Design and Analysis of a Tweet Alert System for Identifying Real Time Traffic Using K-Means Clustering Algorithm

D D D Suribabu, T Hitendra Sarma, B Eswar Reddy

Abstract--- At present, one of the major issues for an individual to meet their requirements is the disordered traffic. In order to resolve this issue, this proposed thesis focuses on designing an application in order to classify each and every individual tweet based on the traffic related keywords and assign a unique label for all tweets. If any message, which contains traffic-related information, it is being sent as an alert to the end users who are following the current user, or else the same tweet will be just posted on the user wall. In the digital era, the social networks have become a fascinating domain for every human for sharing and communicating their recent updates among each other. In order to implement this application, it chooses a compatible social media that is Twitter, for sending traffic related tweets to the followed users. This problem is solved by applying the K-Means algorithm for identifying the traffic-related keywords from the tweet and then clustering the traffic tweets and normal tweets into two separate categories.

Key Words: Traffic Tweets, Tweet Classification, Social Networks, Text Mining Technique, Twitter Stream Analysis,K_Mean

1. INTRODUCTION

Twitter is becoming one of the fabulous blogging services, which has received much attention in recent days. This is one of the OSN the service which attained a variety of individual's interest towards it in terms of updated status information can be shared among the friends, family, and coworkers [1]. In twitter, each and every message is termed as a status update message or simply SUM, which is just a message to wish friends or colleagues. As we see there are lots of research papers published on twitter as of now, about various possible facilities that are present in twitter. All the previous research papers are mainly divided into three groups: Initially if we look in the starting group, we will try to identify maximum researchers analyze the complete network structure of Twitter and they want to calculate the workload that is present for the Twitter application [2]-[4]. In the next group, many researcher staff tries to examine or find out the important characteristics of a Twitter application as one of the social medium [5]-[8]. In the last group we can examine clearly that user's try to create unique apps in order to compete for the twitter. As we all know that, in the twitter there is a separate terminology that was used in order to represent the tweets. When any tweet user try to post a new message or tweet on his/her wall, then such a message is termed as Status Update Message (SUM).This SUM contains not only the information related to tweet in text manner but also contains some additional information like tweet time stamp, geographic coordinates like latitude and longitude of that posted tweet user, username of that corresponding tweet posted user and finally the hash tags that was applied for that appropriate tweet[9]-[14].

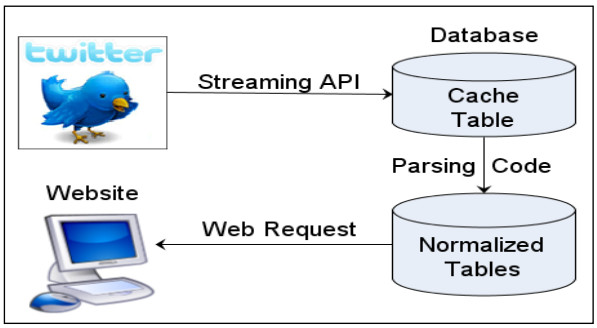

FIGURE 1: DENOTES THE SAMPLE ARCHITECTURE OF A TWITTER

From the above figure 1,we clearly find out the detailed demonstration about the flow of a twitter API, where each and every twitter has a facility to stream the information. For this twitter will connect with stream API, once the twitter sends any tweet it will be saved into the database with the help of cache table. Now the information which is available in the cache table will be parsed, the term parsing indicates that token is generated for the each and every individual tweet. In this parsing approach, it will try to normalize the tweets and eliminate if there are any duplicate tweets that are available in the cache table. Once the tweets are normalized then the information is passed into the website in the form of a web request.

Generally, in the twitter there are a separate terminology that was used in order to represent the tweets, one among them is when a user message is shared in the social networks, it is called as Status Update Message (SUM), and it may contain, apart from the text, it also contains the other information like timestamp, geographic coordinates (i.e. latitude and longitude), username who posted that tweet and also the hash tags that was applied for that appropriate tweet. As we all know that a lot of SUMs which refer to a certain topic or related to a limited geographic area may provide some information about an event or topic [9]- [10].

Revised Manuscript Received on March 02, 2019.

