

A Study on Detecting Patterns in Twitter Intra-topic User and Message Clustering

Marc Cheong
Faculty of IT
Monash University
Clayton, Australia

marc.cheong@infotech.monash.edu.au

Vincent Lee
Faculty of IT
Monash University
Clayton, Australia

vincent.lee@infotech.monash.edu.au

Abstract— Timely detection of hidden patterns is the key for the analysis and estimating of driving determinants for mission critical decision making. This study applies Cheong and Lee’s “context-aware” content analysis framework to extract latent properties from Twitter messages (tweets). In addition, we incorporate an unsupervised Self-organizing Feature Map (SOM) as a machine learning-based clustering tool that has not been investigated in the context of opinion mining and sentimental analysis using microblogging. Our experimental results reveal the detection of interesting patterns for topics of interest which are latent and cannot be easily detected from the observed tweets without the aid of machine learning tools.

Keywords- Online documents; Group interaction: analysis of verbal and non-verbal communication; Pattern recognition systems and applications.

I. INTRODUCTION

Twitter [1] is a popular microblogging platform that has the sole purpose of letting its users express themselves within 140 characters. It is fast gaining momentum across the world. Originally, it was used for the benign reason of sharing information about themselves with friends and family as a form of online ‘presence’ [2] by answering the simple question: *What are you doing?*

Recently, Twitter has evolved from its basic roots to becoming a facilitator to ‘push the message across’. Now, Twitter is used for more serious purposes such as product marketing, political campaigning, citizen journalism, and market research. On the social end of the spectrum, Twitter is used to connect with other people with same interests, spread Internet-based phenomena (*memes*), and communicate with celebrities.

The aforementioned usages of Twitter make it suitable as a source of Web-based collective intelligence that is useful in gathering opinions and information for effective decision making. Aside from the domain of Twitter message contents (*tweets*) and chatter, the Twitter user base itself gives us insight into the collective wisdom of microbloggers.

In this paper, we use a novel approach to discover user demography, habits, and sentiments when contributing to popular topics of discussion on Twitter. We directly use the Twitter-supplied user information and message information for tweets that match a specified topic and attempt to discover pattern commonalities in the user base and their Twitter habits. This allows us to identify niche communities

which contribute to a topic, cluster them according to similarities – in demography, usage habits, and sentiments, and visualize such clusters.

Our research contributes to the knowledge and practice of microblogging; to our best knowledge, there is no prior work done on the discovery of the latent properties of Twitter communities ‘within’ certain topics.

II. RELATED WORK

Work on Twitter in academia has been limited due to Twitter being a relative newcomer in the social media scene. Related work (since 2008) in studying the dynamics of the Twitter community have been in the domain of user intentions and ‘tweeting’ style (Mischaud [2]; Java *et al.* [3]).

Studies on the emergent properties of Twitter have been conducted by Huberman *et al.* [4], and Java *et al.* [3] who mainly cover the aspect of the social networking pattern exhibited by Twitter users. The conclusions derived from these papers indicate Twitter and other such networks are utilized by users to fulfill information needs, foster connections with others, and share knowledge.

Cheong & Lee [5] have studied the emergent properties of users chatting about ‘trending topics’ (*trends*), in terms of demographics which closely relate to the specific ‘trending topic’. They have also proposed a framework for automated extraction and analysis of demographics and usage habits related to any given topic on Twitter [6].

III. METHODOLOGY

This paper applies Cheong & Lee’s framework [6] in detecting and clustering user/messaging patterns in three corpuses of messages, i.e. political activism, world news, and popular technology. This is based on their data-collection framework using a modified method from [5] in conjunction with the Kohonen Self-Organizing Map [7] algorithm.

This paper builds upon the case studies mentioned in [6] and clarifies certain points not evident in those case studies, by evaluating the effectiveness of visual clustering, comparing it to traditional naïve clustering methods, and re-evaluating the accuracy of prediction of banned users (defined in Section III.B.2).

A. Data Gathering

We use a Perl program based on [6] to harvest Twitter data, as illustrated in Fig. 1.

