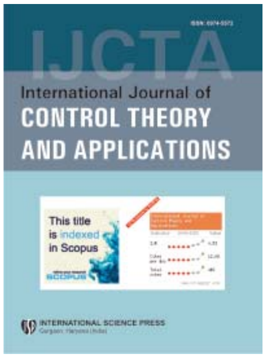

International Journal of Control Theory and Applications

ISSN : 0974–5572

© International Science Press

Volume 9 • Number 42 • 2016

A Novel Approach for Keyword Extraction from Microblog

Tanya Gupta[a] and Navdeep Kumar[a]

[a]Department of Computer Science and Engineering, Chandigarh University, India

E-mail: tanyaguptacgc@gmail.com, navdeep.cb@gmail.com

Abstract: This paper deals with study of different keyword extraction techniques. TF IDF (term frequency – inverse document frequency) and Noun Phrase extraction are two techniques which have been mapped on twitter data. As twitter data is user-generated and hence, uncertain, more advanced techniques other than traditional natural language processing algorithms needs to be developed for keyword extraction from short text. Few better pre-processing strategies and NP Chunk have been embedded into twitter keyword extraction for short text. More meaningful results have been obtained using proposed optimized technique.

Keywords: *Social media analysis, keyword extraction, tf-idf.*

1. INTRODUCTION

The trend detection from social media has been started since topic detection and tracking concept given by James Allan [16] in 2002. After topic detection and tracking, many of the researchers have gained interest and worked on first story line detection, news detection, event detection, keywords analysis, topic modeling etc. Trend is a logical entity and is used for to analyze that about what, when and from where people are discussing. This helps to track human behavior and predicts mood of the citizen. This research service to a community which helps to track user-interest and happening around the world.

The information among different communities is disseminated through social interactions. Users consume and create content on web. The global effect of social media is that it is pushing information on web. In 2010, the amount of Facebook traffic is more than that of Google traffic. To gain insights to human behaviour and marketing analytics, the human generated content can be mined to get some useful knowledge about human behaviour. Marketing communication like what are people thinking about the products, how are people responding to particular brand, what are people discussing about, who are the competitors etc is wide application area which tracks human activity. Industrial talk about political campaigns is another social media application which is important insight. This useful information is also used for real time evaluation of different events happening around. Online forums, reviews and feedback online make decisions about product and purchase. Processing of social media content to identify relevant and useful communities is another important concept.

Modeling of flow of information, predicting the information, where the information needs to be placed to make it popular, what people think about each other, what are common interests of people and how do we collect the information online etc are important challenge for research areas in social media data. The real time social media data undergoes spread of information via different networks. There are multiple data sources for social networks like those of Facebook, Twitter, LinkedIn etc. The piece of information that propagates through nodes (profiles) is still a wide area of search. The relevant information is quite fine and expanded on large scale. Small textual fragments which spread through some underlying networks are viral.

2. LITERATURE SURVEY

They extracted [1] the keyword from the Chinese micro blog and to extract that keywords we use three features graph model, semantic space and location of words. In which they used the methodology In the first step we need to download the micro blog API of a user.In the second step we do pre processing by data cleaning, word segment, POS tagging and stop word removal. In the third step we create a graph model to extract keywords on the basis of co occurrence between the words and we give a sequence number to the words according to the location and find a weight of the words by Score formula. In the fourth step we create a semantic space on the basis of topic detection and compute statistical weight by TFIDF. In the fifth step we consider the another feature that is the location of words and compute the rank value in which we conclude that a number having smaller location will be ranked higher.

Another paper [2] focuses on the structure approach and graph generation.The approach used in this paper is structure based in which we create graph model and we identify the bursty topics and events.In the clustering of topic the tweets of twitter are separated to produce two graph *i.e*. homogeneous graph and heterogeneous graph.For homogeneous graph we use OSLOM algo to find the users interaction.For heterogeneous rank we use rankclusalgo to construct a set of tweets ranked with number.

