Trending Topic Discovery of Twitter Tweets Using Clustering and Topic Modeling Algorithms

Ma. Shiela C. Sapul, Than Htike Aung and Rachsuda Jiamthapthaksin
p5919398@au.edu, g5619309@au.edu, rachsuda@scitech.au.edu
Computer Science Department, Assumption University, Bangkok, Thailand

Abstract— **There is no previous research that compares the results of *k*-means, CLOPE clustering and Latent Dirichlet Allocation (*LDA)* topic modeling algorithms for detecting trending topics on tweets. Since not all tweets contain hashtags, we considered three training data feature sets: hashtags, keywords and keywords + hashtags in this study. Our proposed methodology proved that CLOPE can also be used in a non-transactional database like Twitter data set to answer the trending topic discovery and could provide more topic patterns than *k*-means and LDA. Using additional feature sets has improved the results of *k*-means and LDA, thus, keywords + hashtags can identify more meaningful topics.**

Keywords: Trending Analysis; Twitter Data Mining; Tweets; CLOPE; k-means; Latent Dirichlet Allocation

I. INTRODUCTION

Trending analysis is the process of collecting information and attempting to spot a pattern or a trend on information. It may also be used to verify whether objectives of an organization are achieved. One of the most interesting and important source of information social networking applications used in trending analysis is Twitter. A trend listed on Twitter includes hashtags or taglines that become popular immediately at a particular time and location. Twitter trends are generated based on the tweet counts of a particular hashtag or tag line.

APEC, established in 1989 to address the economic growing needs of the countries in the Asia-Pacific region, consists of twenty one member countries. It aims to promote free trade, a sustainable economic growth and prosperity throughout the Asia-Pacific region. Every year there is an annual APEC Economic Leaders meeting. The leaders of the member countries discuss different issues from counter terrorism to climate change, food security, improving job opportunities for their people and resolving issues among them. People of the member countries may express their opinions through any social media application, but Twitter provides the fastest way of getting opinions or feedback from stakeholders.

The main contribution of this research is to utilize clustering and topic modeling algorithms and three vector sets for discovering trending topic from the APEC 2016 tweets. Several preprocessing steps are introduced for gathering and preparing hashtag and words as featured vectors from the tweets. Then, the clustering algorithms for numerical attributes and categorical attributes, i.e., *k*-means and CLOPE, and the popular topic modeling algorithm, Latent Dirichlet Allocation (LDA) are applied to discover significant patterns. To the best of our knowledge, there is no previous work that performs a comparative study on the three algorithms with the three feature vectors for the trending topic discovery from tweets. Moreover, the results obtained will also reveal the importance of having APEC summit among leaders of the member countries and may provide them substantial information as to the success of the organization in implementing their agendas. We compared the performance of the results of the clustering and topic modeling algorithms to determine which one can effectively perform topic discovery.

The next sections will discuss as follows: Section II describes some of the related works. Section III includes the discussion of the details of the data and the trending topic discovery of tweets using clustering and topic modeling algorithms. Section IV and V present the discussion of the experiment results and conclusion, respectively.

II. RELATED WORKS

Trending of topics in social media provides an important source for current information about events in our world and can also provide us about user's interest, profile and behavior that can be used by any organization in improving or marketing their products and services. Twitter trending list provides the latest topic from across the world. How these trends are determined is very important. Twitter determines its trending topic through the frequency count of hashtags being repeated in the tweets and the period the tweets were posted. In our study, we used two additional feature sets to discover the trending topics. CLOPE and k-means clustering and LDA topic modeling algorithms are used to provide the clusters of the feature vectors sets.

Related studies in utilizing tweets by Abel et al. introduced URL-based and content-based strategies to model user profiling by linking users' tweets with news events from major news websites [1]. They concluded that tweets with URLs that linked to an external news resource are likely related to the linked source and for tweets that do not contain URLS; they utilized the content-based strategy by comparing tweets with every news article, using hashtags and entity based.

Trending analysis in the work of Lu, Yang et al. predicted whether news and tweets is likely to be popular or omitted in the trending lists [2]. They concluded that if more tweets are posted about the news, the longer it will be in the trending list. They used trend momentum technique that can

978-1-5090-4834-2/17/$31.00 ©2017 IEEE

predict the time a topic will be trending; however, they did not consider other factors for a topic to be trending like the number of unique users and the frequency count of the topics over a certain period of time. Our study is not concerned in the changes of the trends or topics; our main goal is rather to determine if we can discover a pattern from the APEC tweets and if the patterns provide the summits agenda. In the work of Lee et.al, understanding the content or what about the trending topic is important [3]. They used trending topics, tweets and topic definition as their datasets and classified tweets into eighteen general categories using Naives Bayes Multinomial (NBM) text based and network based classification method, they want to determine if both classification techniques will have all the five similar topics in their results. In our study, we used clustering and topic modeling algorithms to prove which of the algorithms used can provide more topics and adding additional feature sets to improve the results of *k*-means and LDA models.