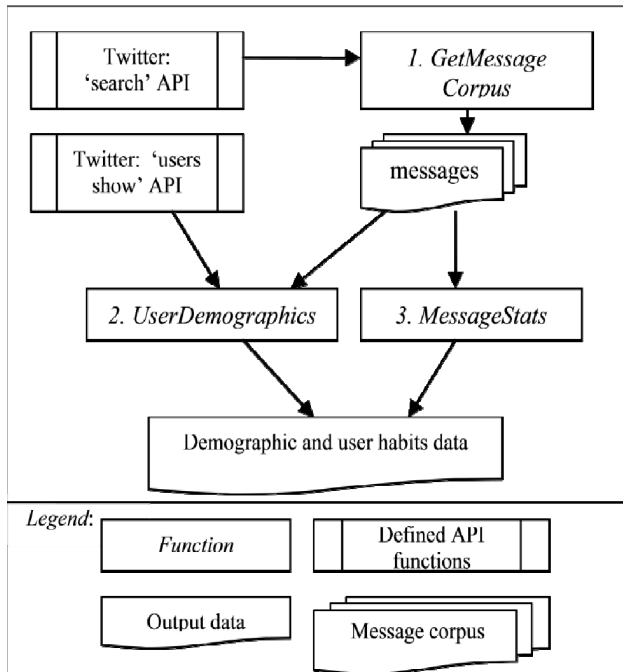


Figure 1. Cheong and Lee's Framework [4].

Twitter's API is searched for messages related to terms of interest (*GetMessageCorpus*) in the figure above. The resulting set of messages obtained is then used to identify users contributing to the topics; these users are then used as queries to the Twitter API to query their user profile information (*UserDemographics* module).

The raw tweets themselves are also processed (via module *MessageStats*) to deduce the style of writing, the type of content shared, and the messages' emergent properties.

B. Attributes of Interest

The Twitter API exposes a rich source of data about both the tweet and its corresponding users; the raw data dump of a tweet as well as its associated metadata is returned as part of an API request.

Certain real-world properties about the users can be inferred from such data. For an extensive discussion about the attributes in the Twitter API, kindly refer to [6].

1) *Message attributes*: The tweet (message) is scanned to determine presence of common message indicators, such as replies, 'retweets' (forwarding messages), and hashtags (keyword tagging).

Presence of information sharing in the form of Uniform Resource Identifier links (URIs) or in the form of pictures is also observed. Finally, the type of device used (e.g. mobile phones) which hint on the usage patterns of Twitter users,

can be derived from the client software used to publish tweets.

2) *User attributes*: Users' web usage habits can be determined by attributes such as the type of website they own, and customizations of their Twitter profiles. Related to this are the attributes of Twitter friends-to-followers ratio, frequency of posts, and account age, which summarizes the user's Twitter usage habits and motivations.

Several important demographic attributes such as specific country and gender can be determined through algorithms proposed by [6] as gender is not stated on Twitter user profiles, and the user-specified locations can be inaccurate or incomplete. In addition, the statuses of rogue users are also listed, as some users have been banned for violating Twitter terms of service such as by spamming and spreading viruses.

C. Clustering, visualization, and analysis

Following prior phases, we obtain corpuses of synthesized and obtained attributes on the user/message base for each of our three case studies (detailed in Section IV).

Each of the corpuses containing the attributes mentioned above are then analyzed via a Self-Organizing Map (SOM)-based clustering and visualizing package: *Eudaptics Viscovery SOMine*.

SOM [7] is a powerful technique based on the artificial neural network concept that projects input from multiple-dimension space into maps of 2-dimensions where similar features are located near each other on the map, which is good for visualization. The resulting SOM clusters are then visualized, allowing us to see what emergent attributes of each cluster of users make it uniquely distinguishable from the other clusters in terms of their contribution to a particular topic

Visualizing a corpus of documents or files helps a user "opportunistically explore" a large amount of information (up to less than a million items) [8], which is another justification of our use of visual clustering to segment our corpus of Twitter messages and metadata.

We then interpret our findings based on the observation criteria found within prior work [5, 9, 10]. Following from that, we compare the result of SOM classification with another clustering algorithm (k-means using minimal Euclidean distance) to see the different ways data can be visually clustered and any similarities between clustering methods; and conduct a follow-up survey of the classification of spam users in our dataset.

IV. EXPERIMENTAL RESULTS

A. Dataset Used

We apply our framework to reveal the emergent properties behind the Twitter user base expressing their views on the abovementioned topics. The following are the topics covered:

- the 2009 Iran Election issue (political activism)
- the iPhone OS 3.0 software launch (popular technology)

- US President Obama's foreign policy stance (an issue of global concern, major world news).

