Transactions in GIS

Research Article Transactions in GIS, 2015, 19(1): 1–18

Mining Trajectory Data and Geotagged Data in Social Media for Road Map Inference

Jun Li,[*] Qiming Qin,[†] Jiawei Han,[‡] Lu-An Tang[‡] and Kin Hou Lei[‡]

[*]*IRSGIS, Peking University, Department of Computer Science, University of Illinois*
[†]*IRSGIS, Peking University*
[‡]*Department of Computer Science, University of Illinois*

Abstract
As mapping is costly and labor-intensive work, government mapping agencies are less and less willing to absorb these costs. In order to reduce the updating cycle and cost, researchers have started to use user generated content (UGC) for updating road maps; however, the existing methods either rely heavily on manual labor or cannot extract enough information for road maps. In view of the above problems, this article proposes a UGC-based automatic road map inference method. In this method, data mining techniques and natural language processing tools are applied to trajectory data and geotagged data in social media to extract not only spatial information – the location of the road network – but also attribute information – road class and road name – in an effort to create a complete road map. A case study using floating car data, collected by the National Commercial Vehicle Monitoring Platform of China, and geotagged text data from Flickr and Google Maps/Earth, validates the effectiveness of this method in inferring road maps.

1 Introduction

Road maps are an indispensable component in basic geographic information, and play an important role in a variety of fields including scientific research, public management, and commercial applications such as navigation and web map services. Road maps are expected to exhaustively reflect the current status of physical roads in the real world; otherwise, the quality of services depending on road information would be directly affected. Therefore, keeping road maps up-to-date is extremely important (Zhang 2004). Mapping roads, however, is costly and labor-intensive, especially in developing countries where road construction and rebuilding projects can be seen almost everywhere. In China, the total length of highways grows at an average annual rate of 13.44%, from 45,300 km in 2006 to 84,900 km in 2011 (Ministry of Transport of China 2012). Likewise in India, the total road length increased more than 11 times during the 60 years between 1951 and 2011 (Ministry of Road Transport and Highways of India 2012). In the past few decades, governments are less and less willing to absorb the costs, and programs of map updating are seriously lagging behind in many countries (Goodchild 2007).

To reduce the cost of the cycle of road surveying and updating, scholars have introduced the "wiki" philosophy into the map updating process. This innovation allows the public to

Address for correspondence: Qiming Qin, Institute of Remote Sensing and Geographic Information Systems, Room 501, Remote Sensing Building, Peking University, HaidianDistrict, Beijing, 100871, China. E-mail: qmqin@pku.edu.cn
Acknowledgements: This research is funded in part by National Science and Technology Pillar Program of China (Grant No. 2012BAC16B04), China Scholarship Council, US National Science Foundation grants IIS-0905215, CNS-0931975, CCF-0905014 and IIS-1017362. The authors would like to thank anonymous reviewers for their valuable comments on this article. Thanks also go to Max Yazhbin and Cassandra Rosado for their assistance in proofreading.

doi: 10.1111/tgis.12072

participate in creating or updating features in maps through websites or mobile applications by sharing their knowledge of the world. The information, shared by ordinary people, is called "user generated content", known as UGC (Krumm 2008).

Currently, UGC is being used for mapping roads in two ways. One way is collaborative mapping programs such as OpenStreetMap. Assisted by online mapping tools, registered users can update road features on the map based on remote sensing images, GPS trajectories, or their empirical knowledge. This provides a new data source for maps rather than from professional surveying practitioners, and the generated information is fairly accurate. But the problem is that both the editing and the creating of map features in these programs are completely manual; moreover, this kind of professional activity hinders more users from joining in. Therefore, without enough users' participation, these collaborative mapping programs are characterized by the problems of limited spatial coverage and low updating speed (Girres and Touya 2010). Another way is to automatically extract road information from trajectories uploaded by user. In this case, UGC contributors need only to upload their trajectory data, and the task of extracting road features is accomplished by computer programs. Compared with collaborative mapping programs, this approach is highly automated and does not rely on labor. For example, Bruntrup et al. (2005), Davies et al. (2006), Cao and Krumm (2009) and Li et al. (2012) use different kinds of vehicle trajectories to generate road networks in Germany, the UK, the US and China. However, the vehicle trajectory data only embody spatial information of roads (the location of roads), and do not provide attribute information (e.g. road name) necessary for road maps.

To the best of our knowledge, no existing work has extracted both spatial and attribute information of roads from UGC in an automated manner. The key to fulfilling this goal is to find a way to automatically extract the attribute information of road features which is often assigned by humans.

The data from social media sites such as Twitter, Facebook, Google+, and Flickr offer a source for obtaining attribute information on roads. According to EMarketer, there will be a massive 1.43 billion social network users in 2012, representing a 19.2% increase over 2011 numbers (EMarketer 2012). During the past five years, social media have become increasingly equipped with location-based features (Sui and Goodchild 2012), and users can share with others geotagged or location-specific media data, including geotagged text, photographs, video or Quick Response (QR) codes. Geographic content embedded in comments shared by social media users is also called Ambient Geographic Information in Stefanidis et al. (2012) and has been used for event detection (Crooks et al. 2013). Likewise, a review or comment about newly built sections of a city such as a new road or a new restaurant that is posted with geographic locations on social media websites can quickly result in updates to the previous data and enhance others' recognition of their environment. More importantly, this data can be updated very fast. Therefore, geotagged data in social media have opened a new way of acquiring map information including attribute information which cannot be obtained from other UGC, e.g. user-generated trajectories.

This article proposes a UGC-based automatic road map inference method that applies data mining techniques and natural language processing tools to trajectory data and geotagged data in social media to extract both spatial and attribute information in an effort to create a complete road map. The next section describes the general workflow of the UGC-based road map inference method, and introduces the implementation of the spatial and attribute information extraction process in detail. Section 3 presents a case study of the proposed method – an application in Beijing's road map updating, and discusses areas for future research. Section 4 concludes the article with a summary of our study.