Most researches in data clustering and topic discovery utilizes some of the well-known clustering and topic modeling algorithms. Hong et al proposed a model based on statistical and sparse modeling techniques for discovering topics based on the users' geographical location and interest that can be used to create the users' profile [4]. The work of Dela Rosa et al. classified and clustered tweets into categories using *k*-means and LDA [5]. They used hashtags as the key indicator in identifying the topic and identified six pre-defined topics. They find out that *k*-means can provide more meaningful results than LDA and both methods provide poor quality clusters results. On the other hand, our proposed methodology for topic identification used five pre-defined topics and three key indicators. For the semantic association of hashtags to tweets the work of Muntean et al. applied *k*-means clustering [6]. They provided a relation that hashtags may have a connection with the top terms of the cluster and categorized hashtags to different datasets. In our study, we applied also *k*-means but using two additional feature sets for the clusters. The additional feature sets provided more number of instances of the predefined topic in the words + hashtags feature set. For transactional database clustering, Yang et al. applied CLOPE clustering technique in which intra-clustering of transactional items can be improved by increasing the height-to-width ratio of the cluster [7]. They used transactional database rather than analyzing tweets. P. Yong and D.Li applied SCALE clustering algorithm for clustering high dimensional real and artificial trading dataset [8]. It efficiently reduced the search space by using weighted coverage density as the clustering criterion. Their method provided a higher clustering quality than CLOPE. Our work applies CLOPE to prove that it can be also applied to non-transactional datasets. The work of Vosecky, et al. used multiple entities; such as location, organization, general terms and time duration to discover topic patterns and used hashtags as topics labels [9]. It is not clear to which of these entities were derived from Twitter or from URL links provided in the tweets. In contrast to using hashtags as topic labels; hashtags already indicates a topic in a tweet. They compared the results of their experiment using three LDA, *k*-means, DBSCAN clustering and topical modeling algorithms with their Multiple-Faceted Topic Model (MfTM) technique; it has better results for clustering tweets than the three clustering algorithms. In our study we used five pre-defined topics which were derived from the objectives of the summit, making use of hashtags and words as part of the topic detection.

Streaming Twitter data is an important factor for real-time estimation. The work of Miller et al. compared the naïve and history sensitive algorithms in terms performance and space efficiency for real-time estimation of the frequencies of hashtags [10]. Both algorithms performed well in getting the trending tweets; but in terms of space efficiency, naïve algorithm uses lesser space and time. In our proposed model, frequency count is not only the basis of trending topic discovery. We generated clusters and evaluate each cluster with the pre-defined topics.

III. TRENDING TOPIC DISCOVERY OF TWITTER TWEETS

A. STEPS FOR TRENDING TOPICS DISCOVERY

There are 6 steps required for trending topic analysis of Twitter Tweets as shown in Fig. 1: Domain Understanding, Data Collection, Pre-Processing of Data, Feature Selection and Extraction, Data Mining and Interpretation and Evaluation. The details of each step are explained in the following subsections.

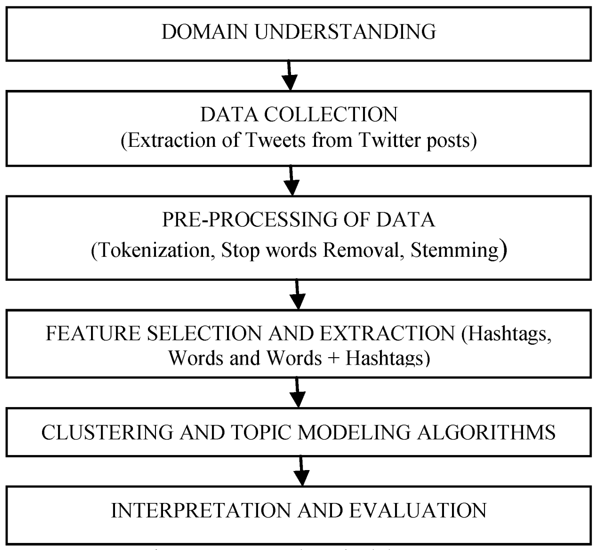

Figure 1. Proposed Methodology

1. DOMAIN UNDERSTANDING

There are two domains used in this study: Twitter data and APEC summit. It initially starts with understanding the purpose of the Twitter application, how messages and opinions from users are posted and how topics are selected to be in the trending list. Second domain understands the purpose of the APEC summit and its agenda. The five pre-defined topics used in this study were derived from the APEC's agenda.