FIGURE I. TOPICS USED AND OVERVIEW STATISTICS

Topic	Messages (excluding bans)	Banned users	Unique users (excluding bans)
Iran Election	4905	0	1953
iPhone	4246	2	3368
Obama	4640	5	3115

Table 1 summarizes the keywords, topics, and vital statistics of the accumulated corpus of data. Note that the occurrences of banned users are as defined in Section III.B.2).

B. Case study on the Iran Election

The sentiments of the users discussing this topic can be broken down demographically into 4 clusters.

In Fig. 2, firstly, we observe the clustered users from multiple countries contributing to chatter about Iran's election aftermath (blue). This user base is relatively new, Iranian web-based Twitter users registered at most for a month, and exhibits frequent patterns of replying (indicating communication).

The second cluster (red) is mainly web users from Iran and other countries with a more 'seasoned' user base with accounts older than 3 months. They contribute sparingly to Twitter, but have a high usage of other social media sites (blogs or social networks).

Thirdly (yellow cluster) are social media users from Iran, the US, and other countries. The high incidents of mobile/social clients and long message sizes lead us to deduce that these users are generating awareness of the Iranian situation via social media, possibly among the younger generation.

Finally, the green cluster identifies users with high variance in Twitter account and nationality, who frequently posted URL links in messages; suggesting that this topic is discussed and shared by a large spectrum of users.

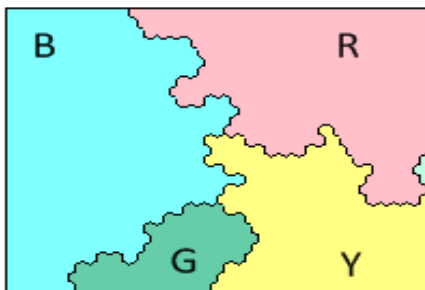


Figure 2. Iran Election aggregated SOM

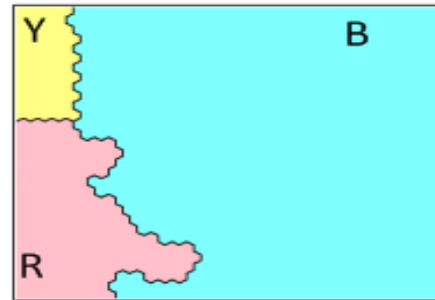


Figure 3. iPhone aggregated SOM

C. Case study on the iPhone software launch

We now map out demographics for people using Twitter as a medium to express their thoughts about a popular technology/consumer product, as clustered in the SOM in Fig. 3.

The blue cluster represents the majority of chatter on Twitter regarding the iPhone. They are male, users with accounts greater than 90 days, coming from countries where the iPhone is marketed. Interesting features of this cluster include usage of mobile devices and high adoption of blogs or social media sites.

The second cluster (red) consists of new accounts and higher ratio of followers to followees with high frequency of Twitter posts per day. Their tweets have links and shared content, and the majority of them have no country and gender specified, which suggest postings by news organizations or news aggregators; an example of this would be the Twitter account of *news.com.au*, an Australian news corporation. A small subset in this second cluster consists of Japanese Twitter users, which is notable as it reflects rather accurately the market sentiment of a new iPhone launching in Japan.

The final cluster (yellow) consists of fresh, one-day-old Twitter accounts with unpopular social connections and lacking in profile customization. These users frequently post more than 50 tweets daily with URIs; suggestive of spam-like behavior [5]. An example of such a suspicious tweet contains a website URI followed by:

"Free iphone's' I just got mine! where's yours!? Huh??"

This illustrates the nature of Twitter spam capitalizing on the popular nature of the iPhone topic.

D. Case study on Obama's foreign policy

The keyword 'Obama', referring to the president of the USA, is tracked on Twitter to study the impact of his June 2009 foreign policy statements on Twitter user sentiment (Fig. 4).

The biggest cluster (blue) consists of American residents genuinely discussing about this topic, as their accounts are mainly more than three months old, their messages are almost always long, and their messaging style is focused towards replies, indicating conversation.

