Computers in Human Behavior xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Feature selection for event discovery in social media: A comparative study

Jie Zhao [a], Xueya Wang [a], Peiquan Jin [b,*]

[a] School of Business, Anhui University, 230039 Hefei, China
[b] School of Computer Science and Technology, University of Science and Technology of China, 230027 Hefei, China

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Microblog
Event discovery
Feature selection
Algorithm comparison

ABSTRACT

Microblog as one kind of typical social media has many research implications in social event discovery and social-media-based e-learning and collaborative learning. At present, researchers usually employ feature-based classification approaches to detect social events in microblogs. However, it is very common to get different results when different features are used in event discovery. Therefore, it has been a critical issue how to select appropriate features for event discovery in microblogs. In this paper, we analyze five different feature selection methods and present an improved method for selecting features for microblog-based event discovery. We compare all the methods on a real microblog dataset in terms of various metrics including precision, recall, and F-measure. And finally we discuss the best feature selection method for the event discovery in microblogs. To the best of our knowledge, there are no such comparative studies on feature selection for event discovery in social media, and this paper is expected to offer some useful references for the future research and applications on the event discovery in microblogs.

1. Introduction

Microblogs have been one of the most important information sources for people to get news information and interact with social media. Recently, researchers begin to study the influence of microblog as well as other types of social media in e-learning and collaborative leaning (Dascalua, Bodea, Lytras, Ordoñez de Pablos, & Burlacua, 2014; Lytras & Ordóñez de Pablos, 2011; Tolosa, Labra Gayo, Martínez Prieto, Méndez Núñez, & Ordóñez de Pablos, 2010). Their studies have shown that social media can be a knowledge base that consists of questions and answers, as well as new information such as fresh social events. A possible solution for social media collaborative learning is to build a system that allows users to post questions on it and automatically extract answers from social media platforms. Here, the key issue to extract answers, which is regarded as topic detection or event discovery in many previous studies (Sakaki, Okazaki, & Matsuo, 2010; Zheng, Jin, Zhao, & Yue, 2014). In this paper, following this background, we focus on event discovery in social media (e.g., microblog), and aim to study one of the fundamental feature-selection problem in this issue.

Microblog has been a research focus in recent years. The rapid development of microblogging services also makes it more complex and difficult to manage public opinions in Internet, due to the opening and real-time features of microblogs (Zheng et al., 2014). Microblogs spread very fast in Internet without any mechanisms in information auditing and credibility identification. This is one of the critical problems that lead to many public emergencies in recent years. For example, as reported in Xie (2012), 22 events among the total 138 public events of China in 2011, occupying about 16 percent, are first initialized from microblogs. Microblogs have dramatically changed the traditional information spreading ways. Both governments and enterprises are facing a new critical problem in microblog age, which is how to adapt themselves to the development of microblogs. Specially, as many public emergencies originate from microblogs, governments and enterprises need some effective ways to detect events from microblogs so that they can conduct efficient counter-behaviors to cope with the influence of microblogs.

Previous studies have shown that public emergent events have very different ways when spreading on microblogging platforms (Zhao, Jin, & Huang, 2011). Therefore, there are some existing works focusing on analyzing the new features of microblogs and further presenting new approaches for event discovery from microblogs. Basically, these works can be classified into three types: specific event discovery, specific person's detection, and

* Corresponding author.
E-mail address: jpq@ustc.edu.cn (P. Jin).

http://dx.doi.org/10.1016/j.chb.2014.11.007
0747-5632/

integrated event discovery. For specific event discovery (Sakaki et al., 2010), people usually use some specific keywords in microblogs such as earthquake and tornado to find new events. As microblog users can post information very quickly when these events happen, it is possible to detect such events timely. For specific person's detection (Popescu & Pennacchiotti, 2010; Popescu, Pennacchiotti, & Paranjpe, 2011), people use the name of a specific person to search his or her relevant events in microblogs. For integrated event discovery, a very popular method is to detect those hot words in microblogs (Long, Wang, Chen, Jin, & Yu, 2011). As location information is important in the Web (Zhao, Jin, Zhang, & Wen, 2014), some researchers also utilize the location information embedded in microblogs, i.e., Geo-tags, to find local events from microblogs (Lee, Wakamiya, & Sumiya, 2011).

