Future Generation Computer Systems 64 (2016) 175–185

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Discovering the core semantics of event from social media

Weidong Liu [a], Xiangfeng Luo [a,*], Zhiguo Gong [b], Junyu Xuan [a], Ngai Meng Kou [b], Zheng Xu [c]

[a] Shanghai University, Shanghai, China
[b] University of Macau, Macau, China
[c] The Third Research Institute of Ministry of Public Security, Shanghai, China

HIGHLIGHTS

- Proposing a Markov random field based method for discovering the core semantics of event.
- Learning the association relation distribution of event by small scale association relations.
- Maximizing the coverage of association relation distribution by the minimum number of short texts.

ARTICLE INFO

Article history:
Received 20 March 2015
Received in revised form
13 November 2015
Accepted 15 November 2015
Available online 2 December 2015

Keywords:
Core semantics
Semantic link network
Information gradient

ABSTRACT

As social media is opening up such as Twitter and Sina Weibo,[1] large volumes of short texts are flooding on the Web. The ocean of short texts dilutes the limited core semantics of event in cyberspace by redundancy, noises and irrelevant content on the web, which make it difficult to discover the core semantics of event. The major challenges include how to efficiently learn the semantic association distribution by small-scale association relations and how to maximize the coverage of the semantic association distribution by the minimum number of redundancy-free short texts. To solve the above issues, we explore a Markov random field based method for discovering the core semantics of event. This 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. We evaluate our method by comparing with two state-of-the-art methods on the TAC dataset and the microblog dataset. The results show our method outperforms other methods in extracting core semantics accurately and efficiently. The proposed method can be applied to short text automatic generation, event discovery and summarization for big data analysis.

1. Introduction

With booming social media, the data explosion of microblog on blogosphere accompanies with hot events. For example, a large volume of microblogs discussed about "*USA Boston Marathon bombing*", "*the US surveillance program PRISM*" and so on. Given the microblogs/short texts[2] about a concrete event, information about the event is unevenly distributed on these "event messages" since some ones might contain much more important and diverse information (e.g., different event time, locations, participants, processes, and opinions) than others (e.g., redundancy and noises in short texts). Besides, these short texts are globally semantic redundant and locally semantic sparse since many short texts contain the same content and local ones only focus on one aspect of the whole event. Understanding the event concisely and thoroughly is impossible when these redundant short texts may crowd out other ones which contain important and diverse information. For example, when we query by keyword "*Ebola*", Sina Weibo returns redundant Chinese microblogs about "*Ebora of salmon*" and Twitter returns redundant tweets about "*A Italian doctor catches Ebola*" as shown in Fig. 1, which crowd out many important microblogs which discuss about outbreak, spreading and control of event "*Ebola*".

* Corresponding author.
E-mail addresses: liuwd@shu.edu.cn (W. Liu), luoxf@shu.edu.cn (X. Luo), fstzgg@umac.mo (Z. Gong), xuanjunyu@shu.edu.cn (J. Xuan), yb27406@umac.mo (N.M. Kou), xuzheng@shu.edu.cn (Z. Xu).

[1] Chinese microblogging website http://weibo.com/.

[2] There is a word limitation of microblog. For example, each tweet in twitter cannot exceed 140 words.

http://dx.doi.org/10.1016/j.future.2015.11.023
0167-739X/

(a) Results returned by Sina Weibo.

(b) Results returned by Twitter.

Fig. 1. Redundant sentences returned by microblogging services when inputting keyword query “Ebola”.

Herein, how to automatically discover the core semantics of event from big social media data is a challenging problem, since it is time-consuming and unpractical to manually find out the core semantics of event from big media data.

Existing methods to solve this problem are summarized as follows:

(1) Feature-based methods. These methods directly use basic statistic technique on features including word frequency, title words, cure words which are considered for selecting sentences as core semantics [1,2]. Structural features of discourse are used to identify core semantics by rhetorical structure analysis, pragmatic analysis, lexical chain, latent semantic analysis [3]. Besides, more features are used in some specified semantics discovery methods whose features include hashtags, timestamps and emotion labels [4].

(2) Graph-based methods. These methods construct graph where short texts as nodes and text-pairwise relations as edges [5–8]. Top k short texts are selected as core semantics by ranking values of graph-based features or values of Markov random walk on the graph [5–8]. Besides, such methods can be extended into conditional random fields which identify core semantics by labeling sentences, where the sentence label influences the labels of nearby sentences [9].

(3) Clustering-based methods. These methods cluster short texts into different clusters, and then select some short texts from each cluster to represent the semantics of the cluster [10–12]. The clustering methods include hierarchical clustering, partitional clustering and semantic-based clustering [13]. Besides, some priori knowledge or constraint conditions in specified domain are considered in clustering [14,15].

(4) Semantic link-based method. Semantic link-based methods have strong abilities in semantics organization, semantic community discovery and emerging semantics learning/reasoning [16]. Such methods have been used in semantic representation [17], semantic organization [18], semantic interaction [19,20], semantic community discovery [21,22] and semantic linking space for Cyber-Physical Society [23,20].

(5) Other methods. These methods include Bayesian topic model-based methods [24], Neural Networks-based methods, Decision tree-based methods and so on [25,26].

However, these methods have the following limitations:

(1) Semantic association loss. The graph-based and cluster-based methods often use vector space model to represent short texts and use vector-based similarity methods. Obviously, these similarity-based methods lost many semantic association relations;

(2) High computational cost. The time complexity of most the above methods [5–8,10–12], which have to compute text-pairwise similarity, is $O(n^2)$. It is unpractical when the text number is large in big data;

(3) Redundancy-prone results. The above methods pay less attention on the issue of redundancy and result in redundant results since these methods assign almost the same values to alike short texts.

To solve the above limitations, we propose a Markov random field based method for discovering the core semantics of event:

(1) To avoid semantic association loss, our method makes semantic collaborative computation to learn the whole association relation distribution of an event by small-scale association relations.

(2) To reduce computation cost, our method makes probabilistic inference in a limited keyword association link network, rather than text-pairwise computation.

(3) To be free of redundancy, our method proposes information gradient computation by maximizing information gradient of k short texts since information gradient decreases when redundancy increases.

Compared with existing methods, the contributions of our method are summarized as follows:

(1) Our method learns association relation distribution by semantic collaborative computation.

(2) Our method is efficient by probabilistic inference on semantic association link network.