitude data of traffic accident location.

Part 3 is information matching. A traffic accident location may correspond with many retrieval results, because all relevant points can be retrieved in part 2. As a consequence, center of gravity is applied to search for the optimal point. Firstly, the center of gravity of all points is calculated. Then, the point which is closest to the center of gravity is regarded as the final matching result.

3 DATA PROCESSING

This section aims to detail the process of data processing, and is organized as four procedures. The first step is data preprocessing of original textual alarm information. Second step presents how to get the location name from textual alarm information. And third step explains the detailed procedure of information retrieval. The final step describes information matching. Through these steps, longitude and latitude data of traffic accident location would be obtained exactly.

3.1 Data Source and Preprocessing

In this paper, textual alarm information was collected from 122 Alarm Reception, a division of the Chinese Public Security Traffic Management authority, which disposes alarm calls of traffic accidents. And the data on October 20, 2014 in Beijing were applied to analyze, 2321 accidents in total. Table 1 presents a part of original alarm information. Among them, the data format includes two columns, and they are call time and description of traffic accident location respectively.

Table1. Original Alarm Information.

CALL_TIME	PLACE
2014/10/20 1:20	涉酒：纪家庙路丰台东路路口
2014/10/20 2:59	花乡汽车博物馆门前
2014/10/20 5:21	望京八十中学校园
2014/10/20 8:35	北京右外医院北门门前 民警到
2014/10/20 11:45	天通苑龙德广场路口
2014/10/20 13:05	丰台南宫市场早市西门前 要求回复
2014/10/20 16:53	通州月亮河小区门前
2014/10/20 21:53	107 国道小十三里路口
2014/10/20 21:56	107 国道小十三里路口 重复报警
2014/10/20 22:41	朝阳西直河仁爱医院门前

Original textual alarm information mainly relies on records by telephone operators. When a traffic accident happens, people will make a call to 122 Alarm Reception for treatment. And operator will transform voice information into textual alarm data. Annotations are always presented additionally, which mainly contain the nature of the traffic accidents, whether needing to reply, alarming times, etc. For example, the annotations in Table1 are "涉酒", "民警到", "要求回复", "重复报警". Therefore, the form of the original information has not unified format. If experiment deals with original alarm information directly, the final result will appear lots of errors because of redundant information and repetitive information. In order to gain accurate results, data preprocessing is particularly needed. Removing redundant information and repetitive information will reduce amounts of further research work and lower the risk of program misreading.

3.2 Location Name Identification

Ansj segmentation tool is applied to identify the name of traffic accident location. Paper uses it to process textual alarm information based on built-in dictionary and user defined dictionary. Because this study has a high-demand of accuracy about location name keywords in Chinese sentences, the integrity of user defined dictionary has a direct impact on segmentation results. As a consequence, building a comprehensive location name database as the user defined dictionary is the primary task before segmenting Chinese sentences. Based on electronic maps of Beijing, paper uses GIS technique to collect all names of point layer as the content of user defined dictionary. Furthermore, each name corresponds to a part-of-speech, and there are 22 classes 99 kinds of part-of-speech in total. Some part-of-speechs which relate with location name are: 'ns' represents location, 'nst' represents transliterating location, 'nw' represents new word and 'nt' represents organization. When the user inputs a piece of textual alarm information, if the location name in this data is contained in user defined dictionary, the Ansj tool will distinguish this location name.

Table 2 proposes the algorithm of location name identification in Java programming environment. Ansj can realize different segmentation effect through different interface. This study uses interface of Natural Language Processing (NLP) to achieve segmentation of Chinese sentences, because it can load user defined dictionary and achieve segmentation at high-precision. However, the process of Chinese word segmentation only transforms sentence into sequence made up of a series of words, there is no increase or decrease in the context of textual alarm information. To gain location name merely, 'endsWith' statement is applied to distinguish location name according to the part-of-speech postfix. If a word ends up with part-of-speech of location, it will be output in this process.

Table2. Algorithm of Location Name Identification.

Input: Textual alarm information **Output:** Traffic accident location names
Procedure: • Load user defined dictionary. • Call the file of textual alarm information and read the content line by line. • If line l_i (i=1, 2,⋯ n) is null, terminate algorithm. Else, segment textual alarm information by NLP segmentation interface. • If a word end up with 'ns' or 'nw' ⋯, this word will be output as a location name. • Getting all the traffic accident location names.

3.3 Spatial Information Retrieval

Structuring a spatial information database is the basis of information retrieval. Table 3 is presenting a part of spatial information database in MySQL. It mainly contains ID, location name, longitude and latitude, address, and type. Among them, ID as the data table's primary key determines the unique value of each piece pf data. In addition, 'DISPLAY_X' represents longitude and 'DISPLAY_Y' represents latitude. All of this spatial data is acquired from electronic map of Beijing using ArcGIS application. Specially, the function of geometric calculation is used to get data of longitude and latitude. And in this spatial database, there are 48 classes of location including government agencies, railway station, subway stations, commercial buildings, retail industry, etc. 664988 location names in total.

Table3. Part of Spatial Database in MySQL

ID	LOCATION	DISPLAY_X	DISPLAY_Y	ADDRESS	TYPE
1	团结小学附属幼儿园	116.37317	39.73888	NA	A701
2	新世纪双语启蒙幼儿园	116.42748	39.72514	光明路	A701
3	大兴区团结实验学校	116.38905	39.74811	NA	A701
4	北京市大兴区海迪学校	116.40949	39.78439	西三路	A701
5	汇源佳儿童艺术培训中心	116.40787	39.76658	S329-黄亦路	A701

On the basis of spatial database, this study retrieved spatial information in the Java programming environment with SQL statements. Table 4 is presenting the algorithm of spatial information retrieval. The main work of this algorithm is achieving fuzzy search with SQL statements. And all of related spatial data corresponding to each traffic accident location name could be obtained by these steps.

Table4. Algorithm of spatial information retrieval.

Input: Traffic accident location names **Output:** Longitude and latitude of traffic accident locations $\{(\varphi_1, \lambda_1), (\varphi_2, \lambda_2),\cdots,(\varphi_n, \lambda_n),\}$
Procedure: • Connecting MySQL database with Java environment. • Call the file of traffic accident location names and read the content line by line. • If the words of line l_i (i=1, 2,⋯ n) is null, terminate algorithm. Else, retrieve spatial information using SQL statement which is *SELECT * FROM 'DATABASE' WHERE 'PLACE' LIKE '%KEYWORD%'*. • Getting all the longitude and latitude data of traffic accident locations.

3.4 Information Matching

After the process of spatial information retrieval, each location name might get more than one corresponding results. How to choose the most optimal point from multiple data as the final result become the key to this study. So the calculation of center of gravity is applied to information matching, which is regarded as the standard.

It could be formulated in the following form:

$$G_0(\varphi_0,\lambda_0)=\begin{cases}\varphi_0=\dfrac{\sum_{i=1}^{n}\varphi_i}{n}\\ \lambda_0=\dfrac{\sum_{i=1}^{n}\lambda_i}{n}\end{cases} \quad (1)$$

Where φ_0 is longitude of center of gravity, λ_0 is latitude of center of gravity, φ_i and λ_i are longitude and latitude of a point of retrieval results respectively.

Then, the study calculated distance between the center of gravity and each point from retrieval results. The distance can be calculated using the following equations:

$$d(G_0,X_i)=R\left\{\arccos\left[\begin{array}{l}\cos\lambda_i\cos\lambda_0\cos(\varphi_i-\varphi_0)\\ +\sin\lambda_i\sin\lambda_0\end{array}\right]\right\} \quad (2)$$

Where X_i is a point of retrieval results, R is a constant which presents earth radius and equals 6371km.

And the final step is comparing all the distance and getting the point which is closest to the center of gravity as the final result.

4 RESULTS AND ANALYSIS

After data processing, location name identification, spatial information retrieval and information matching, textual alarm information is transformed into longitude and latitude data of traffic accident locations. This structured data plays a significant role in transportation researches. In this section, paper analysis results from two aspects, they are accuracy and practical application respectively.