D D D Suribabu, Research Scholar, Dept of CSE, JNTUA, Ananthapur, AP, India

Dr.T Hitendra Sarma, Professor, Dept of CSE, SRIT, Ananthapur, India

Dr.B Eswar Reddy, Professor & Principal, Dept of CSE, JNTUACE, Kalikiri, AP, India

2. RELATED WORK

In this section we mainly discuss about the knowledge regarding the working of twitter and its various services. Now let us discuss about those in detail as follows:

Motivation

Now a day's twitter is the main source for spreading the ideas and information throughout the Web from one location to other. The message which was passed through twitter is known as the tweet, where it will mainly discuss various trends, innovative ideas, upcoming or ongoing events, and a lot more. This is one of the main reason which laid a way for data mining community to show more interest on this twitter application. In recent days all the researchers try to assume twitter as the good resource provider in order to detect all the new events in the real-time. In this current paper author mainly presented 4 challenges of tweets like:

1. Health Related Symptoms Identification,
2. Natural Disasters Detection,
3. Latest Trends and Emerging Topics Detection,
4. Rating Analysis.

All the above 4 challenges are mainly based on Data Mining principles such as data clustering and data classification. The researcher try to review all these pre-defined approaches by providing a detailed description for each and every challenge individually. As we know that Twitter is one type of micro blog, their rates of activity reached for a high level without the need of precedent factor[11]. By conducting a survey on the micro blogging service, we got an estimation report that over a hundred of millions of users are registered in this micro blogs as tweet users for communicating and sharing their mutual interest one among other with their last thoughts, different moods or activities in various words. This can be clearly explained in above figure 2, where all the four challenges of the tweet are shown in the tweet event detection model.

3. KNOWLEDGE DISCOVERY

In this section we mainly discuss about the knowledge discovery that was used in proposing and implementing this current paper

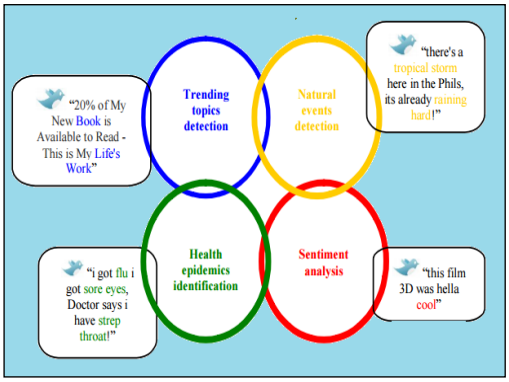

FIGURE 2: DENOTES THE 4 CHALLENGES FOR TWEET EVENT DETECTION

1. A Novel Graph Mining technique to show the communication details with the help of exchanged messages.

2. A Novel Text Mining technique calculated based on type of message content that is available during communication.

From the above two contexts, first one is nothing but graph mining based on the analysis of links that are having between each and every tweet messages. Next if u go with an alternate way, there u can find out the text mining approach based on the analysis of the message content what that was messaged by the tweet users. In general, almost all the users try to show their interest on the tweets content using text mining. For describing the event or a tweet, we have four main dimensions' like

a. **Event Type**: what is happening in that tweet?
b. **Time:** when an event is happening?
c. **Location:** where an event is happening?
d. **Entities:** who is involved in an event?

All the above four dimensions are needed for classification or clustering of any tweet[12]. As we told from figure 2 there are four events which take place for a tweet, all the four events mainly deal with the above dimensions for identifying the tweet and its usage.

4. A NOVEL PROPOSED TRAFFIC DETECTION SYSTEM BY USING TWITTER STREAM APPLICATION

In this section we will find out the Novel proposed traffic detection system by using twitter as the main application including the service like Service Oriented Architecture (SOA) and based on the Status Update Message (SUM) that was used to identify the status of any tweet.

4.1 Preliminaries

In this section, we mainly discuss about the proposed twitter streams for identifying the real time traffic. The proposed system is mainly divided into 3 modules like:

I) To identify the Status Update Message (SUM) and do pre-processing.
II) To elaborate the Status Update Message (SUM) and find its importance.
III) To Classify the Status Update Message (SUM).

From the above three main modules we can observe one thing common in all of the 3 modules like Status Update Message (SUM). These can be shown clearly in figure 3 and this was applied on an example of traffic event identification.

