Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6213–6219
Marseille, 11–16 May 2020

Twitter Trend Extraction: A Graph-based Approach for Tweet and Hashtag Ranking, Utilizing *No-Hashtag* Tweets

Zahra Majdabadi, Behnam Sabeti, Preni Golazizian, Seyed Arad Ashrafi Asli
Omid Momenzadeh and Reza Fahmi
Miras Technologies International
Number 92, Movahed Danesh St., Tehran, Iran
{zahra, behnam, preni, arad, omid, reza}@miras-tech.com

Abstract

Twitter has become a major platform for users to express their opinions on any topic and engage in debates. User debates and interactions usually lead to massive content regarding a specific topic which is called a *Trend*. Twitter trend extraction aims at finding these relevant groups of content that are generated in a short period. The most straightforward approach for this problem is using *Hashtags*, however, tweets without hashtags are not considered this way. In order to overcome this issue and extract trends using all tweets, we propose a graph-based approach where graph nodes represent tweets as well as words and hashtags. More specifically, we propose a modified version of *RankClus* algorithm to extract trends from the constructed tweets graph. The proposed approach is also capable of ranking tweets, words and hashtags in each trend with respect to their importance and relevance to the topic. The proposed algorithm is used to extract trends from several twitter datasets, where it produced consistent and coherent results.

Keywords: Twitter Trend Extraction, Tweet Clustering, Graph Representation, EM Algorithm

1. Introduction

Today several platforms are offering users the opportunity to express their opinions on emerging topics related to ongoing events in their lives as well as national and international matters. Among these platforms, Twitter has become a major player, in which users tend to express their opinions on any topics and engage in debates posting millions of tweets every day. User debates and interactions usually lead to massive content regarding a specific topic which is called a *Trend*. Twitter trends are specifically important due to their influence on public opinion and their ability to affect politicians, governments, organizations, companies and almost any other entities. Twitter trend extraction, as the name suggests, is aimed at detecting ongoing trends among millions of tweets that are posted rapidly. Being able to automatically detect and extract these trends is appealing to anyone interested in analyzing hot topics and ongoing online debates.

There are two main approaches for automatic trend extraction: embedding based and graph-based. Embedding based approaches try to find a vector representation of each tweet and detect groups of similar tweets that could indicate trends. On the other hand, in graph-based methods, the goal is to represent tweets as a graph and employ graph clustering algorithms in order to extract similar sub-graphs as trends.

Many twitter users tend to use hashtags, which indicate the main topic of their tweets. Hashtags can also be very useful to find similar tweets and follow debates related to specific topics. The issue with this strategy is that not all users tend to use hashtags and this will lead to ignoring all no-hashtag tweets for detecting trends.

In this research, tweets are represented as a graph that includes three types of nodes: tweet, word, and hashtag. Tweet nodes are connected to their corresponding words and hashtags with the weight defined as *TF-IDF* scores. Word and hashtag nodes have also inter-connections based on their co-occurrence scores. After constructing the graph representation of tweets, a graph clustering algorithm is employed to extract similar tweets based on their neighborhood. This process will lead to finding similar tweets that have some equivalent words and hashtags in common. The proposed method is inspired by the *RankClus* algorithm which was initially proposed to cluster graph-structured data (Sun et al., 2009). This method is also capable of ranking nodes in each extracted cluster which is especially useful in this research line since it provides the most important tweets, words, and hashtags in each trend.

The proposed method is extensively evaluated on two Persian and English sets of tweets, where it produced consistent and coherent trends as they were subjectively assessed. We have demonstrated the extracted trends using top tweets, words, and hashtags in the Results section. Our contributions in this research are as follows:

- We used a graph representation employing three types of nodes; namely tweet, word, and hashtag. We also employed *TF-IDF* and co-occurrence scores to construct the graph edges.
- We employed a graph clustering method inspired by *RankClus* algorithm which is capable of clustering nodes as well as ranking them in each cluster.
- We also proposed a post-processing step that utilizes cluster scores in order to prune the extracted clusters and keep the consistent ones.
- The proposed method is evaluated on Persian as well as English tweets, in order to demonstrate the language independence property of our approach.

