The Journal of Systems and Software 102 (2015) 226–236

Contents lists available at ScienceDirect

The Journal of Systems and Software

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

Measuring the veracity of web event via uncertainty

Xinzhi Wang [a], Xiangfeng Luo [a,b,*], Huiming Liu [a]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai, China
[b] State Key Laboratory of Mathematical Engineering and Advanced Computing, Wuxi, China

ARTICLE INFO

Article history:
Received 25 February 2014
Revised 16 June 2014
Accepted 10 July 2014
Available online 19 July 2014

Keywords:
Web event veracity
Topic detection and tracking
Big data

ABSTRACT

Web events, whose data occur as one kind of big data, have attracted considerable interests during the past years. However, most existing related works fail to measure the veracity of web events. In this research, we propose an approach to measure the veracity of web event via its uncertainty based on its features distribution on different kind of confident websites. Firstly, the proposed approach mines various event features from the data of web event which may influence on the measuring process of uncertainty. Secondly, one computational model is introduced to simulate the influence process of the above features on the evolution process of web event. Thirdly, matrix operations are managed to facilitate practice. Finally, experiments are made based on the analysis above, and the results proved that the proposed uncertainty measuring algorithm is promising to measure the veracity of web event for big data.

1. Introduction

Today, big data has been attracting more and more attention. Challenges and opportunities of big data era are defined as being five Vs, i.e., volume (amount of data), velocity (speed of data in and out), variety (range of data types and sources), value (desirable quality of data) and veracity (trustworthiness of various data).[1] The volume of big data is massive, and main technologies to store big data include distributed cache, distributed database and distributed file system (Zhang et al., 2014; Zhang X and Liu C et al., 2014). The big data is also time-sensitive, defined as velocity, referring to the speed at which new data is generated and the speed at which data[1] moves around, and main methods to face this challenge include MapReduce (Dean and Ghemawat, 2008) and concurrency control. Moreover, types of big data consist of text, video, audio, webpage, stream and even aggregation of above, which is the meaning of variety. Another V to take into account when looking at big data is value, which is important as the result that users can make a benefit for any attempt to collect and leverage big data. The last V, veracity, refers to the trustworthiness of the big data, with large amount, high speed, many forms and uneven quality. We can safely argue that 'veracity' is the most important V of big data to gain ultimate and valid value. This paper aims to measure the veracity of web data which occurs as one challenge bridging users and big data.

Web event is a story or a scandal occurred in the society or on the web reflected by a series of associated web pages with time. It is hard to measure web event veracity as the result that volume and various web events happen all the time. However, veracity of web event has to be detected in order to get trustworthy and valuable information. In its evolution process, social event is deeply affected by corresponding web information, whose one performance is uncertainty. In other words, monitoring veracity of web event via uncertainty can help user understand social events. Generally, event with high uncertainty are more likely to turn into popular or emergent event (Haddow et al., 2010). Here uncertainty of web event is determines by its features distribution on different kind of confident websites instead of entropy used in informatics. For instance, if an event has high uncertain distributed in low confident website, then the event is more likely to be faked; namely, its veracity is low; and vice versa. So, this research employs multi-factor based uncertainty to measure the veracity of web event. However, it is difficult to manually analyze the veracity of the volume web events in their evolution process on the web, because it is a killing of time and energy. So, to understand and quickly respond to the volume web events, it is necessary to measure the web event veracity automatically.

For instance, a Chinese girl named MeiLing Guo showed off her luxurious life on the web, and claimed to be one of the managers of Red Across in China. This message was detected by netizens who suspected the corruption of Red Across. The event ended up with administrative recombination in Red Across, and resulted in sharp

* Corresponding author at: School of Computer Engineering and Science, Shanghai University, 333 Nanchen Road, Baoshan District, Shanghai 200444, China. Tel.: +86 13817505710.
E-mail addresses: wxz8912@163.com (X. Wang), luoxf@shu.edu.cn (X. Luo), a60695420@163.com (H. Liu).

[1] http://staff.science.uva.nl/~demch/////presentations/sne2013-01-03-bigdata-5V-infra-v04.pdf.

http://dx.doi.org/10.1016/j.jss.2014.07.023
0164-1212/

decline of donations which had bad influence on the Red Across of China. This web event impacted greatly on our society. The veracity of this event is low with just one doubtful piece of microblog at first, but went high at last with various guesses occurred.

Veracity of web event varies in the evolution process. One of veracity performances of web event is its uncertainty, and the other is website confidence, two of which is determined by various factors such as webpages distribution on websites, attributes distribution on webpages, etc. In this paper, we firstly proposed an approach mining several event features from webpages of web event. Then, one computational model is introduced to simulate the influence process of the above features on the evolution process of web event. Furthermore, matrix operations are managed to facilitate practice. At last but not at least, experiments are made based on the analysis above, and the results proved that the proposed model is promising when measuring the veracity of web event via its uncertainty.

The organization of this paper is as following. Related work is introduced in Section 2. Basic terms and event features are defined in Section 3. The veracity measurement of web event via uncertainty is introduced in Section 4. The relation of the measuring processes with matrix is discussed in Section 5. Section 6 shows some experiments and analysis. And in the last section, conclusions are given.

