Building the Profile of Web Events Based on Website Measurement

Zheng Xu, Junyu Xuan, Yiwei Zhu and Xiao Wei

Abstract Nowadays, Web makes it possible to study emergencies from web information due to its real-time, open, and dynamic features. After the emergence of a web event, there will be numerous websites publishing webpages to cover this web event. Measuring temporal features in evolution course of web events can help people timely know and understand which events are emergencies, so harms to the society caused by emergencies can be reduced. In this paper, website preference is formally defined and mined by three proposed strategies which are all explicitly or implicitly based on the three-level networks: website-level, webpage-level and keyword-level. An iterative algorithm is firstly introduced to calculate outbreak power of web events, and increased web pages of events, increased attributes of events, distribution of attributes in web pages and the relationships of attributes are embedded into this iterative algorithm as the variables. By means of prior knowledge, membership grade of web events belong to each type can be calculated, and then the type of web events can be discriminated. Experiments on real data set demonstrate the proposed algorithm is both efficient and effective, and it is capable of providing accurate results of discrimination.

Keywords Website preference · Web mining · Web events

Z. Xu (✉)
The Third Research Institute of the Ministry of Public Security, Shanghai, China
e-mail: xuzheng@shu.edu.cn

J. Xuan
Shanghai University, Shanghai, China

Y. Zhu
Zhejiang Business Technology Institute, Ningbo, China

X. Wei
Shanghai Institute of Technology, Shanghai, China

N.Y. Yen and J.C. Hung (eds.), *Frontier Computing*, Lecture Notes
in Electrical Engineering 422, DOI 10.1007/978-981-10-3187-8_1

1 Introduction

Web event is what social Medias (i.e., BBS, blog, and news sites) discuss via cyber and influence on our real society. People can discuss web event in various forms, such as commenting news, posting and replying in forum, or recording and messaging in blog, etc. A web event could be a hot news story, like a scandal, or the cover of a social event in the real world on the web, like an earthquake. In general, the content of a web event is not stationary, but it will change with the evolution of this web event. This evolution may come from the development of a news story and the change of interests of its web followers. At a given time, a web event is generally composed by some sub-topics which focus on different aspects of this web event. These discussions, which describe lots web events, have an impact on the evolution of web event. In return, our society will be influenced by the information in web. So the detection and prediction of web events evolution is a meaningful work. To get this goal, we put our hands to measure and analyse evolution features of web events.

The identification of website preferences web events is of significant. The merits are: (1) instead of the most visited websites, the specified websites can be recommended to users and organizations who interest on the aspects of web events covered by these websites. Then, users can follow their interested aspects of web events by following the recommended websites; (2) identify the malicious websites which only publish malicious aspect of web events. If the slander information of a web event is only come from one website, it is possible that this website is spreading slander information. From above incidents, the tapping phone is one kind of social event happened in our society but mapped on the web. By the mapping, social events spread, evolve and mutate in the web along with interaction with real world. And we call such events as social events mapped on web. The latter incident is caused by message on web and impact on real world. In other words, this kind of event happened in virtual world but evolve with human interference. We call such events as web sentiment events. All of these two kinds of events are called web event. Some web events have much bad influence on society. To avoid these bad influences, it is necessary to monitor and predict the evaluative tendency of web events. Therefore, how to collect and organize web events in the intelligent and automatic way, and how to track and measure dynamic evolution of web events are becoming an important subject in the field of information processing.

In this paper, website preference is formally defined and mined by three proposed strategies which are all explicitly or implicitly based on the three-level networks: website-level, webpage-level and keyword-level. An iterative algorithm is firstly introduced to calculate outbreak power of web events, and increased web pages of events, increased attributes of events, distribution of attributes in web pages and the relationships of attributes are embedded into this iterative algorithm as the variables. By means of prior knowledge, membership grade of web events belong to each type can be calculated, and then the type of web events can be dis-criminated. Experiments on real data set demonstrate the proposed algorithm is both efficient and effective, and it is capable of providing accurate results of discrimination.

2 Related Work

The evolution is a basic feature of web events and is also a part of studies on Topic Detecting and Tracking (TDT) [1–3]. Traditional TDT involves detecting unknown events, gathering and segmenting information, detecting when the event first reported, detecting follow-up reports of events and tracking events' tendency. Generally, TDT technology attempts to detect unknown web events and make related news pages clustered. Although TDT tracks development of web events, it does not measure the dynamic evolution process of web events. So we cannot have a global and clear understanding of web events. Qi [4] suggests that a website should be evaluated from three aspects: usefulness, service quality and physical accessibility. The qualities of content and structure of websites will impact on their usage preferences which means the efficiency of using these websites. And the content is more important than structure in the long run [5]. The content and structure of website is evaluated to fit better the needs of visitors by reorganizing the documents [6]. There are also many extensions of LDA which have considered different aspects of documents. There are also some works trying to release the independent of documents and discovered topics by considering the citation relations between documents [7] and relations of topics. However, all these works are still based on 'bag-of-words' assumption and the relations of keywords within documents are ignored. Some researchers were ware of this gap. In our previous work, in order to detect and describe the real time urban emergency event, the 5W (What, Where, When, Who, and Why) model is proposed by Xu [8]. Xuan [9] proposed a framework to identify the different underlying levels of semantic uncertainty in terms of Web events, and then utilize these for Webpage recommendations. The basic idea is to consider a Web event as a system composed of different keywords, and the uncertainty of this keyword system is related to the uncertainty of the particular Web event. Liu [10] explored a Markov random field based method for discovering the core semantics of event. The method makes semantics collaborative computation for learning association relation distribution and makes information gradient computation for discovering k redundancy-free texts as the core semantics of event. A crowdsourcing based burst computation algorithm of an urban emergency event is developed in order to convey information about the event clearly and to help particular social groups or governments to process events effectively [11–15].

3 Iterative Method

In this paper, the communities of keyword level network are adopted to represent subtopics of a given web event. Since the each keyword is a semantic unit of a web event, the community of a number of keywords, which have relative close relation with each others, can be seen as a sub-topic of a web event. The most straightforward method to get the preferences of websites would be to detect the communities of keyword level network and then these detected communities could be seen as the different sub-topics of a web event. The preferences of websites could be computed as the membership degree on each community. The procedure of this method can be described as (Fig. 1),

(1) Construct keyword network of a number of webpages published on a number of websites;

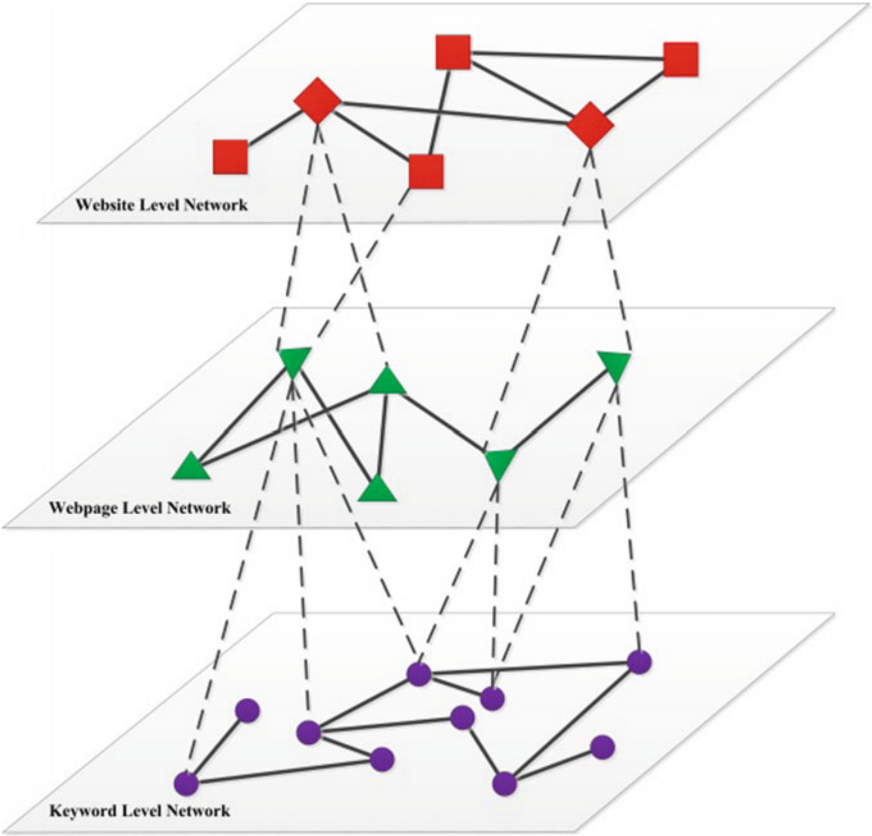

Fig. 1 The proposed method

(2) Do community detection on this keyword network;
(3) Compute the membership degree of each website on the detected communities.