The rest of the paper is organized as follows. Some previous researches and studies on tweet clustering and trend extraction are reviewed in Section 2. Then in section 3, the

proposed method is introduced in detail. Section 4, 5 are dedicated to experiments and results of our method. Finally section 6 concludes the paper.

2. Related Work

Two main approaches have been proposed for tweet clustering and twitter trend extraction. One relies on constructing a graph to represent tweets and then employs graph clustering algorithms for trend extraction. The other approach is based on converting each tweet to an embedding vector and clusters the generated embedding vectors to detect trends.

A summarization method was proposed by *Duta et. al.*, which constructs a graph of tweets by generating semantic similarity between tweets utilizing *WordNet* anthology (Miller, 1995). Community detection algorithms are then employed for tweet clustering and summary generation (Dutta et al., 2015). Co-occurrence score can also be used as weights in the constructed graph. *Kim et. al.* utilized this approach by constructing a graph using only frequent words as nodes (Kim et al., 2014). Strongly related groups of words are then extracted by applying maximum *k-clique* algorithms. The generated words represent twitter trends and can be used to generate trend summary. Another method for tweet graph construction is to use cosine similarity for each pair of tweets based on their *TF-IDF* vector representations. *Manaskasemsak et. al.* proposed to use this approach for graph construction and then employed *markov* clustering algorithm for detecting similar tweets (Manaskasemsak et al., 2016). The same methodology was also employed by *Kim et. al.* where they extracted core topics using graph clustering algorithms (Kim et al., 2012).

The introduced papers, so far, only used tweets to construct the graph representation. Another approach is to also include users and their interactions in the final representation. This method was employed by *Cataldi et. al.* where they proposed to construct a directed graph of twitter users and apply *Page Rank* (Page et al., 1999) algorithm to calculate the ranking of users (Cataldi et al., 2010). They model the life cycle of words based on users' ranking and utilized this life cycle to temporally analyze the emerging topics and trends. The same ranking policy was also employed on a heterogeneous network of tweets and users, where the interactions between users are utilized for ranking and clustering the tweets (Prangnawarat et al., 2015).

A Hashtag Graph-based Topic Model was proposed by *Wang et. al.*, where tweets are projected into a weighted hashtag graph (Wang et al., 2014). In this model, tweets are linked directly to their associated hashtags which also produces an indirect relation to other similar hashtags (that did not appear in tweets). The model then jointly estimates the probability distribution of hashtags over topics and the distribution of topics over words. The final model is then used to detect hot emerging topics from tweets. *Wang* also suggested a two-stage hierarchical topic modeling for tweet topic extraction (Wang et al., 2017). In this model, Gibbs Sampling algorithm is applied to find the first level clusters. Tweets in each cluster are then merged to form virtual documents, which are then fed to another topic modeling algorithm to find new and meaningful tweet clusters. The same policy is also proposed by *Ifrim et. al.* which uses *dendogram* cutting on *tweet-by-term* matrix to extract clusters (Ifrim et al., 2014).

One problem with using hashtags for trend extraction is that users usually tend to use different hashtags regarding the same topic. This phenomenon will lead to several hashtags representing the same trend, which need to be merged in order to produce coherent trends. *Muntean et. al.* addressed this issue by making virtual documents containing all tweets with the same hashtags (Muntean et al., 2012). *K-Means* algorithm is then applied to cluster these virtual documents and find semantically similar hashtags. The other issue with hashtags is that several users do not use them in their tweets and this will lead to ignoring all *no-hashtag* tweets if one is only using hashtags for trend detection. *Rosa et. al.* proposed a supervised method utilizing hashtags as labels in order to also classify *no-hashtag* tweets (Rosa et al., 2011). This approach will also designate hashtags to these kinds of tweets addressing the mentioned issue.

