Event Detection Based on Interactive Communication Streams in Social Network

Yadong Zhou
Ministry of Education Key Lab for Intelligent Networks and Network Security
Xi'an Jiaotong University
Xi'an, China
ydzhou@xjtu.edu.cn

Hong Xu
Ministry of Education Key Lab for Intelligent Networks and Network Security
Xi'an Jiaotong University
Xi'an, China
rushant@stu.xjtu.edu.cn

Lei Lei
Xi'an University of Technology
Xi'an, China
leilei@xaut.edu.cn

ABSTRACT
With the rapid development of social network, users are used to discuss and plan activities with their online friends by the form of interactive communication streams in mobile social networks. The information of these activities can be applied to track the prospective behavior and following demand of users for smart service supporting systems. In this paper, we describe a prospective activity as an event that happens at scheduled location and time, and propose a method to detect the event. The data of interactive communication streams are divided into five types, including *request*, *question*, *confirmation*, *denying* and *uncertainty*. We employ a combined multi-classifier based on D-S evidence theory to classify interactive streams into the five types, and extract the information of event, location and time in each text. Then a reasoning model is proposed to deduce the user's final intention of prospective activity through the series of different types of interactive streams. Based on the real data collected from social network, the experimental results show that our method could detect the information of events effectively.

CCS Concepts
• **Information systems → Data stream mining.**

Keywords
event detection; interactive stream; combined classifier; reasoning model; social network.

1. INTRODUCTION
With the rapid development of social network, e-commerce and mobile Internet, a large number of interactive dynamic texts expand in a geometric rate. As a result, a large amount of data is spreading over social networks, which provide important clues about specific situations. Detecting events over social streams has many applications such as smart service support, scheduling management and decision making [1]. How to efficiently access the integrated information resource through information fusion has become the research focus [2]. In recent years, micro-blogging swept the world at an alarming rate, and have produced large amounts of texts with short length and different data structure, which contains much useful information and allows users to share their future plans to attend events or go to specific locations [3].

At present, effective method of event detection based on interactive communication streams is still relatively challengeable. Mining such social streams is more challenging than traditional text streams, because of the existence of both text content and implicit network structure within the stream [4]. At the same time, similar to micro-blogging, the reply messages from forum, feedbacks, SMS, instant messaging and e-mail have less content ranged from a few dozen words to a hundred words or so. Compared to the feature selection of long texts, the main problem of short texts feature selection is: in short text the feature space is sparse and difficult to exploit the correlation between features, and different features impact on the classification results much different [5].

It is not easy to accomplish the goal of mining and detecting through interactive streams in social network due to the following challenges: 1) Processing stream texts of social network. Stream texts of social network present the characters of singularity, sparse feature and non-standard expression. The text is limited-length short message, which contributes to the serious data sparseness problem. The texts contain amount of colloquial speech and lots of omissions and coreference. This increase the difficulty when processing social network text. 2) Processing stream texts in different domains. In different domain, users have different linguistic habits when communicate with others in social network. For example, in travel forums, users may talk freely, but in science and technology forums, users talk seriously. It may make the size of feature space too large when handling the texts. 3) Detecting prospective behavior of individual user. It is difficult to extract valuable information to detect user's prospective behavior and demand in the immediate future. Most researches on user behavior focus on group behavior analysis in this area. In the aspect of individual behavior detection, to the best of our knowledge, there is not efficient method.

In this paper, we define an event as something that happens at a scheduled location and time. We propose an intuitive method to determine the classification labels of stream texts and feature sets with a focus on users' intention in interactive social network. We classify stream texts into five types, including *request* (R), *question* (Q), *confirmation* (C), *denying* (D), *uncertainty* (U), and employ a combined multi-classifier based on D-S evidence theory to improve classification accuracy. After labeling stream texts, we get logical relations between the user and reviewers. The replies logically relevant to the thread title of the master's message are extracted from all the replies and organized into conversation. A reasoning model based on stream text type and conversation logistic structure is proposed to detect the event that will happen in users' schedule. The detection results return as a vector containing location, time and activity in an acceptable accuracy. The main contributions of this work are summarized as follows:

MOBIMEDIA 2016, June 18-19, Xi'an, People's Republic of China

DOI 10.4108/eai.18-6-2016.2264140

- We propose a framework to extract users' prospective activities from interactive communication streams. These valuable information provides some basis for other studies.
- We introduce a multi-classifier combined with D-S theory to improve the accuracy of determining sentence type.
- We conduct experiments over stream dataset collected from Sina micro-blogging site. The experimental results demonstrate the effectiveness of our approach.

