International Journal of Hybrid Information Technology
Vol.7, No.1 (2014), pp.201-210
http://dx.doi.org/10.14257/ijhit.2014.7.1.16

Towards Events Detection from Microblog Messages

Jie Zhao, Xueya Wang and Zheng Ma
School of Business, Anhui University
zj_teacher@126.com

Abstract

Microblogs have dramatically changed the mechanism of information propagation. It has been an inevitable issue for governments and enterprises to face the challenges that microblogs bring to public safety management. Since 2010, a lot of emergent events were firstly reported in microblogs, and this trend is even becoming more and more prominent. As microblogs have shown much influence in public safety and sentiment management, governments and enterprises have to employ effective ways to monitor and predict emergent events in microblogs; otherwise when emergent events in microblogs happen, they will be in a passive situation and even the society's stability will be affected. Based on the properties of microblog information and its diffusion, this paper presents a survey on the events detection from microblogs, and recent advances on some key related issues are especially focused and discussed. Finally, we give a framework for events detection on microblog platforms and its key issues are discussed, which are expected to bring valuable references for the researches in microblog information extraction.

Keywords: *microblog, event detection, survey*

1. Introduction

Microblog has been one of the important media for web users to express their opinions and spread interested information on the Internet. According to a recent report posted by the Chinese Network Information Center, the total number of Chinese microblog users has been increased to over 250 million, and the increasing rate is highly over 290% [1]. The opening and real-time features of microblog bring new challenges in the network information management. As a result, a lot of public events have been initialized on the microblog platforms and finally become social hot spots. For example, as reported in the literature [2], in 2010 there are 22 events originally posted in microblogs, which occupies 16% of the whole 138 events. Another study showed that in 2011 this percentage was extremely increased into 36% [3], which means 36% of the whole hot events were originally reported on microblog platforms. As a consequence, microblog has thoroughly changed the traditional information spreading way. How to suit for the special properties of microblog has been an urgent and important issue for both government and business.

In this paper, we give a survey on events detection from microblog data stream. We emphasize the traditional event detection methods as well as the state-of-the-art approaches in event finding on microblog platforms. And on the basis of the survey, we finally present some future research directions towards this area.

The remainder of the paper is structured as follows. Section 2 analyzes the properties of microblog, as well as the challenges in microblog events detection. In Section 3, we discuss the traditional methods for event detection from web. Section 4 explores the framework for events detection from microblog messages, and the conclusions are in Section 6.

ISSN: 1738-9968 IJHIT

2. Challenges of Events Detections on Microblog Platforms

Microblog is one of the new social network platforms boosted in recent years. Generally, microblog has the following properties:

(1) There are a great number of microblog users in the Internet. This number, compared with other types of social communication platforms, is rather a huge one. For example, Twitter has over 100 million users and so do other microblog platforms such as Sina Weibo and Tecent Weibo in China. Those users can generate a large amount of information every day on the microblog platform.

(2) Microblog messages contain rich social network information. This is much different from other types of information, which only present their content. On the contrast, microblog users are connected each other on the microblog platforms via following, reviewing, and reposting behaviors.

(3) Microblog messages are usually very fresh as many users will post microblog message at the first time when they see or meet some special events. Another reason for the real-time property of microblog is that nowadays smart phones are very popular for people to post microblogs in time.

The special characteristics of microblog bring new opportunities and challenges for event detection on the microblog platforms. Those opportunities can be summarized as follows:

(1) As microblogs have the real-time property, events detected from microblog messages are usually more fresh and useful than those from traditional information sources such as web pages. Users are enabled to post event information at a very early stage when the event occurs. This behavior is usually faster than other media such as news portal website or newspaper.

(2) As there are a great number of active users on the microblog platform, who may distribute among a large scope of geographical area, we are able to detect more useful events from microblogs. On the contrast, traditional events detection in new portal websites depends on the engagement of news workers.

(3) As microblog has the social network property, we are able to detect some special events from microblogs, such as events in a specific domain, events in a specific group, or events in a specific area.

3. Events Detection from Web

3.1 Emergent Events Detection from Web Pages

With the rapid development of web technologies, web information has shown a dramatic increasing on its data volume and data creation speed. How to utilize web information to detect burst events has been a hot topic in recent years [9]. It has been validated that people can get more predictive information from web than from traditional ways. And a lot of real systems have been designed according to this idea, including GPHIN [10], EMM [11], and InSTEDD [12].

