2016 10th International Conference on Complex, Intelligent, and Software Intensive Systems

Clusters of Trends Detection in Microblogging: Simple Natural Language Processing vs Hashtags – Which is More Informative?

Tomasz Hachaj
Pedagogical University of Krakow
Institute of Computer Science
2 Podchorazych Ave, 30-084
Krakow, Poland
e-mail: tomekhachaj@o2.pl

Marek R. Ogiela[1, 2]
[1] AGH University of Science and Technology
Cryptography and Cognitive Informatics Research Group
30 Mickiewicza Ave, 30-059, Krakow, Poland

[2] Pedagogical University of Krakow
Institute of Computer Science
2 Podchorazych Ave, 30-084, Krakow, Poland
e-mail: mogiela@agh.edu.pl

***Abstract*— In this paper we introduce the initial proposition and evaluation of the method that enables detection of clusters of trends among microblogging posts gathered from a given social graph. By the cluster of trends we mean the trending words that are popular among same group of people and which describes their common interests. The information about shared interests of group of people in the social network is very important for business. Knowing it we can for example perform directed advertising campaign aimed at single community of people. We validate our approach on large datasets that contains 22 030 252 tweets posted by 20 130 followers of the world-known actress. We found that clusters of trends detection in microblogging with simple natural language processing (namely lemmatization) did not give any valuable information for business. For the other side hashtags frequency filtering and probability conditional probabilities graph clustering resulted in valuable informative about structure of interest in social network.**

Keywords: Microblogging; community detection; social graph; social media; hashtags; lemmatization; natural language processing

I. INTRODUCTION

Microblogging is nowadays a very popular method of sharing people's opinion. The social portals like Twitter or Facebook enables rapid broadcasting the most popular posts using an opinion sharing capability when a user can re-post the message that he or she was interested in. In microblogging the length of the single post is limited to certain maximal number of characters (for example 140 for Twitter). Those posts can also contain special hashtags. Those are single words that begin with # and contain only alphanumeric symbols. The role of hashtags is describing the content of a post. The most common hashtags are often called a trending hashtags and posts that contain those tags are amount posts in trending topics.

Because social networks are used commonly by a very large population of people the massive data stored in historical posts and incoming data streams is very useful for social trends analysis. This potential has yet been spotted and various method that take advantage of social media big data has been reported. The most notable are those that are connecting to marketing. Those are for example analysis of consumers customs [1], [2] or brand popularity modelling [3]. Among most popular method that are used for this type of analysis we can mention sentiment analyze, keywords search, trends analysis [4], association rule mining [5], various social graph analysis metrics [6], [7], text mining and social metrics analysis (shares, likes, follows etc.) [8].

In this paper we introduce the initial proposition and evaluation of the method that enables detection of clusters of trends among microblogging posts gathered from a given social graph. By the cluster of trends we mean the trending words that are popular among same group of people and which describes their common interests. Based on our state-of-the-art revision that type of method has not been yet proposed. The information about shared interests of group of people in the social network is very important information for the business. Knowing it we can for example perform directed advertising campaign aimed at single community of people. When those people are followers (people who have subscription) of a given person we can use the brand of this person to promote our product among this group of people. Provided that this product has strong connection with interests of this group we can anticipate the success of our investment. We validate our approach on large datasets that contains 22 030 252 tweets posted by 20 130 followers of the world-known actress. We have evaluated natural language processing approach and hashtags-based approach for trending (popular) topics detection.

II. MATERIALS AND METHODS

Microblogging posts supply us with very large amount of data on different subjects. We can however assume that people who are somehow connected in the social graph (are friends / followers of each other) share similar interest. Knowing that we can also assume that some words, phrases or hashtags that are connected with that shared interest appears more often than the other in microblogging posts. To detect clusters of topics in this paper we propose two-stage approach. At first we filter the posts to detect trending topics among network. Than we generate the condition – probability based graph that model the connection between those topics. The last step is detecting communities in that graph. The community

978-1-5090-0987-9/16 $31.00 © 2016 IEEE
DOI 10.1109/CISIS.2016.44

detection is optimization processes that maximize so called modularity function [9].

To detect words that describes trends we have used simple natural language processing (NLP) operation namely lemmatization. Lemmatization is a procedure in which words with the same morphological root are identified, despite their surface differences. Lemmatization considers the syntactic category of words, presenting, for instance, different lemmas for a noun or a verb (in the same word family) [10]. Before lemmatization we have removed all hashtags (words that begins with #), all mentions (words that begins with @), common words and links. We have used the list of a 186 of the most commonly used words in every day English vocabulary [11]. Also not alphanumeric letters has been removed. We have used only lemmatized words with length equals or grater 2. We have used OpenNLP algorithms for word tokenizing and lemmatization [12].

