Automatic Event Detection and Storytelling in Social Media

Feng Chen

Carnegie Mellon University

Social media such as Twitter and Weibo are experiencing an explosive growth with billions of global users sharing their daily observations and thoughts. When traditional media is monopolized by close groups, or even sometimes under threat from criminal organizations, people shift to social media as their main information platform. Beyond public interests, event-related microblog can provide highly detailed information for those interested in public health, homeland security, financial analysis, and political study. Social media provide an open platform where people can publish and share their sentiments and opinions online and in real time. Due to the nature of social interactions, information spreads in a speed where traditional media will be difficult to catch up. There is the realty that some messages that propagate through social media may not be reliable and carry a lot of false alarms, but there are measurable differences in the way messages propagate that can be used to classified them as credible or not credible [2]. Therefore, social media could be best suited as a real-time "sensor network" of human activities in cyberspace [1]. By sensing information from social media, it becomes possible to apply machine learning techniques to automatically detect ongoing significant societal activities, trace the origins for storytelling, and even forecast the future developments.

However, there are a number of technical challenges to be addressed. First, the language used in social media is heavily informal, ungrammatical, and dynamic. Given a targeted area of interests, such as public health and transportation, it is challenging to develop an automatic robust content filter that can extract the related information from social media, in which the majority of information is about people's daily activities. Traditional document classification techniques cannot be directly applied because it is difficult to collect sufficient training labels and to adapt to the high degree of dynamicity. For example, for the applications related to civil unrests, it is difficult to find a static vocabulary of related terms. Each specific unrest event may relate to the specific context of background, such as elections, financial crises, crimes, and etc. The related terms will be context dependent and highly dynamic. To address this challenge, we develop a transfer learning based content filter, which transfers labels from news reports to the social media space (e.g., twitter), considers topic modeling to capture the event context, and explores social network relationships for knowledge propagation. Using this approach, it is possible to build a robust content filter by using only a limited amount of labels from news reports, and can be incrementally updated by collecting new labels from news reports. In order to make the approach practical to process the huge volume of social media data, we further develop a stochastic vibrational inference algorithm for the proposed model with a close to linear time cost.

The second challenge is the design of a unified data structure to integrate heterogeneous social media data sources, such as twitter, blogs, news reports, and Facebook. Social media is heterogeneous in nature, since different types of entities are involved, such as user, document, location, time, image, video, and etc. In addition, the data structure should also support the modeling of prior knowledge related to the targeted area of interest. For example, for the study of civil unrest events, we already

have rich domain knowledge about different types of societal activities that have relationships to unrests (e.g., there are high chances of protests near election dates; financial crises may lead to riots, riots, and revolution). To address this challenge, we develop a preliminary framework that has two layers. The top layer is organized by topics, organizations, locations that are observed from social media, and the second and latent layer is organized using an entity or concept graph that models potential events or activities based on our domain knowledge.

The third but not the last challenge is the design of a framework for event storytelling. Based on the above proposed solutions, we consider a design of sliding window based approach. In each sliding window, we first build the unified data structure to model heterogeneous social media data and domain knowledge related the targeted area of interests. After that, we consider context based similarity metrics to connect similar topics, organizations, participants, and geographic locations in adjacent sliding window. The third step is to find pieces of context segments which have high occurrences in historical data or are consistent with the latent entity or concept graph. The visualization of the discovered context segments with high confidence will potentially provide a storyline to explain the origin, evolution pattern, and even future developments of the ongoing events.

The above three challenges and the proposed solutions will be studied in three applications, including disease outbreaks detection, transportation safety analysis, and civil unrests detection using twitter, news reports, and blogs.

References

[1] Ahlqvist, Toni; Bäck, A.; Halonen, M.; Heinonen, S. "Social media road maps exploring the futures triggered by social media". VTT Tiedotteita - Valtion Teknillinen Tutkimuskeskus (2454): 13, 2008.

[2] How Fast the News Spreads Through Social Media, http://blog.sysomos.com/2011/05/02/how-fast-the-news-spreads-through-social-media/

[3] Castillo, Carlos; Mendoza, Marcelo; Poblete, Barbara, "Information Credibility on Twitter," WWW, 2011, pp. 675-684.

[4] Hossain, M. S.; Butler, P.; Boedihardjo, A. P.; Ramakrishnan N.. Storytelling in Entity Networks to Support Intelligence Analysts. In KDD '12, 2012, pp. 1375-1383

[5] Shahaf, D.; Guestrin, C., "Connecting the Dots between News Articles," in KDD '10, 2010, pp. 623–632.