GPHIN (Global Public Health Intelligence Network) was first initialized by the Canada Ministry of Hygiene [10]. GPHIN is well-known because of its prediction on the SARS event in 2003. At that time, GPHIN detected that there were surprisingly increasing of respiratory disease related queries on the web, and they finally reported this fact to the World Hygiene Organization (WHO). GPHIN employs many information technologies to detect and track

events related with hygiene. Those technologies include web search technologies, data mining, automatic translation, information filtering, and so on. The typical working process of GPHIN is semi-automatic which consists of an automatic machine processing stage and a manual evaluating and analyzing stage. Particularly, it first obtains information from the Internet based on some pre-defined search queries and performs filtering and refinement. After that, it delivers the filtered information to experts for further evaluation and analysis. At present, GPHIN is only focused on hygiene-related events.

EMM (Europe Media Monitor) was started by European Union in 2006 [11]. Its working process is similar to GPHIN, except that it can monitor different types of media information including terrorist, forest fire, hygiene provenance, and so on. The major technologies used in EMM are information extraction and exploratory search. In detail, it first extracts information elements from web pages and clusters them into topics, and then presents them in a news exploring framework. Its major strong point is that it can support multiple types of languages such as English, Chinese, and French. Currently, it can support sixteen languages. Basically, EMM has not employed complicated algorithms in its implementation. The key technique in EMM is the keyword-based textual search, and it maintains about 1,600 keywords which are used to match the crawled web pages during the information extraction procedure.

InSTEDD (Innovative Support to Emergencies, Diseases and Disasters) was first suggested by the famous doctor Larry Brilliant in 2008 [12]. Its initial goal is to establish an information sharing and communication platform for early-warning information focusing on hygiene and diseases. InSTEDD was sponsored by Google.org, a non-profitable organization occupied by Google. Also it got technical support from GPHIN. The basic idea of InSTEDD is to offer a information communication network for doctors community so as to realize information interaction and resource sharing among the community. A simple example is as follows. A doctor in South Asia reported a very special heart disease on the platform, and other experts can quickly see this information and post their comments and give suggestion for possible treatment. Similar to GPHIN, InSTEDD is also towards hygiene area. As its technical support is mainly from GPHIN, the implementation of InSTEDD is also similar to GPHIN.

In summary, all the above mentioned systems have some limitations regarding the events detection issue. First, they all depended on the keyword-based textual search technology. Second, they all need manual interactions. And finally, they offer little support for Chinese information processing and lack the deep understanding and modeling for events in Chinese web pages.

3.2 Topic Detection and Tracking

Topic detection and tracking (TDT) is a new research topic originated from 1990s. TDT is heavily related with web information retrieval and web information extraction. Presently, TDT is mainly focused on news webpages and its main task is to automatically recognize the border of news reports, to detect and track news topics, and to perform multilingual analysis [4]. The tasks of TDT are formulated by several institutes including DARPA, CMU, and Dragon Systems. The researchers from those institutes also conducted early studies in this area [13].

Currently, most of TDT algorithms employ the clustering idea and different clustering methods are used, such as Single-Pass Clustering, KNN, K-Means, and Layered Clustering [13, 14]. A topic is usually represented as a centroid and organized using a layered structure. Although there have been a lot of works focusing on TDT, the effectiveness of TDT is still far to the requirements of real applications. For instance, the Layered Clustering based TDT algorithm has been demonstrated to have the best performance among the proposed algorithms, but previous studies showed that its

precision cannot satisfy the needs in real applications [15]. One practical system using the TDT technologies is Google News, in which the non-layered approaches for TDT are used. However, the error rate of Google News when detecting topics is over 50% [16].

Traditional TDT techniques cannot be applied directly for microblog messages. The major reason is due to the length limitation of microblog messages, *i.e.*, a microblog message cannot exceed 140 characters. Traditional TDT methods are generally based on some machine learning model which needs training on a corpus consisting long texts. However, the short messages on the microblog platform will result in a poor model.

3.3 Event Detections from Microblogs

Recently, with the popularity of microblog service, there are also some research works focusing on extracting events from microblog messages. Those works can be classified into three types, namely specific events detection, specific person related events detection, and general events detection. For the specific events detection [5, 6], researchers used some special keywords such as earthquake and storm to search microblog messages to find interested events. For the specific person related events detection [7, 17], previous studies were similar to the specific events detection, i.e., it issued special person names to detect person related events. For the general events detection, in the literature [8] the researchers proposed to find the hot keywords in microblogs and then to further detect events. In [18, 19], the TDT methods were introduced to microblog event detection. However, those methods all used textual words to represent events, and lacked a formal and precise description of events.