Another aspect of trend detection is tweet ranking which is not addressed before. More specifically, one needs to know the most important tweets and hashtags in each trend in order to fully understand and analyze the extracted trends. In this research, we propose a method for trend detection as well as tweet and hashtag ranking. Our proposed approach is also capable of utilizing *no-hashtag* tweets.

3. Proposed Approach

In our proposed model, tweets are represented as a heterogeneous graph, in which there are three types of nodes: tweets, words, and hashtags. After representing tweets using a graph structure, a ranking based clustering algorithm is employed for detecting relevant nodes in the graph which are considered as the trends. In this approach, since the graph is constructed using only tweets, there is no need for user data collection. This algorithm finds clusters of tweets that have some words and hashtags in common. We assume that tweets with shared words and hashtags are about a specific trend. Therefore, the algorithm utilizes all tweets: with or without hashtags. Finally, a scoring algorithm evaluates each cluster and finds the most coherent trends.

The proposed model is described in three sections:

- Graph Construction: describes the graph nodes and edges.
- Graph Clustering: describes the ranking and clustering algorithm.
- Cluster Scoring: describes the scoring algorithm used for finding the most coherent trends.

3.1. Graph Construction

Tweets are represented as a tri-type weighted graph where node types are: tweets, words, and hashtags. As a result, there are six types of edges between nodes: tweet-tweet, tweet-word, tweet-hashtag, word-word, word-hashtag, hashtag-hashtag. In this research, we only consider edges between words and hashtags as well as edges

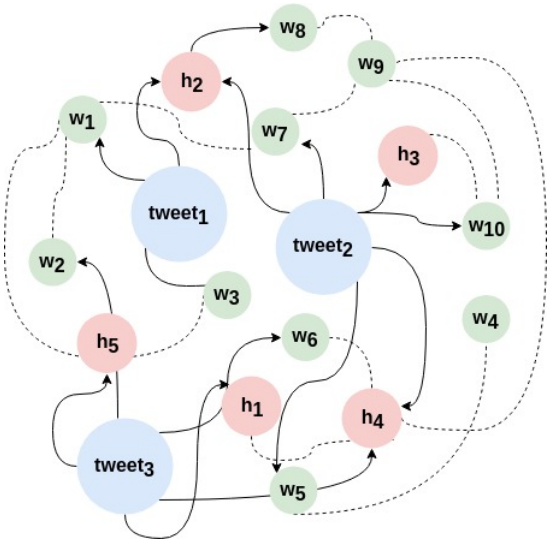

Figure 1: Graph representation of tweets: the graph contains three node types representing tweets, words and hashtags respectively. Graph edges are weighted based on co-occurrence scores.

between tweets and words and tweets and hashtags. A sample graph is illustrated in Figure 1.

Weights between tweet nodes and word nodes are calculated using their corresponding Term Frequency-Inverse Document Frequency (TF-IDF) scores. Also, Weights between two words, two hashtags, and a word and a hashtag are calculated by their co-occurrence score which means the number of times they appear in the same tweet. In order to emphasize on hashtags, the weight between a hashtag and its corresponding tweet is set to the maximum value, in this case, 1. This way the algorithm pays more attention to the hashtags in tweets since they carry helpful information about trends. The weighting policy for graph construction is summarized in Equation 1 (where t means tweet, h means hashtag and w means word).

$$e_{ij} = \begin{cases} Co-Occurrence(i,j) & i,j \in \{w,h\} \\ TF-IDF(i,j) & i \to t, j \to w \\ 1 & i \to t, j \to h \\ 0 & O.W. \end{cases} \quad (1)$$

3.2. Graph Clustering

In order to cluster nodes in the graph, *RankClus*, which is a ranking based graph clustering algorithm, is utilized (Sun et al., 2009). *RankClus* works on heterogeneous graphs and integrates ranking and clustering together.

