Information Processing and Management 52 (2016) 1227–1246

Contents lists available at ScienceDirect

Information Processing and Management

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

Evidential estimation of event locations in microblogs using the Dempster–Shafer theory

Ozer Ozdikis*, Halit Oğuztüzün, Pinar Karagoz

Department of Computer Engineering, Middle East Technical University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 24 July 2015
Revised 23 April 2016
Accepted 10 June 2016
Available online 4 July 2016

Keywords:
location estimation
Microblogs
Event location
Dempster–Shafer theory
Evidential reasoning

ABSTRACT

Detecting real-world events by following posts in microblogs has been the motivation of numerous recent studies. In this work, we focus on the spatio-temporal characteristics of events detected in microblogs, and propose a method to estimate their locations using the Dempster–Shafer theory. We utilize three basic location-related features of the posts, namely the latitude-longitude metadata provided by the GPS sensor of the user's device, the textual content of the post, and the location attribute in the user profile, as three independent sources of evidence. Considering this evidence in a complementary way, we apply combination rules in the Dempster–Shafer theory to fuse them into a single model, and estimate the whereabouts of a detected event. Locations are treated at two levels of granularity, namely, city and town. Using the Dempster–Shafer theory to solve this problem allows uncertainty and missing data to be tolerated, and estimations to be made for sets of locations in terms of upper and lower probabilities. We demonstrate our solution using public tweets on Twitter posted in Turkey. The experimental evaluations conducted on a wide range of events including earthquakes, sports, weather, and street protests indicate higher success rates than the existing state of the art methods.

1. Introduction

Social networks, particularly microblogs, are increasingly used as communication platforms by millions of people worldwide (Java, Song, Finin, & Tseng, 2007). The availability of Internet access and affordable smart phones allows people to easily access online services to share photos, videos, and text messages with others in their network. With its innovative microblogging concept, Twitter is among the most popular social networking platforms enabling users to post 140 character text messages as tweets, and share them with their followers.

Tweets are usually very responsive to the real-world events, sometimes even more immediate than the relay of traditional news resources (Imran, Castillo, Diaz, & Vieweg, 2015). This has led to Topic Detection and Tracking (TDT) techniques that were previously proposed for newspaper articles and blog posts being extended and adapted to perform event detection on Twitter (Allan, 2002; Atefeh & Khreich, 2015; Fiscus & Doddington, 2002; Yang, Pierce, & Carbonell, 1998). In these studies, an event is defined as an activity that happens at a specific time and place. Although state of the art solutions detect events and their times with satisfactory accuracy, we believe that the current solutions about the "place" are still in infancy. In this work, we particularly focus on the location aspects of events in microblogs. We use Twitter as a representative of

* Corresponding author.
E-mail addresses: ozer.ozdikis@ceng.metu.edu.tr, oozdikis@gmail.com (O. Ozdikis).

http://dx.doi.org/10.1016/j.ipm.2016.06.001
0306-4573/

microblogs, due to its widespread usage and easily accessible public data. However, we believe our proposed solution is also adaptable to other microblogs.

Tweets can contain geographical footprints (Stefanidis, Crooks, & Radzikowski, 2013). A tweet becomes geotagged in terms of latitude and longitude if the user posts it using a GPS-enabled device and allows the sharing of this geographic information. Moreover, users may mention places in the content of their tweets, or state their home location in their Twitter profiles. These three attributes of tweets are widely used for their spatial analysis. However, each attribute poses its own challenge. For example, although GPS coordinates provide a precise geographic position on earth in terms of latitude and longitude, this location and that of the event mentioned in the tweet may not be the same. Once an event is broadcast in the media or shared among people via phone calls or retweets, the GPS coordinates of the recently posted tweets may quickly spread across locations that are further away, making the event location more difficult to detect (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013; Sakaki, Okazaki, & Matsuo, 2013).

Challenges related to the tweet content and location attribute of user profiles mostly concern text processing and geoparsing, i.e., relating a given text to spatial locations (Hill, 2006). These attributes are uncontrolled free-text fields. Their quality of content is not as good as that contained in news articles due to the idiosyncratic spellings, unusual writing conventions and abbreviations. Therefore, state of the art NLP parsers do not perform as accurately on tweets (Lingad, Karimi, & Yin, 2013).