At last from both the graph results,the meaning to tweet is done using python and then we join the tweets with the same name [17]. Infuture different graph models can be used for different types of events and to construct a method that can define the events.Authors [3] constructed a method for keyword extraction and to find the solution for the problem such as high variance and lexical variants and to compare the current technique with already existing technique. For the problem of lexical variant i.ethe words although speeling different have same meaning but the technique which we use does not know about it.so to extract that word as keyword we used two methods Brown clustering:-in this we cluster the words with the same meaning such as ‘no’ ‘noo’ etc and then find the feature for the individual cluster.

Continuous word vector:-in this method we define a layer by finding its probability and then the word is changed into continuous word vector.In the last we predict the length of keyword and to define it we find the ratio between number of keywords and the total number of words in the tweets.furthermore more method such as linear regression can be used to predict the number of keywords.In future the keyword extraction technique can be used in document summarization.In this paper [4] we implement a system that detect the popular keyword and the bursty keyword in which it detect the abbreviations, any typing or spacing error. The first step used in detecting trend and bursty keywords is to collect the candidate keywords *i.e*. the first word starting with the capital letter or the word enclosed in quotation mark is considered as candidate keyword.The second step is to merge the keywords and to merge that we consider acronyms, typo and spacing and then we find the term frequency accordingly.The next step is to detect and select the popular keywords from the candidate keywords that were merged and to select the bursty keywords we use burst ratio. A prototype system can be build that can detect more bursty keywords.

Next, a paper [5] proposes a technique called TOPOL which identifies the irrelevant noisy data from the useful data.The first step used is pre processing step in which elimination of hastags,URL,non textual symbols from the tweet is done.Second step consist of mapping in which a matrix is generated by applying SVD technique. Third and the last step used is the topic detection step in which the topic are selected based on the interest.Finally the results are computed based on the parameters such as topic recall, keyword precision, keyword recall.In future many other algorithms and techniques can be used for detecting the bursty topics.

In this paper [6] different methods and technique and approaches are discussed which are used in keyword extraction.Further it defines graph based method which are based on the extraction of nodes.This paper also discusses the extraction for the Croatian language.selectivity based keyword extraction method is used in which in future we can consider different length text,different languages based on different dataset,new techniques for evaluation,to find whether entities are extracted and in text summarization.This paper [7] presents different methods and approaches for keyword extraction..paper also focuses on the graph types in which vertex and edge representation is considered..further selectivity based keyword extraction is used in which text is represented in the form of vertex and edges. The result is computed on in degree, out degree, closeness,selectivity.The paper [8] proposes a keyword extraction method that represents the graph for the text and applies the centrality measure and finds the relevant vertices.This paper proposes a technique called TKG(Twitter keyword graph) in which three steps are performed. The first is pre-processing in which stop words etc are removed. In the second step graph is represented in which nearest neighbour and all neighbours are considered. The results are computed based on the precision, recall, F-measure as well as scalability is also accessed. the future scope can be to use other centrality measure and defining more structure for graph using heuristics and elimination of noisy data.Another paper [9] summarizes the data based on particular keyword.Two algorithms are used TDA(topic detection using AGF)and TCTR(topic clustering and tweet retrieval).The methodology used is first to extract tweets from twitter,then tfidf is applied which gives weight to the words along with the frequency. AGF is evaluated using keyword rating and concept for the imitation of the mental ability of word association. The results are calculated based on the class entropy,purity,cluster entropy.The future of this paper is to consider the sentiments and emotions of tweets and the number of retweets will also be taken.

3. PROPOSED METHODOLOGY

The proposed methodology is a hybrid technique of noun phrase chunk and twitter keyword graph. The noun phrase chunk is the named entities as obtained using part of speech tagging. The named entities are then fed into twitter keyword graph to create network among co-occurring words and hence, betweenness centrality is obtained. Fig 1 shows the architecture of proposed technique NETKG (Named entity based twitter keyword graph).