The idea behind this algorithm is that better clustering leads to better ranking and vice versa. The node type that we want to cluster is called the target type and the other ones are attribute types. For each arbitrary cluster, two ranking functions are defined: one is conditional rank which determines how much the attribute type nodes are scored, the other one is within-cluster rank which is calculated by ranking scores of target type nodes in the cluster.

The algorithm works as follows: for each cluster, conditional ranks of attribute typed nodes are represented as a rank distribution. Then, K rank distributions can be used to build a mixture model whose goal is to find the best component coefficient score for each target node to be in a certain cluster. Next, each target node will be represented in a K dimensional vector, in which each value shows the component coefficient score of each cluster. Then the distance between all target nodes and clusters will be calculated and each node will be assigned to the nearest cluster. This procedure will be repeated until clusters converge. The readers are referred to the original *RankClus* paper for an in-depth description of the algorithm (Sun et al., 2009).

In this research tweet is the target type, since we are interested in clustering tweets. We also consider words and hashtags as attribute type nodes in the graph. In our model, conditional rank for a given tweet is defined according to Equation 2 (t: given tweet, t': other tweets, w: word in tweet). Within-cluster rank is also calculated using Equation 3 (w: given word, t: tweets which contain tweet, t': other tweets).

$$Conditional-rank(t) = \frac{\sum_w weight(t,w)}{\sum_{t'}\sum_w weight(t',w)} \quad (2)$$

$$Within-cluster-rank(w) = \frac{\sum_t weight(t,w)}{\sum_t \sum_{w'} weight(t,w')} \quad (3)$$

3.3. Cluster Scoring

After extracting clusters, we need to evaluate them and remove inconsistent trends. In order to evaluate and prune clusters, a scoring algorithm is proposed here.

First, all unique words in all tweets are extracted. Then the score for each word in each cluster is calculated. These scores are then normalized and form the score matrix: S. This matrix represents the scores of each word in all clusters as illustrated in Figure 2. Now the intuition behind our proposed algorithm is simple: if a word has a high score in cluster i and low scores in all other clusters, then this particular word is important for cluster i. However, words with high scores in all clusters are not important. This brings up the concept of *Entropy*. Hence we can calculate the entropy for each word based on its scores on all clusters according to Equation 4. The calculated entropy is then employed as a weight for scoring each cluster based on Equation 5.

$$Entropy(w_i) = -\sum_c S(c,w_i)\log(S(c,w_i)) \quad (4)$$

$$Score(cluster_i) = \sum_w \frac{S(c_i,w)}{Entropy(w)} \quad (5)$$

4. Experiments

In order to evaluate the proposed approach, we first need to gather tweets posted in a short period of time. We considered daily tweets as trends usually change each day. We have collected tweets in English as well as Persian to also illustrate the language-independent property of the proposed algorithm.

Language	Date	#tweets	#hashtags	#tweets with hashtags	#trends
En	2019-08-22	13689	2154	3873	2
	2019-08-29	13662	2040	3358	3
Per	2019-06-25	3281	2345	7506	4
	2019-06-26	30753	2310	6961	3
	2019-08-01	31386	2333	5983	4
	2019-09-12	29097	2047	5485	2
	2019-09-13	29254	2083	5267	1
	2019-09-17	29095	2163	5152	4
	2019-09-19	28429	2135	5495	4

Table 1: Twitter data statistics: we have collected several days' tweet in English (En) and Persian (Per), the number of trends for each day are manually extracted.