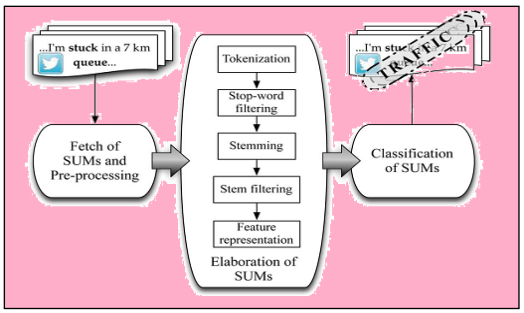

Figure 3: Illustrates the Working Flow Of Proposed Tweet Alert System

I) To Identify The Status Update Message (Sum) And Do Pre-Processing.

In this module we mainly discuss about the status update messages (SUM) and try to identify all the SUM's from all users and try to apply pre-processing for those tweets. Here the term pre-processing in data mining indicates cleaning of messages if there is any abused content, spelling mistakes, extra values, and special characters and so on. The term pre-processing will identify all these things and try to ignore un-necessary constants that are available in the messages and try to identify the important characteristic like traffic related keywords. Here each and every tweet contains a lot of information which defines the complete description about that tweet.

II). To Elaborate the Status Update Message (Sum) and Find Its Importance

In this module we mainly discuss about the "Status Update Message (SUM) and try to find its importance by reading the SUM". Here we applied text mining technique in which all repetitive words along with stop words are removed and only the words which contain useful information is only kept available in the SUM. Now we try to arrange all the available words in a set and then identify each and every word with the pre-defined set that is available in the database. If the available word is matched with the set of words, then such a tweet is known as traffic related tweet and if the same words are not matched with any of the set, then the tweets is treated as non traffic tweet. Here we divide each and every tweet into two tokens one is traffic related tweet token and other one is non traffic related tweet token.

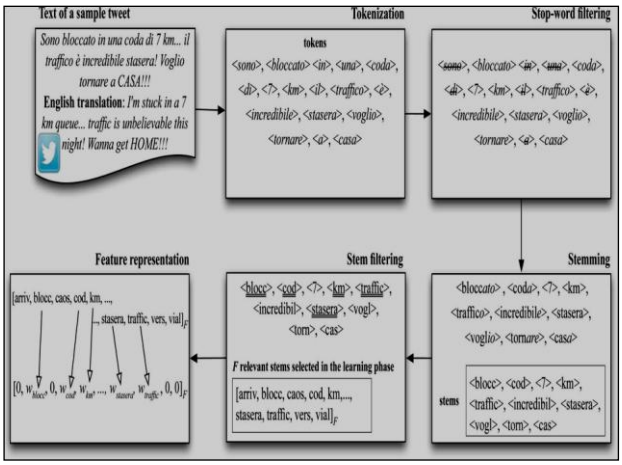

Figure.4: Denotes the Flow of a Sample Tweet with Several Data Mining Elaboration Techniques

III). To Classify the Status Update message (Sum)

In this module we try to classify the tweets based on the semantic meaning of each and every tweet, basically tweets are classified into two types: One is Traffic Tweet and other is Normal Tweet. So based on the semantic meaning of each and every user posted tweet, the tweets are classified into either any one of the two categories. If the message contains one or more than one traffic related keyword, then such a tweets are classified as Traffic related tweets and if the same message contains no single word related to traffic then such a tweets are labeled or classified as Normal tweets.

1. PROCEDURAL APPROACH OF OUR PROPOSED MODEL

In this current section we try to describe about our proposed model which is analyzed in this present thesis.

Let A is defined as a Novel system which contains the following attributes like :

A= {IP, PR, OP}

IP = Input

PR= Procedure.

OP= Output.

Now We assume IP = {U, T, TS, URL, Tk}.

1. Initially we try to consider U as a set of twitter users who are present in this current system.

U = {u1, u2,……un}.

Here the tweets are termed with U1,U2 and so on till Un.

2. Next we try to consider about the 'T',which is nothing but a set of Twitt's of a specified twitter user who try to login into the application.

T= {t1, t2, t3…tn}.

Here the set of twitt's or SUM's are termed with letters like t1,t2,t3….tn.

3. Now we try to identify TS as the twitter streamer which can analyze the twits.

4. Now we try to find the URL

5. Finally we try to identify the Tk parameter as the tokenization of SUM.

PR = Procedure.

5. K-MEANS ALGORITHM FOR CLUSTERING THE TWEETS

In this section we mainly describe about the k-means algorithm for clustering the tweets based on the keywords.

MATHEMATICAL REPRESENTATION

Let us assume a set of N tweets with letters't' and they are initialized to T = {t1, t2, t3,..., tN }.

Here each and every tweet t1,t2 and so on is represented with a set of words from google wiki or word frequency data sets.

Let us assume that WI is word index for set of google wiki pages from where these tweets are going to be matched

Assume |W I| = n1

Let us assume that WO as the words present in a tweet message.

Assume |W O| = n2.

If we want to calculate the matrix for the set of words classified from the tweet message, we need to map the matrix representation with Mt_Wiki.

As we all know that for any matrix there will be a representation, so for this matrix we assume the order as N × n1.

Finally the Mt_Wiki is treated as a Boolean sparse matrix and is defined as follows:

Mt_Wiki $_{i,j}$ =

$$\begin{cases} 1 & \text{if (Wiki}_j \text{ is neighbor of corresponding tweet ti using a bidirectional approach)} \\ 0 & \text{If it is False} \end{cases}$$

In a same way let MWord be the second matrix which holds all the words that are identified from a tweet. The order of MWord matrix is N × n2 and is defined as:

MWord $_{i,j}$ =

$$\begin{cases} \text{Frequency(WOj)} & \text{if (W Oj} \in \text{ti)} \\ 0 & \text{otherwise} \end{cases}$$

Now the distance between two tweets ti and tj is measured as follows:

$dist(ti, tj) = \alpha * Gdist(ti, tj) + \beta * Wdist(ti, tj)$

Here α and β are two sample parameter and we assume that $\alpha + \beta = 1$. The value of the parameters make biased towards one measure. Here in order to form the value equal to 1,we assume α to be 0.75 and β to be 0.25.Here in order to find the cosine similarity for the values ti and tj,we try to assume G_{dist} as the graph distance for the google wiki words and also.Now the minimum distance for tweets and corresponding wiki words are identified as follows:

$$Gdist(t_i, t_j) = \min_p \min_q \left\{ SPL(M\_WIKI_{i,p} * WI_p, M\_WIKI_{j,q} * WI_q) \right\}$$

Here the term SPL represents for Shortest Path Length. Here from the below figure 5,we can find some resultant words which are highlighted bold and formed as a cluster based on some traffic tweets. We can see in Figure 5 that, all the bright words (i.e Most frequently occurred words) belong to some events/topics related to traffic, which clearly signifies that tweets in this cluster are semantically related to traffic on a road. All the tweets related to traffic (Due to jam,accident,heavy vehicles etc.). The small errors in the cluster are due to the sparseness and noisiness of some tweets.

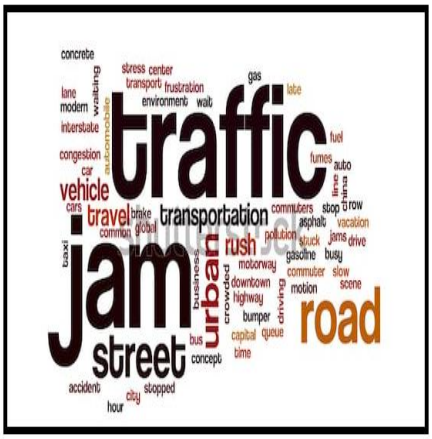

FIGURE.5: REPRESENTS THE BAG OF WORDS RELATED TO TRAFFIC TWEET

From the above figure 5, we can identify a bag of words related to traffic and if any tweet is matched with a set of words from the MWord matrix then that tweet message is classified as traffic tweet and they will be sent as a warning to the corresponding end user. And if the same tweet has no word matched from the bag of words related to traffic, then such a tweet is identified as normal tweet and this will be posted on user wall.

6. RESULT ANALYSIS

As we all are facing much road traffic and in turn a number of accidents while we try to travel from one place to another place. Although there are number of traffic regulatory devices and rules, still there was a heavy traffic jam in current days through which wastage of fuel during waiting in traffic. So in this paper we implement a new tweet system which can monitor the current road traffic and it can send that information to others who are directly linked with that tweet account. This is mainly done with the help of a Twitter stream analysis.