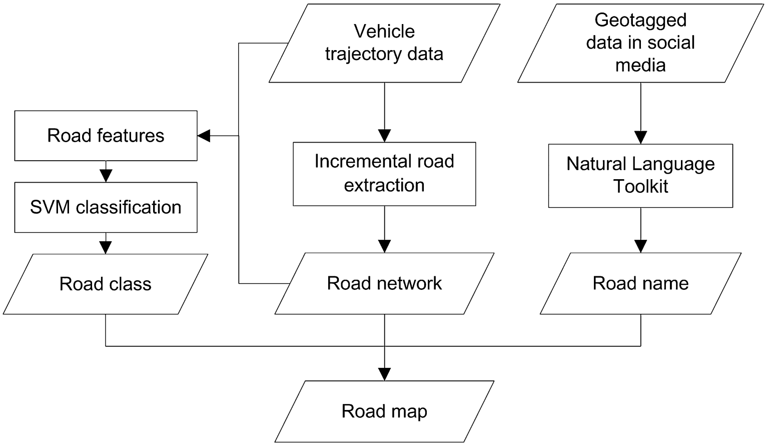


Figure 1 General workflow of road map inference

2 Methodology

A complete digital road map should be composed of at least two types of information: spatial information showing where a road is and attribute information telling which road it is. Vehicle trajectory data can tell us where roads are because vehicles are driven on roads most of the time, and their trajectories are consistent with the geometric information of roads. Attribute information, road names in particular, can be obtained from geotagged data in social media. Social media users sometimes point out place names when they share their feelings, experiences, comments, or photos, e.g. a restaurant comment "Salmon in Yotsuba Japanese restaurant on Yiheyuan Road is delicious". Combined with the geographic coordinates embedded in the comment, we can infer that in the locations represented by the geographic coordinates there is a road called Yiheyuan Road. In addition, the road class can be derived from the movement characteristics of vehicles.

The general workflow of the road map inference method is shown by Figure 1. As mentioned before, the two data sources are: vehicle trajectory data and geotagged data in social media. First, vehicle trajectory data are used to extract the location of the road network (the location of road centerlines) with the method proposed by Li et al. (2012). After the road network is obtained, the road features for each road segment are calculated based on the associated trajectory points, and then the Support Vector Machine (SVM) is used to classify every road segment into different types according to feature values. In the meantime, the Natural Language Toolkit (NLTK) is applied to extracting road names from geotagged text data in social media. At the last step, road network, road class, and road name are integrated together, based on the spatial location to generate road maps with both spatial and attribute information.

2.1 *Spatial Information Extraction*

In many cases, the sampling frequency of vehicle trajectory data is relatively low, and the shape of trajectories cannot reflect the real shape of roads making the road network extraction difficult. Also the time complexity of the road extraction algorithm should be scalable to be able to process a large amount of trajectory data. In view of the above problems, the road extraction method proposed by Li et al. (2012) is used here, which is applicable to both high- and low-frequency sampled data through integrated use of both spatial and semantic relationships of trajectory points. For clarity, a schematic description is provided here, and more details can be found in the original paper.

The general philosophy of this method is that the more trajectories pass through a place, the larger the probability of this place being a road. Each trajectory can be thought of as a driver's observation of a road. By integrating many drivers' observations about the same road, relatively accurate information about this road can be gathered. At the technical level, initial vehicle trajectories are assumed to be centerlines of roads and used to construct a candidate road network. Then the other trajectories are used to modify the corresponding roads in the candidate road network by an insert and merging process. The candidate road network increases gradually from scratch. After all trajectories have been processed, some post-processing such as confidence filtering, road smoothing, and road linking is conducted to refine the candidate road network. The time complexity of the whole process is $O(n \times m + m \times m)$, where n is the number of trajectory points, and m is the number of candidate roads.

In the ultimate road network extracted from vehicle trajectories by the above procedures, a road is a collection of road segments. A road segment is a polyline feature linking two intersections, and this polyline feature describes the location of road centerlines. A bidirectional road is represented by two polyline features, each of which describes a direction.

2.2 Attribute Information Extraction

2.2.1 Road class inference

Road class is an important type of attribute information for roads, and it is often assigned by transportation authority according to road characteristics, such as whether the opposite directions are separated and how many lanes are there. Different countries have different road type classification systems. For example, in China, urban roads are classified into four types: expressway, major road, minor road, and local road (Ministry of Construction of China 1991).

The key to automatically inferring road class is to obtain information about the road characteristics which play an important role when the transportation authority determines road types. The vehicle trajectory data can provide us with information about the traffic flow, speed distribution of vehicles, and distance distribution of vehicles from road centerlines. All these are defining factors for road class, and are called "road features". The supervised learning method is adopted to derive road class from these impacting features. Road class inference is composed of three major parts: feature calculation, feature selection and classification.

A. Feature calculation. The features which can be obtained from vehicle trajectory data are categorized into three groups: traffic-, speed- and width-related, as shown in Table 1. Due to the unknown relationship between road type and the speed distribution or the distance distribution of trajectory data, multiple statistical measures of these two distributions are selected and calculated. In later steps the measure most related with the road type is checked. In the traffic-related group, there is only one feature: vehicle density, which is defined as the number of vehicles per unit length of road per unit time. In the speed-related group, there are three features: average speed, third quartile speed and 90th percentile speed. The average speed is the mean value of moving speeds of all vehicle trajectory points associated with one road segment.

Table 1 Road type classification features

Traffic-related	Speed-related	Width-related
Vehicle density	Average speed Third quartile speed 90th percentile speed	Average width Third quartile width 90th percentile width

The third quartile speed is the speed value below which 75% of observations fall. Likewise, the 90th percentile speed is the speed value below which 90% of observations fall. The width-related group is similar to the speed-related group and also contains three features.

For each road segment, the above seven features are calculated. First, two thresholds d_1 and α_1 are used to pick out the associated trajectory points of each trajectory segment. d_1 is the distance threshold between a trajectory point and a road segment, and α_1 is the angle difference threshold between the road segment and the moving direction of the trajectory point. Only associated points will be used to calculate feature values. Then, according to the definition of all features, seven feature values are computed and stored in a 9-tuple:{*sid*, *loc*, *density*, *speend_ave*, *speed_75*, *speed_90*, *width_75*, *width_90*} where *sid* is the identifier number of road segment, *loc* stores the geographic extent of the road segment and the remaining seven attributes correspond to the seven features.

B. Feature selection. In supervised learning, feature selection is a critical procedure which searches through the feature set and finds the subset with the least classification error. As can be seen from Table 1, all features in the same group contain the same kind of information, thus there is no need to use all features for classification. Only one feature is selected from each group, and there would be nine combinations (1×3×3) of feature sets. In addition, in order to compare with the nine combinations, a tenth candidate which contains all the features will be used.

According to Blum and Langley (1997), there are two feature selection methods: the filter and the wrapper method. The filter method is defined as a preprocessing step to induction that can remove irrelevant attributes before induction occurs. The wrapper method, on the other hand, is defined as a search through the feature sets using the estimated accuracy from an induction algorithm as a measure of goodness of a particular feature subset (Weston et al. 2000). Because the number of feature subsets is small and the training data set would not be large either, the more computationally expensive and accurate wrapper method is chosen for feature selection. For each feature subset, the same training data are used to learn a classification model, and then the model is evaluated with test data. Each feature subset will correspond to a classification accuracy. The feature subset with the largest classification accuracy is selected as the ultimate feature set for classification of road types.