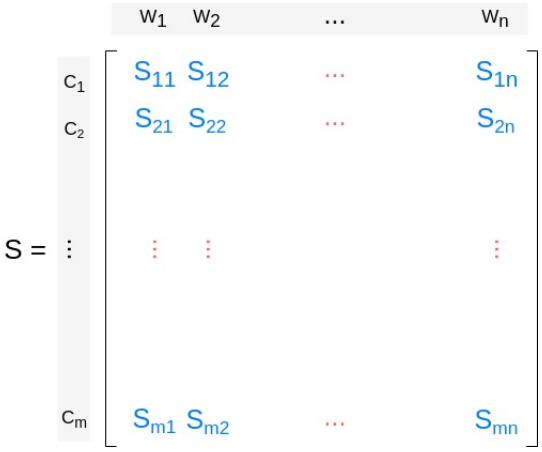

Figure 2: Matrix S: this matrix contains the scores for each word in each cluster. This matrix is utilized to find the most coherent clusters or trends.

Tweets are collected using *Twitter Firehouse API*. For Persian tweets, we provided a list of common keywords to the API to collect any Persian tweets regardless of the topic. This way, tweets can be about somebody's daily life or a certain trend and we expect the model to detect trends and ignore others. For English tweets, on the other hand, we only used crypto-currency related keywords to only collect topic-specific tweets. We aim at testing the model's ability to detect trends in focused datasets in this scenario. The dataset statistics is provided in Table 1.

Baselines: We compared our model with a well-known topic modeling algorithm, Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Another approach for comparison is to represent tweets using TF-IDF vectors and employing *k-means* algorithm for clustering and trend extraction.

The Evaluation Scenario is as follows:

1. Tweets for a certain day are collected.
2. The collected tweets are manually analyzed to extract trends (as ground truth trends).
3. The proposed algorithm is employed to extract trends from tweets.
4. Baseline methods are also employed for trend extraction.
5. Extracted trends are manually investigated to determine how many trends have been successfully extracted by each approach.

Implementation Notes: Some pre-processing steps and hyper-parameters are described here. These specifications are necessary to reproduce the results:

- pictures, emojis, mentions, numbers, and URLs are removed.
- Each tweet is normalized and all words are replaced by their stemmed version (using Hazm[1]).
- Edge Weights are calculated as described in Equation 1.
- The *RankClus* algorithm is applied by K initial random clusters. The number of Expectation–Maximization iterations and maximum clustering iterations are set to 6 and 11 respectively.

5. Results and Discussion

Cluster scores on two sample dates on the English and Persian datasets are presented in Table 2. As illustrated in this table, for the first one, the clustering algorithm emitted 2 clusters in its process and our scoring algorithm also assigned 0 scores to two clusters. Hence only two trends are extracted from this date: C_1 and C_3. The latter has a higher score which indicates a more consistent and coherent trend. Extracted trends and sample tweets from each trend for the first date in Table 2 are illustrated in Table 3. For each trend, top tweets with hashtags, top tweets without hashtags, top hashtags and top words are presented. The proposed algorithm is capable of providing a ranking for all node types in the graph. This feature is very useful for analyzing trends and illustrating trends for users. Also in Table 5, two extracted trends for second date in Persian dataset is demonstrated.

Table 4 provides a comparison between our proposed algorithm and baseline models. The comparison is based on the

[1]https://github.com/sobhe/hazm

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10
$2019-08-22$	1.37	0.0	5.88	0.0	-	-	-	-	-	-
$2019-09-17$	5.056	0.529	0.543	0.511	0.769	0.521	0.506	0.512	3.124	0.513

Table 2: Clusters' scores calculated using Equation 5 for two days.

