Crowdsensing-based Web Crawler for Emergency Event Analysis

Wei Wu
Shanghai Institute of Technology
Shanghai, China
weiwu@sit.edu.cn

Xiao Wei*
Shanghai Institute of Technology
Shanghai, China
shawnwei@outlook.com

Bin Pan, Xiaokang Xu
Shanghai Institute of Technology
Shanghai, China
bin.pan@foxmail.com

ABSTRACT
In the domain of emergency event analysis, it is still a difficult issue to acquire the event information from the Web efficiently. To solve the problem, this paper proposes a crowdsensing-based Web crawler for emergency event analysis. When an emergency event occurs, web users post event information on the Web with geographical position, which can be regarded as crowd sensors. In the proposed method, the crawler takes advantage of the information from these crowd sensors, such as semantic information, geographical information, sentiment information, etc. to get the information of event efficiently. Experimental results show that the proposed method can improve the efficiency of crawler when compared with common crawlers.

CCS Concepts
• World Wide Web→Web searching and information discovery • World Wide Web→Web mining.

Keywords
Emergency Event Management; Web Crawler; Crowd Sensor; Social Computing; Semantic Link Network.

1. INTRODUCTION
An emergency event is a sudden, urgent, usually unexpected incident or occurrence that requires an immediate reaction or assistance for an emergency situation faced by a social group (e.g., a corporation) or recipients of emergency assistance [1,3]. During the evolution and management of an emergence event, the Web plays a very important role with the development of the Internet and mobile devices. The reasons fall into two aspects. In one respect, the Web is the most effective way to spread the information of emergence which may promote the evolution of the event. In another respect, the Web records the information of the event when transferring it, which becomes the information resource of emergency event management. Therefore, it is a basic issue for the emergency event analysis to get the available information from Web efficiently. However, the Web information of emergency event has different characteristics in the varying periods of event evolution, which make it difficult to acquire related information for the event analysis from the Web. In the preliminary period of the event, the related information is rare, which is difficult to discover and acquire. While, in the outbreak period of the event, the related information is too much to deal with. That is to say that it is still a problem to be solved for emergency event analysis to get the event information from the Web high efficiently.

* Corresponding Author: Xiao Wei, shawnwei@outlook.com.

To solve the problem, this paper proposed a crowdsensing-based Web crawler for emergency event analysis. In fact, a web user can be seen as a sensor of an emergency event. For example, if a user makes a post in micro blogs or BBS about an earthquake occurrence, then she/he can be seen as an earthquake sensor. The web can therefore be seen as a sensor receiver [2]. Specially, when an emergency event occurs, the web users who are nearby the place may post messages on the Web by mobile device. Some of these posted messages are marked with geographical position. At the moment, these sensors are crowd in geographical position which is also important information for event analysis. We also can get the related information according to the crowd relation among sensors. In our method, the crawler takes advantage of these crowd sensors to get the information of event efficiently.

The rests of the paper are organized as follows. In section 2, we define the crowd sensor and discuss the features of crowd sensor. In section 3, we propose the crowdsensing-based crawler. In section 4, we evaluate the proposed crowdsensing-based crawler by several experiments. In the end, we conclude the paper in section 5.

2. CROWD SENSORS

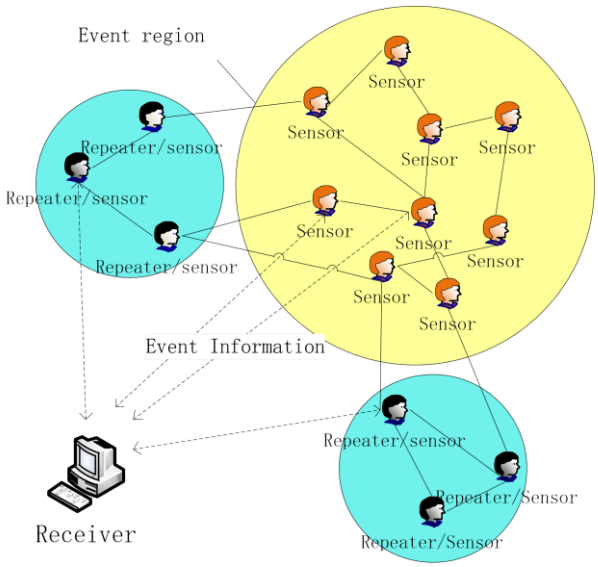

Figure 1. Crowdsensing-based event information acquisition

MOBIMEDIA 2016, June 18-19, Xi'an, People's Republic of China
Copyright © 2016 EAI
DOI 10.4108/eai.18-6-2016.2264141

When an emergence event occurs, the web user, who is related to the event and posts information on the Web, actually acts as a sensor. The process of crowdsensing-based event information acquisition is shown in Figure 1. In the figure, the yellow circle is the event region. The users in the event region undergo the event and may post messages on the Web, these users act as sensors, get the information from the event and wait for the receiver to gather the information. The users who are outside the yellow circle act as repeater or sensors. When they just repeat the information from sensor users, they are repeaters. Otherwise, if they post new web information about the event, they are sensors. These sensors /repeaters are not isolated. There are kinds of relations among these nodes, such as geographical association, semantic association, sentiment association, etc. All these relations connect the sensors/repeaters into a crowd sensor network, which is helpful to the crawler to acquire the event information more efficiently.

2.1 Event Sensor

Definition 1: Event Sensor (ES) is a Web user who focuses on an emergency event and generates or spreads event information on the Web, which is denoted as

$$ES = \{I, S, G\} \ , \tag{1}$$

in which, I is the set of information that the sensor generates and defined as

$$I = \{< i,t > | \ i \ is \ the \ information \ of \ the \ event \ at \ time \ t\}; \tag{2}$$

S is the set of sentiment of the sensor about the event and defined as

$$S = \{< s,t > | \ s \ is \ the \ sentiment \ of \ the \ event \ at \ time \ t\}; \tag{3}$$

G is the set of geographical position of the sensor and defined as

$$G = \{< g,t > | \ g \ is \ the \ geographical \ position \ of \ the \ sensor \ at \ time \ t\} \tag{4}$$

According to definition 1, an event sensor can continuously provide the information about the event, the sensor's sentiment about the event, and the geographical positions of the sensor.

2.2 Crowd Event Sensor Network

Definition 2: Crowd Event Sensor Network (CESN) is a semantic link network of event sensors, in which the network nodes are event sensors and the edges denotes semantic relations among nodes. CESN is denoted as

$$CESN = \{ES, E\}, \tag{5}$$

in which, ES is the set of event sensors which is defined by Equation (1). E is the set of edges, which is defined as

$$ES = \{< sem, sen, geo >_{i,j}\} \ . \tag{6}$$

In (6), $< sem, sen, geo >_{i,j}$ denotes an edge from sensor i to sensor j. *sem*, *sen* and *geo* are the weights of the relations of different relations among this pair of sensors respectively. *sem* denotes the strength of the semantic relation between the sensor i and sensor j. *sen* denotes the strength of the sentiment relation between sensor i and sensor j. *geo* denotes the strength of the geographical position relation between sensor i and sensor j.

The Crowd Event Sensor Network has the following characteristics:

1) Variety

The number of sensors changes with the evolution of the emergence event. Generally, at the beginning of event, there are few users who focus on the event, that is to say, the number of sensors is small. When the event enters breakout period, a large amount of Web users take part in the event and the number of sensors increases accordingly.

2) Dynamic

The sensors acted by Web users are mobile as time goes on, which makes the geographical position change. Moreover, the sentiment of the Web users are also able to change as time goes on. For example, with the evolving of the event, the sentiment of Web users may change from happy to sad. The dynamic of Crowd Event Sensor Network includes the changing of sensor values, the changing of relations among sensors, and the changing of topology of network.

3) Crowd

In the emergence event sensor network, the sensors are crowd according to different relations. The direct crowd is geographical position, which means sensors are close together. Semantic crowd means that a large amount of sensors post similar contents on the Web. Sentiment crowd means that a large amount of sensors have same sentiment on the event. All these kinds of crowd are shown as community structure in the event sensors network.

4) Duplication

When an emergence event occurs, some Web users generate messages about the event. Moreover, there are more users who just spread the messages. Therefore, there are many duplication sensors in the event sensor network. Although these duplication sensors don't generate new information about the event, their participations enhance messages from some aspects which lead to the crowd and are helpful to crawler.

According to the analysis on the above four characteristics, the crowd event sensor network consists of abundant knowledge about the event both in the sensor content and in the topology of the network, which can be used for the crawler to crawl the event information efficiently.

Compared with common sensor network, such as wireless sensor network, both event sensor network and sensors are virtual and needs to be constructed when they are used by the crawlers.

3. CROWDSENSING-BASED CRAWLER

A traditional crawler does a common task, while the event crawler does a special task. In this section, we propose a crowdsensing-based crawler for the event information acquisition. These two kinds of crawlers are compared in Table 1.

Table 1. Comparison between normal Web crawler and crowdsensing-based crawler

	Traditional crawler	**Crowdsensing-based crawler**
Specific goal	No	Yes
Knowledge base	No	Yes
Work mode	Single, Passive	Compound, Active
Direction	No	bi-direction

Without the supporting of knowledge base, traditional crawlers work based on the hyperlink on the Web, generally they download all web contents without content analysis. The crawler just does a

single task to crawl webpages. All the downloaded webpages are analyzed by some event analysis system. In fact, the flow of the information has no direction, which is from the Web to the crawler.

The proposed crowdsensing-based crawler has clear purpose that is to acquire Web information of an emergence event. With the supporting of knowledge base, the crawler works in the compound mode, that is to say, the crawler does the analysis when it gets the web content and makes decision based on analysis result. The information flow is bi-direction. The crawler not only get the web content passive but also detect and analysis the Web actively.

3.1 Knowledge base

Compared with the traditional crawler, the crowdsensing-based crawler has some intelligence, such as, sensor detection, sensor network construction, web content semantic analysis, etc. which need the support of knowledge base.

The knowledge base should consist of the following aspects of knowledge to support the crawler to work in an intelligence way.

1) Emergence event knowledge

The crawler needs to know the characteristics of emergence event, such as, the feature of Web information of an emergence event, the features of the event when it is in different periods. All these features help the crawl to detect event information more accurately.

2) Sensor knowledge

Event sensors are acted as Web users. The crawler needs to detect these event sensors among a large amount of Web users. This kind of knowledge describes the characteristics of event sensor, which helps the crawler to detect sensors for the emergence event.

3) Sensor network knowledge

This kind of knowledge helps the crawler to construct the sensor network for the emergence event, which should include the characteristics of emergence event sensor network and the rules to optimize the network.

3.2 Work process of the crowdsensing-based crawler

The work process of the crowdsensing-based crawler for emergence event information acquisition is shown in Figure 2, which is described as follows:

1) Define emergence event. The crawl task starts with emergence event definition. In this step, based on event knowledge base, an event is described by keywords and some other features.

2) Detect sensosr. In this step, the crawler analysis web information and detect sensors from it based on sensor knowledge base.

3) Construct sensor network. According to the existed event sensor network, the detected sensors may be a new sensor or an old one. The new sensor is added to the sensor network and the old sensor is used to update the duplicated one in the sensor network.

4) Acquire event information. Based on the sensor network, the web information is gathered though the sensors and filtrated by the event sensor network. As a result, only the needed event information is crawled and transmitted to the event analysis component.