C. Classification. Choosing a good and suitable classifier is very important. Until now, many classification methods have been proposed, including decision tree classifications, Bayesian classifications, rule-based classifications, artificial neural networks (ANNs), and Support Vector Machines (SVMs) (Han et al. 2011). It is generally accepted that SVM is a good classifier with a good ability for anti-noise, so we choose it as the classification method.

SVM (Vapnik, 1998) uses a nonlinear mapping to transform the original training data $X \in R^n$ into a higher dimensional space $\Phi(X)$, and within this space it searches for the linear optimal separating hyperplane to be used for separating data into different types. The mapping is performed by a kernel function K. The decision function for SVM is:

$$f(x) = \omega \cdot \Phi(X) + b = \sum_{i=1}^{l} y_i \alpha_i K(X_i, X) + b \tag{1}$$

where X_i, y_i is the support vectors and their labels; X is a test tuple; α_i and b are numerical parameters that were determined automatically by the optimization or SVM algorithm; and l is the number of support vectors. Given a test tuple, X, we put it into the above decision function, and then check the sign of the result. If the sign is positive, then X falls on or above the maximum marginal hyperplane, and will be predicted to belong to class +1. Otherwise if the sign is negative, then X is predicted to belong to class –1.

In this research, LIBSVM (A Library for Support Vector Machines), developed by Chih-Chung Chang and Chih-Jen Lin at National Taiwan University, is used to implement SVM classification (Chang and Lin 2011). The radial basis function is used as the kernel: $K(x, y) = e^{-\lambda\|x-y\|^2}$. Given a set of training set, LIBSVM is used to classify the road segments with unknown types.

SVM is a supervised learning method, meaning that labeled samples (data with known types) are needed to train the classification model. Therefore, the type information of road segments is gathered through various sources, such as online sources and field survey. According to the instructions in the feature selection section, these labeled data are partitioned into two categories: training data and testing data. After determining the effective feature combination by the wrapper method, a learned classifier is used to predict the types of other unlabeled road segments.

2.2.2 Road name extraction

Geotagged data in social media contains many kinds of information. In this work, only two types of information are used: text content and geographic location. Text content is written by users and can be a few words, one sentence or a paragraph. Geographic location usually exists in the form of latitude and longitude. A 3-tuple is used to store the data: {*pid*, *loc*, *text*} where *pid* is the identifier of a geotagged record, *loc* is the geographic coordinate representing where the record is uploaded, *text* is the content written and uploaded by the user. With respect to social media data with geotags, *text* is likely related with *loc*, and *text* could be thought of as users' understanding about the place represented by *loc*. Particularly if some road names appear in *text*, usually either the road is at that location or not far from it. Thus geotagged data in social media serve as an abundant source for road name information.

The key to automatically extracting road names is how to process the text written by a person. In computer science, these data are called natural language, which (unlike artificial or constructed languages), is not organized in a fixed structure. Another feature of text data in social media is that it covers a broad range of topics, and it is hard to tell which text contains road names. In order to identify correct road names, not only the key word matching but also the lexical category of words are used. NLTK (Steven et al. 2009) is a library for processing human language and it is used as the tool for processing texts in social media. The road name extraction process is composed of three steps: filtering effective data, identifying road name and assigning names to road segments.

Table 2 Road related keywords

Road-related keywords
"highway", "expressway", "avenue", "ring", "road", "street"

Table 3 Simplified part-of-speech tagset (Steven et al. 2009)

Tag	Meaning	Examples
ADJ	adjective	*new, good, high, special, big, local*
ADV	adverb	*really, already, still, early, now*
CNJ	conjunction	*and, or, but, if, while, although*
DET	determiner	*the, a, some, most, every, no*
N	noun	*year, home, costs, time, education*
NP	proper noun	*Alison, Africa, April, Washington*
NUM	number	*twenty-four, fourth, 1991, 14:24*
PRO	pronoun	*he, their, her, its, my, I, us*
P	preposition	*on, of, at, with, by, into, under*
V	verb	*is, has, get, do, make, see, run*

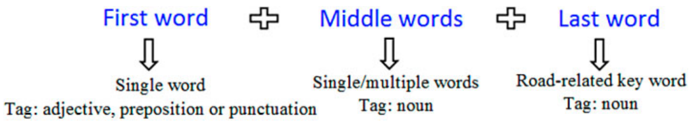


Figure 2 Model for identifying road names

First, the geotagged data which are more than d_2 away from any road segments are removed. Another filter is conducted using the road-related keywords, and only those data whose text contains at least one of the key words in Table 2 are left for further processing.

Second, the lexical category of words is chosen to refine the results in the first step, and to identify road names. A text with road related keywords does not necessarily guarantee that there are road names in it. One reason is that the keyword might not be used as a noun, e.g. "ring" could also be used as a verb. Another reason is that it might not contain an explicit road name, e.g. "on the road". In order to avoid this problem, the tokenization function of NLTK is used to break up sentences into words and punctuation, and the tagging function is used to classify words into their part of speech (Table 3 shows a simplified part-of-speech tagset). Then the model shown by Figure 2 is used to identify road names. The model consists of three parts: the first word should be a single word and its lexical category should be either adjective, preposition or punctuation; the middle words could be one or more noun words; the last word should be one of the road-related key words and its lexical category should be noun. If part of the text matches well with the model, then it is thought that a road name exists in the text, and the name is the middle words plus the last word. When a geotagged text contains road names, it is transformed into a road name point represented by a 3-tuple: {*rid*, *loc*, *name*}

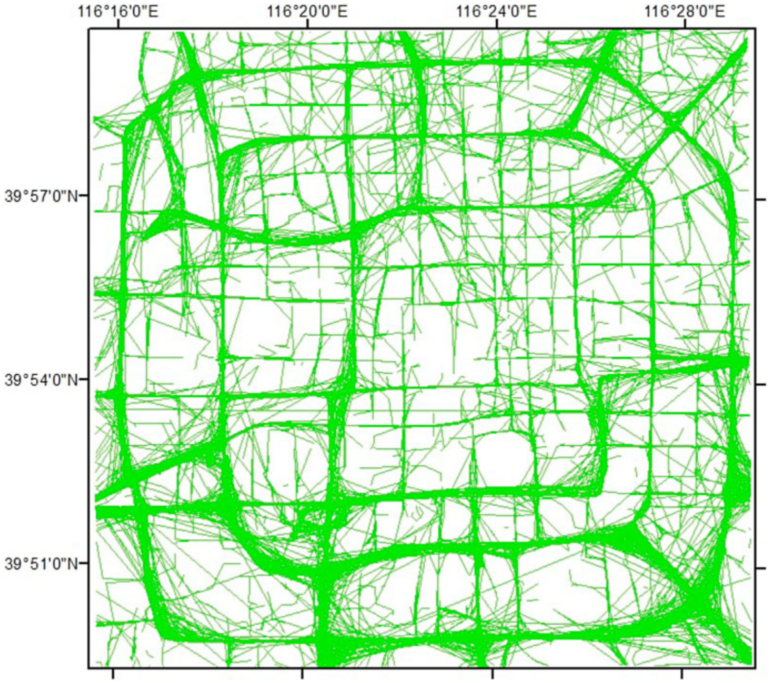


Figure 3 Vehicle trajectories in the study area

where *rid* is the identifier of a road name, *loc* is the geographic coordinate of the road name point, *name* is the road name.