In case of hashtags-based approach we simply extracted hashtags that were present in tweets dataset. We have used the same trending words extraction algorithm described in next section both for lemmatized words and hashtags.

A. *Trending Words Extraction*

The trending words extraction can be described by following filtration process:

$$F = \left(\begin{array}{c} f: \exists \mathcal{J} = (j,k,\dots,l): \forall i \in \mathcal{J}, f \in \mathcal{S}_i \wedge \frac{\#(f \in \mathcal{S}_i)}{\#\mathcal{S}_i} >_{T1} 0.01 \\ \wedge \\ \#\mathcal{S}_i >_{T2} 100 \wedge \frac{\#\mathcal{J}}{\#\mathcal{N}} >_{T3} 0.025 \wedge l(f) >_{T4} 4 \end{array} \right) \quad (1)$$

Where:

$\mathcal{J} = (j,k,\dots,l)$ – is a list of indices of followers that have word/hashtag *f* among their tweets;

$\mathcal{S}_i, i \in (j,k,\dots,l)$ – unordered list of words/hashtags that belongs to influencer i;

$\#(h \in \mathcal{S}_i)$ – number of words/hashtags of type f in unordered list that belongs to influencer i;

$\#\mathcal{S}_i$ – number of all words/hashtags in unordered lists that belongs to influencer i;

$\frac{\#\mathcal{J}}{\#\mathcal{N}}$ – number of users followers that have used word/hashtag *f* divided by a number of all followers;

$l(h)$ - the length (in characters) of word/hashtag *f*, excluding #;

$T1, T2, T3, T4$ – threshold values of the model.

First threshold remains only those hashtags that are present in more than $T1$ percent of all posts of that person. The second threshold eliminates users that do not use enough hashtags to use them for statistical computation. They have to post more than $T2$ hashtags. A next threshold governs if a hashtag is present in posts of more than $T3$ percentage of distinct users. The last threshold checks if length of tagging word is greater than $T4$.

B. *Graph Structure Generation*

The trending words discovered in previous section might coincidence with each other. We can present this relation with a graph. The graph we used is directed and fully connected. Its vertices represent the words we found with (1). Each vertex stands for a separate word or phrase. The edge between vertex A and B has weight that equals the conditional probability that word / phrase A is present in the tweet provided that B is also present in a tweet (P(A/B)). Community extraction can be done for example with an algorithm [9] and its implementation [13].

III. RESULTS

We have implemented the prototype of our method in JAVA SE 1.8. We have stored our data in PostgreSQL 9.4. The community structures detection and visualization we used has been implemented in Gephi 0.8.2. For the network layout calculation we used algorithm [14]. After applying that method the length of edge that links to vertices is inversely proportional to conditional probability of occurrences of both tags together in users' blog posts. We have used a followers community of the official twitter accounts of an actress that recently played a role in the newest silver screen production of very famous film series. This person is known worldwide however her profile and fan base is written mostly in English. The data was downloaded by crawler application written in JAVA using Twitter REST API version 1.1. The crawler has operated in parallel requesting Twitter server with several dozens of accounts. The dataset consisted of 22 030 252 tweets posted by 20 130 followers. The total hashtags count was 10 398 879 with 1 389 414 distinct hashtags. Total count of lemmatized words was 160 583 861 with 1 638 939 distinct.

At first we have performed filtration of all words from tweets that was previously lemmatized with the method described in section 2. The words that remain after filtration are (value in brackets is count of words): great (696), happy (424), watch (365), thank (327), please (365), follow (733), thanks (408), think (250), person (599), today (512), video (293), check (305), photo (410). The hashtags that remain after filtration are: nowplaying (384), thewalkingdead (275), spectre (478), soundcloud (270), London (343), starwars (245), fashion (324), scandal (328), xfactor (436), oscars (342).

Despite filtering-out common words as it was described in previous section the remaining phrases do not tell us much about context of the sentence. Due to this fact the further analysis was performed on hashtags set. We have used those filtered hashtags set to generate the graph presented in Figure 1. We have run the community detection algorithm [9] on this graph.

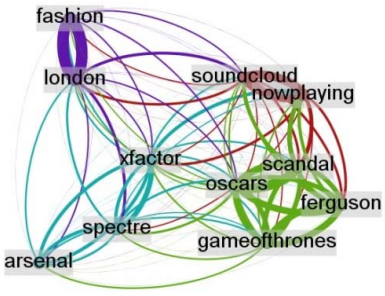

Figure 1. This figure present the result of community detection algorithm on filtered hashtag set. Detected communities are colored-coded.

IV. Discussion

As can be seen in previous section simple NLP based on lemmatization and frequent words occurrence is not capable to provide much useful information about the subject of users posts. Despite common words filtration the words that remains in result set are also common words like great, happy, thanks etc. that might be useful rather for sentiment analysis than automatic posts content tagging.