4.1 Accuracy of Retrieval Results

First of all, some standards should be confirmed to measure the accuracy of retrieval results. In the field of electronic map, both Baidu Map and Auto Navi Map (AMAP) have mature systems and on a large scale, therefore, choosing these two maps as standards is authoritative and appropriate.

Then, according to the same textual alarm information, paper searched spatial data both in Baidu Map and AMAP. As a result, two sets of standard longitude and latitude data were obtained. In this step, Application Programming Interface (API) is applied to assist in acquiring structured data. And the function mainly used was Address Resolution which can search longitude and latitude data according to the textual information. Especially, different from electronic map used in this paper is referencing World Geodetic System-1984 Coordinate System (WGS-84), Baidu Map is using its own BD-09 Coordinate System and AMAP is referencing GCJ-02 Coordinate System. Therefore, in order to be compared together, deviations of three sets of results were rectified.

Finally, an evaluation index is needed necessarily. Paper defined an index of similarity as follow:

$$sim(X_i,Y_i)=e^{-d(X_i,Y_i)} \quad (3)$$

where X_i is a point generated from the ith textual alarm information and acquired by one retrieval method, Y_i is a point generated from the same textual alarm information and acquired by another retrieval method, $d(X_i,Y_i)$ is distance between point X_i and point Y_i. Table 5 presents accuracy analysis results.

Table5. Accuracy Analysis Results

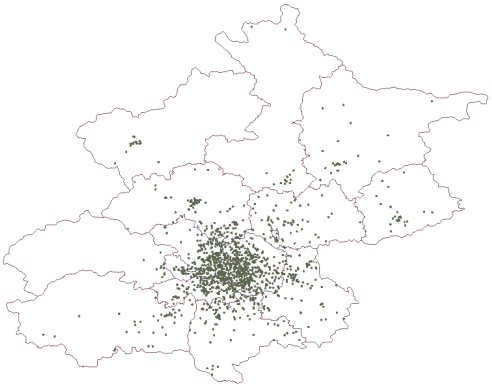
	Similarity	Distance (km)	Number	Proportion
Baidu Map results & retrieval results	0.385	d≤1.5	1536	66.18%
		1.5<d≤5	260	11.20%
		5<d≤10	167	7.20%
		d>10	358	15.42%
Auto Navi Map results & retrieval results	0.404	d≤1.5	1600	68.94%
		1.5<d≤5	287	12.36%
		5<d≤10	111	4.78%
		d>10	323	13.92%
Baidu Map results & Auto Navi Map results	0.562	d≤1.5	1637	70.53%

From Table 5, some important conclusions can be found. Firstly, the similarity in the Table 5 is arithmetic mean value of all pair of points from two sets of data and it is generally small. Even between two authoritative electronic maps, similarity is only 0.562. This is because diffident retrieval method has different base map and principle of searching. Secondly, the accuracy of retrieval results in this study can be reflected by distance. Considering the actual situation, paper defines that if a distance less than or equal 1.5 kilometres, two points will be considered much consistent with each other. In each pair of methods, about 70% of the data are consistent. And it follows that, most of results acquired by retrieval method used in this study are accurate.

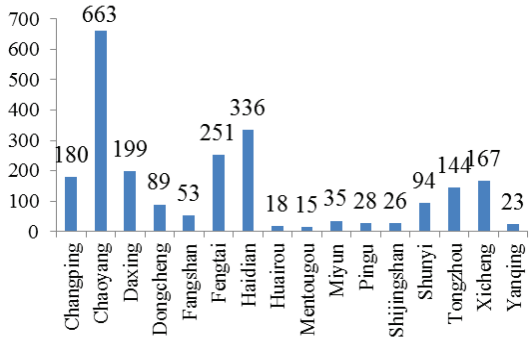
4.2 Retrieval Results in Practical Application

Longitude and latitude of traffic accident location contain a great value for transportation researches. Therefore, this study visualized and quantified retrieval results using GIS technology. FIGURE 2 shows the spatial distribution features of traffic accidents on October 20, 2014 in Beijing.

There are two auxiliary charts, one is scatter diagram which describes distribution of the traffic accident visually, and the other one is histogram which presents the distribution quantifiably. In the histogram, abscissa represents district, 16 in total, ordinate is the number of traffic accidents.