After the second step, a large amount of road name points are obtained. Then they are associated to the road segments by the nearest neighbor method. The situation could happen that a road segment is associated with multiple different names. The voting strategy is used here to solve this problem, which means the name appearing most often is assigned to its associated road segment.

3 Results and Discussion

3.1 Study Data

The vehicle trajectory data were collected by the National Commercial Vehicle Monitoring Platform of China (NCVMP). The NCVMP is a nationwide floating car management platform which is tracking nearly one million special vehicles all over China. The dataset collected by this platform contained the following fields: license plate number, latitude, longitude, travel speed, travel direction, and positioning time. We selected part of Beijing City as our study area which covered 39°49′26″ – 39°59′39″N and 116°15′27″E – 116°29′29″E, and the trajectory data within this area during the period from 11th to 15th September 2010 were gathered for extracting road networks. This trajectory dataset was composed of 1,048,576 records, as shown in Figure 3.

Many social media sites can be a data source for this research; however, due to data access policies, we only obtain geotagged data from Flickr and Google Earth/Maps. Flickr (Figure 4)

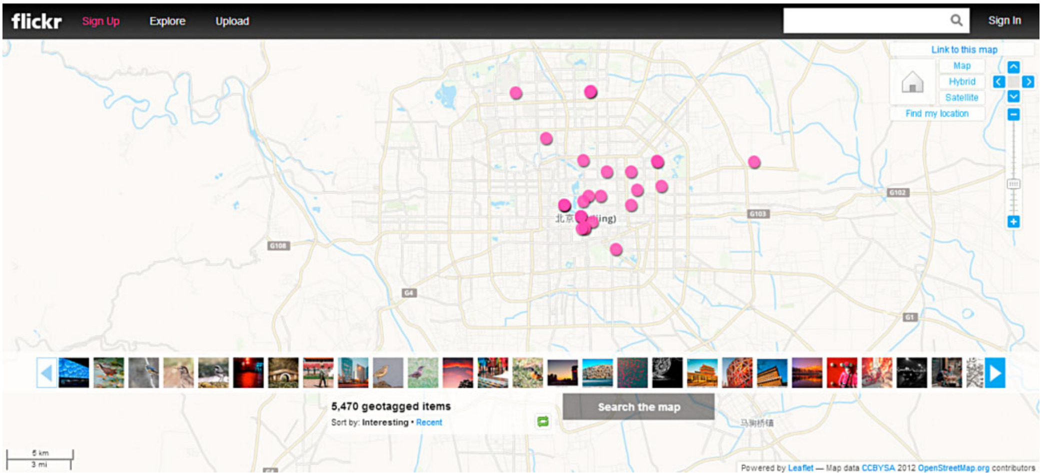


Figure 4 Geotagged photos on Flickr.com

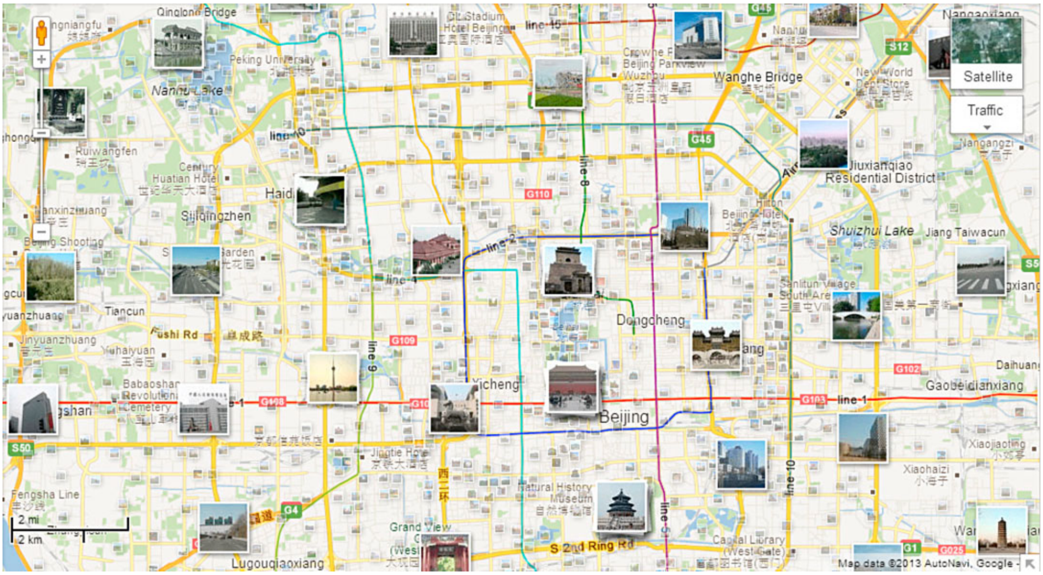


Figure 5 Geotagged photos on Google Maps

is an image and video hosting website where people can share personal photographs and videos with others. Each photo can be assigned a title, description, tags, upload time, taken time, etc. Likewise, in Google Maps/Earth (Figure 5), users can also write a description of their photos. Both Flickr and Google Maps allow users to geotag photographs to show where these photos were taken. The texts about photos written by users as well as geographic locations were the geotagged text data we used. The data within the study area were uploaded by users during the period from 3rd November to 3rd December 2012 and were chosen as our second study dataset. The dataset consists of 5,847 records. Figure 6 shows the spatial distribution of

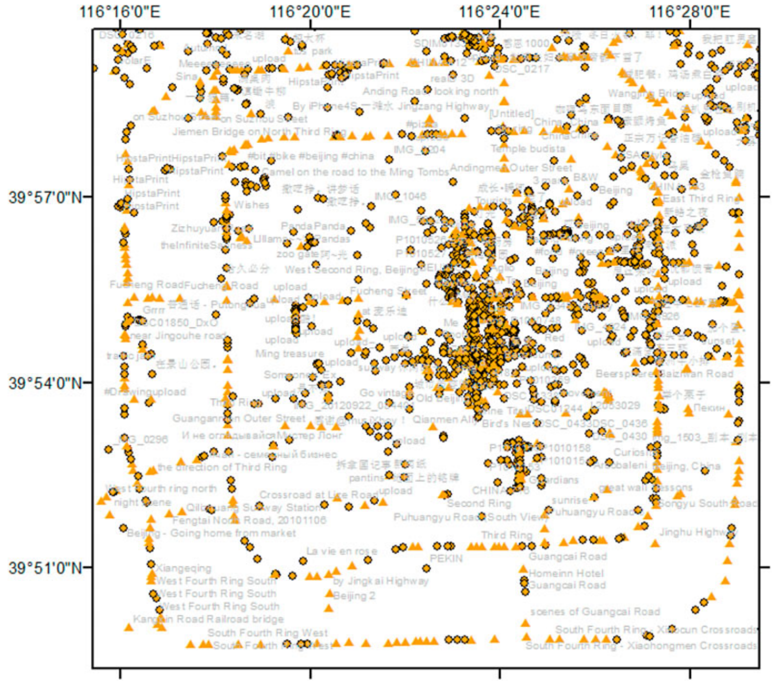


Figure 6 Geotagged data in the study area (triangles represent the geotagged photos from Google Maps, and circles represent the geotagged photos from Flickr)

this dataset. The mismatch of data collection time between these two datasets is due to data access limitations, but this would not affect the demonstration of the proposed method.