There are also some researches concentrating on microblog opinions extraction [20-23]. In general, it consists of two steps to detect opinions from microblogs [20]. The first step is to detect topic and recognize opinioned sentences, and the second step is to classify opinions into different polarities, *e.g.*, positive or negative. At the first step, we need to detect the specific topic in microblogs and associate the topic with opinioned sentences. In the past, researchers used machine learning methods or statistical approaches to analyze the sentiment information hiding in web information [21, 22], while on the microblog platform the traditional machine learning methods are often with a low precision in sentiment analysis. In addition, previous studies used hot keywords to identify events and topics [23], which could introduce inconsistence between events and opinioned sentences. For example, if we only use the keyword "Olympic Games" to identify events, we may associate the following sentences with the event, but those sentences reported different topics.

S1: Beijing 2008 Olympic Games was successful! (positive)

S2: Barcelona Olympic Games was awful. (negative)

In summary, there are some problems existing in current microblog events detection, which can be summarized as follows:

(1) A formal description of events on microblog platforms is lacked. Therefore, it is difficult to detect different types of events from microblogs through a general framework.

(2) Previous works focusing on specific events detection can reach a high precision, while no effective ways were proposed for general events detection. As

microblog messages are very short, traditional machine learning based methods usually fail to extract useful textual properties from microblog messages.

(3) Existing microblog opinions extraction only used hot words to identify events, which will introduce the inconsistence between events and opinioned sentences. Some works only considered the textual contents in microblog messages and neglected the social properties of microblog platforms. Generally, the reviewing and reposting behaviors of users also have rich sentimental information.

4. A Framework for Events Detections from Microblogs

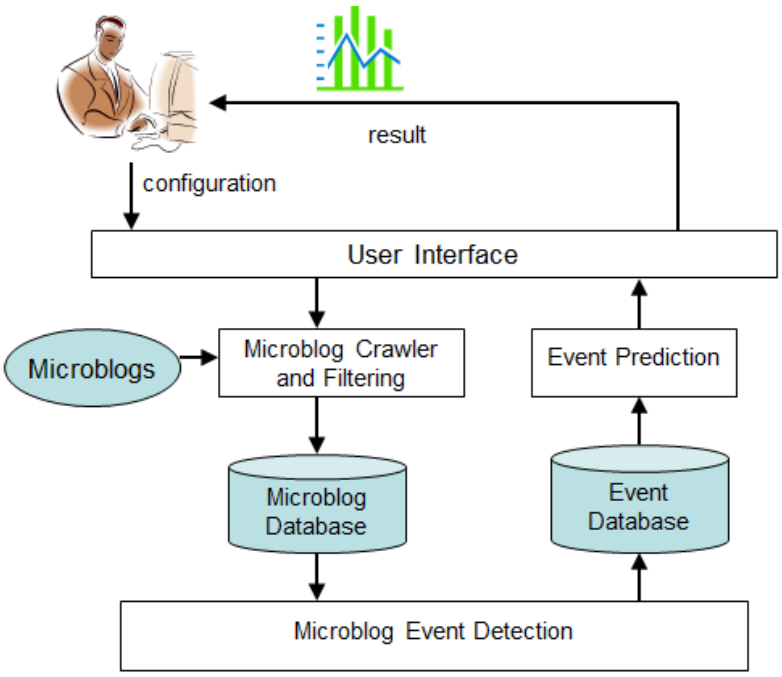

Figure 1. The framework for events detection from microblogs

Based on the analysis on the related work, we propose a framework for events detection and sentiment analysis from microblog messages, as shown in Figure 1. The framework mainly consists of three modules, namely microblog crawler and filtering, microblog events detection, and event prediction.

4.1 Microblog Crawler and Filtering

Microblogs can be crawled by the search API provided by microblog platforms. For example, Twitter as well Sina Weibo provides such APIs for users to get microblogs freely.

For the filtering of microblogs, we can employ a measuring algorithm to evaluate the quality of microblogs. Microblog quality evaluation can be done in terms of user aspect and content aspect. According to the user-oriented microblog quality evaluation, the user behaviors are used to tell whether a user is a spamming robot. According to the content-oriented viewpoint, the microblog messages are used in the evaluation process,

and different metrics can be applied for this purpose. Table 1 lists some basic metrics for the quality evaluation of microblog content.

