Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 35 (2014) 456 – 463

Procedia Computer Science

18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014

Incident related tweet extraction with density ratio estimation

Hidekazu Yanagimoto[a,*], Suguru Isaji[b]

[a] *Osaka Prefecture University, 1-1, Gakuen-cho, Naka-ku, Sakai, Osaka, 599-8531, Japan*
[b] *NTT docomo, 4-1-22, Onogara-dori, Chuo-ku, Kobe, Hyogo, 651-0088, Japan*

Abstract

These days social media services are widespread and are infrastructure of communication in the Internet. Since Twitter is one of the most famous real-time communication services, we can grasp opinions of crowds in the real world analyzing tweets. There are usually various kind of opinions in Twitter and you need to deal with the opinions carefully. In this paper we focus on tweets on an incident and extract tweets reflecting sufferers' opinions. When a incident happens, vast amount of tweets are created by many Twitter users. We compare tweets by sufferers with ones by others and extract tweets unique to the sufferers with density ratio estimation. In experiments we confirm that our proposed method can extract tweets including sufferers' opinions.

Peer-review under responsibility of KES International.
Keywords: Natural language processing, Social media mining, Density ratio estimation

1. Introduction

These days social media services are widespread and are infrastructure of communication in the Internet. Since Twitter is one of the most famous real-time communication services, we can grasp opinions of crowds in the real world analyzing tweets. There are many non-informative tweets because of limitation of 140 characters. And there are usually various kind of opinions in Twitter and you need to deal with the opinions carefully. We assume such difference among opinions resulted from communities constructed in Twitter based on their homophily. For example, the communities are constructed in the same generation or from persons living in near area and opinions are shared inside the communities. We compare tweets between communities and extract tweets that represent their opinions.

In this paper we focus on tweets on an incident and extract tweets reflecting sufferers' opinions. When a incident happens, vast amount of tweets are created by many Twitter users. Since the users consist of sufferers and non-sufferers, it is difficult to find only sufferers' needs and you can not support them appropriately. Moreover, even if we can extract only tweets generated by sufferers, it is difficult to find only tweets that describe their needs and interests in detail. To achieve the goal we compare tweets by sufferers with ones by others and extract tweets unique to sufferers with density ratio estimation. The density ratio estimation estimates ratio between two probability density distributions

* Corresponding author. Tel.: +81-72-254-9279 ; fax: +81-72-254-9279.
E-mail address: hidekazu@cs.osakafu-u.ac.jp

Peer-review under responsibility of KES International.
doi:10.1016/j.procs.2014.08.126

with linear regression using a kernel function. We assume that tweets by sufferers and non-sufferers are generated under similar but different probability density distributions. In experiments we confirm that our proposed method can extract tweets including sufferers' opinions. Our proposed method contributes to natural language processing and social media mining.

In Section 2 related works are introduced covering researches using Twitter as a social sensor system. In Section 3 the proposed method is described. Especially we explain architecture of the proposed method and density ratio estimation. In Section 4 some results of evaluation experiments are explained. In the experiments we use tweets that shows heavy rain in Yamaguchi and Shimane on July 28, 2013. In Section 5 we describe conclusion and future work.

2. Related works

Our proposed method analyzes tweets and extracts tweets which represent opinions in a target group. The proposed method is one of researches on social media mining using Twitter. In this section we describe researches related to our proposed method. Our proposed method is similar to researches using Twitter as a social sensor system to detect real incidents since both of researches extract opinions from many tweets. In the researches many researchers regard tweets as temporal and spacial data since tweets have a time stamp, location where they are generated, and text describing incidents. Especially in Twitter outputs of sensors are written with natural language although usual sensors output numerical data. To analyze them we have to use natural language processing technique and statistical technique since they are written with a natural language and the amount of them are very large.

Sakaki et al.[1] proposed a system that estimated the center of earthquake and movement of typhoon using Twitter. Since Twitter users generally make tweets describing real-time events, they can estimate the location where they happen dealing with tweets as temporal and spacial data. Speaking concretely, the system used Support Vector Machine to select tweets related to target events and Kalman filter to predict the center of the events. The research used Twitter users as sensors, analyzed outputs written with natural languages and detect real-time events. Hence, it defined that Twitter had the ability to act as a sensor system including many users.

Aramaki et al.[2] reported that influenza epidemics detection using Twitter. They analyzed tweets that mentioned influenza patients and detected influenza epidemics. First they extracted related tweets from a tremendous amount of tweets from Twitter. To realize it they used Support Vector Machine which inputs were some words included in tweets. As a result the approach had advantages in early stage detection. Achrekar et al.[3] reported that the number of influenza related tweets highly correlated with influenza-like illness (ILI) activity in The Center for Disease Control and Prevention (CDC). To predict influenza epidemics they constructed a auto-regression model which inputs were the percentage of physician visits due to ILI in a week and the number of unique Twitter users with influenza related tweets in a week.

Using Twitter, you can detect many kinds of situations in the world regarding it as a social sensor system. The approaches' achieving high performance, they assume that users cover target areas exhaustively and their tweets include reliable information. First twitter users are not distributed uniformly in the world. For example, in big cities there are many Twitter users and they usually generate more tweets related to target events than in countrysides. Hence, tweets generated in countrysides are less reliable than ones in big cities since the number of tweets in countrysides are small and noise easily distorts information. Additionally the number of informative tweets related to an event are small since Twitter users' sensitivity depends on topics. Isaji et al.[4] reported that they analyzed snow related tweets and the number of the tweets depended on how often it snowed in the area. For example, in a snowy area there are less snow related tweets than in a less snowy area although it snowed in both of the areas. Hence, it is desirable that you analyze twitter not using the frequency of tweets simply to avoid such problems.

Isaji et al.[5] proposed an approach to extract tweets unique to a community comparing two communities' generating tweets. They extracted tweets comparing two probability distributions estimated with tweets that the two communities generated. However, the results included many noise data and it needed many human efforts. Moreover, based on term frequency, tweets extracted with the approach are sensitive to a corpus. To extract more appropriate tweets a new approach are developed not depending on the term frequency.

In this paper we propose a new approach to extract tweets unique to a community with density ratio estimation[6]. The density ratio estimation is used in anomaly detection comparing normal data with data including anomaly. It does not need to estimate each density distribution generated from different communities and estimate density ratio directly