Social Media Understanding by Word Cloud Timeline

Virach Sornlertlamvanich
Sirindhorn International Institute of Technology (SIIT), Thammasat University
Pathum Thani 12121, Thailand
virach@gmail.com

Abstract

Text from social media is significant key information to understand social movement. However, the length of the social media text is typically short and concise with a lot of absent words. Our task is to identify the proper keyword representing the message content that we are accounting for. Instead of training the model for keyword extraction directly from the Twitter messages, we propose a new method to fine-tune the model trained from some known documents containing richer context information. We conducted the experiment on Twitter messages and expressed in word cloud timeline. It shows a promising result.

1 Credits

We adopted general Thai word segmentation module to extract the words and generate the key words for a specific domain based on the texts from Wikipedia[1]. The list of key words is then used to query the related tweets through the Twitter search API[2] to collect the related tweets. In this study we propose an effective method to fine-tune the key words extracted from the document texts of Wikipedia to suit the relatively short texts from Twitter. The experiment and implementation have been conducted by Kobkrit Viriyayudhakorn.

2 Introduction

Social media is a massive communication data for understanding the social behavior as well as sensing network is a massive monitoring data for observing the global environment (Gundecha and Liu, 2012). Both are the generated data that reflecting the real-time current situation of society and environment. In the rapid change of the current world, it is necessary to understand the situation and make a suitable response timely. The effect of happening or disaster nowadays has a trend to cause tremendous and pervasive damages. Since Great Hanshin earthquake in 1995, Indian Ocean earthquake and tsunami in 2004, Illinois hurricane Katrina in 2005, Arab spring a series of anti-government protests in 2011 uprising in Tunisia spread out to Yemen, Egypt, Syria, Libya and most of Arab countries, Tohoku earthquake and tsunami in 2011, Occupy Wall Street in 2011, until the recent Thailand coup d'etat in 2014, it is wondered whether we can learn something about these historical events. Focusing on social happenings, it is efficient enough to collect the social media data from the widely used social media applications such as Facebook, Twitter, Whatsapp, Line, or WeChat. Social media are actively used in most of the recent cases (Kaplan and Haenlein, 2010). If we ever view them in a proper dimension it is no doubt that we can somehow forecast, prevent, avoid the happenings by warning or influencing the communities to relief the disaster or the undesirable social situation development. In reality, social media data are vast, noisy, distributed, unstructured, and dynamic.

[1] http://th.wikipedia.org/
[2] https://dev.twitter.com/docs/streaming-api

28th Pacific Asia Conference on Language, Information and Computation pages 2–4

To study the evolution of social behavior on a happening, we analyze the time series of tweets related to the topic of the recent Thailand coup d'etat in 2014. In the 2013 survey[3], there are 12 million twitter users in Thailand with 200,000 active users/day. This means that if we can screen for the related tweets we can observe the movement of the community tie-up.

In our experiment, we estimate the topic related keywords from the target document that we can simply collected from the Internet news. Tweet is a short 140-character text, which is more likely to be a conversational text comparing to the written document, which is a kind of political news or review. There is a difference in the extracted keyword. We therefore apply a technique in GETA (Generic Engine for Transposable Association) called WAM (Word Article Matrix) to expand the set of keyword reflecting the nature of the text from Twitter (Murakami et al., 2004).

The transition of word cloud in a time series can express the social interest at the moment. From the set of related tweets, we extract keywords and express them in a word cloud manner. We then put the word cloud on the time series to create a word cloud timeline. Word cloud (Trant and Wyman, 2006; Kipp and Campbell, 2006) at each moment expresses the social interest, which significantly changes at the time of happening.

3 Keyword expansion

WAM (Word Article Matrix) is a table of weighted relation between document and keyword. Keywords in a document are counted to fill in the table.

WAM is created in Figure 1 (a) when the input documents are word segmented (in case of non-segmented language such as Thai) or lemmatized, and the corresponding keywords are counted. The matrix is used to operate dot matrix with the input of training set of tweets shown in Figure 1 (b). As a result, table of the most associated documents to the training set is obtained. The ranked documents can be cut off by setting up a threshold for the associated value as shown in Figure 1 (c). With another dot matrix in Figure 1 (d) the expanded associated keyword can be obtained with the weight. By training through the set of targeted tweets, the associated keywords in the target domain can be created. Now we can rank the keyword by its associated weight to retrieve the topic related tweets from Twitter.