Table 1. Metrics used to evaluate the quality of microblog messages

No.	Metric	Description
1	Posting Date	Microblog emphasizes on freshness, and outdated microblogs usually contain duplicated contents.
2	Part of Speech	Many microblogs contain only some modal words, which are useless for events detection.
3	Originality	Reposted microblogs are usually duplicated data.
4	Text Length	Long text may contain more information.
5	Hashtag	If a microblog message contains too many hashtags, we usually cannot understand its meaning.

4.2 Modeling Evolutional Events

Evolutional events consist of two kinds of information, which are event information and evolutional information. For this reason, we conduct an object-oriented approach in this paper to establish the representation of evolutional events. An evolutional event is modeled as a triple: $E = \{EID, AD, DD\}$, where *EID* is the identifier of the event, *AD (Attribute Descriptor)* describes the static properties of the event, *i.e.*, those properties that are not changing with time, *DD(Development Descriptor)* represents the dynamic evolutional properties of the event.

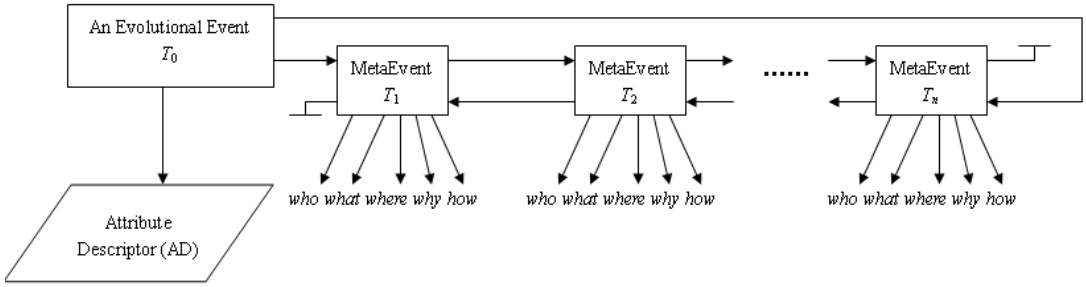

Figure 2. The model for evolutional events

According to this representation, the dynamic properties of an evolutional event can be regarded as the temporal changes of the event. On the other hand, most evolutional events are related with locations. So we can model an evolutional event as a spatiotemporal object, and construct a spatiotemporal model to represent the dynamic evolutional properties of the evolutional events.

The spatial dimension in evolutional events usually refers to a location in a large scale, *e.g.*, Beijing or Shanghai. Besides, the spatiotemporal changes in the evolutional event model are also different from those in traditional spatiotemporal data model. That is, traditional spatiotemporal changes generally refer to the split, merge, or shape change of a geographical object, while the evolution of event usually refers to the beginning, development, end, and influence of an event.

In this paper, we use an event-based spatiotemporal model to describe the evolutional properties of public emergencies. An event-based spatiotemporal model looks each spatiotemporal change as an event and constructs a series of events according to the time dimension to represent the whole spatiotemporal changing history of the object. In this paper, we define each evolutional event, which is detected from Web pages, as a Meta Event, and build the Meta Event List to represent the evolutional process of the event (as shown in Figure 2). Each meat event represents a specific state of the event. The spatial attribute is defined in the meta event using a where element. We also maintain the temporal relationship between a meta event with its previous and succeed state.

4.3 Microblog Events Detection

As described in Figure 2, we have to solve two key issues for events detection from microblog messages: meta event detection and evolution model formulation.

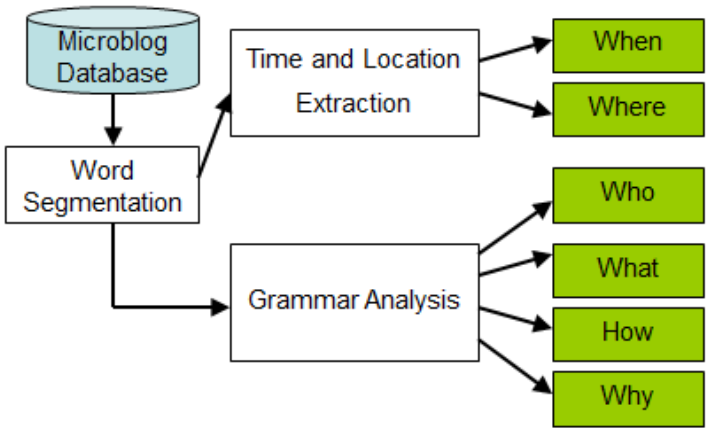

Figure 3. Extracting meta events

Figure 3 shows the process of extracting meta events. As microblogs are very short, we directly use the POS (Part of Speech) tagging and grammar analysis to extract the description of WHO/WHY/HOW/WHAT. For time and location extraction, we can use previous works on time and location extraction to obtain the information about WHEN/WHERE [24-27].

After extracting the meta events, we first cluster the meta events according to topic words, and further organize them into the evolutional model shown in Figure 2. The critical issue is how to cluster the meta events. We propose a topic gene based algorithm for this purpose. The algorithm is shown in Figure 4. A topic gene refers to the WHO/WHAT/WHERE/WHEN elements in meta events. We found that WHO/WHAT is more important in describing the topics of events. Moreover, the WHO/WHAT words are usually nouns, verbs, adjectives, time words, and numbers. Therefore, we represent the meta events in terms of topic genes, and then compute the cosine similarity among meta events to get the final clusters.

Algorithm: *EventDetection*

Input: the set of meta events, *M*

Output: a set of clustered events, $E = [E1, E2, \ldots, En]$, in which $Ei = <P, Mi>$. *P* is the topic of the cluster, *Mi* is the set of clustered meta events.

Procedure:

For each Mi in M **do**

1. Determine the topic genes of *Mi*, i.e., WHO/WHAT/WHEN/WHERE
2. POS tagging for the topic genes
3. Merge topic genes and form the topic strings
4. Compute the weights of topic strings
5. Normalize the weights and output the topic strings for *Mi*

End For

Compute the cosine similarity between the meta events based on the topic strings

Classify meta events into $E = \{E1, E2, \ldots, En\}$, the meat events are ordered in *Ei*

Ei.P = The WHO and WHAT words in Ei with highest term frequency

Return *E*

End *Event Detection*

Figure 4. Events detection based meta events

4.3 Microblog Events Prediction

Event prediction aims to provide decision support for real applications. It has to solve two key problems: event selection and prediction model.

For event selection, we can classify all the events into three types, namely macro economical events, micro economical events, and social events. The macro economical events refer to those ones influencing macroeconomics, such as consumers' confidence and industry developing trend. The micro economical events are those related with specific business organizations, such as sales of cars, prices of merchants, and so on. The social events are those involving social community, such as educational events, touring, and so on.

For the prediction models, the regression model can be used to predict the future events. This is mainly because the prediction model has to be widely evaluated in real applications. Therefore, in real world we usually use some classical models in statistics.

5. Conclusions

Microblogging service has been a hot topic in recent years. In this paper, we analyzed the challenges of microblog events detection, and gave a survey on the recent advances on microblog events detection, based on which a framework for microblog events detection was proposed. The detailed architecture as well as the key modules was discussed in the paper.

In the future works, we will concentrate on the implementation of the proposed system, especially on the realization and performance evaluation of the key algorithms. Another work is to apply the system into some real business area to demonstrate its use in real applications.

Acknowledgements

This work is supported by the National Science Foundation of Anhui Province (no. 1208085MG117), the National Science Foundation of China (no. 71273010), and the Soft Science Research Program of Anhui Province (grant no. 11020503056).

References

[1] CNNIC, "The 29th Report of the Internet Development in China", http://www.cnnic.cn/dtygg/dtgg/201201/W020120116337628870651.pdf.
[2] Y. Xie, "Report on the Public Opinions and Crisis Management in China", Social Science Literature Press, (**2011**).
[3] IRI, "Report on the Network Public Index in China (3rd Quarter, 2011)", http://blog.sina.com.cn/s/blog_6da2050f0100ycwt.html.
[4] J. Makkonen, H. Ahonen-Myka and M. Salmenkivi, "Simple Semantics in Topic Detection and Tracking", Information Retrieval, vol. 7, no. 3-4, (**2004**), pp. 347-368.
[5] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors", In Proc. of WWW, (**2010)**.
[6] J. Huang and M. Iwaihara, "Realtime Social Sensing of Support Rate for Microblogging", In Proc. Of DASFAA, Hong Kong, (**2011**).
[7] A. M. Popescu, M. Pennacchiotti and D. A. Paranjpe, "Extracting Events and Event Descriptions from Twitter", In Proc. of WWW, India, (**2011**).
[8] R. Long, H. Wang, Y. Chen, *et al*., "Towards Effective Event Detection", Tracking and Summarization on Microblog Data", In Proc. of WAIM, Wuhan, (**2011**).
[9] J. Zhao and P. Jin, "Towards the Extraction of Intelligence about Competitor from the Web", Proc. of Second World Summit on the Knowledge Society (WSKS'09), LNAI 5736, Chania, Crete, Greece, (**2009**), pp. 118-127.
[10] E. Mykhalovskiy and L. Weir, "The Global Public Health Intelligence Network and Early Warning Outbreak Detection: a Canadian Contribution to Global Public Health", Canadian Journal on Public Health, vol. 97, no. 1, (**2006**).
[11] European Commission Joint Research Centre, Europe Media Monitor, http://emm.jrc.it/.
[12] InSTEDD, In: http://instedd.org/.
[13] J. Zhao, X. Li and P. Jin, "A Time-Enhanced Topic Clustering Approach for News Web Search", International Journal of Database Theory and Application, vol. 5, no. 4, (**2012**), pp. 1-10.
[14] M. Masaki, M. Takao and S. Isamu, "Topic Detection and Tracking for News Web Pages", In Proc. Of Web Intelligence (WI), **(2006)**.
[15] D. Trieschnigg and W. Kraaij, "TNO Hierarchical Topic Detection Report at TDT 2004", Topic Detection and Tracking Workshop Report, (**2004**).
[16] M. Connell, A. Feng, G. Kumaran, *et al*., "UMass at TDT 2004", Topic Detection and Tracking Workshop Report, (**2004**).
[17] A. M. Popescu and M. Pennacchiotti, "Detecting Controversial Events from Twitter", In Proc. of CIKM, Toronto, Ontario, Canada, (**2010**).
[18] K. Watanabe, M. Ochi, M. Okabe, *et al*., "Jasmine: A Real-time Local-event Detection System based on Geolocation Information Propagated to Microblogs", In Proc. of CIKM, Glasgow, UK, (**2011**).
[19] R. Lee, S. Wakamiya and K. Sumiya, "Discovery of Unusual Regional Social Activities using Geo-tagged Microblogs", In Proc. Of WWW, India, (**2011**).
[20] M. Tsytsarau and T. Palpanas, "Survey on Mining Subjective Data on the Web", Data Mining and Knowledge Discovery (DMKD), vol. 24, no. 3, (**2012**), pp. 478-514.
[21] J. Zhao, "Towards Privacy-Preserved Query Optimization on Microblog Data", International Journal of Hybrid Information Technology, vol. 5, no. 4, (**2012**), pp. 157-170.
[22] D. Choi and P. Kim, "Sentiment Analysis for Tracking Breaking Events: A Case Study on Twitter", Proc. Of ACIIDS, (**2013)**, pp. 285-294

[23] G. Li and F. Liu, "A Clustering-based Approach on Sentiment Analysis", In Proc. Of 2010 International Conference on Intelligent Systems and Knowledge Engineering (ISKE), IEEE CS press, (**2010**), pp. 331-337
[24] P. Jin, J. Lian, X. Zhao and S. Wan, "TISE: a Temporal Search Engine for Web Content", 2008 International Symposium on Intelligent Information Technology Application (IITA'08), IEEE CS, Press, Shanghai, China, vol.3, (**2008**), pp. 220-224
[25] P. Jin, X. Li, H. Chen and L. Yu, "CT-Rank: A Time-aware Ranking Algorithm for Web Search", Journal of Convergence Information Technology, vol. 5, no. 6, (**2010**), pp. 99-111.
[26] X. Zhao, P. Jin and L. Yue, "Automatic Temporal Expression Normalization with Reference Time Dynamic-Choosing", The 23rd International Conference on Computational Linguistics (COLING), (**2010**), pp. 1498-1506
[27] Q. Zhang, P. Jin and L. "Yue, Extracting Focused Locations for Web Pages", First International Workshop on Web-based Geographic Information Management (WGIM) (in conjunction with WAIM'11), L. Wang *et al.*, (Eds.): WAIM 2011 Workshops, LNCS 7142, Springer, (**2011**), pp. 76-89