2. DATA COLLECTION

10,668 Twitter tweets related to APEC were collected using Twitter for Google Apps and Twitter Archiver. Twitter for Google Apps connects Twitter with Google apps for accessing the Google sheets. Twitter Archiver saves the tweets retrieved from Twitter in Google sheets.

3. PRE PROCESSING OF DATA

Punctuation marks, @username, RT marked, URLS, slang or Non-English terms, outliers and other non-relevant symbols were removed. Stop words such as is, are, am, he, she, has, its, on, that, were, and etc., were removed because they appear to have of little value in the selection of clusters. Stemming was also applied to reduce inflected words to their word stem, base or root form so as to minimize term duplication.

4. FEATURE SELECTION AND EXTRACTION

Extractions of the terms in the tweets are categorized into three features: Hashtags, Words and Words + Hashtags. These categories will serve as the data sets in this study.

5. DATA MINING

k-means and CLOPE clustering and LDA topic modeling algorithms are applied to the three feature sets used in this study in order to extract patterns for trending topic discovery. More details of the algorithms are provided in Section B.

6. INTERPRETATION AND EVALUATION

The last phase is to determine whether our models provides relevant results based on our objectives by comparing the results derived from *k*-means, CLOPE and LDA.

B. ALGORITHMS

In machine learning, *k*-means and CLOPE clustering analysis is an example of unsupervised learning that do not rely on pre-defined classes and class-labeled training examples. *k-means* is a traditional clustering algorithm that aims to look for each observation of each attribute values and compare to the nearest mean and is used to define a cluster that are similar to one another. CLOPE is a categorical data clustering algorithm considered to be a fast and effective clustering algorithm for transactional data. LDA is a topic modeling algorithm that allows automatic discovery of topics in a document. The *k-means,* CLOPE and LDA pseudocode are given below.

1. Initialize the center of the clusters.
2. Attribute the closest cluster to each data point.
3. Set the position of each cluster to the mean of all data points belonging to that cluster.
4. REPEAT steps 2-3 UNTIL convergence

Figure 2. *k*-means

Phase 1 – Initialization
1. WHILE NOT end of the database file
2. Read the next transaction (***t***, unknown);
3. Put ***t*** in an existing cluster of a new cluster C_i that maximize profit
4. Write **(*t,i*)** *back to the database*

Phase 2 - Iteration

5. REPEAT
6. Rewind the database file
7. *moved* = false
8. WHILE NOT end of the database file read (*t,i*)
9. Move ***t*** to an existing cluster or new cluster C_j that maximize profit
10. IF C_i is not equal to C_j thenTHEN
    write (*t,j*)
    moved = true
11. UNTIL not move

Figure 3. CLOPE clustering

1. FOR all topics in $k \in [1,K]$ DO
   Sample mixture components $\vec{\varphi}_k \sim \mathrm{Dir}(\vec{\beta})$
2. FOR all documents $\boldsymbol{m} \in [1,M]$ DO
   Sample mixture proportion $\vec{\varphi}_m \sim \mathrm{Dir}(\alpha)$
   Sample document length $N_m \sim \mathrm{Poiss}(\epsilon)$
3. FOR all words $\boldsymbol{n} \in [1, N_{m}]$ do
   Sample topic index $z_{m,n} \sim \mathrm{Mul}(\vec{\vartheta}_m)$
   Sample term for word $w_{m,n} \sim \mathrm{Mul}(\vec{\varphi}_{z_{m,n}})$

Figure 4. Latent Dirichlet Allocation

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Tweets about APEC 2016 were gathered during the APEC summit meeting using Twitter for Google Apps and Twitter Archiver. There are three types of tweets: reply, retweet and normal or fresh tweets. All of these types of tweets were considered in the study. Pre-defined list of topics are identified and based on the summit's objectives. These pre-defined topics are: *human, business, security, economy and president*. The collected tweet sets includes the following 16 fields: Date of the Tweet, Number of Follows, Screen Name or Username, Retweets per tweet, Full Name, Number of Favorites or like per tweets, Tweet Text or Tweet Title, Verified or not, Tweet ID, User membership, Application Platform, Location, Number of followers, Bio, Profile image, and Picture.