The rest of this paper is organized as follows. In section 2, the architecture of the event detection system is illustrated and the detailed method of social network data processing is presented. In section 3, the combination of multi-classifiers combined with D-S evidence theory is put forward to reduce misclassification error. In section 4, the reasoning method is designed and explained in detail. The experiment results are shown and analyzed in section 5. A brief conclusion is presented in section 6.

2. METHOD FRAMEWORK

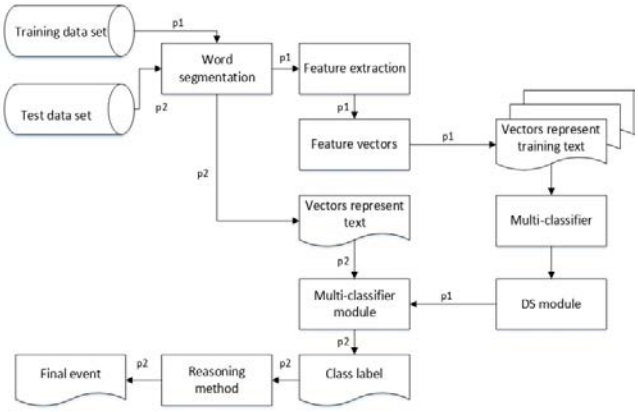

Figure 1. The framework of event detection

The architecture of the proposed method is present in Fig.1. P1 denotes the training process and P2 denotes the test process. The training data is labeled with R, Q, C, D or U. Firstly, we use the word segmentation method to convert text content from social network to vector space. Then we select features from each types of sentence. The selected feature vectors are applied to represent the original text content. In order to obtain a proper classifier model, we combine these five features into one dimensional feature vector. A multi-classifier model combined by D-S evidence theory is proposed to improve the classification accuracy. Then, the sentences are classified to the five types by the multi-classifier model. In this way, we get the relationships between each pair of sentences that from one conversation. Finally, the final scheduled events are detected by a proposed reasoning method.

3. STREAM TEXT PROCESSING AND CLASSIFICATION

3.1 Feature Extraction and Weight Calculation

3.1.1 Feature extraction

Firstly, we use Chinese words segmentation method to accomplish named entity recognition and part-of-speech tagging. For more details about the Chinese words segmentation method, please visit the website 'https://github.com/hankcs/HanLP'. While tagging the training data set, we build four sentiment word dictionaries. These four dictionaries also have their own part-of-speech tags which represent *question*, *confirmation*, *denying*, *uncertainty* respectively. Considering words with same part-of-speech tag are different from others in other type of sentences, we select their part-of-speech as feature set. In this way we can reduce the feature space dimension, time complexity and improve the accuracy.

We propose an intuitive approach to determine the set of features with focusing on user intention in social network, such as asking for travel routes or recruiting travelers. For type *request*, master lists his trip plan contains time, location and current condition. some of his/her followers may want to join the trip. So reply of followers can contain personal contact information and some other colloquial terms to express their agreement. Some followers may have question about the master's trip plan. These followers' contents contain interrogative pronoun about time, address or other things. The reply contents of followers who don't want to join would contain some emotional words. The reply contents of followers who are not sure whether to join the trip would contain uncertain terms like *think it over*, *give attention* and so on. These colloquial terms and special words mentioned above are collected when labeling training data and consolidated into four dictionaries with a same part-of-speech respectively.