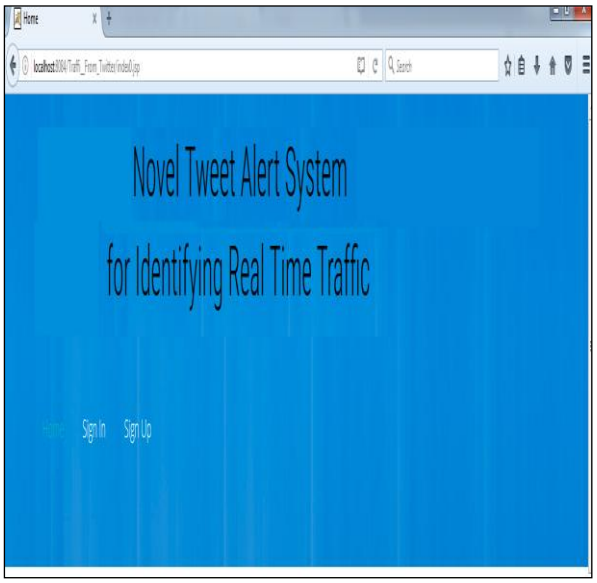

FIG: 6.1 SHOWS THE HOME PAGE

The above figure represents the home page of twitter alert system where we need to signup if we are a new user and login if we already exists as a twitter user.

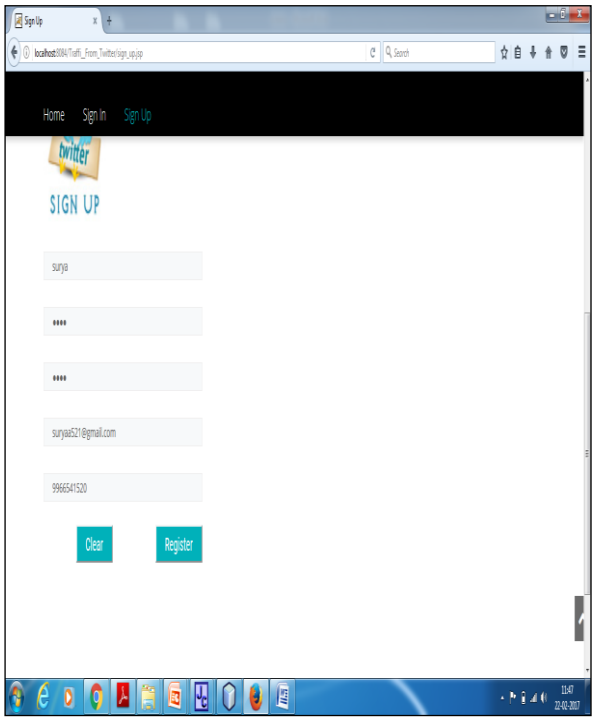

FIG: 6.2 SHOWS THE USER LOGIN PAGE

The below figure shows how friends are being followed and added into the twitter account for updates related to traffic.