A. Data Preprocessing and Parameter Setting of the Algorithms

A search rule was used to gather tweets containing the terms #APEC, #APEC2016, APEC and APEC2016. English tweets or words were used as additional qualifiers for filtering tweets. English tweets were the only ones considered in the study due to the language barrier. In the first run, it provided around 8,000 tweets and for every hour, updates on the number of tweets increased by a hundred. The total number of tweets collected before pre-processing is conducted is 10,668 tweets. The tweet text or tweet field was considered for the data processing. The rest of the fields were ignored. Tweet text field contains noise and outliers. A need for data cleaning or text processing was performed. The researcher used R programming tool to remove punctuation marks, URLS, “RT” marker, tags to users, and numbers by replacing it with blank space. Stop words and stemming were also applied. Duplicates in hashtags and words were removed also. WEKA machine learning software is used for data analysis and predictive modeling of the datasets. Java is also used to enhance the feature of WEKA to retrieve the attributes belonging to a cluster.

At the beginning of the experiment, out of the 10,668 tweets gathered, there are 8,271 retweets. A frequency count of greater than or equal to ten is used as the threshold value for each dataset. After the threshold size was identified, duplicates were removed and a total of 733 unique terms (hashtags and words) were used to create three datasets: Hashtag, Words and Words plus Hashtags. The details of the three dataset are presented in Table I. The summit’s agenda was the basis for identifying the five pre-defined topics. If the number of pre-defined topics is increased, there is a possibility that a cluster may discover more topics.

Table I. Details of the Three Datasets

Datasets	No. of Terms	No. of Terms Threshold of 10 (frequency count)	No. of Terms with no Duplicates
Hashtags	881	94	82
Words	4,255	813	651
Words +Hashtags	5,136	907	733

For the *k*-means and LDA experiments were run using *k* equals to 5 and 10, where *k parameter* denotes the number of clusters. On the other hand, CLOPE does not require specifying this parameter; the number of clusters is automatically generated by the algorithm. In addition, the numbers of terms per group of clusters were set to 5, 10 and 20.

B. Experimental Results

In order to validate the results obtained by the algorithms, three research candidates manually scanned through the tweets datasets and are able to identify five trending topics as shown in Table II. This information will be used as a ground truth to be compared with the probable themes or topics discovered by the algorithms later on.

Table II. Trending Topic List

Topic no.	Trending Topic Title	Remarks
1	Human	About human capital development, job and skills
2	Business/SME	About business or fostering small and medium enterprise
3	Security	About security, climate change, sustainable and resilient community
4	Economy	About enhancing the regional economy
5	President	About topics or issues among leaders

Table III. Using Hashtag in *k*-means, CLOPE and LDA to Generate Five Clusters (Trending Topics)

k-means		CLOPE		LDA	
Hashtags	**Trending Topics**	**Hashtags**	**Trending Topics**	**Hashtags**	**Trending Topics**
#Obama #Putin #cdnpoli #Peru #XiJinping	-	#Putin #XiJinping #Lima #Duterte #Obama	-	#XiJinping #Periscope #Lima #Xi #Abe	-
#Putin #Duterte #Lima #Trump #XiJinping	-	#XiJinping #Obama #Putin #Lima #Peru	-	#Lima #Trump #Putin #TTIP #Brexit	Economy
#XiJinping #Lima #Peru #Duterte #China	-	#FTAAP #XiJinping #Putin #ImpactY-ourWorld #India	Economy	#XiJinping #Peru #China #TPP #FTAAP	Economy
#Putin #Russia #US #Russian #Trump	-	#WeHadE-nough #China #EAEU #SilkRoad #trade	Economy	#Putin #Duterte #AbuDhab-i #COP22 #AskStat	Security
#Lima #XiJinping #Periscope #Abe #Obama	-	#China #Freetrade #FTAAP #TPP #trade	Economy	#Obama #Putin #US #Russian #Russia	-

Due to space limitation, we only show details of five clusters with 5 terms to illustrate the results of the experiments. Table III compares the result of the three algorithms using the hashtag dataset and list five terms for each cluster. Both *k*-means and LDA applied *k* equal 5, where *k* is the number of clusters. For CLOPE, the algorithm itself generates the number of clusters, and result of *k* equal 10. We used only the top 5 clusters for CLOPE for the comparison in this table. Results show that CLOPE and LDA have the same number of trending topics and economy is the prevalent topic. For CLOPE result, three of its clusters discovered a pattern for the topic “economy.” In LDA, two of its clusters discovered a pattern for the topic “economy” and one pattern for “security.”

