Examining Information on Social Media: Topic Modelling, Trend Prediction and Community Classification

Anjie Fang
Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, UK
a.fang.1@research.gla.ac.uk

ABSTRACT

In the past decade, the use of social media networks (e.g. Twitter) increased dramatically becoming the main channels for the mass public to express their opinions, ideas and preferences, especially during an election or a referendum [5, 7]. Both researchers and the public are interested in understanding what topics are discussed during a real social event [10], what are the trends of the discussed topics [8] and what is the future topical trend [9]. Indeed, modelling such topics as well as trends offer opportunities for social scientists to continue a long-standing research, i.e. examine the information exchange between people in different communities (e.g. in [6]).

We argue that computing science approaches can adequately assist social scientists to extract topics from social media data, to predict their topical trends, or to classify a social media user (e.g. a Twitter user) into a community. However, while topic modelling approaches and classification techniques have been widely used, challenges still exist, such as 1) existing topic modelling approaches can generate topics lacking of coherence for social media data [4, 10]; 2) it is not easy to evaluate the coherence of topics [2, 3]; 3) it can be challenging to generate a large training dataset for developing a social media user classifier. Hence, we identify four tasks to solve these problems and assist social scientists.

Initially, we aim to propose topic coherence metrics that effectively evaluate the coherence of topics generated by topic modelling approaches. Such metrics are required to align with human judgements. Since topic modelling approaches cannot always generate useful topics [1], it is necessary to present users with the most coherent topics using the coherence metrics. Moreover, an effective coherence metric helps us evaluate the performance of our proposed topic modelling approaches.

The second task is to propose a topic modelling approach that generates more coherent topics for social media data. We argue that the use of time dimension of social media posts helps a topic modelling approach to distinguish the word usage differences over time, and thus allows to generate topics with higher coherence as well as their trends. A more coherent topic with its trend allows social scientists to quickly identify the topic subject and to focus on analysing the connections between the extracted topics with the social events, e.g. an election.

Third, we aim to model and predict the topical trend. Given the timestamps of social media posts within topics, a topical trend can be modelled as a continuous distribution over time. Therefore, we argue that the future trends of topics can be predicted by estimating the density function of their continuous time distribution. By examining the future topical trend, social scientists can ensure the timeliness of their focused events. Politicians and policymakers can keep abreast of the topics that remain salient over time.

Finally, we aim to offer a general method that can quickly obtain a large training dataset for constructing a social media user classifier. A social media post contains hashtags and entities. These hashtags (e.g. "#YesScot" in Scottish Independence Referendum) and entities (e.g. job title or parties' name) can reflect the community affiliation of a social media user. We argue that a large and reliable training dataset can be obtained by distinguishing the usage of these hashtags and entities. Using the obtained training dataset, a social media user community classifier can be quickly achieved, and then used as input to assist in examining the different topics discussed in communities.

In conclusion, we have identified four aspects for assisting social scientists to better understand the discussed topics on social media networks. We believe that the proposed tools and approaches can help to examine the exchanges of topics among communities on social media networks.

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
 978-1-4503-5022-8/17/08
DOI: http://dx.doi.org/10.1145/3077136.3084156

REFERENCES

[1] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Examining the Coherence of the Top Ranked Tweet Topics. In *Proc. of SIGIR*.
[2] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Topics in Tweets: A User Study of Topic Coherence Metrics for Twitter Data. In *Proc. of ECIR*.
[3] Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data. In *Proc. of SIGIR*.
[4] Anjie Fang, Craig Macdonald, Iadh Ounis, Philip Habel, and Xiao Yang. 2017. Exploring Time-Sensitive Variational Bayesian Inference LDA for Social Media Data. In *Proc. of ECIR*.
[5] Anjie Fang, Iadh Ounis, Philip Habel, Craig Macdonald, and Nut Limsopatham. 2015. Topic-centric Classification of Twitter User's Political Orientation. In *Proc. of SIGIR*.
[6] Philip D Habel. 2012. Following the opinion leaders? The dynamics of influence among media opinion, the public, and politicians. *Political Communication* 29, 3 (2012), 257–277.
[7] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proc. of WWW*.
[8] Oren Tsur and Ari Rappoport. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proc. of ICWSM*.
[9] Xingqi Wang, Lei Qi, Chan Chen, Jingfan Tang, and Ming Jiang. 2014. Grey System Theory based prediction for topic trend on Internet. *Engineering Applications of Artificial Intelligence* 29 (2014), 191–200.
[10] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Proc. of ECIR*.