After feature selection for each types, we combine the five feature vectors into one-dimensional feature vector. The process is described as following.

$$F_R = \{ f_{R_1}, f_{R_2}, \cdots, f_{R_l} \} \quad (1)$$

$$F_Q = \{ f_{Q_1}, f_{Q_2}, \cdots, f_{Q_j} \} \quad (2)$$

$$F_C = \{ f_{C_1}, f_{C_2}, \cdots, f_{C_k} \} \quad (3)$$

$$F_D = \{ f_{D_1}, f_{D_2}, \cdots, f_{D_l} \} \quad (4)$$

$$F_U = \{ f_{U_1}, f_{U_2}, \cdots, f_{U_m} \} \quad (5)$$

F_R, F_Q, F_C, F_D, F_U represent the feature vector of each type sentence. For multi-classifier, the feature vector F is their serial combination $F = \{ f_{R_1}, f_{R_2}, \cdots, f_{U_{m-1}}, f_{U_m} \}$.

3.1.2 Weight calculation

Given a sentence *S*, after word segmentation system, the sentence *S* is shown by part of speech. We use TF-IDF algorithm [6] to compute the weight of each element in *F*. Then the sentence *S* is shown by space vector model just like numeric vector.

$$S_i = \{ w_{i_1}, w_{i_2}, \cdots\cdots, w_{i_n} \} \quad (6)$$

W_i means the weight of the *i*th feature item calculated by TF-IDF algorithm and *n* denotes the total number of features.

3.2 Sentence Classification and D-S Theory

3.2.1 Multi-classifier combination

Due to the characters of singularity， sparse feature and non-standard expression in social network text, we select multiple classifier, including SVM, Logistic and Naïve Bayes, to reduce the misclassification error. Each classifier assigns a probability to each type and outputs the result according to the highest probability.

We notice that the probability distribution of each classifier sometimes differs from each other on the five types. They classify the same sentence into different types in some case. So in order to deal with this case reasonably and reduce the error rates, D-S evidence theory is employed.

3.2.2 Dempster-Shafer (D-S) evidence theory

This session can be seen as a problem of information fusion and probabilistic reasoning. As different classifiers have different classification results facing same sentence. When a single classifier classifies the sentence, it assigns BPA values on five types. Then the classifier simply selects the type with the highest BPA value among the five types. D-S theory coordinates different evidence and uses incomplete factors synthetically to reduce the redundancy and conflict and get consistency description of things. In this section, we select classifiers of SVM, Naïve Bayes, and Logistic to classify the same sentence and fuse their BPA values to get the final result. This effect can't be attained through any single classifier. For more information about D-S theory, please refer to [9, 10]. The utilization process of D-S evidence theory is shown in Fig. 2.

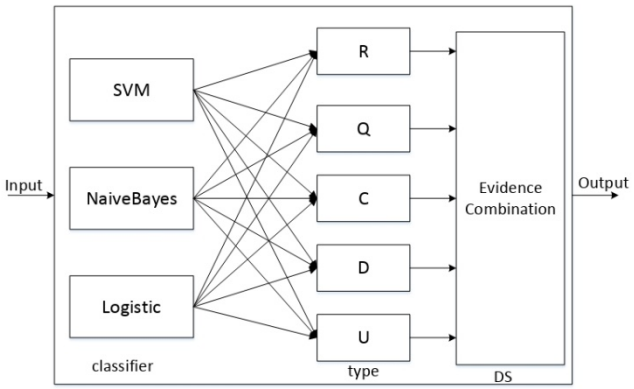

Figure 2. The utilization process of D-S theory

4. REASONING MODEL

In this paper, experimental data is collected from Sina micro-blogging. Fig. 3 is a simple example of user's Sina micro-blogging messages (original Chinese content have been translated into English).

