A Survey on Detecting Public Emergencies from Web Pages

Jie Zhao 1, Peiquan Jin 2, Guorui Huang 3
1 School of Business, Anhui University, Hefei, 230039, China
2 School of Computer Science and Technology, University of Science and Technology of China, Hefei, 230027, China
3 New Star Research Institute of Applied Technology, Hefei, 230028, China
jpq@ustc.edu.cn
doi:10.4156/aiiss.vol3. issue3.8

Abstract

Detecting public emergencies from Web pages has been one of the focuses in the research community, as Web contains plenty of valuable information. In particular, Web information usually exhibits some spatiotemporal features, which can be utilized to analyze the pattern of public emergencies. In this survey paper, we review the state-of-the-arts of Web-based detection of public emergencies, including current prototypes and academic research results. After that, we emphasize some research challenges for the detection of public emergencies from Web pages. Finally, a framework to detect public emergencies from Web is presented. Its architecture as well as basic solutions for some critical issues are investigated. This paper is expected to provide some new prospective for the study on public crisis management and emergency decision in the Web and information era.

Keywords: Public Emergency, Web Pages, Detection, Survey

1. Introduction

In recent years, SARS, H1N1, and other social events have gathered a lot of attentions from both government and research communities. For example, China government has set up many policies to deal with social emergencies [1]. Recent research and practical experiences show that it has been one of the most important tasks for the world to study the detection and early-warning technologies of public emergencies.

In general, public emergencies can be classified as four types, which are natural disasters, accidents, public healthy events, and social security events [2]. Among those four types of public emergencies, we found that some events evolve with time. We called such events as “evolutional events” or “evolutional public emergencies”. Typical evolutional events include some prevailing diseases and social security events. Based on the review on the SARS and other events, we note that most of evolutional events have specific evolutional characteristics in the Internet. Therefore, if we can predict the happening of public emergencies according their evolutional specialty in the Web, it will bring many benefits for both the government and people.

Current treatment on public emergencies mostly uses the “after event” strategy, namely the government deal with public emergencies only after the events have happened. Such policies are helpful in the treatment of natural disasters and accidents, but for evolutional events, we have better solutions. A recent survey shows that about 90% of decision information can be acquired from the Web [3]. Thus if we can build a framework to detect evolutional events in the Web and then make early-warning for those events, it will improve the effectiveness and efficiency of the government on the treatment of public emergencies.

There are some previous work on the treatment of public emergencies, most of which are focused on the public health area [4, 5]. For example, China government has built a system to monitor public healthy events. It makes important achievements in China’s processing SARS, H1N1, and other affairs. However, this system is limited in the scope of hospitals and some related government agencies. The information sources are very limited, and the efficiency is very low. So it is not suitable for early-warning. In other aspects, many people choose the Web as information publisher, such as Blog and BBS, so many recent research work concentrates on
this topic. The most famous works are the Global Public Health Intelligence Network (GPHIN)[6, 7], the Europe Media Monitor (EMM) [8] which is built by the European Commission Joint Research Centre, and Google sponsored Innovative Support to Emergencies, Diseases and Disasters (InSTEDD) [9]. The GPHIN system was developed in 1998 and now it has been one of the main information resources in WHO. All those systems adopt the similar techniques as traditional search engines, which focus on the search of Web pages. Hence, the effectiveness mainly depends on the level of intelligence analysts.

In this paper, we present a survey on the detection and early-warning of evolutional events in the Web. Traditional approaches depend on the keyword-based Web search and text mining techniques and have low efficiency when producing intelligence. We analyze the necessity of the detection and early-warning of evolutional emergencies, as well as the state-of-the-art of this area. We focus on the two dimensions of Web information, the temporal dimension and spatial dimension, to study the description of evolitional events, and build a spatiotemporal model to describe the evolutional events.

The remainder of the paper is organized as follows. Section 2 discusses the state-of-the-art of the related work. Section 3 describes the research issues for the detection of evolutional events. In Section 4, the framework to detect evolutional emergencies is presented. And conclusions are in Section 5.

2. Related work

Traditional emergencies detection mainly rely on some management mechanisms, such as periodically reporting. As Web contains timely information with a broad range, it has been a hot topic in recent years to detect emergencies from Web. Compared with existing detecting mechanisms for emergencies, Web-based emergencies detection has the following unique properties. First, the evolution of emergencies in Web has shown a particular feature related with time and locations. Second, emergencies in Web are usually associated with a series of topics in Web. Third, emergencies in Web are typically evolved with time and locations. Those special features of emergencies in Web make it difficult to directly use current approaches in Web information extraction and topic detection to find emergencies in Web. So far, there are some related works focused on this issue, which are described in the following parts.

2.1. Systems for Emergencies Detection in Web

There are several systems aiming at detecting emergencies from Web, including the GPHIN system in Canada, EMM in Europe Union, InSTEDD sponsored by Google.

The GPHIN system (Global Public Health Intelligence Network) was first established by the Ministry of Health, Canada [6]. GPHIN contributed a lot for the SARS in 2003. Three months before the bursting of SARS, GPHIN detected from Web that the population suffered from respiratory diseases in the world had significantly increased. Then they reported this fact to WHO. It is because the early-warning from GPHIN that we are able to avoid the spread of SARS all over the world [7]. GPHIN adopted many technologies to detect and track influenced events from Web, including Web search, data mining, automatic translation, and information filtering. It used a semi-automatic way to perform its tasks, which consisted of a first step retrieving information from Web automatically and a second manual step to analyze and evaluate the information, which is conducted by experts. However, GPHIN only considered health information so far.

The EMM system, started from 2006, was first initiated by Europe Union, with similar motivation to GPHIN [8]. The difference of EMM from GPHIN is that it can track many types of Web information, not only health information, but also terrorist attacks, forest fire, or other emergencies. Its main idea is to extract information elements from Web and then cluster them in certain ways. For example, it can put all the news reports related with a specific event in one page, and provides a news explorer for experts to analyze the information. The most specific feature of EMM is that it supports information extraction from sources with sixteen languages,
such as English and Chinese. The main technique it adopted is the keyword-based search, which is focused on news web pages. Particularly, it used about 1,600 keywords in its current version to find interested events.

The InSTEDD system is an information sharing system, which was proposed by Larry Brilliant in 2008. InSTEDD was sponsored by Google and also got help from the technology groups in GPHIN [9]. The goal of InSTEDD is to build a network for information sharing and communication so as to realize the cooperation among the communities all over the world. A simple example can be described as follows. A doctor found a strange disease in South-East Asia. He then put this message in InSTEDD and other doctors can get this information and pose their opinions on this event. Similar to GPHIN, InSTEDD first aimed at tracking infectious diseases. As its technologies were mainly come from GPHIN, the implementation of InSTEDD is very similar to GPHIN. Or in one word, InSTEDD is an open-sourced GPHIN.

However, there are several limitations of the systems discussed above. First, they all relied on the keyword-based search technology and can not effectively find the evolutional properties of emergencies in Web. Second, they all needed certain intervention from experts.

2.2. Topic Detection and Tracking in Web

A topic typically contains a seed event and some relevant activities [10]. Topic detection and tracking, known as TDT, was developed in 1990s on the basis of Web information retrieval and Web information extraction. It concentrates on the processing of news Web pages, with the main task on identifying news boundary and detecting and tracking emergency news. Since 1996, researchers from DARPA and CMU conducted a lot of studies in this field [10]. Although current techniques in information retrieval and extraction can be applied in basic TDT, there are also many issues remained unsolved in this area.

The current topic detection algorithms are mainly based on the clustering technique, such as two-phase clustering algorithm [11], K-NN [12], K-means [13], and the layered approach [10]. A topic is usually represented by its kernel and as a layered form. Though there are a few studies in topic detection [14, 15], those works are not feasible for practical applications. For example, the layered approach as proved as an efficient one among the sample-based algorithms has been argued that its low precision can not satisfy the requirements from real applications [16]. One real system of topic detection developed by Google is Google News, which adopts the single-layered techniques rather than the conventional layered approach. However, the error rate in Google News is relatively high, i.e. more than 50% as reported in the literature [14].

2.3. Models for Public Emergencies

Modeling public emergencies has been one of the important issues in public security and emergency management. A model is regarded as an abstraction of the real world, so a public emergency model aims at representing the characters of public emergencies in real world at the largest extent. To exploit this problem, in [17] the authors modeled the happening, dispersing, and elimination of public emergencies as some evolving process in physics. They finally proposed a series of models for public emergencies, including the acceleration model, speed computation model, strength computation model, and cost model. This approach used the dynamics theory and presented a possible solution for the modeling of public emergencies. However, how to implement such a model is still a problem. Other people tried to modeling public emergencies on the basis of the factors influencing public emergencies [18-20]. For example, in [19] they utilized the transferring information of bank checks to detect the activity pattern of people and furthermore the abnormal events. Another example reported in [20] used the information about the base stations for cell phones to find out public emergencies. Those models only addressed the factors that may influence public emergencies, but ignored the evolution rules of public emergencies.
3. Research Issues for Web-based Public Emergencies

3.1. Modeling Evolutional Emergencies in Web Environment

For this issue, we need to study the correlation between Web information sources and evolutional emergencies on the basis of the characteristics of Web pages and features of evolutional emergencies. Then we can further investigate the conceptual modeling approach to describing evolutional emergencies in Web environment, probably from the static and dynamic properties of evolutional emergencies. The final solution to this issue is to design an ontology-based data model for Web-based evolutional emergencies, which is able to capture different features of evolutional emergencies and provide a uniform representation for them.

3.2. Approaches to detecting evolutional emergencies from Web pages

The intrinsic feature of evolutional emergencies is that they are all time-evolving, thus how to detect time-evolving events becomes the most critical task in the detection of evolutional emergencies. In particular, we need to develop efficient and effective algorithms to extract time-evolving events from different Web information sources, such as News pages, Blog, and BBS. Also the organization and storage structure for detected evolutional emergencies should be studied. The most important issues in this topic can be summarized as follows:

(1) Detection of time-evolving topics
Time-evolving topics are those topics changing with time. A typical topic usually has a start point, a developing period, and an end instant. That time information forms the foundation of the detection of evolutional events. Basic tasks on detecting time-evolving topics consist of keywords extraction, time information extraction, location information extraction, and construction of time-evolving topics, while different Web sources are needed to be considered.

(2) Ontology-based construction of evolutional events
Based on the detected results of evolutional topics, we are able to further construct evolutional events according to the ontology developed in 3.1. Other works are also needed to be considered, for example, how to find fraud events, and how to eliminate collisions among detected events.

3.3. Tracking Evolutional Events

It is very important to track the spatiotemporal features of evolutional events, e.g., to find the spread trend of a specific event in a certain geographical range. The main task on tracking evolutional events is to represent the evolutional styles of events in the time dimension as well as in the space dimension. Based on this representation, spatiotemporal analysis and mining approaches can be applied to conduct some reasoning work on emergencies events. Another issue is how to evaluate the crisis of evolutional events.

3.4. Theme-based System Implementation

On the basis of theoretical studies, a theme-based prototype system is expected to be implemented to realize the detection of evolutional emergencies from Web pages. The themes are from real demands in different applications. There are some system-related issues in order to build such a prototype, including the architecture design, theme database, experimental dataset, user interface, and visualization tools.

4. A Framework to Detect Web-based Public Emergencies

In this section, we will discuss a basic framework to detect Web-based public emergencies. The framework is proposed in our current research project, which provides basic solutions to the issues discussed in Section 3.
4.1. Modeling Evolutional Web 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.

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 Fig.1). Each meta 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.

![Fig.1 The Spatiotemporal model for evolutional events](image)

4.2. Detection of Web-based Evolutional Emergencies

We propose to use a four-step procedure to perform the detection of Web-based evolutional emergencies.

(1) Topic detection based on Web page title
Topic detection is the first step to find evolutional emergencies. As most topics can be found in the title of Web page, we prefer to scan the Web page title to find possible topics.

(2) Extracting time and temporal relationships from Web pages
Time plays important role in the detection of evolutional emergencies. For the topics extracted in the previous step, next we will determine the related time of each topic extracted. Current research in Web time extraction mainly focus on the update time of Web pages, or in other words, the crawled time, which is widely used in Web archive system. However, for
evolutional events, we pay more attention on the content time of Web pages. In particular, we will use the following algorithm to extract the content time related with a specific topic. First, we process a Web page into a DOM tree, in which each leaf node represents a sentence in the Web page. Second, we annotate the leaf nodes in the DOM tree using the TIMEX2 standard. Finally, we determine the time related with the given topic on the basis of the properties of annotated time and some heuristic rules. For example, for a topic in a news report, the most related time usually appears in the title or in the first line of the news report.

(3) Constructing time-evolving topics
In this step, we integrate the extracted topics and content time and form a set of meta-events, which are then be processed using clustering approach to produce some groups. This aims at gathering all the events related with a specific topic into one group, which will be further used to generate evolutional events.

(4) Model-based construction of evolutional emergencies
Based on the extracted time-evolving topics and the results of clustering, we construct evolutional emergencies according to the predefined ontology model.

4.3. System Architecture

Based on the analysis on the state-of-the-art, we present a framework for the detection and early-warning of evolutional events in the Web. The system consists of five modules, as shown in Fig.2, which are the time-changing topic detection module, event construction module, visualization of event module, query & tracking module, and evaluation & early-warning module. The basic process of the system is as follows. We first extract the topics and time information and form time-changing topics. This procedure will use the thematic corpus, TIMEX2 specification [21], and predefined extracting rules. Next, we normalize the time-changing topics and construct evolutional events according the event model. The results will be written into the event database. Then, when we post a request to query and track evolutional events, the query and tracking module will search the event database and return the results to the visualization module. The visualization model will present the retrieved evolutional events in the user interface using a time-based ranking approach [22]. When we issue a request about the evaluation and early-warning, the evaluation and early-warning module will analyze the event database and obtain some spatiotemporal patterns. After that, it will evaluate the events and provide the early-warning output according to the RCA rules. Also the early-warning results can be presented in the user interface on the support of the visualization module.

![Fig.2. The System Framework for the Detection and Early-warning of Evolutional Events](image-url)

(1) Time-Changing Topics Detection
The main function of this module is to extract the time-changing topics from Web pages and send them to the event construction module. Evolutional events are those events that evolve
with time, so the time-changing topics detection is the foundation of the detection of evolitional events.

The difficulties in the time-changing topics detection include the extraction and description of the topic keywords, time information, and location information. Different Web pages have different content structures and other properties. Meanwhile, the time and location representation styles in Web pages are much different. For example, in some Web pages, the Christmas day is expressed as December 25, while other ones may use the “Christmas Day”.

(2) Event Construction

In this procedure, we integrate the extracted time-changing topics and construct meta events. We use the clustering approach to aggregate the similar meta events into one category. Then, we generate normal data structures for events according the event model.

(3) Visualization of Events

This module provides the static visualization and dynamic visualization for evolitional events. The static visualization can show the status of an event, such as theme, level, and so on. The dynamic style can show the changing process of the event locations in a time line.

(4) Query and Tracking

This module provides the data analysis and query function for related managers. It consists of two procedures, which are the database query procedure and the results clustering procedure. The results’ clustering supports different clustering ways, including time-based clustering, location-based clustering, and keyword-based clustering.

(5) Evaluation and Early-warning

In this module, we use spatiotemporal analysis approaches and perform data mining work on the events dataset. The time dimension and location dimension will be used in the spatiotemporal data mining procedure. At the same time, we build the crisis evaluation model to evaluate the events and provide early-warning information. The basic idea of this module is as follows:

(a) First, we use the visualization module to present spatiotemporal representation of the events. This representation will exhibit the evolution of an event, as well as the tracking of the event.

(b) Next, we use clustering, statistics, and other methods to analyze the spatiotemporal distribution, spatiotemporal classification, spatiotemporal change patterns on the events dataset in order to obtain some implicit patterns and knowledge.

(c) At the same time, we use a fuzzy evaluation method to evaluate the crisis level of the detected evolitional events.

(d) Finally, we set up the RCA (Rule-Condition-Action) rules for the early-warning purpose, which is based on the spatiotemporal analysis and crisis evaluation.

5. Conclusions

Detecting public emergencies from Web pages plays important roles in public security and emergency management, as Web contains plenty of valuable information and intelligence. In this paper, we made a survey on this field and discussed the related techniques. We also pointed out some research challenges in this research area, and presented a Web-oriented framework for the detection of evolitional public emergencies.

Currently we are working on a system to detect public emergencies from Web pages, which is based on the architecture shown in Fig.2. Another future work is the case studies of public emergencies detection, as different application domains have different definitions on emergency.

6. Acknowledgments

This work is supported by the National Natural Science Foundation of China under the grant no. 70803001, the Open Projects Program of National Laboratory of Pattern Recognition (no.20090029), the Key Laboratory of Advanced Information Science and Network Technology of Beijing (no. xdxx1005), and the USTC Youth Innovation Foundation.
7. References

[9] InSTEDD [EB/OL], In: http://instedd.org/, 2009
[21] TIMEX2 [EB/OL], In http://timex2.mitre.org/