		2019-08-22
C_1	Tweets with Hashtag	Lightning network is a complete waste of time and the trolls #bitcoin #crypto #cryptocurency #blockchain #btc
		All you need to know is do you have patience to buy now on hold till then #ltc #litecoin #bitcoin #btc
	Tweets w/o Hashtag	Guys you really need to invest in bitcoin It's gotten me so much money and the market is only going to make it better
		"Amrita weren't you really in to bitcoin for a long time "Me" What is bitcoin"
	Top Hashtags	#binance, #bitcoin, #cryptocurency, #btc
	Top Words	bitcoin, price, btc, market
C_3	Tweets with Hashtag	So a credit card you can't put in your jeans or a leather wallet Got it #boldstrategycoton
		That new Apple Card is so #highmaintenance like it can't even be put in your jean pockets or leather wallet #AppleCard
	Tweets w/o Hashtag	Here's a card that you can't put into your wallet so Apple will make a wallet for £ and get more money
		So it's a credit card that you can't put in your pocket or your wallet
	Top Hashtags	#AppleCard, #boldstrategycoton, #highmaintenance
	Top Words	wallet, apple, card, leather, credit

Table 3: Extracted trend samples on English tweets. Only two top clusters are illustrated. For each cluster two sets of tweets are presented: top tweets containing hashtags and top *no-hashtag* tweets. Top hashtags and words are also presented for each trend.

Lang	Date	K-Means	LDA	Ours
En	2019-08-22	50%	100%	**100%**
	2019-08-29	66%	100%	**100%**
Per	2019-06-25	25%	50%	**75%**
	2019-06-26	66%	66%	**66%**
	2019-08-01	59%	25%	**100%**
	2019-09-12	50%	50%	**100%**
	2019-09-13	100%	0%	**100%**
	2019-09-17	50%	25%	**100%**
	2019-09-19	100%	0%	**75%**

Table 4: Accuracy of our model compared to baselines. For each approach, extracted trends are manually checked to find out how many trends have been successfully extracted. Values are the *Accuracy* of the models.

model accuracy in finding the trends that have been manually extracted as the gold standard. As the results suggest, our proposed model outperforms the baselines in most cases.

6. Conclusion

User debates and interactions on Twitter usually lead to massive content regarding a specific topic which is called a *Trend*. Twitter trend extraction aims at finding trends in a short time period. In this paper, we introduced a novel approach for twitter trend extraction which utilizes tweets without hashtags and also produces a ranking for tweets in each trend. The proposed model utilizes graph clustering techniques for trend extraction. Moreover, tweets are represented as a graph, in which nodes represent tweets, words, and hashtags. Edges between nodes are constructed using a weighting policy utilizing co-occurrence scores. *RankClus* algorithm is then employed for ranking and clustering tweet nodes. We also proposed a scoring algorithm to find the most relevant and coherent trends. The proposed approach is capable of utilizing *no-hashtag* tweets as well as tweets with hashtags. Also, this model can capture multiple hashtags that correspond to the same trend.

In order to evaluate the proposed approach, several tweet sets have been collected. These tweets were first manually analyzed to find the trends in each set. Then the proposed algorithm alongside two baseline methods were employed for automatic trend extraction. The extracted trends were then manually compared with true trends. The results suggest a superior performance in our proposed model.

In order to improve the model, one can try different ranking functions. Also, the evaluation can be extended to analyze all tweets systematically and produce more reliable results.

7. References

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Cataldi, M., Di Caro, L., and Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the*