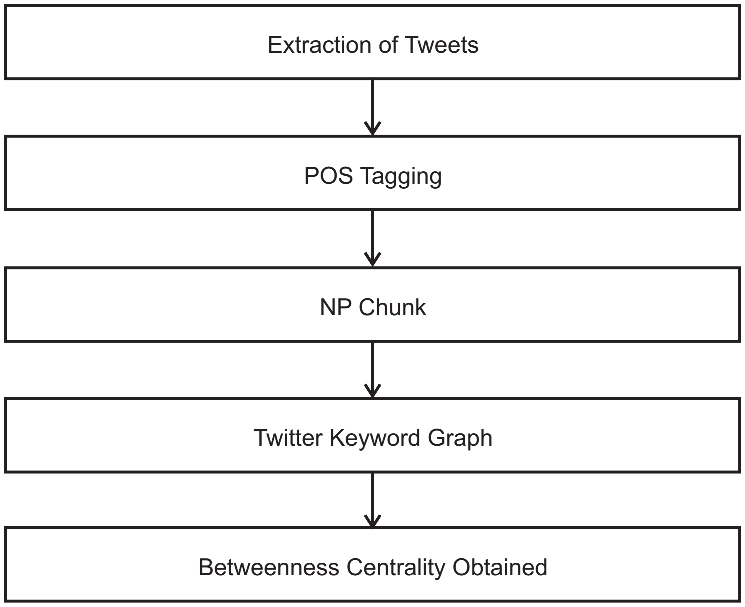

Figure 1: Architecture of NETKG

4. RESULTS AND DISCUSSION

The results obtained after applying hybrid keyword extraction techniques in order to obtain meaningful results are mentioned in table 1 for top 10 keywords obtained, in table 2 for top 20 keywords obtained and in table 3 for top 30 keywords obtained. The precision and recall are calculated as per given results. Fig 2, Fig 3 and Fig 4 shows the performance measure of recall, precision and *f* measure respectively.

Table 1
Top 10 entities obtained after applying different techniques

TF IDF	*NP Chunk*	*TKG*	*NETKG*	*Bursty Keywords*
‘all’, ‘now’, ‘art’, ‘plant’, ‘day’, ‘and’, ‘week’, ‘eat’, ‘follow’, ‘try’	‘art’, ‘day’, ‘move’, ‘mom’, ‘gon’, ‘tell’, ‘week’, ‘plant’, ‘purpose’, ‘monte’	Leave wait california suaveserg jokes well competition pulpo lakes freedom	artfino follow choice art competition pulpo california wait freedom gallery	Art, damfunk, denniscoughling, monte, artefino, venice, hornets, monte, wedding, plant, wonder, mom, app, gallery, book, weather, offer, meat, refuse, brain, orlando, hotrod, competition, pulpo, casino, California, deluxpromo, Gucci, imawarchild, rustypritchard, Mcleancromer
R = 1/32	R = 4/32	R = 4/32	R = 6/32	
P = 1/10	P = 4/10	P = 4/10	P = 6/10	

Table 2
Top 20 entities obtained after applying different techniques

TF IDF	*NP Chunk*	*TKG*	*NETKG*	*Bursty Keywords*
‘all’, ‘now’, ‘art’, ‘plant’, ‘day’, ‘and’, ‘week’, ‘eat’, ‘follow’, ‘try’, ‘that’, ‘mom’, ‘every’, ‘let’, ‘thats’, ‘app’, ‘tell’, ‘dec’, ‘date’, ‘keep’	‘art’, ‘day’, ‘move’, ‘mom’, ‘gon’, ‘tell’, ‘week’, ‘plant’, ‘purpose’, ‘monte’, ‘damfunk’, ‘text’, ‘denniscoughlin’, ‘foul’, ‘gravity’, ‘press’, ‘artfino’, ‘mcleancromer’, ‘shoe’, ‘chiaroscuro’	All, follow Leave, wait California, Sad, Suaveserg, Deepening, Jokes, Art, last, ready, well, press, Competition, pulpo, lakes, imawarchild, freedom, starting	Press, artfino, follow, imam, choice, art, competition, pulpo, California, wait, imawarchild, freedom, casino, catch, deluxpromo, Gucci, trip, winter, gallery, youre	Art, damfunk, denniscoughling, monte, artefino, venice, hornets, monte, wedding, plant, wonder, mom, app, gallery, book, weather, offer, meat, refuse, brain, orlando, hotrod, competition, pulpo, casino, California, deluxpromo, Gucci, imawarchild, rustypritchard, Mcleancromer
R = 4/32	R = 8/32	R = 6/32	R = 10/32	
P = 4/20	P = 8/20	P = 6/20	P = 10/20	