In addition to these challenges, users do not have to reveal their GPS locations, or their city of residence in their profiles. They also do not have to mention any location name in their tweets. As a result, uncertainty and a lack of rich and reliable data is a major common problem to be overcome.

Problem definition: Spatial analysis in social networks may focus on several aspects. Some studies aim to infer the location where a photo or tweet has been posted even if the user did not share the GPS data of the mobile device (Lee, Ganti, Srivatsa, & Liu, 2014; Van Laere, Schockaert, & Dhoedt, 2012; Watanabe, Ochi, Okabe, & Onai, 2011). Similarly, there are efforts to assign geographical coordinates to textual resources, such as Wikipedia articles (Van Laere, Schockaert, Tanasescu, Dhoedt, & Jones, 2014). Estimating locations of users by utilizing the content in social networks is another active research area (Chang, Lee, Eltaher, & Lee, 2012; Cheng, Caverlee, & Lee, 2010). In this work, our geospatial analysis focuses on estimating locations of events that can be detected by examining microposts in social networks. Therefore, we describe the problem as follows; given a set of tweets that are presumably about an event, estimate the location of this event by exploiting the available evidence in the tweets and in the profiles of the users who posted the tweets.

This problem is also referred to as *event localization* by Giridhar, Abdelzaher, George, and Kaplan (2015a); Giridhar et al. (2015b). As the definition of an event, we adopt the terminology in the literature, and define it as an activity that happens at a specific time and place (Allan, 2002; Atefeh & Khreich, 2015; Fiscus & Doddington, 2002). Moreover, in this work, we assume that an event might be happening at multiple locations. For example, a hurricane hitting several cities on the coast can be considered as an event with multiple locations.

Contributions: According to our observation, in most of the similar previous studies, the three location-related attributes of the tweets are considered alternatives to each other. If tweets are expected to have been sent from places closer to the event location, then the GPS coordinates of the geotagged tweets might be a fruitful source of information (Crooks et al., 2013; Sakaki, Okazaki, & Matsuo, 2010). For the tweets that are not geotagged, the location attribute in the user profile can give an insight into the location of the event (Achrekar, Gandhe, Lazarus, Yu, & Liu, 2013; Sakaki et al., 2013). Alternatively, if we postulate references to the event location in tweets, then the tweet content can be useful for location estimation (Unankard, Li, & Sharaf, 2015). We believe that the location-related features in tweets are not alternatives for each other, but rather, they complement each other. In this work, we use all three spatial attributes in tweets as evidence sources and combine them systematically in a single model in order to estimate the locations of events, represented by collections of clustered tweets. To this end, we propose applying the theory of belief functions, also called the Dempster–Shafer (DS) theory (Dempster, 1967; 1968; Shafer, 1976). The contributions of this work can be summarized as follows:

- The problem of location estimation for events is investigated using DS theory, which allows us to use the existing evidence pertaining to the location of event in a complementary way and extract belief intervals for the candidate locations.
- The proposed location estimation method is not specific to the event type thus, it does not require any prior event annotation for training. It is experimentally evaluated on a set of tweets posted in Turkey about events of different types, including concerts, sports, street protests, accidents, and earthquakes. The results show that the proposed method can estimate the location of events with higher accuracy in comparison to the existing state of the art methods.
- Estimations are made for locations at multiple granularities, namely at the city and town levels. Accordingly, we define an association of evidence between coarse-grained and fine-grained data based on the mixed class hypothesis in DS theory.
- We demonstrate that the contribution of each attribute in the location estimation problem may change temporally depending on the event type. For some events, GPS coordinates are very reliable for the first few tweets, but they diffuse over time as more tweets are received. For some other types of events, the location references in the tweet content turns out to be more accurate source of evidence over time.
- Since DS theory yields probability intervals for each discrete geographical entity in the domain, all the locations related with a given event are marked accordingly on a map. This view offers an intuitive graphic representation of the geography of the event.