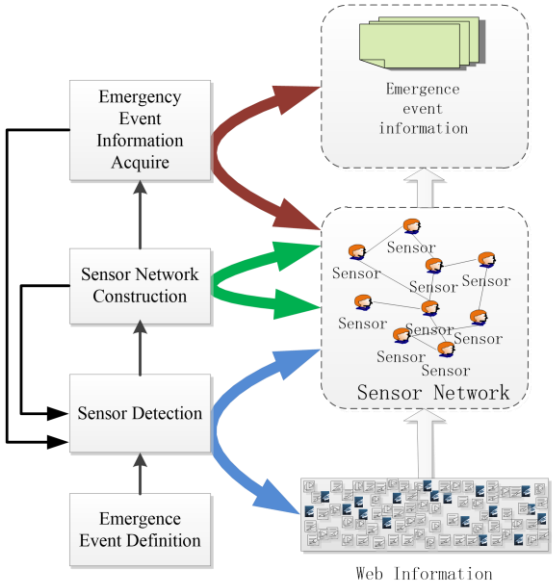

Figure 2. The work process of the crowdsensing-based crawler for emergence event

4. EVALUATION

We design two experiments to evaluate the ability of proposed crowdsensing-based crawler to get web information of emergence event.

1) Dataset

We select a news website (sina news, http://news.sina.com.cn/) and a micro blog (sina micro blog http://weibo.com) as the data source, all the webpages on the website and all the messages on the blog form the dataset used in the experiments.

2) Evaluation indexes

In the experiment, three evaluation indexes are used to evaluate the effect of the crawlers.

Recall ratio is the fraction of the web information that is relevant to the emergence event that is successfully crawled.

Semantic recall ratio is the updated recall ratio after the duplicated

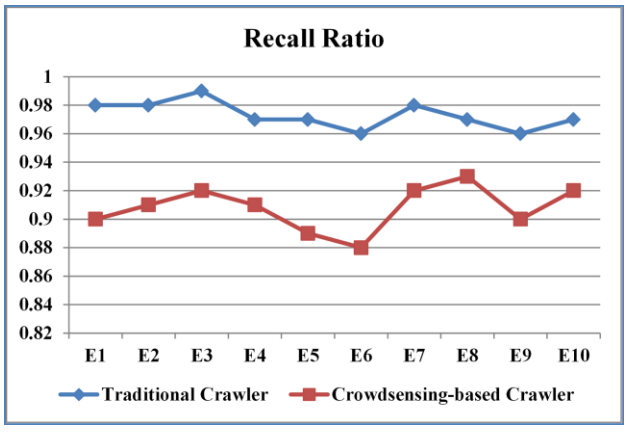

Figure 3. The comparison of recall ratio between the traditional crawler and the proposed crawler.

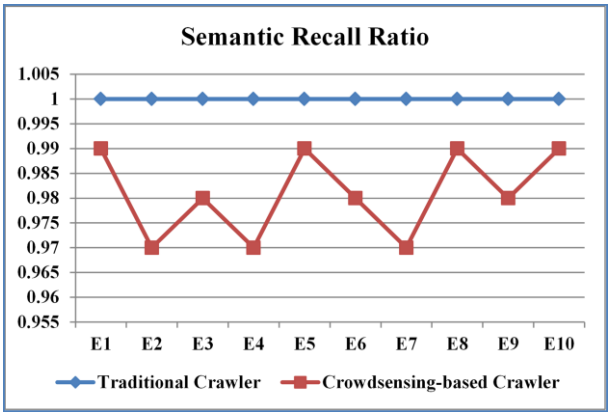

Figure 4. The comparison of semantic recall ratio between the traditional crawler and the proposed crawler.

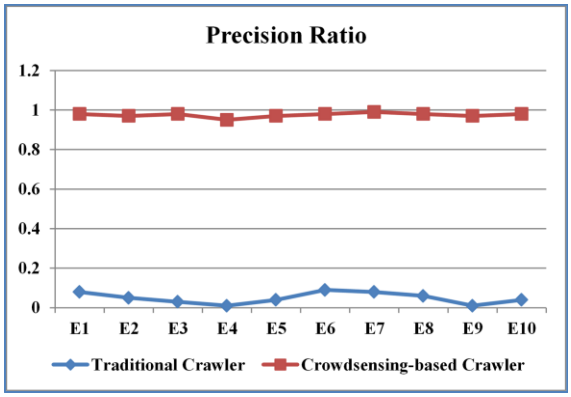

Figure 5. The comparison of precision ratio between the traditional crawler and the proposed crawler.

web information is removed from the result set.