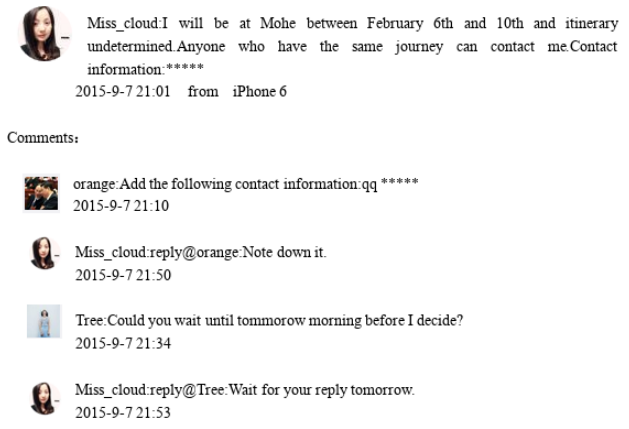

Figure 3. An example of social network stream text

In Fig. 3, there are two threads: Miss_cloud and orange, Miss_cloud and Tree. We organize the original contents into the form of two conversations between user and respondents. For example, in the conversation between user Miss_cloud and respondent orange, we obtain the request type of user's content and the confirmation type of orange content in a certain accuracy by means of method introduced in section 3. Then we extract the logical relationship between user and commenter in content. The process is same to other logical conversations that extracted from user's Sina micro-blogging messages. In order to get the final event automatically, we take advantage of reasoning method to simulate the human's logical thinking processes. We treat the combination of two types as the minimum unit. These minimum units are shown in Table 1.

Table 1. Minimum units

Type	R	Q	C	D	U
R	RR(↓)	RQ(↓)	RC(1)	RD(1)	RU(1)
Q	QR(↓)	QQ(↓)	QC(1)	QD(1)	QU(1)
C	CR(↓)	CQ(↓)	CC(1)	CD(1)	CU(1)
D	DR(↓)	DQ(↓)	DC(1)	DD(1)	DU(1)
U	UR(↓)	UQ(↓)	UC(1)	UD(1)	UU(1)

As shown above, the sign "1" denotes result is in current unit, and the sign " ↓ " denotes current unit needs further reasoning.

Assuming the conversation start with types of request or question and end with types of confirmation, denying or uncertainty. When the length of conversation is two, the number of reasoning rules is $C_{10}^{1} \cdot C_{15}^{1} = 150$. When the length of conversation is four, the number of rules will reach up to $C_{10}^{1} \cdot C_{25}^{1} \cdot C_{25}^{1} \cdot C_{15}^{1} = 93750$. Considering the length of conversation can't be predicted and rule stocks of reasoning rules would be very large, we cut the conversation into pairs of combined sentences which are close to each other. For example, in a conversation, it contains several sentences $S_1, S_2 \cdots S_n$, we organize these sentences into child dialogs just like following sequence $[S_1,S_2]$, $[S_2,S_3]$, $\cdots$ $[S_{n-1},S_n]$.

5. EXPERIMENT RESULT AND ANALYSIS

The experiment data is composed of 7,596 micro-bloggings messages from 2,046 users. These micro-bloggings messages are manually labeled with the best matching type without considering their contextual information. We choose 3189 micro-blogging messages as training data and the rest as test data. The part-of-speech tags select as features are shown and explained in Table 2.

Table 2. Part-of-speech tags

Tags	Explanation
ns	Location
nt	Organization
t	Time
qt	Time measure word
ry	Interrogative pronouns

ryt	Time interrogative pronoun
rys	Place interrogative pronoun
ryv	Predicate interrogative pronouns
m	Number
con	Defined in confirmation dictionary
den	Defined in denying dictionary
un	Defined in uncertainty dictionary

Table 3. Comparison of classification results

Classifier	Accuracy
SVM	80.01%
Logistic	79.86%
Naïve Bayes	79.07%
Combined classifier	82.03%

Based on the features in Table 2, comparison of classification results for sentence types is present in Table 3. It is observed that the accuracy of combined classifier based on D-S evidence theory performs much better than other classifiers. The main reason of misclassification is that the features of different types appear in the same sentence in some special circumstances. For instance, the sentence 'Good plan, take me with you. How can I get in touch with you?'. The first half of the sentence denotes the type of confirmation, and the last half denotes question. It is difficult to determine the type of sentence in such a situation by the classifiers. Other important reasons of misclassification include that the features in messages are sparse after Chinese text segmentation and removing stop words, and appearance of colloquial words in messages also disturbs the accuracy of classification.