Table 3
Top 30 entities obtained after applying different techniques

TF IDF	*NP Chunk*	*TKG*	*NETKG*	*Bursty Keywords*
‘all’, ‘now’, ‘art’, ‘plant’, ‘day’, ‘and’, ‘week’, ‘eat’, ‘follow’, ‘try’, ‘that’, ‘mom’, ‘every’, ‘let’, ‘thats’, ‘app’, ‘tell’, ‘dec’, ‘date’, ‘keep’, ‘bet’, ‘sad’, ‘also’, ‘wonder’, ‘wedding’, ‘ready’, ‘make’, ‘purpose’, ‘add’, ‘book’	‘art’, ‘day’, ‘move’, ‘mom’, ‘gon’, ‘tell’, ‘week’, ‘plant’, ‘purpose’, ‘monte’, ‘damfunk’, ‘text’, ‘denniscoughlin’, ‘foul’, ‘gravity’, ‘press’, ‘artfino’, ‘mcleancromer’, ‘shoe’, ‘chiaroscuro’, ‘follow’, ‘venice’, ‘resisting’, ‘daymeon’, ‘fit’, ‘imma’, ‘menu’, ‘tweet’, ‘cuz’, ‘rico’	All follow imma every leave wait casino winter california sad suaveserg deepening hit jokes art last tons catch ready well press competition pulpo lakes imawarchild freedom was deluxpromo trip starting	press artfino follow imma choice semi truth art competition warmer pulpo california article wait imawarchild freedom casino rustypritchard mcleancromer tell catch deluxpromo lmao gucci trip winter week weekend gallery youre	Art, damfunk, denniscoughling, monte, artefino, venice, hornets, monte, wedding, plant, wonder, mom, app, purpose, book, weather, offer, meat, refuse, brain, orlando, hotrod, competition, pulpo, casino, California, deluxpromo, Gucci, imawarchild, rustypritchard, Mcleancromer
R = 8/32	R = 10/32	R = 8/32	R = 12/32	
P = 8/30	P = 10/30	P = 8/30	P = 12/30	

4.1. Recall

Table 4
Recall obtained after applying different techniques

Recall	*TF IDF*	*NP Chunk*	*TKG*	*NETKG*
Top 10	1/32	4/32	4/32	6/32
Top 20	4/32	8/32	6/32	10/32
Top 30	8/32	10/32	8/32	12/32

Recall	*TF IDF*	*NP Chunk*	*TKG*	*NETKG*
Top 10	0.031	0.125	0.125	0.1875
Top 20	0.125	0.250	0.1875	0.3125
Top 30	0.250	0.3125	0.250	0.375