3.2 Experiment Results

3.2.1 Road network

Figure 7 shows the road network extracted from vehicle trajectories. One hundred and fifty-three road segments in total were generated, which formed the skeleton of Beijing's roads. Figure 8 shows the road extraction result within a subregion of the study area (marked by a red rectangle in Figure 7) in detail. It can be seen from Figure 8c that the spatial location of extracted roads match well with the physical roads displayed in remote sensing images. The effectiveness of this method has been validated by comparing the extracted road network with both a business road map and remote sensing images in Google Earth in Li et al. (2012).

3.2.2 Road class

The vehicles tracked by the NCVMP were special vehicles including vehicles transporting hazardous materials, coaches, and tourist shuttles, so they were driven on arterial roads most of time. Because of this, we categorized the extracted road segments into two classes: expressway and major roads. We chose 16 road segments as a training set and another 30 segments as a

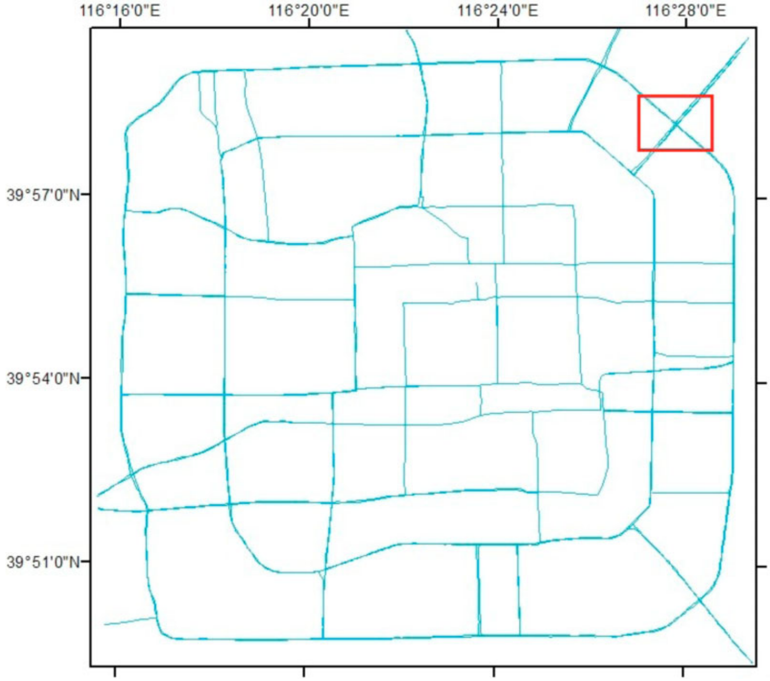
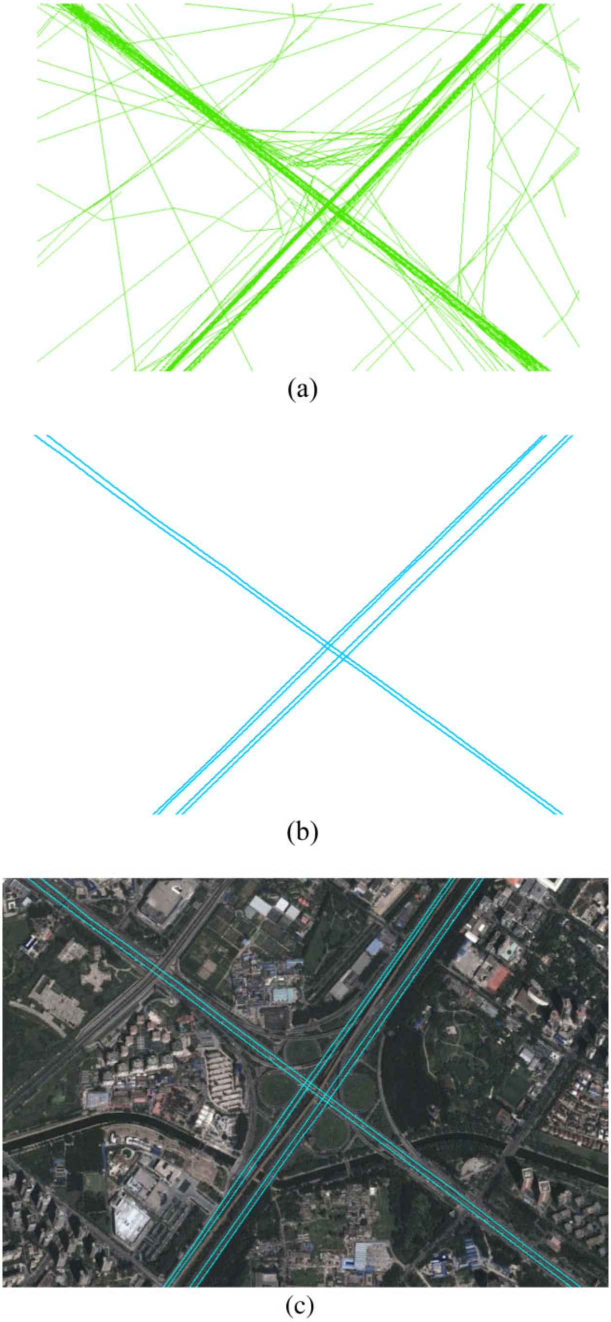


Figure 7 Extracted road network

testing set, and obtained the true values for their classes. For the training set, we chose eight expressways and eight major roads in order to make the classification model learn the defining characteristics of both types of roads, while testing data were randomly chosen.

In the first step, d_1 was set to 30 m based on the maximum road width and the location error of GPS, and α_1 was set to 30° based on general driving rules. During feature selection, we run the SVM algorithm on the training set for each of the 10 combinations of features and evaluated the quality of every feature combination with the test set. By comparing the predicted classes of 30 road segments given by the classification model with their true class labels, we calculated the proportion of the correctly classified tuples to the total number of test tuples and took this value as the measure of goodness of feature subsets. Table 4 shows the classification accuracy of SVM for all feature combinations.