The communities of keyword level network are only based on the keyword relations between each other, a horizontal relation in Fig. 1. This relation implies that the keywords, which have close association relations with each others, will be more likely to describe same sub-topic of a web event. Actually, the webpage level network will also influence the formation of communities at keyword level. When all the keywords are in the same webpage by the mapping relations between keywords and webpages, it is also possible that they are talking the same sub-topic of a web event. However, the relations in the keyword level network, ALNK, does not take the mapping relations into consideration, which only consider the statistical values of co-occurrence relations on all the webpages. For example, two keywords, ki and kj, have a small co-occurrence relation which means that they do not usually show in the webpages simultaneously. However, if two webpages which contain keywords, ki and kj, respectively and they are in the same community of webpage level network, keywords, ki and kj, are also have big probability to talk about same sub-topic of a web event. Similarly, the communities of webpage level are also influenced by the mapping relations between webpages and websites. Inspired by their inter-dependency and inter-limitation relations of websites, webpages and keywords, a iterative algorithm is proposed to optimize the formation of keyword communities/sub-topics.

4 Fuzzy Based Algorithm for Type Discrimination of Web Events

With the time changing, the emergent degree of web events changes is in dynamic change. One event in different segments has different emergent degree, so for a web event, it may go through three states: general state, hot state, and emergent state. Fewer domestic and foreign scholars study on emergent level classification of web events in different segments, so that the lack of a prior knowledge of type discrimination of web events in different segments. Therefore, we study the changes of features and emergent degree of web events in evolution course, and we can obtain the relationship between emergent degree and outbreak power, fluctuation power. Then by studying these relationships, we extract features of different emergent degree, establish evolution model of web events, and construct the membership model for type discrimination of web events as prior knowledge. Thereby to provide effective guidance for the type prediction of web event in later section.

For the result of algorithm, it describes the emergent degree of web events and it is called outbreak power. In this paper, “day” is the minimum time granularity. Source data of temporal features of web events is collected from different news sites daily, Algorithm 1 calculate the daily outbreak power of web events based on these source data, and then time series data of outbreak power of web events in a certain time interval are obtained, as shown in Fig. 2. The outbreak power of web events, which is calculated by Algorithm 1, combines the increased webpages, increased attributes of events, and distribution of attributes in webpages. The Algorithm 1 considers the physical attributes of web events, semantic content, and distribution of web events on web. So the outbreak power we get can comprehensively describe the evolution course of web events.

Herein, 100 web events were selected as the experimental object. And 60 web events among experimental object as training set to establish prior knowledge of web events, and the remaining 40 web events were as test set of type discrimination.

In experiment, we first trained 60 web events in training set, annotated the web events according to their emergent degree, so these 60 web events were labelled as emergent event, hot event or general event. By statistics on the training set, we calculated the membership frequency of temporal features belonging to each type when temporal features took different values, and combined with prior knowledge of our cognition on web events, we got the membership distribution of each temporal feature belonging to different types of web events. Here, Fig. 3 shows the membership distribution of average outbreak power belonging to different types of web events.

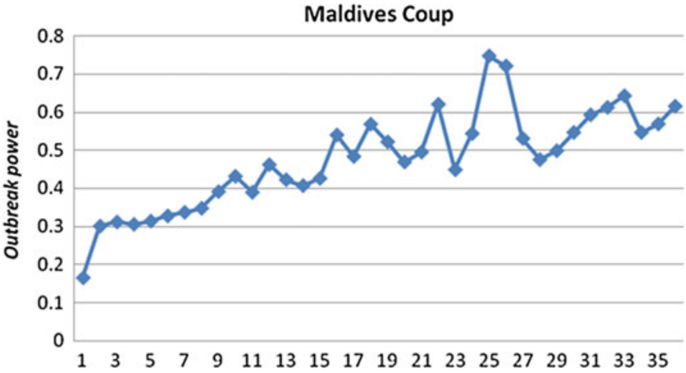

Fig. 2 The outbreak power of web events in a certain time

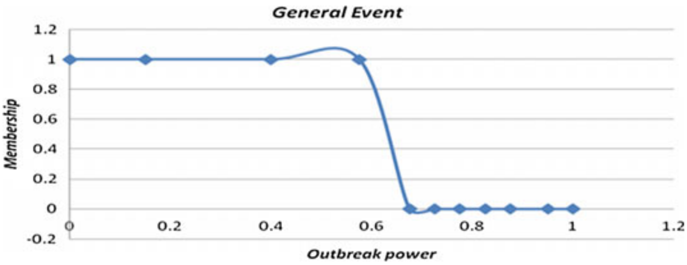

Fig. 3 The membership distribution of average outbreak power

5 Conclusions

In this paper, website preference is formally defined and mined by three proposed strategies which are all explicitly or implicitly based on the three-level networks: website-level, webpage-level and keyword-level. An iterative algorithm is firstly introduced to calculate outbreak power of web events, and increased web pages of events, increased attributes of events, distribution of attributes in web pages and the relationships of attributes are embedded into this iterative algorithm as the variables. By means of prior knowledge, membership grade of web events be-long to each type can be calculated, and then the type of web events can be dis-criminated. Experiments on real data set demonstrate the proposed algorithm is both efficient and effective, and it is capable of providing accurate results of discrimination.

Acknowledgements This work was supported in part by the National Science and Technology Major Project under Grant 2013ZX01033002-003, in part by the National High Technology Research and Development Program of China (863 Program) under Grant 2013AA014603, in part by the National Science Foundation of China under Grant 61300202, in part by the China Post-doctoral Science Foundation under Grant 2014M560085, and in part by the Science Foundation of Shanghai under Grant 13ZR1452900.

References

1. C. Yang, X. Shi, and C. Wei. Discovering Event Evolution Graphs from News Corpora. IEEE Trans. On Systems, Man and Cybernetics—Part A: 39(4):850–863, 2009.
2. Juha Makkonen. Investigation on event evolution in TDT. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language, pp. 43–48, 2003.
3. J. Allan, G. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic Detection and Tracking Pilot Study Final Report. In Proceedings of the Broadcast News Transcription and Understanding Workshop, 1998.
4. Shanshan Qi, Crystal Ip, Rosanna Leung, and Rob Law. 2010. A new framework on website evaluation. In E-Business and E-Government (ICEE), 2010 International Conference on. IEEE, 78–81.
5. Michael J Davern, Dov Te'eni, and Jae Yun Moon. 2000. Content versus structure in information environments: A longitudinal analysis of Website preferences. In Proceedings of the twenty first international conference on Information systems. Association for Information Systems, 564–570.
6. Barbara Poblete and Ricardo Baeza-Yates. 2006. A content and structure website mining model. In Proceedings of the 15th international conference on World Wide Web. ACM, 957–958.
7. Jonathan Chang and David M Blei. 2010. Hierarchical relational models for document networks. The Annals of Applied Statistics 4, 1 (2010), 124–150.
8. Z. Xu et al. Crowdsourcing based Description of Urban Emergency Events using Social Media Big Data. IEEE Transactions on Cloud Computing. doi:10.1109/TCC.2016.2517638.
9. J. Xuan, X. Luo, G. Zhang, J. Lu, and Z. Xu. Uncertainty Analysis for the Keyword System of Web Events. IEEE Transactions on Systems, Man, and Cybernetics: Systems. doi:10.1109/TSMC.2015.2470645.

10. Z. Xu et al. The Semantic Analysis of Knowledge Map for the Traffic Violations from the Surveillance Video Big Data. Computer Systems Science and Engineering, 30(5):403–410, 2015.
11. Z. Xu et al. Crowdsourcing based Social Media Data Analysis of Urban Emergency Events. Multimedia Tools And Applications, doi:10.1007/s11042-015-2731-1.
12. Z. Xu et al. Incremental building association link network. Computer systems science and engineering, 26(3):153–162, 2011.
13. X. Luo, Zheng Xu, J. Yu, and X. Chen. Building Association Link Network for Semantic Link on Web Resources. IEEE transactions on automation science and engineering, 2011, 8 (3), 482–494.
14. C. Hu, Zheng Xu, et al. Semantic Link Network based Model for Organizing Multimedia Big Data. IEEE Transactions on Emerging Topics in Computing, 2014, 2(3), 376–387.
15. Z. Xu et al. Knowle: a Semantic Link Network based System for Organizing Large Scale Online News Events. Future Generation Computer Systems, 2015, 43–44, 40–50.