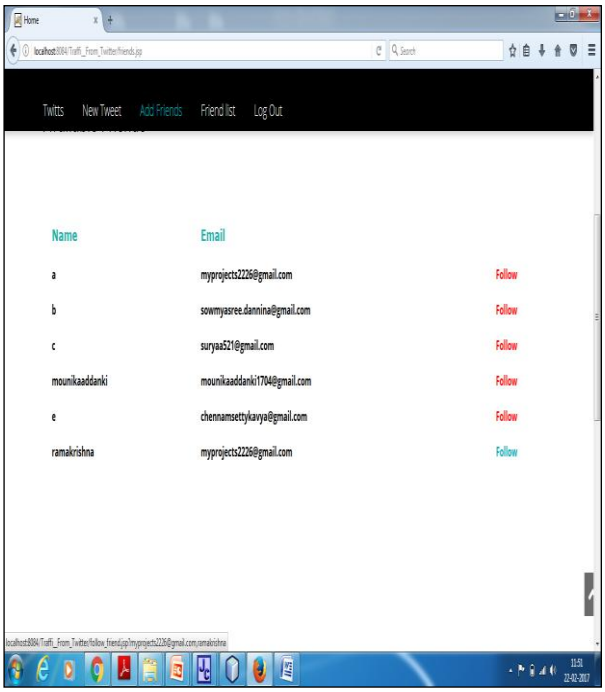

FIG:6.3 SHOWS THE FRIENDS ADDED INTO THE ACCOUNT

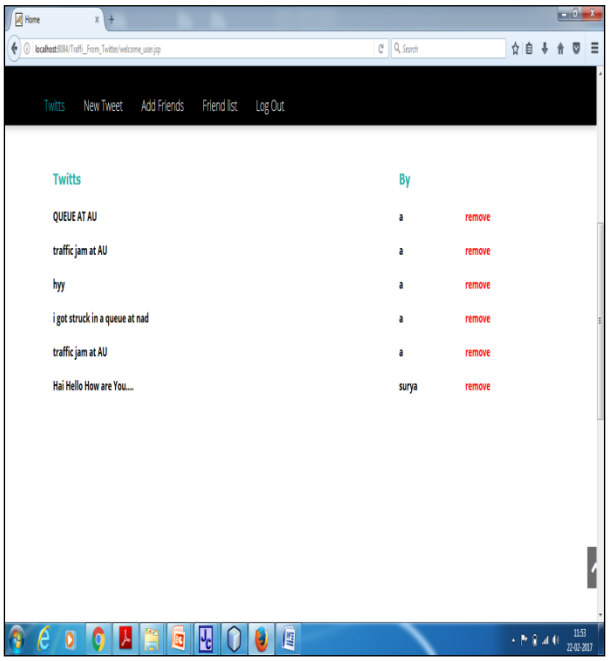

FIG: 6.4 SHOW THE FINAL TWEETS SHOWING THE ALERT MESSAGES OF TRAFFIC

Figure shows the final tweets and the result about final traffic alerts in the city where you can clearly find the areas with traffic problem being notified as an alert.

7. CONCLUSION

This paper, described for the first time a novel system for identifying the traffic using twitter stream analysis. This proposed system allows us to build a traffic alert system using K-Means algorithm and considering twitter as a main source for updating the status of traffic to each and every one who come in that region. Here the traffic event can be occurred based on different regions like cricket or football match, procession and manifestation, or not. By conducting various experiments on our proposed model we finally came to a conclusion that our proposed approach is best in providing security for the current society in terms of alerting when there is traffic near that way and avoiding huge maintenance for the end users who wish to travel in the same direction.

8. REFERENCES

1. A well known authors like B. Huberman, and F. Wang, "Social Networks that Matter: Twitter Under the Microscope.
2. A well known author A. Tumasjan, wrote a paper on Predicting Elections with Twitter:" Proc. Fourth Int'l AAAI Conf. Weblogs and Social Media (ICWSM), 2010.
3. M. Sarah a well known author who proposed a paper on "Twitter and the Micro-Messaging Revolution," 2008.
4. A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," Proc. Ninth WebKDD.
5. Two well known authors like H. Kwak and C. Lee, described about the "What is Twitter, A Social Network or A News Media.

6. A well known author like R.Ady proposed a paper on "Twitter as a Learning Tool. Really," ASTD Learning Circuits, p. 13, 2009.
7. K. Borau, and R. Shen, "Microblogging for Language Learning: Using Twitter to Train Communicative and Cultural Competence," Proc. Eighth Int'l Conf. Advances in Web Based Learning.
8. J. Hightower and G. Borriello, "Location Systems for Ubiquitous Computing," Computer, vol. 34, no. 8, pp. 57-66, 2001.
9. A. Gonzalez, L. M. Bergasa, and J. J. Yebes, "Text detection and recognition on traffic panels from street-level imagery using visual appearance," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 228–238,Feb. 2014.
10. N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, and P. Chaovalit, "Social-based traffic information extraction and classification,"in *Proc. 11th Int. Conf. ITST*, St. Petersburg, Russia, 2011,pp. 107–112.
11. P. M. d'Orey and M. Ferreira, "ITS for sustainable mobility: A survey on applications and impact assessment tools," *IEEE Trans. Intell. Transp.Syst.*, vol. 15, no. 2, pp. 477–493, Apr. 2014.
12. K. Boriboonsomsin, M. Barth, W. Zhu, and A. Vu, "Eco-routing navigation system based on multisource historical and real-time traffic information," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4,pp. 1694–1704, Dec. 2012.