Finally, we extract 190 samples to check the accuracy of event detection after classification of sentence type. The accuracy of event detection by the reasoning model mentioned in section 4 is 74.72%. Through analyzing the reasons of incorrect detection of prospective events, we find that some of users usually mention relative descriptions of time or location more than once in a micro-blogging message. They would declare both the location they will visit and the location where they won't visit. Whether we regard this sentence as request type of confirmation or denying, we will extract different information of event from it. These comprehensive factors lead the accuracy of final event detection be lower than the accuracy of classification of sentence type.

6. CONCLUSION

In this paper we propose a method to identify the type of sentences and introduce D-S evidence theory to improve the classification accuracy. Then we put forward a reasoning model to get the information of final event from users' interactive messages automatically. We take five sentence types into account in our experiments, and the methods are also applicable for other interactive social network circumstance like SMS and WeChat.

Experimental results show that our method performs better than single classifiers. However, the features in sentences are too sparse to be extract in some circumstances and respondents express their suggestions casually most of the time. In the future we will study circumstances that a sentence is of two types or more. We will also improve our dictionaries and extract more effective features to release feature sparsity to improve classification accuracy.

The method proposed in this paper is helpful to set up a smart home system through interactive communication streams in social network. In the system, users' prospective behavior can be predicted through their interactive messages in social network and predict their demands automatically. The system could improve the comfort and convince of users without manual control.

7. ACKNOWLEDGMENTS

The research presented in this paper is supported in part by the National Natural Science Foundation (61572397, 61502383, 61375040, 61571360), Fundamental Research Project of Natural Science in Shaanxi Province (2015JM6298, 2015JM6299), Specialized Research Plan Funded Project of Shaanxi Province Department of Education (15JK1505) and Specialized Research Fund for the Doctoral Program of Higher Education (20120201120023).

8. REFERENCES

[1] Zhou X, Chen L. 2014.Event detection over twitter social media streams. *The VLDB journal*. 23(3): 381-400.

[2] Sun Q, Wang Q, Qiao H. 2009.The algorithm of short message hot topic detection based on feature association. *Information Technology Journal*. 8(2): 236-240.

[3] Shen Y, Tian C, Li S, et al. 2009.The grand information flows in micro-blog. *Journal of Information & Computational Science*. 6(2): 683-690.

[4] Aggarwal C C, Subbian K. 2012.Event Detection in Social Streams. *SDM*. 12: 624-635.

[5] Liu, Z., Yu, W., Chen, W., Wang, S., & Wu, F. 2010. Short text feature selection for micro-blog mining. In *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on* (pp. 1-4). IEEE.

[6] Zadeh, L. A. 1986. A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination. *AI magazine*, 7(2), 85.

[7] Wu H C, Luk R W P, Wong K F, et al. 2008. Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems*. 26(3):55-59.

[8] Yang Y, Pedersen J O. 1997. A comparative study on feature selection in text categorization. *ICML*. 97: 412-420.

[9] Wu, H., Siegel, M., Stiefelhagen, R., & Yang, J.2002. Sensor fusion using Dempster-Shafer theory [for context-aware HCI]. In *Instrumentation and Measurement Technology Conference, 2002. IMTC/2002. Proceedings of the 19th IEEE* (Vol. 1, pp. 7-12). IEEE.

[10] Shafer, G. 1992. The Dempster-Shafer theory. Encyclopedia of Artificial Intelligence, SC Shapiro.

[11] Abdelhaq, H., Sengstock, C., & Gertz, M. 2013. Eventweet: Online localized event detection from twitter. Proceedings of the VLDB Endowment,6(12), 1326-1329.

[12] Liu T, Chen S, Liu Y, et al. 2014.SHE: smart home energy management system for appliance identification and personalized scheduling. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*: Adjunct Publication. ACM: 247-250.