As can be seen from Table 4, the feature combination of vehicle density, 90th percentile speed and 90th percentile width, shows the best performance for the SVM algorithm to characterize the road class. As a result, we chose the model learned based on this feature subset as the classification model. Then the unknown road type segments were classified into two types using the SVM classification model. Among the 153 road segments, 52 were classified as major roads, while the other 101 were classified as expressways. Figure 9 shows the comparison of feature value distribution between expressways and major roads. Figure 10a shows road classes of the road network with different style lines: double line for expressways and single line for major roads. Compared with Google Maps, only 10 out of 153 road segments did not match, so the classification accuracy was 93.5%. Figure 10b highlights the unmatched road segments with marked lines. The classification result showed that vehicle trajectory data is effective for inferring road classes.



(a)

(b)

(c)

Figure 8 Road extraction result in a subregion of the study area: (a) Vehicle trajectories; (b) Extracted roads from trajectories; and (c) Extracted roads overlaid on remote sensing images

Table 4 Classification accuracy of SVM for all feature combinations (Numbers are used to represent different features: 1-vehicle density, 2-average speed, 3-third quartile speed, 4–90th percentile speed, 5-average width, 6-third quartile width, 7–90th percentilewidth)

Feature combination	No. of correctly classified tuples	Classification accuracy
125	27	90.00%
126	24	80.00%
127	25	83.33%
135	24	80.00%
136	25	83.33%
137	28	93.33%
145	26	86.67%
146	27	90.00%
147	29	96.67%
All	21	70.00%

3.2.3 Road name

The raw geotagged text data in Flickr and Google Maps were first filtered by the distance threshold (45 m) and road-related keywords, and then further processed by NLTK. The reason why the distance threshold for geotagged texts is larger than that for trajectory points was due to the geotagged texts posted on roads and because these data posted along roads can be used for extracting road names. The effective road name points were generated which are shown in Figure 11. Through the voting strategy, every road segment was assigned the name which repeats most often in the road name points near to this road segment. Figure 12 shows the extracted road network with names. The names of 20 road segments were not identified because no road name points can be found nearby. Also by comparing the result with Google Maps, we found that 10 names were wrongly identified. This was caused by the geotagged texts in which the alias names of roads were used. Overall, 30 road names were missed or wrongly identified, so the accuracy for road name extraction was 80.4%.

3.2.4 Final road map

The vector features for roads, road classes, and road names were linked together through their locations. The first component as the spatial information, and the last two components as the attribute information, constituted a road map with comprehensive information. Figure 13 shows the final road map extracted from the vehicle trajectory data and the geotagged data in social media, which constitutes both spatial and attribute information.

3.3 Discussion

By comparing our result with those of Bruntrup et al. (2005), Davies et al. (2006), Cao and Krumm (2009), and Li et al. (2012), we can see that our method not only obtains road locations which can be extracted by their methods, but also has identified road classes and names.

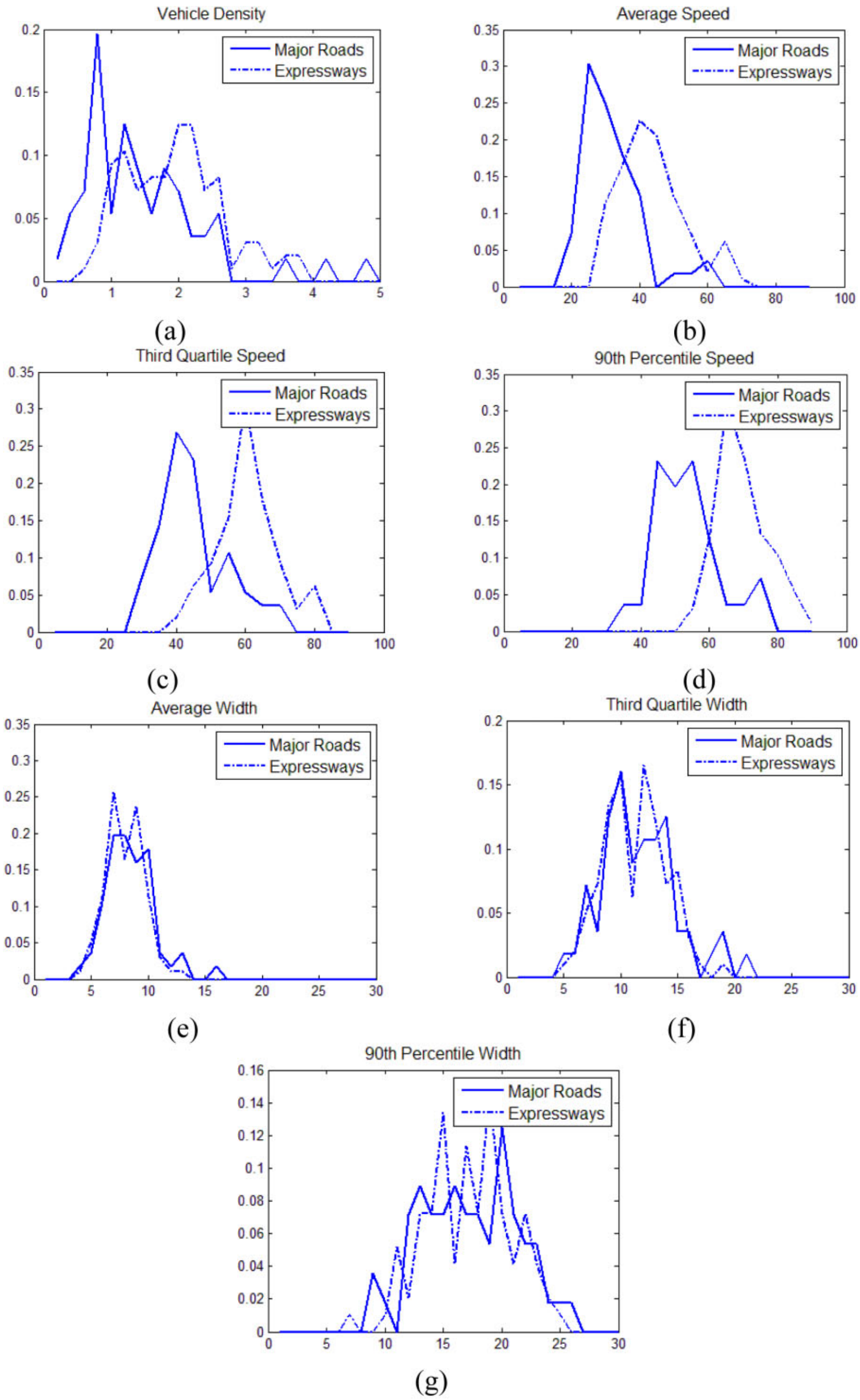
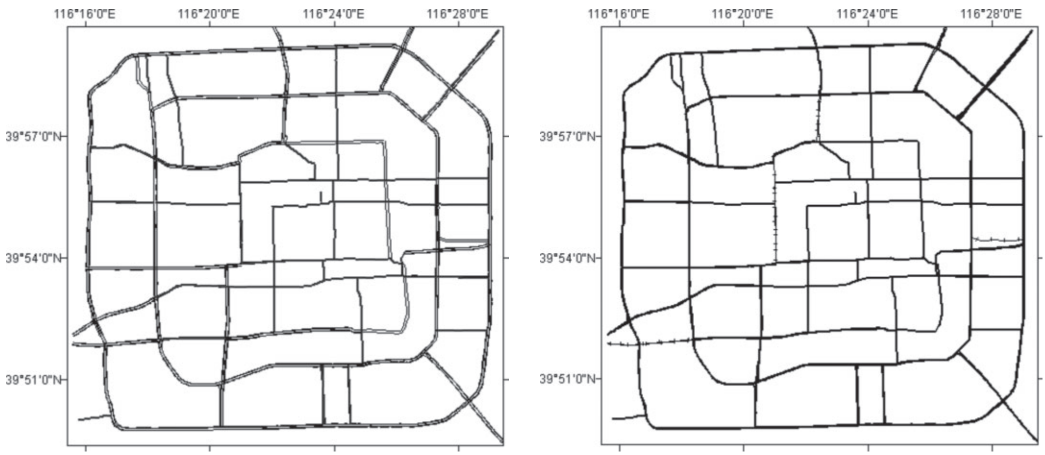