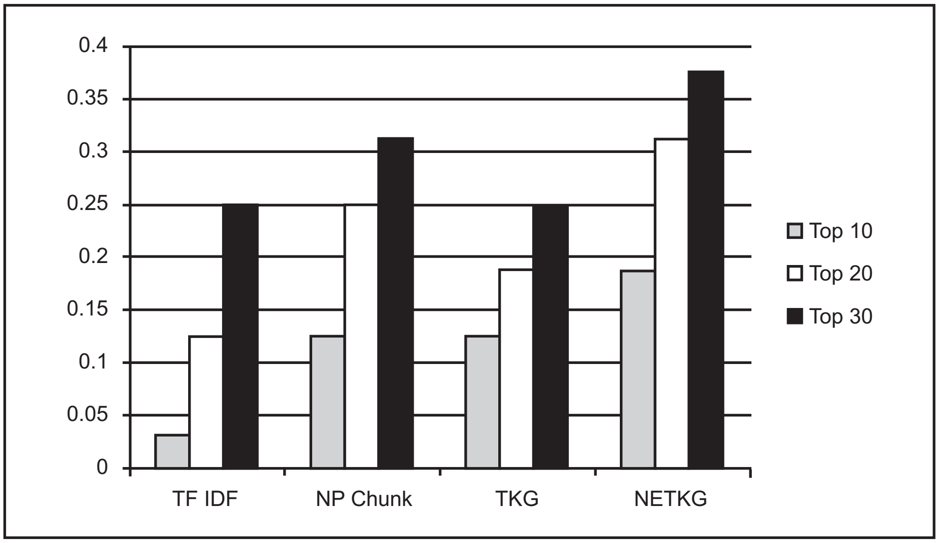

Figure 2: Recall performance of different techniques

4.2. Precision

Table 5
Precision obtained after applying different techniques

Precision	*TF IDF*	*NP Chunk*	*TKG*	*NETKG*
Top 10	1/10	4/10	4/10	6/10
Top 20	4/20	8/20	6/20	10/20
Top 30	8/30	10/30	8/30	12/30

Precision	*TF IDF*	*NP Chunk*	*TKG*	*NETKG*
Top 10	0.1	0.4	0.4	0.6
Top 20	0.2	0.4	0.3	0.5
Top 30	0.266	0.333	0.266	0.40

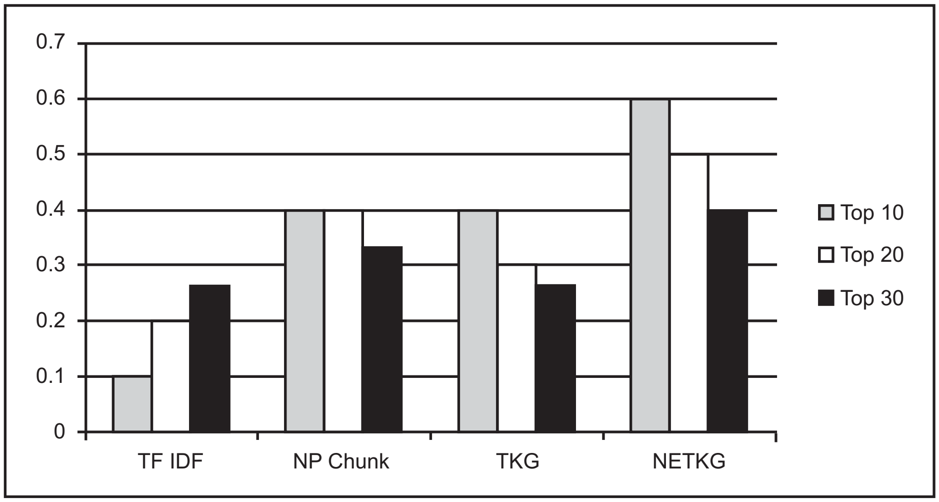

Figure 3: Precision performance of different techniques

4.3. F Measure

Table 6
F-Measure obtained after applying different techniques

F Measure	*TF IDF*	*NP Chunk*	*TKG*	*NETKG*
Top 10	0.047328	0.190476	0.190476	0.190476
Top 20	0.153846	0.307692	0.230769	0.384615
Top 30	0.257752	0.322424	0.257752	0.387097