Table IV. Using Words in *k*-means, CLOPE and LDA to Generate Five Clusters (Trending Topics)

k-means		CLOPE		LDA	
Words	**Trending Topics**	**Words**	**Trending Topics**	**Words**	**Trending Topics**
trade Asia pacific Leader free	**Economy**	**summit President Leader Meeting fight**	**Security**	**Meeting China during POTUS final**	-
Meeting President Leader summit Chinese	-	**Meeting japan Abe Shinzo Minister**	-	**President Chinese Peru summit CEO**	-
Leader connect protectioni -sm focus amid	**Security**	**President Meeting China peru Chinese**	-	**Japan says Abe Shinzo PM**	-
say Leader world forward want	-	**Meeting Leader economic Family during**	**Economy Security**	**Leaders Economic photo talk briefing**	**Economy**
firm confirm russian relations willingness	-	**Meeting President courteous China Leader**	-	**summit trade global fight Minister**	**Economy Security**

Table IV compares the result of the three algorithms using the Words dataset and list five terms for each cluster. Both *k*-means and LDA applied *k* equal 5, where *k* is the number of clusters. For CLOPE the algorithm itself generates the number of clusters, and result of *k* equal 16. We used only the top 5 clusters for CLOPE for the comparison in this table. Results show that CLOPE and LDA have the same number of trending topics. Both CLOPE and LDA have one cluster that discovered two topics, which are "economy" and "security." Our result shows that a cluster may discover more than one topic even if fewer terms are used.

Table V. Using Words + Hashtags in *k*-means, CLOPE and LDA to Generate Five Clusters (Trending Topics)

k-means		CLOPE		LDA	
Words + Hashtags	**Trending Topics**	**Words + Hashtags**	**Trending Topics**	**Words + Hashtags**	**Trending Topics**
#Obama Meeting #Putin Barack President	-	**#XiJinping China global economic advocates**	**Economy**	**#APEC #XiJinping President #Lima XHNews**	-
President #Putin Meeting #XiJinping #Lima	-	**Leader Meeting economic summit #Putin**	**Economy**	**#APEC Economic photo China Meeting**	**Economy**
fight protection- sm Leader vow trade	**Economy Security**	**#Lima global opportunity issue discuss**	**Human**	**#APEC final Obama POTUS says**	-
Leader say forward pacific partnershi- p	**Economy**	**economic Leader Meeting Family #XiJinping**	**Economy Security**	**Meeting #Putin #APEC Putin summit**	-
firm #Putin #Trump relations confirm	**President**	**#XiJinping President trade Leader Meeting**	**Economy**	**#APEC Leaders Meeting Duterte trade**	**Economy**

Table V compares the result of the three algorithms using the Words + Hashtags dataset and list five terms for each cluster. Both *k*-means and LDA applied *k* equal 5, where *k* is the number of clusters. For CLOPE the algorithm itself generates the number of clusters, and result of *k* equal 15. We used only the top 5 clusters for CLOPE for the comparison in this table. Results show that *k*-means discovered more patterns than LDA. CLOPE has the highest number of discovered topics. Economy is the most prevalent topic in the three algorithms. Using Words + Hashtags improved the results both in *k*-means and CLOPE, more clusters discovered topic patterns compared in the previous results that use only Hashtags or Words.

Results in table III, IV and V shows that more topics can be discovered in the Words and Words + Hashtags datasets. It also proved that words may further explain the abstract concept that a hashtag may represent.

Table VI. Comparisons of Five Clusters Generated by Using *k*-means, CLOPE and LDA

		KMeans	CLOPE	LDA
No. Trending Topics	Hashtags –five terms	-	3	3
	Hashtags –ten terms	3	7	8
	Hashtags –twenty terms	6	12	10
	Words –five terms	2	3	3
	Words –Ten terms	6	7	4
	Words –Twenty terms	4	10	6
	Words + Hashtags – Five terms	4	6	2
	Words + Hashtags - Ten terms	5	10	6
	Words + Hashtags - Twenty terms	5	14	6

Table VII. Comparisons of Ten Clusters Generated by Using *k*-means, CLOPE and LDA