Figure 9 Comparison of feature value distribution between expressways and major roads



(a) (b)

Figure 10 (a) Road network with classes; and (b) Comparison between predicted values for road class and true values

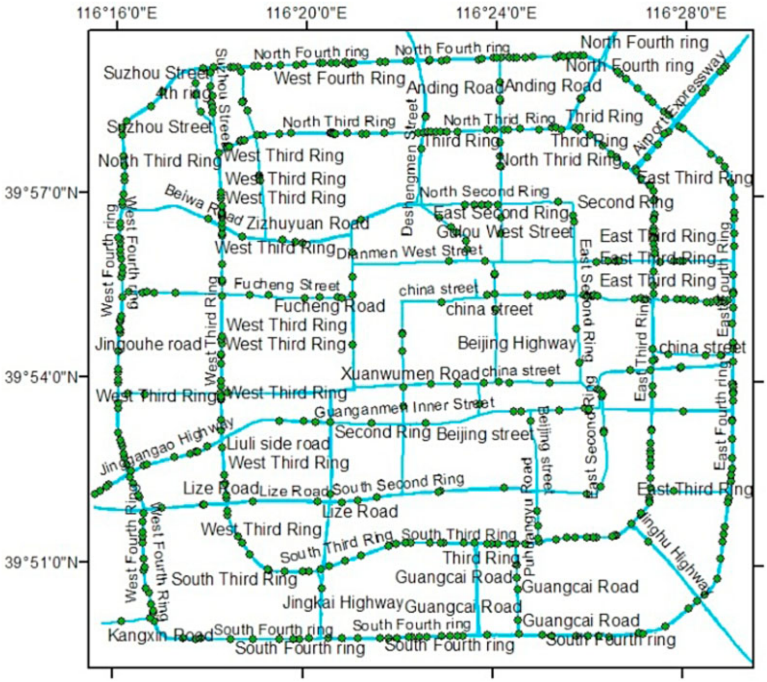


Figure 11 Road name points

Compared with collaborative mapping programs, our method requires less manual work and extracts information from social media data which users have stronger incentives to provide.

The major contribution of this article was to propose an automatic way of deriving both spatial and attribute information of roads by mining trajectory data and geotagged data in

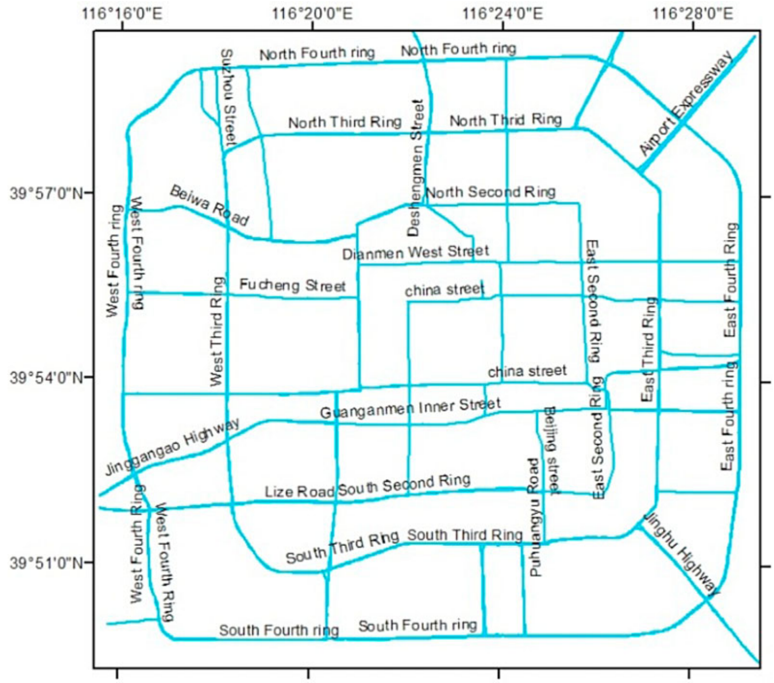


Figure 12 Road network with names

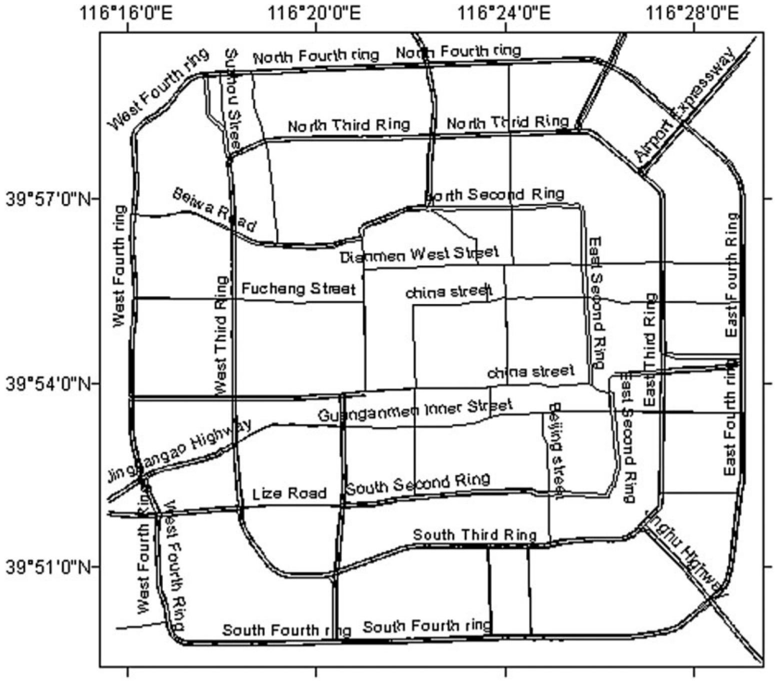


Figure 13 Road map with names and classes

social media because this method requires less manual work for contributors than existing collaborative mapping programs. However, it is clear that as both trajectory data and text data are complicated, more work needs to be conducted to improve the accuracy of road map extraction. Future work should focus on the following aspects:

1. The trajectory data used in this work came from one platform, and these vehicles are mainly driven on arterial roads. If more types of roads need to be updated in maps, multiple sources of trajectory data (e.g. taxi trajectory data) need to be used, and this will raise two questions: one is how to accurately extract roads from different sources of trajectory data since the positioning accuracy or sampling frequency might be variable in different datasets; another is what would the classification accuracy be if the number of types to be classified is three, four, or even more? Could it still stay high?
2. The analysis and understanding of natural language is a significant challenge. The text data in social media are casual, and instead of using official road name, users might use similar sounding names (Lincoln Street and Lincoln Avenue), or abbreviations (Green St.). Further work is needed to improve the road name identification method. In addition, with respect to a road that is assigned different names by different users, some trust analysis methods for crowdsourcing data can be introduced in the voting strategy procedure.
3. Social media data covers a broad range of topics, and users share comments, reviews, feelings, or experiences about everything. We can also focus on mapping other geographic features, sidewalks for instance. By analyzing the related keywords, sidewalk maps might also be found.
4. Unlike conventional geographic data, volunteered geographic information (VGI) is collected using different methods by different individuals with different motivations and preferences (van Exel et al. 2010). Although some research work has been conducted on quality assessment of VGI (van Exel et al. 2010; Begin et al. 2013), further research on operational assessment indicators is still needed.

4 Conclusions

With the development of location-aware technologies and Web 2.0, an unprecedentedly large amount of moving object and geotagged media data are generated, providing a new comprehensive way of obtaining knowledge about the Earth and its human and natural complexities. More specifically, the moving object data collected by the pervasive location-acquisition devices allow us to know where things are. In addition, geotagged messages used for sharing knowledge among people in social media can serve as a way to know what is here. These two types of data are closely related, and mining them together can result in more abundant knowledge about the world.

This article explores the integrated use of these data in the field of road map inference. The road network is extracted from vehicle trajectory data by an incremental generation method, and road classes are obtained based on the SVM algorithm. By means of NLTK, road names are identified from geotagged data in social media. The results provide both spatial and attribute information for road map inference. Although further work needs to be conducted to improve the accuracy, this work makes a positive first step on the problem that we still do not have the tools to automatically discover relevant information for a particular application over the massive data, as stated in Sui and Goodchild (2012). This work lays a foundation for automatically inferring road maps from low-cost UGC, and can also be applied in road map updating when an existing road map is available.

References

Begin D, Devillers R, and Roche S 2013 Assessing volunteered geographic information quality based on contributors' mapping behaviors. In *Proceedings of the Eighth International Symposium on Spatial Data Quality*, Hong Kong

Blum A and Langley P 1997 Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97: 245–71

Bruntrup R, Edelkamp S, Jabbar S, and Scholz B 2005 Incremental map generation with GPS traces. In *Proceedings of the Ninth IEEE International Conference on Intelligent Transportation Systems*, Las Vegas, Nevada: 574–79

Cao L L and Krumm J 2009 From GPS traces to a routable road map. In *Proceedings of the Seventeenth ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2009)*, Seattle, Washington: 3–12

Chang C C and Lin C J 2011 LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2(3): 27.1–27.27

Crooks A, Croitoru A, Stefanidis A, and Radzikowski J 2013 #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS* 17: 124–47

Davies J, Beresford A, and Hopper A 2006 Scalable, distributed, real-time map generation. *IEEE Transactions on Pervasive Computing* 5(4): 47–54

EMarketer 2012 Facebook Helps Get One in Five People Worldwide Socializing on Online Networks. WWW document, http://www.emarketer.com/Article/Facebook-Helps-One-Five-People-Worldwide-Socializing-on-Online-Networks/1008903

Girres J F and Touya G 2010 Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS* 14: 435–59

Goodchild M F 2007 Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructure Research* 2: 24–32

Han J W, Kamber M, and Pei J 2011 *Data Mining: Concepts and Techniques* (Third Edition). San Francisco, CA, Morgan Kaufmann

Krumm J 2008 User-generated content. *IEEE Pervasive Computing* 7(4): 10–1

Li J, Qin Q M, Xie C, and Zhao Y 2012 Integrated use of spatial and semantic relationships for extracting road networks from floating car data. *International Journal of Applied Earth Observation and Geoinformation* 19: 238–47

Ministry of Construction of China 1991 The urban road design specification. *CJJ*: 37–90

Ministry of Road Transport and Highways of India 2012 Basic Road Statistics of India, 2010–2011. WWW document, http://www.morth.nic.in/writereaddata/mainlinkFile/File839.pdf

Ministry of Transport of China 2012 Statistical Bulletin for the Industry Development of Highway and Waterway Transportation in China. WWW document, http://www.mot.gov.cn/zizhan/siju/guihuasi/tongjixinxi/niandubaogao/201204/t20120425l231653.html

Sui D and Goodchild M F 2012 The convergence of GIS and social media: challenges for GIScience. *International Journal of Geographical Information Science* 25: 1737–48

Stefanidis A, Crooks A, and Radzikowski J 2012 Harvesting ambient geospatial information from social media feeds. *GeoJournal* 78: 319–38

Steven B, Loper E, and Klein E 2009 *Natural Language Processing with Python*. Sebastopol, CA, O'Reilly Media, Inc.

van Exel M, Dias E, and Fruijtier S 2010 The impact of crowdsourcing on spatial data quality indicators. In *Proceedings of the Sixth International Conference on Geographic Information Science (GIScience 2010)*, Zurich, Switzerland

Vapnik V 1998 *Statistical Learning Theory*. New York, John Wiley and Sons

Weston J, Mukherjee S, Chapelle O, Pontil M, Poggio T, and Vapnik V 2000 Feature selection for SVMs. In Leen T K, Dietterich T G, and Tresp V (eds) *Advances in Information Processing Systems 13*. Cambridge, MA, MIT Press: 668–74

Zhang C 2004 Towards an operational system for automated updating of road databases by integration of imagery and geodata. *ISPRS Journal of Photogrammetry and Remote Sensing* 58: 166–86