C_1	Tweets with Hashtag	کارگر #هپکو بودن جرم است کارگر #هفت_تپه بودن جرم است اصلا #کارگر بودن جرم است. *Working in #Hepco is crime; Wroking in #HaftTappe is crime; even #working is crime in here.*
		ملت ایران کارگر شجاع #هپکو را تنها نگذارید و یاریگر در برابر ستمگر باشید *Brave Iranian don't leave #Hepko workers. Help them toward cruelty.*
	Tweets w/o Hashtag	الان باید صدا کارگر هپکو باشیم وقت آن نیست که درستی شیوه اعتراضاتشون رو نقد کنیم قطعا مطالباتشون به حق است و همین به تنها کافی است *It is the time to be the Hepko workers' voice. It is not the time to criticize them. Their demand is righteous, and that is the whole story.*
		داستان تلخ و تکراری و تمام نشدنی خصوصی سازی و سرگذشت شرکت هپکو اراک؛ از اوج تا سقوط را در اینجا بخوانید *Read the bitter and repetitive story of privatization of Arak Hepko company here.*
	Top Hashtags	#کارگران_هپکو ,#هپکو_اراک ,#هپکو
	Top Words	هپکو، کارگر، حق، اراک
C_9	Tweets with Hashtag	#جمهوری_اسلامی، #ایران را تبدیل به کرم قلاب ماهیگیری #آمریکا برای گرفتن ماهی #عربستان #سعودی کرده است *I.R #Iran transforms Iran to fishing bait of #USA to catch #SaudiArabia.*
		آیت الله خامنه ای رهبر انقلاب #ایران با #آمریکا در هیچ سطحی مذاکره نخواهد شد *Ayatollah khamenei: #Iran won't negotiate with #USA in any level.*
	Tweets w/o Hashtag	به نظرم اگه عربستان حمله ایران به تاسیسات نفتی رو پاسخ نظامی داد این دفعه ایران کعبه رو بزنه *What I am saying is if Saudi Arabia responds Iran's militarily attack to its petroleum infrastructures, Iran Attacks Mecca this time!*
		یک مقام آمریکا حمله موشک به آرامکو در سعودی از خاک ایران صورت گرفته است *An American official said that the rocket attack* to Aramco in Saudi Arabia is initiated from Iran.
	Top Hashtags	#عربستان,#ایران ,#آرامکو
	Top Words	ایران، آمریکا، عربستان، آرامکو، حمله

Table 5: Extracted trend samples on Persian tweets.

tenth international workshop on multimedia data mining, page 4. ACM.

Dutta, S., Ghatak, S., Roy, M., Ghosh, S., and Das, A. K. (2015). A graph based clustering technique for tweet summarization. In *2015 4th international conference on reliability, infocom technologies and optimization (ICRITO)(trends and future directions)*, pages 1–6. IEEE.

Ifrim, G., Shi, B., and Brigadir, I. (2014). Event detection in twitter using aggressive filtering and hierarchical tweet clustering. In *SNOW-DC@ WWW*, pages 33–40.

Kim, S., Jeon, S., Kim, J., Park, Y.-H., and Yu, H. (2012). Finding core topics: Topic extraction with clustering on tweet. In *2012 Second International Conference on Cloud and Green Computing*, pages 777–782. IEEE.

Kim, T.-Y., Kim, J., Lee, J., and Lee, J.-H. (2014). A tweet summarization method based on a keyword graph. In *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, page 96. ACM.

Manaskasemsak, B., Chinthanet, B., and Rungsawang, A. (2016). Graph clustering-based emerging event detection from twitter data stream. In *Proceedings of the Fifth International Conference on Network, Communication and Computing*, pages 37–41. ACM.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Muntean, C. I., Morar, G. A., and Moldovan, D. (2012). Exploring the meaning behind twitter hashtags through clustering. In *International Conference on Business Information Systems*, pages 231–242. Springer.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Prangnawarat, N., Hulpus, I., and Hayes, C. (2015). Event analysis in social media using clustering of heterogeneous information networks. In *The Twenty-Eighth International Flairs Conference*.

Rosa, K. D., Shah, R., Lin, B., Gershman, A., and Frederking, R. (2011). Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*, 63.

Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., and Wu,

T. (2009). Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 565–576. ACM.

Wang, Y., Liu, J., Qu, J., Huang, Y., Chen, J., and Feng, X. (2014). Hashtag graph based topic model for tweet mining. In *2014 IEEE International Conference on Data Mining*, pages 1025–1030. IEEE.

Wang, B., Liakata, M., Zubiaga, A., and Procter, R. (2017). A hierarchical topic modelling approach for tweet clustering. In *International Conference on Social Informatics*, pages 378–390. Springer.