		KMeans	CLOPE	LDA
No. Trending Topics	Hashtags –five terms	7	6	5
	Hashtags –ten terms	9	9	7
	Hashtags –twenty terms	9	18	18
	Words –five terms	5	7	4
	Words –Ten terms	6	14	7
	Words –Twenty terms	6	27	9
	Words + Hashtags – Five terms	9	8	6
	Words + Hashtags - Ten terms	10	18	12
	Words + Hashtags - Twenty terms	13	29	13

Table VI summarizes the comparison of using five, ten and twenty terms for *k* equal 5, where *k* is the number of clusters. Result shows that if we add more terms, our model can discover more patterns. For example in CLOPE, using hashtag dataset, the number of topics discovered almost doubled for the ten and twenty terms. We also performed other experiments to generate 3 and 10 clusters. We also perform other experiments to generate 10 clusters with 5, 10 and 20 terms. However, due to space limitation, we only present the summary of the results as shown in Table VII. Our results in both tables VI and VII coincides in such a that if more terms are considered then it can discover more topics and in the three algorithms used CLOPE discovered more topics compared with the K-Means and LDA both in the words and words + hashtags datasets. Applying more feature sets for *k*-means also improves its performance, where it can identify more topics in some instances compared with LDA. It should be noted that for the results generated by using CLOPE with Words and Words + Hashtags in table VII, the total number of topics is more than the number of terms because there are clusters that discovered more than one topic. Table VII also show that adding more clusters increases also the number of patterns discovered.

V. CONCLUSION

This paper shows the potential of applying *k*-means, CLOPE clustering and LDA topic modelling algorithms in discovering trending topics in tweets. Previous studies results can be improved by adding additional feature sets like the Words and Words + Hashtags. CLOPE has also the potential in clustering non-transactional data and can identify more topics compared with the two other algorithms. Our results showed that if more terms are used for *k*-means CLOPE, and LDA, more pre-defined topics can be discovered. Further refinement in the pre-processing steps is to be considered to identify more vague and generic hashtags or words so the number of attributes can be lessened and only hashtags and words that are specific will be used, and to consider the use of test set to assess further the data mining models used in this study. Hashtags are used by Twitter as a criterion for the trending list, but words may further explain the abstract concept that a hashtag may represent. Using Words + Hashtags gives more meaning than using only Hashtags or Words. For the stakeholders' opinion, the most prevalent topic they felt that were given emphasis by world leaders are economy and security. Our proposed methodology can be applied for other contents in Twitter and has the potential to detect topic discovery in any domain.

VI. REFERENCES

[1] Abel, Fabian, et al. "Semantic enrichment of twitter posts for user profile construction on the social web." *Extended Semantic Web Conference*. Springer Berlin Heidelberg, 2011.

[2] Lu, Rong, and Qing Yang. "Trend analysis of news topics on twitter." *International Journal of Machine Learning and Computing* 2.3 (2012): 327

[3] Lee, Kathy, et al. "Twitter trending topic classification." *Data Mining Workshops (ICDMW), 2011 IEEE 11th international conference on. IEEE*, 2011.

[4] Hong, Liangjie, et al. "Discovering geographical topics in the twitter stream." *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012

[5] K. Dela Rosa, R. Shah, B. Lin, A. Gershman, R. Frederking, "Topical Clustering of Tweets," SWSM'10, July 28, 2011, Beijing, China. Copyright 2011 ACM 1-58113-000-0/00/0010

[6] Muntean, Cristina Ioana, Gabriela Andreea Morar, and Darie Moldovan. "Exploring the meaning behind twitter hashtags through clustering." *Business Information Systems Workshops*. Springer Berlin Heidelberg, 2012.

[7] Yang, Yiling, Xudong Guan, and Jinyuan You. "CLOPE: a fast and effective clustering algorithm for transactional data." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002.

[8] Yong, Peng, and Di Li. "Clustering method study on high-dimensional trading data." *ISSN 1796-2048 Volume 9, Number 3, March 2014* 9.3 (2014): 340

[9] Vosecky, Jan, et al. "Dynamic multi-faceted topic discovery in twitter." *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 2013

[10] Miller, Evan, Kiran Vodrahalli, and Albert Lee. "Estimating Trending Topics on Twitter with Small Subsets of the Total Data." (2015).

[11] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3. Jan (2003): 993-1022.

[12] Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi. "Learning to classify short and sparse text & web with hidden topics from large-scale data collections." *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008.

[13] Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 11 November 2016]

[14] Twitter Archiver, https://chrome.google.com/webstore/detail/twitterarchiver/pkanpfekacaojdncfgbjadedbggbbphi?hl [Accessed 19 November 2016]