The hashtags that tweeter bloggers have used in their posts seems to be much more useful. As can be seen in Figure 1 with proposed approach we were able to detect four clusters of topics the followers of actress are talking about. The first cluster consists of two hashtags that coincidence mostly often (fashion and London) and are connected with fashion. The second (soundcloud, nowplaying) is a community of people interested in music. The third (scandal, oscars, ferguson, gameofthrones) consists of people interested in films and serials. The last (xfactor, spectre, arsenal) describe people that posts about current activate of actress and popular music show. All those information reveals what are interests of most important groups of followers of analyzed person. This data might be utilized to plan the projects that the actress would like to get involved to develop the career. Also her managers might contact the advertising companies (or vice versa) to promote the brands that are connected with interest of fans to increase the sale.

V. Conclusions

Concluding, clusters of trends detection in microblogging with simple natural language processing (namely lemmatization) did not give any valuable information for business. For the other side hashtags frequency filtering and conditional probabilities graph clustering resulted in valuable informative about structure of interest in social network. The result we reported in this paper is very important starting point for further researches. In our opinion the further goals of research should be determination what is a minimal subset (random sample) of original dataset for which we get similar results as for whole dataset. Some microblogging social networks have too many participants for analyses (mainly because of bandwidth limitation of API for number of requests) and it has to be sampled. Also analyses of smaller dataset is cheaper (requires less computational power) than complete one. We have also to validate our method on more datasets to prove its usability in practice. We anticipate that with our method it is also possible to evaluate the response of social network on marketing campaigns. After successfully applying this type of campaigning the structure of social network connected with particular person should visualize increase of count of hashtags that describes the promoted product. Also the graph should polarize around that hashtag informing about possible new directions of advertising. Next most probably useful subjects and phrases should be visualized as hashtags that occurs in same cluster as the advertised hashtag.

References

[1] Leticia Vidal, Gastón Ares, Leandro Machín, Sara R. Jaeger, Using Twitter data for food-related consumer research: A case study on "what people say when tweeting about different eating situations", Food Quality and Preference, Volume 45, October 2015, Pages 58–69

[2] Ling Liu, Jing Wu, Ping Li, Qing Li, A social-media-based approach to predicting stock comovement, Expert Systems with Applications, Volume 42, Issue 8, 15 May 2015, Pages 3893–3901

[3] Amir Hassan Zadeh, Ramesh Sharda, Modeling brand post popularity dynamics in online social networks, Decision Support Systems 65 (2014) 59–68

[4] Marc Cheong, Vincent C. S. Lee, A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter, Information Systems Frontiers (2011) 13:45–59, DOI 10.1007/s10796-010-9273-x

[5] Ahmed Abdeen Hamed, Xindong Wu, Alan Rubin, A twitter recruitment intelligent system: association rule mining for smoking cessation, Social Network Analysis and Mining (2014) 4:212, DOI 10.1007/s13278-014-0212-6

[6] Xanat Vargas Meza, Han Woo Park, Globalization of cultural products: a webometric analysis of Kpop in Spanish-speaking countries, Quality and Quantity (2015) 49:1345–1360, DOI 10.1007/s11135-014-0047-2

[7] Yuichi Sasaki, Daisuke Kawai, Satoshi Kitamura, The anatomy of tweet overload: How number of tweets received, number of friends, and egocentric network density affect perceived information overload, Telematics and Informatics 32 (2015) 853–861

[8] Wu Hea, Shenghua Zhab, Ling Li, Social media competitive analysis and text mining: A case study in the pizza industry, International Journal of Information Management 33 (2013) 464–472

[9] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000

[10] Rodrigues, R., Gonçalo Oliveira, H., Gomes, P.: LemPORT: a High-Accuracy Cross-Platform Lemmatizer for Portuguese. In: Pereira, M.J.V., Leal, J.P., Simões, A. (eds.) Proceedings of the 3rd Symposium on Languages, Applications and Technologies (SLATE'14). pp. 267-274. OpenAccess Series in Informatics, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany (June 2014)

[11] Common English words dictionary, http://www.textfixer.com/resources/common-english-words.php (access date: 2015-12-20)

[12] Oficial website of OpenNLP toolkit https://opennlp.apache.org (acces date 2015-12-20)

[13] R. Lambiotte, J.-C. Delvenne, M. Barahona Laplacian Dynamics and Multiscale Modular Structure in Networks 2009

[14] Mathieu Jacomy , Tommaso Venturini, Sebastien Heymann, Mathieu Bastian, ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, PLoS One. 2014 Jun 10;9(6):e98679. doi: 10.1371/journal.pone.0098679. eCollection 2014