Most studies in event discovery from microblogs employ a feature-based classification method such as SVM (Support Vector Machine) and Naïve Bayes Model. A classification method discovers some rules from certain training sets with knows classes, and then use these rules to predict new classes. Classification as a common machine-learning approach has been widely used in many areas including document classification, information retrieval, intrusion detection, object recognition, information extraction, etc. However, the performance of a classification method is much related with the selected features. Thus, how to effectively select features for event discovery from microblogs has been a research focus in recent years.

In this paper, we concentrate on the feature selection methods for microblog-based event discovery. Particularly, we employ an experimental study on real microblog data to compare the different performance of various feature selection methods and aim to find the best one. In summary, we make the following contributions in this paper:

(1) We analyze five existing methods for feature selection and present an improved method, which is called *Limited Feature Selection* method.
(2) We conduct comprehensive experiments on real microblog data sets and compare the performance of the existing five methods of feature selection and our *Limited Feature Selection* method. Based on the experimental results, we propose some suggestions for choosing feature selection methods in microblog-based event discovery.

The remainder of the paper is organized as follows. In Section 2, we survey the related work. In Section 3 we discuss the feature selection methods for microblog-based event discovery. In Section 4, the comparative experiments and the performance evaluation results are discussed. Finally, Section 5 concludes the paper.

2. Related work

2.1. Microblog-based event discovery

Microblog and its related researches have been a hot topic in recent years, due to the large number of users and the large amount of data in microblogging platforms. Consequently, microblogging platforms are becoming new social interaction places in the Web age.

Compared with traditional Web systems, microblogging platforms have the following unique features:

(1) Microblogs are real-time information. This is mainly because users can post new short microblogs about a certain event easily and conveniently if they are at the event place and happen to know the event. On the contrary, Web pages have to be carefully prepared and reviewed which is more time-consuming. Therefore, the events detected from microblogs are more fresh and valuable for users than traditional Web systems.
(2) Microblogs usually contain more new events than traditional Web pages because of the large number of microblogging users living in a wide geographical range.
(3) Microblogs contain more social-interaction features than traditional Web pages. Thus it is possible to detect more specific events from microblogs, such as events about a specific domain and events in a specific place.

Detecting events from microblogs is somehow similar with some previous work on TDT (Topic Detection and Tracking) (Zhao, Li, & Jin, 2012). These works on TDT typically employ some machine learning models, which are not suitable for microblogging platforms. This is because microblogs are very short (<140 characters) and machine learning models perform poorly on short texts (Zhao et al., 2012). There are also some existing works focusing on microblog-based event discovery (Long et al., 2011; Popescu et al., 2011; Sakaki et al., 2010). However, they only use a term-frequency-based approach, which simply use the hot words to represent events (Long et al., 2011). This approach ignores the other features of events, such as the happening time of events and places of events, and thus cannot form completed description about events. For instance, if we simply use a hot word "Olympic Game" and do not consider other information such as the year of an Olympic Game, we are not able to get the exact and detailed information about Olympic Game. On the other hand, many applications in emergency treatment and public security management have urgent needs on obtaining the details about events.

2.2. Event classification

Presently, microblog-based event discovery is commonly based on some classification models. A classification model aims to determine the exact class of a given object given that there are some pre-defined classes. The input of a classification task is a set of records, each of which is a tuple (x, y). Here, x is a feature set, and y is a set of target classes. A classification model is to obtain a target function f mapping x to y through a learning stage (Field, 2005). Classification models usually utilize some learning algorithms to determine a final function that can describe the relationship between x and y perfectly. This function is not only used to explain the relationship between x and y in the given data set, but also used to predict the possible y of a new x input (Tan, Steinbach, & Kumar, 2006). Thus, we usually divide the entire data set into a training set and a testing set. The training set is used to determine the right function f between x and y, and the testing set is used for performance evaluation.

When dividing the entire data set into a training set and a testing set, if the features of these two sets vary a lot, we can imply that the function learned from the training set will not suit for the testing set. Consequently, we will get poor performance. Therefore, in real applications we usually use cross-validation to measure the performance of a classification model. Cross-validation is a popular method in the areas of statistics and machine learning (Kohavi, 1995). In practical experiments, the most commonly used cross-validation method is the *k-Fold Cross-Validation*. According to this method, the entire data set is partitioned into k groups. Each group is taken as a testing set and the remaining groups are used as the training set. Thus, we will get k functions as well as k precision results. Finally, the mean of the k results is used as the final precision metric to evaluate the overall performance of the classification model.

Please cite this article in press as: Zhao, J., et al. Feature selection for event discovery in social media: A comparative study. *Computers in Human Behavior* (2014), http://dx.doi.org/10.1016/j.chb.2014.11.007