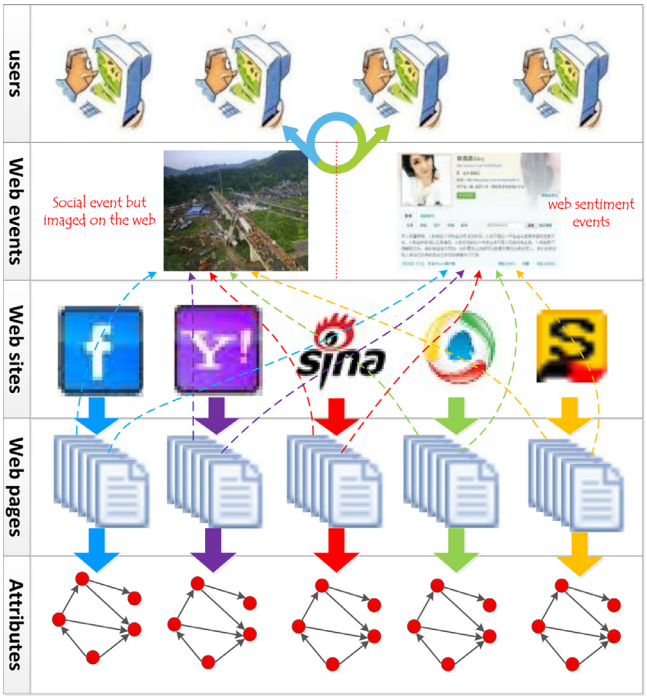

Fig. 1. Interaction among users, web events, websites, webpages and event attributes.

2. Related work

Veracity calculation of web event is one tough work since the data of web event is volume, variety and velocity. Moreover, veracity, one of whose performances is uncertainty, of web event varies during its evolution process. However, few researchers have contributed to measuring web event veracity or uncertainty directly. In the research area of topic detection and tracking (TDT) (Jin et al., 2010; Yin et al., 2008; Yang et al., 2009; Makkonen, 2003), some aspects of web event are measured during the evolution process, such as detecting unknown event and segmenting event information, without considering web event uncertainty. But technologies of TDT lay some foundations for our work, namely measuring the veracity of web event via uncertainty.

In the research area of TDT, Allan et al. (1998a,b) calculated similarity among news and classified the news to corresponding events. If the event of a piece of news is not similar to any previous events, then it is considered as a new event. Chen et al. (2007) put forward the time-hardening theory, which modeled event life cycle based on the time relations among documents. Wei and Chang (2007) proposed a kind of technology to build an event evolution model, which found frequent events section and temporal relations of one event to construct frequent pattern. Among them, the events section refers to a sub-event or events stage. Yang and Shi (2006) proposed a method to extract evolutionary relationship diagram from news reports, and the evolutionary relationship diagram reflects relationship among web events or between events and its sub-events, which furthermore display the structure of web events. Some other aspects (Li et al., 2009; Li and Lau, 2011; Ng et al., 2003, 2005) are also been discussed. In general, TDT technique (Allan, 2000; Mei and Zhai, 2005; Jo et al., 2007; Nallapati et al., 2004) attempts to detect unknown event and cluster corresponding news reports into its event; TDT also involves tracking events, but do not involve the veracity measuring of web event during the evolution process of web event. So it is hard to provide users global and clear cognitions for the web event. This paper came up with a method to calculate the veracity of web event taking the advantage of web event uncertainty.

This paper also discussed what factors may influence on the uncertainty of web event, such as attribute uncertainty, webpage uncertainty, website confidence, event classified webpage uncertainty, attribute distribution on webpage, and event classified webpage distribution on websites. On the other hand, Salton and Yang (1973), Kenneth and Yang (2007) analyze many factors affecting website confidence. In addition, different types of website have various confidences (Nadine and Burke, 2002). The traditional media websites always have higher confidence compared with other websites, such as CNN, New York Times, and the Wall Street and so on. Metzger (2000) indicates that website confidences are dynamic and uncertain; this made it hard to evaluate website confidence clearly. Attribute distribution on webpage and classified webpage distribution on website have great influence to the web event uncertainty as well as the website confidence. In this paper, we define a lot web event features, then establish mathematical model for computing web event veracity via its uncertainty, furthermore gain web event veracity through iterative computational model, and finally matrix operations and experiments are made which proved the iterative algorithm is promising when computing web event uncertainty.

3. Basic terms and event features

As we have discussed in Section 1, to measure the veracity of web event via uncertainty, it is necessary to study the event features of web event in its evolution process. Before the discussion of web event features, we define some basic terms which can help the understanding of veracity measuring process.

3.1. Terms definitions

Normally, users take care of web events when they surf on the internet. And these web events are provided by webpages, which distribute on different websites and contain abundant event attributes. The whole chain process that influences on the measuring process of event veracity is depicted in Fig. 1 . The five levels are users, web events, websites, webpages and event attributes, respectively. And the arrows describe the relations among different levels. This section