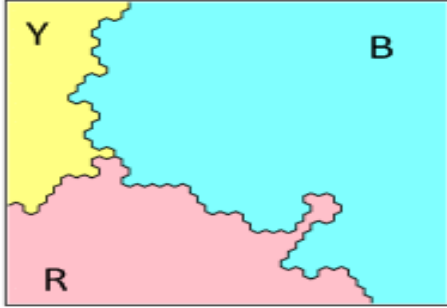


Figure 4. Obama aggregated SOM

The second largest (red) cluster belongs to news sources and opinion leaders. The demographics reveal that users in this cluster have many followers (indicating heavy popularity), predominantly US males, sometimes contributing data from feeds – such as the Really Simple Syndication (RSS) feed format – and frequently publishing URI links in their messages.

Finally, the yellow cluster comprises of mainly new accounts, from indiscernible countries and genders, leading us to suspect that marketing spam or opinion-spam are involved.

An example tweet in the corpus contains a number of hashtags (to gain visibility among users who are searching for particular tags) preceded by the text: “the obamacare news center.” This is clearly an advertising tweet for a website that supports a particular view of Obama; this can be interpreted as a form of opinion-spam.

V. DISCUSSION

Two follow-up studies have been conducted on the case study data set from above, after our original experiments.

A. Revisiting the Efficacy of Novel Spam Detection, via Visual Clustering of Suspect Attributes

First, a random survey (at time of press) on the spam clusters in the case studies above shows that the suspicious user profiles associated with the tweets have been missing or cannot be found.

We surmise that these profiles have been removed by Twitter Inc. due to them identified as being spam users. This is confirmed in some cases where the error pages returned by the Twitter servers state that suspicious activity has been detected. These results show promise in the efficacy of our method in detecting spammers.

B. Visual Clustering and Twitter User Base Pattern Detection with Other Clustering Algorithms

Secondly, we experiment with the visual representation of k-means clustering (using minimum Euclidean distance)

as a side-by-side comparison to our visual clustering for the above case studies. Due to space constraints, we only provide a quick summary of our observation.

The visual representation of cluster assignments observed from k-means clustering approximate those of the SOMs for the case studies mentioned above. This indicates a possibility of using the visual representation of clusters in other clustering algorithms as another way of visualizing and studying patterns in Twitter metadata, users, and tweets.

VI. CONCLUSIONS

Whilst the detection of hidden patterns for topics of interest is also possible based on qualitative subjective judgment, the study has adopted an analytically more objective methodology. Experimental results reveal that the methodology when implemented on various tweets can provide us with a meaningful interpretation.

Future research will be directed to use evolutionary algorithms and visual clustering techniques, for detection of latent patterns and forecasting short- to medium-term trends embedded in the tweets for topics of specific interest.

REFERENCES

- [1] Twitter Inc., “*Twitter*,” Available from <http://twitter.com/>, 2009.
- [2] E. Mischaud, “Twitter: Expressions of the whole self,” Master’s thesis, London School of Economics and Political Science, 2007.
- [3] A. Java, X. Song, T. Finin, and B. Tsen, “Why we Twitter: An analysis of a microblogging community,” in *Proc. 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. Springer-Verlag, 2009, pp. 118–138.
- [4] B. Huberman, D. Romero, and F. Wu, “Social networks that matter: Twitter under the microscope,” 2008, available from <http://ssrn.com/abstract=1313405>.
- [5] M. Cheong and V. Lee, “Integrating web-based intelligence retrieval and decision-making from the Twitter Trends knowledge base,” in *Proc. CIKM 2009 Co-Located Workshops: SWSM 2009*, 2009, pp. 1–8.
- [6] —, “Twitmographics: Learning the emergent properties of the Twitter community,” in *Social Networks Analysis and Mining: foundations and applications*. Springer-Verlag, 2010, vol. (forthcoming).
- [7] T. Kohonen, *Self-Organization and Associative Memory*. Berlin: Springer, 1984.
- [8] S. Lehmann, U. Schwanecke, and R. Dörner, “Interactive visualization for opportunistic exploration of large document collections,” *Inf. Syst.*, vol. 35, no. 2, pp. 260–269, 2010.
- [9] A. Hughes and L. Palen, “Twitter adoption and use in mass convergence and emergency events,” in *Proc. 6th International ISCRAM Conference*, 2009.
- [10] D. Zhao and M. Rosson, “How and why people Twitter: the role that micro-blogging plays in informal communication at work,” in *Proc. GROUP’04*, 2009, pp. 243–252.