Precision ratio is the fraction of crawled event information that is relevant to the emergence event.

3) Experiment process

In the experiment, we select 10 emergence events as the crawler tasks. After the events are defined, the traditional crawler and the proposed crawler do each task respectively. All the crawl actions are limited in the target website and micro blog. When each crawl task is done by the two types of crawlers, the recall ratio is calculated first and the results are show in Figure 3. Then the duplicated information is removed and the recall ratio is calculated for the second time based on the crawl result that has been dealt with. This recall ratio is called as semantic recall ratio and the results are shown in Figure 4. In the end, the Precision ratio is calculated based on the initial crawl result, which is shown in Figure 5.

4) Result and analysis

Figure 3 shows that both the traditional crawler and the proposed crawler have high recall ratio. Comparatively speaking, the traditional crawler works well than the proposed crawler. The reason is the traditional crawler downloads almost the information from the two target website, which is surely to have a higher recall ratio.

Figure 4 shows that the recall ratio when the duplicated information is removed. In event analysis application, the duplicated information doesn't provide new semantic for event. If the duplicated information is removed, it will not influence the result. Figure 4 shows that the recall ratio is almost equal together.

Figure 5 shows that the proposed crawler has very high precision ratio than the traditional crawler. The reason is also that the traditional crawler downloads too much irrelevant information.

Comprehensive considering the Figure 4 and Figure 5, the proposed crawler is similar with the traditional crawler in the semantic recall ratio. Furthermore, it has high precision than the traditional crawler.

5. CONCLUSION

In the domain of emergency event analysis, it is still a difficult issue to acquire the event information from the Web high efficiently. To solve the problem, this paper proposed a crowdsensing-based Web crawler for emergency event analysis. When an emergency event occurs, the web users post event information on the Web with geographical position, which can be regarded as crowd sensors. In the proposed method, the crawler takes advantage of the information from these crowd sensors, such as semantic information, geographical information, sentiment information, etc. to get the information of event efficiently. Experimental results show that the proposed method can improve the efficiency of crawl task when compared with common crawlers.

6. ACKNOWLEDGMENTS

Research work reported in this paper was supported by the Science Foundation of Shanghai under grant no.16ZR1435500.

7. REFERENCES

[1] Boldi, P., Codenotti, B., Santini, M., and Vigna, S. 2004. Ubicrawler: a scalable fully distributed web crawler. *Software Practice & Experience*, 34,8, 711–726.

[2] Edwards-Winslow, F. 2002. An introduction to emergency management. *Public Administration Review*, 62,5, 632-633.

[3] Heydon, A., and Najork, M. 1999. Mercator: a scalable, extensible web crawler. World Wide Web-internet & Web Information Systems, 2,4, 219-229.

[4] Nandagaonkar, S. S., Hanchate, D. B., and Deshmukh, S. N. 2012. Survey on event tracking and event evolution. *International Journal of Computer Applications in Technology*, 3,1,1-4.

[5] Wei, X., Luo, X., et al. 2015. Online Comment-based Hotel Quality Automatic Assessment using Improved Fuzzy Comprehensive Evaluation and Fuzzy Cognitive Map. *IEEE Transactions on Fuzzy Systems*, 23,1, 72-84.

[6] Xu, Z., Liu, Y., Yen, N., et al. 2016. Crowdsourcing based Description of Urban Emergency Events using Social Media Big Data. *IEEE Transactions on Cloud Computing*. DOI= http://dx.doi.org/10.1109/TCC.2016.2517638.

[7] Xu, Z., Liu, Y., Xuan J, et al. 2015. Crowdsourcing based social media data analysis of urban emergency events. *Multimedia Tools & Applications*, 2015:1-18. DOI= http://dx.doi.org/10.1007/s11042-015-2731-1.