(a) Scatter Diagram



(b) Histogram

Fig 2. Spatial Distribution Features.

Amounts of information can be figured out from Figure 2. First of all, Chaoyang district had the largest number of traffic accidents and it is twice the size of Haidian district which was second only to the Chaoyang district. Apart from this, Fengtai, Daxing, Changping, etc. also had a lot of traffic accidents. On the contrary, there were little accidents in Mentougou, Huairou, Yanqing, etc.

The main reasons for these spatial distribution features as follows. First, there are a large number of residence communities and science-and-technology enterprises both in the Chaoyang and Haidian district, particularly the former. As a result, the probability of traffic accident is increased because high traffic volume caused by a plenty of people who work there but live in other districts. Second, as the 'vegetable basket' of Beijing, Daxing district assembles many agricultural production and farming enterprises. Therefore, lots of transport vehicles are bound to increase the occurrence of traffic accidents. Third, Changping district is the largest 'sleeping city' of Beijing, and lots of residents bring a high risk of traffic accident. Finally, Mentougou, Huairou and Yanqing districts, there are less people and vehicles, so the number of accidents is little.

According to these spatial distribution features, some hot spots where happened accidents frequently can be recognized. On the basis of this, department of traffic management could propose some targeted prevention measures to reduce the incidence of traffic accidents. Moreover, spatial distribution features can also reveal road structure and environmental conditions indirectly. Some improvements could be proceeding based on this information. The construction of transport facilities and an improvement in transport conditions will increase the accessibility of some urban regions, increasing the impact of human activities availably [7].

5 CONCLUSIONS

This paper presents a new method based on spatial information retrieval for exploring traffic accident location. Textual alarm information is transformed into spatial data including longitude and latitude. Three techniques, location name identification, spatial information retrieval and information matching, are used to process data. In order to measure the accuracy of this study, Baidu Map and AMAP are chosen to be standards. After a series of comparisons, analysis results show that the retrieval accuracy is consistent with authoritative maps. Specially, this new method realizes the same accuracy with few work supporting. And practical application of retrieval results such as traffic accident spatial distribution features is obtained.

For future research, the authors plan to improve the algorithm of location name identification, the algorithm of spatial information retrieval and the method of information matching. Moreover, the current experiment only considers data of one day, in the future, more data should be processed.

REFERENCES

[1] Ömür Kaygisiz, Şebnem Düzgün, Ahmet Yildiz, Metin Senbil, Spatio-temporal accident analysis for accident prevention in relation to behavioral factors in driving: The case of South Anatolian Motorway, Transportation Research Part F: Traffic Psychology & Behaviour, Vol.33, 128-140, 2015.

[2] M. J. Corby, F. F. Saccomanno, Analysis of Freeway Accident Detection, In Transportation Research Record: Journal of the Transportation Research Board, Transportation Research Board of the National Academies, Washington, D.C., No. 1603, 80-89, 1997.

[3] Miler Miler, Filip Todić, Marko Ševrović, Extracting accurate location information from a highly inaccurate traffic accident dataset: A methodology based on a string matching technique, Transportation Research Part C: Emerging Technologies, VOL. 68, 185-193, 2016.

[4] Yiming Gu, Zhen (Sean) Qian, Feng Chen, From Twitter to detector: Real-time traffic incident detection using social media data, Transportation Research Part C: Emerging Technologies, VOL. 67, 321-342, 2016.

[5] Xu Sun, Yaozhong Zhang, Takuya Matsuzak, Yoshimasa Tsuruoka, Jun'ichi Tsujiif. Probabilistic Chinese word segmentation with non-local information and stochastic training. Information Processing & Management, VOL.49, 626-636, 2013.

[6] Ansj segmentation tool, Github NLPchina, https://github.com/NLPchina/ansj_seg. Accessed June 5, 2016.

[7] Wei Ji, Yong Wang, Dafang Zhuang, Daping Song, Wei Wang, Gang Li, Spatial and temporal distribution of expressway and its relationships to land cover and population: A case study of Beijing, Transportation Research Part D: Transport & Environment, China, VOL. 32, 86-96, 2014.