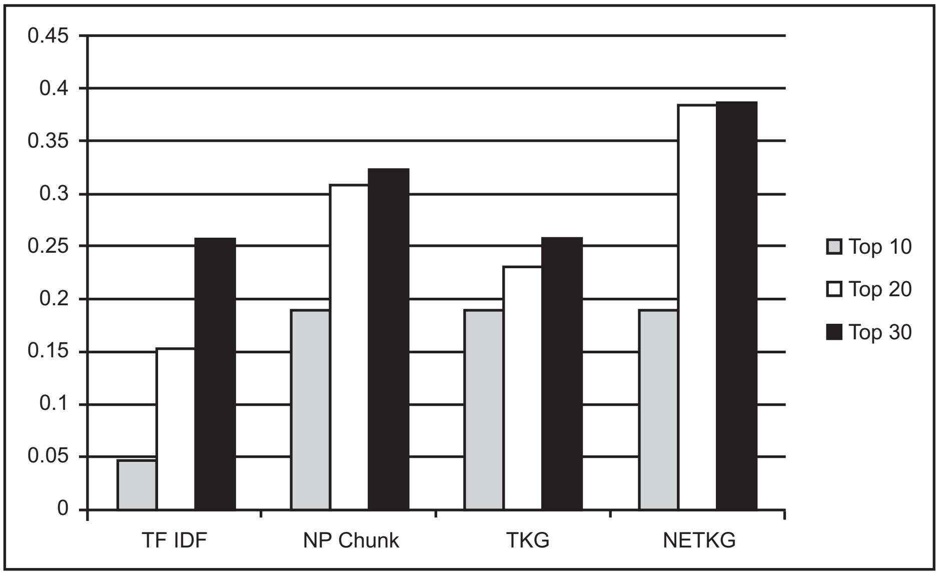

Figure 4: F Measure performance of different techniques

5. CONCLUSION

The proposed work can be summarized as new technique for keyword extraction technique of short text uncertain data. It has been observed that the named entities provide more useful information than the random words. Thus, named entities should be considered for named entities more than that of bursty keywords. The keywords obtained from TF IDF are based on weights which are given to each of the word as mentioned in dataset. More the weight, better is the keyword. However, local residents tend to speak more about local places and entities than that of general words. This gives topic of discussion when named entities are considered exclusively. It has been observed that Precision, Recall and F measure are better for NETKG. In future, noise removal techniques can be used to obtain better results on user-generated data.

REFERENCES

[1] Zhao, H., &Zeng, Q. (2013). Micro-blog keyword extraction method based on graph model and semantic space. *Journal of Multimedia*, *8*(5), 611-617.

[2] Hromic, H., Prangnawarat, N., Hulpuş, I., Karnstedt, M., & Hayes, C. (2015). Graph-based methods for clustering topics of interest in twitter. In*Engineering the Web in the Big Data Era* (pp. 701-704). Springer International Publishing.

[3] Marujo, L., Ling, W., Trancoso, I., Dyer, C., Black, A. W., Gershman, A., &Carbonell, J. Automatic Keyword Extraction on Twitter. *Volume 2: Short Papers*, 637.

[4] Kim, D., Kim, D., Rho, S., & Hwang, E. J. (2013). Detecting trend and bursty keywords using characteristics of Twitter stream data. *International Journal of Smart Home*, *7*(1), 209-220.

[5] Torres-Tramón, P., Hromic, H., &Heravi, B. R. (2015). Topic Detection in Twitter Using Topology Data Analysis. In *Current Trends in Web Engineering*(pp. 186-197). Springer International Publishing.

[6] Beliga, S. (2014). Keyword extraction: a review of methods and approaches.*University of Rijeka, Department of Informatics, Rijeka*.

[7] Beliga, S., Meštrović, A., &Martinčić-Ipšić, S. (2015). An Overview of Graph-Based Keyword Extraction Methods and Approaches. *Journal of Information and Organizational Sciences*, *39*(1), 1-20.

[8] Abilhoa, W. D., & de Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, *240*, 308-325.

[9] Benny, A., & Philip, M. (2015). Keyword Based Tweet Extraction and Detection of Related Topics. *Procedia Computer Science*, *46*, 364-371.

[10] Hong, B., & Zhen, D. (2012). An extended keyword extraction method.*PhysicsProcedia*, *24*, 1120-1127.

[11] https://www.airpair.com/nlp/keyword-extraction-tutorial.

[12] Y. Matsuo andM. Ishizuka(2004). Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools, 13:2004,

[13] Rafea, Ahmed, et al.(2013) *Topic extraction in social media*. Collaboration Technologies and Systems (CTS), 2013 International Conference on IEEE;

[14] K. Zhang, H. Xu, J. Tang, J.Z. Li, (2006)Keyword extraction using support vector machine, in: Proceedings of the Seventh International Conference on Web-Age, Information Management (Waim2006),

[15] S. Rose, D. Engel, N. Cramer, W. Cowley,(2010) Automatic Keyword Extraction from Individual Documents, Text Mining: Applications and Theory.

[16] Allan, J. (Ed.). (2002). *Topic detection and tracking: event-based information organization*. Kluwer Academic Publisher.

[17] Muskan& Kumar M. Review on Event Detection Techniques in social multimedia. *Online Information Review.* Vol. 40, issue 3. Emerald Insight. 2016.