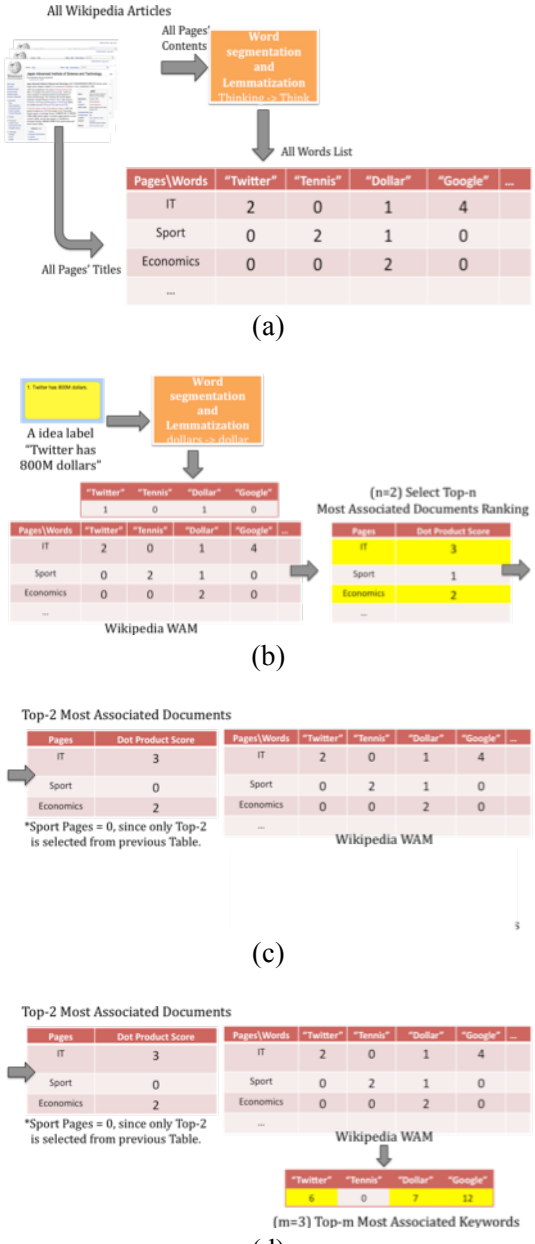


Figure 1: WAM and keyword expansion

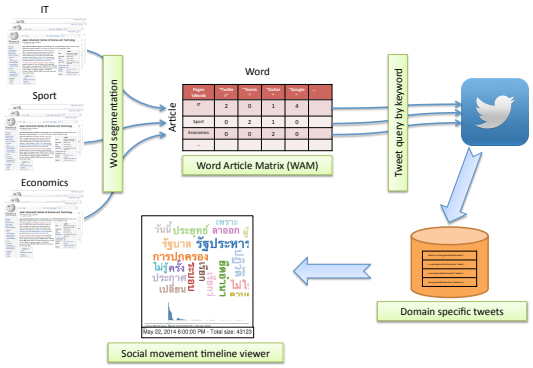
4 Word Cloud Timeline

Figure 2 (a) shows the process in creating Twitter word cloud. A set of topic related documents are collected to create WAM. The WAM is used to expand the keyword from the initial set of tweets. The iterative operation in expanding the keyword

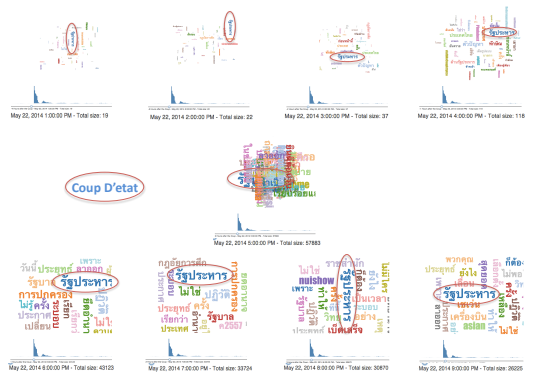
[3] http://www.techinasia.com/thailand-18-million-social-media-users-in-2013/

allows us to query Twitter for better coverage of the tweets. Under the constraint of 100 tweets/query and 7 days search back, we repeatedly issue the query using Twitter search API with the set of keywords (Kumar et al., 2011). As a result, 339,148 tweets centering on the date of coup d'etat on May 22, 2014 are collected. On each day the word cloud is generated to compare on hourly basis.

Investigating the happening that the National Peace Keeping Committee seized power on May 22, 2014 at 4.30 p.m., Figure 2 (b) shows the transition of word cloud around the target time. Significantly the word "coup d'etat" occur in every hour as the most focusing topic. Shortly before the moment of the announcement of seizing the power by the military, it is obvious that the Twitter community is already alert to the possibility of coup d'etat. The density of the keyword increases significantly along the climax moment. The word cloud timeline explicitly shows the critical change point of the happening. Strategic planning can be considered to handle the happening by observing the effectiveness of the timeline of the word cloud.



(a)



(b)

Figure 2: Word cloud and its timeline

5 Conclusion

Word cloud timeline is an effective instrument to monitor the social behavior since the community tie-up of the social media users is reliable. In the modern Internet use, the growth of social media as well as the sensing network is not ignorable. Understanding the movement of the interest in the social media community can be beneficial in the process of strategic planning or decision-making. In coming future, spatial-temporal information can be inclusively considered to create a wider dimension in monitoring the movement and the happening can be understood in a more precise manner.

Acknowledgments

Specially thanks to Kobkrit Viriyayudhakorn for experiment and implementation for this study.

References

Gundecha P., and Liu H. 2012. Mining social media: a brief introduction. Tutorials in Operations Research, Informs, 1(4).

Kaplan A. M. and Haenlein M. 2010. Users of the world, unite! The challenges and opportunities of social media. Business Horizons 53(1):59–68.

Kipp M.E.I., and Campbell D.G. 2006. Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. Proceedings of the ASIST2006.

Kumar S., Zafarani R., and Liu H. 2011. Understanding user migration patterns across social media. Twenty-Fifth International Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence, Palo Alto, CA.

Murakami T., Hu Z., Nishioka S., Takano A., and Takeichi M. 2004. An Algebraic Interface for GETA Search Engine. Proceedings of Program and Programming Language Workshop, Japan.

Trant J., and Wyman B. 2006. Investigating social tagging and folksonomy in art museums with steve. museum. Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland.