Comprehensive Study of Keyphrase Extraction Metrics for Uncertain User-Generated Data

Muskan Garg and Mukesh Kumar

Abstract Social media analysis has received much attention from academic researchers and practitioners. Twitter is one of the most widely used social media platforms and contains ill-formed, uncertain and user-generated information. The word co-occurrence networks represent the statistical and computational evaluation of contextual data which are also called textual networks. Identifying influential nodes as keywords in textual networks of given data is of theoretical and practical significance. Different network metrics can be used to understand the salient features of textual networks. To overcome this research gap, the comprehensive study of different keyword extraction and keyphrase extraction techniques has been explored which gives useful insight into patterns in word co-occurrence networks. The significant insights obtained from textual networks can be used for different applications for social media analytics. The 'First Story Detection' data set has been used for experimental evaluation to identify significant traditional measures.

Keywords Keyword extraction algorithm · Social media analysis · Word co-occurrence model · Textual network metrics

1 Introduction

Social media analysis is one of the most widely studied areas of research. The field of topic detection and tracking has been introduced [1]. The key idea for social media analysis is keyword extraction and keyphrase extraction. The words which are important in terms of any text are said to be keywords. Similarly, the set of words which occur as a phrase and represent the text are known as keyphrase. The keyphrase represents the topic of discussion which summarizes the textual information.

M. Garg (✉) · M. Kumar
UIET, Panjab University, Chandigarh, India
e-mail: muskanphd@gmail.com

S. Mishra et al. (eds.), *Applications of Computing, Automation and Wireless Systems in Electrical Engineering*, Lecture Notes in Electrical Engineering 553,
https://doi.org/10.1007/978-981-13-6772-4_104

Keyphrase Extraction

The keyphrase extraction is the process of identifying important phrases from the given textual data. It is used to summarize the short text to identify topic of discussion from a set of Twitter posts which are related to the same topic or domain. The topic or domain is very subjective decision to decide if the extracted keyphrase should be about main topic or sub-topics under that topic. For instance, 'floods in Kerala 2018' is a main topic but 'having nothing to eat in XYZ village', 'save us! Water all around' and 'our house drowned in flood' are sub-events. These sub-events help us to determine the emergency conditions. Such topic and emergency conditions can be deduced by identifying keyphrases from many Twitter posts. Tackling this kind of information shall provide useful insights into social media data. In this study, short text is being used for keyphrase extraction; the keyphrase extraction and text summarization are used interchangeably.

Recent progress on word co-occurrence network has added statistical and computational significance over information processing. Graph-based keyphrase extraction measures are usually based on the global network metrics which have been used for keyword extraction. The network metrics have been studied for well-formed data and social media data. It has been observed that the keywords obtained using the random walk-based measures are ranked, and using this information keyphrase is extracted. The application domains for keyword extraction and keyphrase extraction techniques are sentiment analysis, topic and trend detection, event detection, disaster management, outbreak detection and many other applications. The major challenges for identifying keywords in social media data are that the data is unstructured. Due to graphical and statistical approach, the random walk-based keyword extraction from word co-occurrence network has proved to be useful for well-formed data. However, for ill-formed data, traditional network metrics for word co-occurrence network have not been explored.

Word co-occurrence networks are graphical networks which are generated from contextual information. These are also referred to as word adjacency model [2]. There are different types of textual networks which can be framed from given data. For each graph G, node is considered as word and edge is considered as a link connecting two words co-occurring in the given document. In this research work, the document is considered as Twitter feed. Based on the co-occurrence and associativity, construction of graph is decided on three parameters as follows. This paper is organized as follows. In Sect. 2, existing metrics have been implemented on standard Twitter feeds and results have been obtained. In Sect. 3, evaluations and results have been analysed and discussed. Conclusion and future work have been presented in Sect. 4.

2 Evolution of Different Approaches

Graph-based keyword extraction techniques can be both supervised and unsupervised, context dependent and context independent. In this research work, many context-independent unsupervised graph-based keyword extraction techniques have been explored. *KeyWorld* is an automatic indexing system which has been proposed by Matsuo et al. [3] which extracts candidate keywords by measuring their influence on small-world properties. It captures characteristic path length and extended characteristic path length. This algorithm has been inspired by small-world phenomenon and KeyGraph algorithm proposed by Ohsawa et al. [4]. Thereafter, Erkan et al. [5] proposed LexRank which is insensitive to noise in text and calculates importance of sentence (or word) using eigenvector centrality. Mihalcea et al. [6] proposed graph-based TextRank model which has been originated from the concept of PageRank. The author further improved TextRank for text summarization. In 2007, Palshikar [7] proposed hybrid and statistics-based approach for keyword extraction using co-occurrence frequency measure. The author described eccentricity-based keyword identification, other centrality measure-based keyword extractions and proximity-based keyword identification. Litvak et al. [8] proposed HITS-based algorithm for keyword extraction. In 2009, for event detection and tracking in social streams, Sayyadi [9] used KeyGraph algorithm which was proposed earlier by Ohsawa et al. [4]. Later, in 2011, the author introduced DegExt, a graph-based language-independent keyphrase extractor. The author used degree centrality for keyword extraction. In 2013, Boudin et al. [10] compared various centrality measures for graph-based keyphrase extraction from short documents. Abilhoa et al. [11] proposed Twitter Keyword Graph (TKG) algorithm to extract keywords from Twitter data. The author introduced all neighbour edging and nearest neighbour edging for constructing graph, frequency-based and inverse-frequency-based weights in graph and different centrality measures. Besides this, another algorithm named selectivity-based keyword extraction (SBKE) has been proposed by Beliga et al. [12]. The author used degree and strength for each node. There are many such algorithms which have been improved on semantic and linguistic features on textual networks including SingleRank, ExpandRank, word co-occurrence statistical information [13], noun phrase-based keyword extraction, semantic relationships using Wikipedia texts [14], weighted lexical complex network-based keyword extraction Bollen et al. [15], keyword and keyphrase extraction algorithm [16] for word co-occurrence network structure, word topic network model [17] and many more. However, such algorithms have not been considered for keyword extraction from Twitter.

3 Experiments and Evaluation

The traditional techniques for keyword extraction and keyphrase extraction using random walk-based measures and other network metrics have been implemented for Twitter data. The data set which has been used for this study is First Story Detection Petrovic et al. [18] data set in which 27 topics have been mentioned as ground truth. The ground truth topic contains topic id among all the 27 topics in front of every tweet id which are given in data set. Among all the tweets, 3034 tweets have been marked against the topic id correspondingly. The data set has been used by extracting all the tweets for one topic. All the tweets are summarized using keyword and keyphrase extraction technique. Python and NetworkX module have been used for implementation of existing keyphrase extraction technique over Twitter data. For every experimental evaluation, Twitter data has been given as input and top values and bottom values for keywords have been obtained as output.

Word co-occurrence networks have been generated using word co-occurrence architecture. Different basic and derived keyword extraction metrics have been studied for textual networks and evaluated after implementation on Twitter feeds of standard data set of FSD. As per the scores of different metrics, top ten keywords (words with highest value) and bottom ten keywords (words with lowest values) have been collected. Significance of measuring values for words varies from one metric to another. Out of given 18 metrics, one metrics is non-beneficiary (lower the value, better is the significance), fourteen metrics are beneficiary (higher the value, better is the significance), and three metrics have been observed as non-significant. The metrics may have overlapping significance. For each topic, corresponding results have been obtained as shown in Table 1 for first topic 'Death of Amy Winehouse'. Italic font indicates that results obtained are meaningful. The topic is given as ground truth in data set FSD. Each word of topic is considered as keyword.

The given values of recall have been obtained by automatic word-to-word matching. However, due to the presence of ill-formed words, the precision measure

Table 1 Data set considered for keyword extraction

Topic no.	Topic name	Keywords selected
1	Amy Winehouse dead	Amy Winehouse dead
2	Space shuttle Atlantis lands safely, ending NASA's space shuttle programme	Space shuttle Atlantis lands safely ending NASA's programme
4	Richard Bowes, victim of London riots, dies in hospital	Richard Bowes victim London riots dies hospital
5	Flight Noar Linhas Aereas 4896 crashes, all 16 passengers dead	Flight Noar Linhas Aereas 4896 crashes all 16 passengers dead
13	Plane carrying Russian hockey team Lokomotiv crashes, 44 dead	Plane carrying Russian hockey team Lokomotiv crashes 44 dead
18	Gunman opens fire in children's camp on Utoya island, Norway	Gunman opens fire children's camp Utoya island Norway

Table 2 Performance measures obtained for topics using different keyword extraction metrics

S. no.	Keyword extraction algorithm	Precision		Recall		*F*-measure	
		Top	Bottom	Top	Bottom	Top	Bottom
1	Degree	0.44	0.06	0.532619	0.073571	0.481900	0.066096
2	Strength	0.46	0.06	0.552619	0.073571	0.502074	0.066096
3	Betweenness centrality	0.44	0.04	0.532619	0.045000	0.481900	0.042353
4	Closeness centrality	0.44	0.08	0.532619	0.102143	0.481900	0.089725
5	Eigenvector centrality	0.42	0.06	0.512619	0.073571	0.461710	0.066096
6	Clustering coefficient	0.04	0.42	0.048571	0.512619	0.043871	0.461710
7	Influence	0.44	0.08	0.529841	0.098571	0.480759	0.088320
8	PageRank	0.44	0.02	0.527619	0.025000	0.479843	0.022222
9	TF-IDF	0.08	0.32	0.098571	0.378254	0.08832	0.346697
10	KeyWorld	0.38	0.02	0.457619	0.028571	0.415213	0.023529
11	LexRank	0.44	0.06	0.532619	0.073571	0.481900	0.066096
12	TextRank	0.42	0.06	0.495476	0.073571	0.454627	0.066096
13	Eccentricity	0.06	0.12	0.073571	0.139365	0.066096	0.128960
14	HITS (avg. (H, A))	0.46	0.04	0.552619	0.053571	0.502074	0.045802
15	HITS (max (H, A))	0.48	0.02	0.581190	0.028571	0.525771	0.023529
16	DegExt	0.44	0.06	0.532619	0.073571	0.481900	0.066096
17	TKG	0.46	0.08	0.552619	0.102143	0.502074	0.089725
18	SBKE	0.12	0.06	0.140794	0.065000	0.129568	0.062400

obtained is low. However, manual intervention of results gives meaningful results and better value for precision. For instance, the topic 'Death of Amy Winehouse' may contain two words out of three for degree precision measure, but *dead* and *died* give clear indication for death and hence, precision may be recorded as one. However, in order to keep results unbiased, automatic evaluation has been preferred. But topic 1 has only three words as keywords which may result in poor analysis of precision and recall. Five topics were selected for experiments on the basis of number of keywords obtained from given topic for evaluation. For better examination for precision and recall values, topics having about ten keywords were selected. Further, precision, recall and *F*-measure have been obtained for each experiment and averaged as shown in Table 2.

4 Discussion

Different performance measures for extracting keywords have been analysed on the basis of experimental evaluation. It has been observed that strength measure and HITS algorithm outperform all the existing techniques on short text as shown in Fig. 1. Also, DegExt, TKG, LexRank, Degree, closeness centrality, eigenvector centrality, betweenness centrality, clustering coefficient and influence measure

Table 3 Inference for different keyword extraction metrics for textual networks

Keyword extraction algorithm	Inference
Degree	Gives the number of words with which word w occurs. Measure of degree centrality for each word
DegExt	
SBKE	Measures the repeated occurrence of word w with its neighbouring words with respect to number of word it is co-occurring with
Strength	Calculated word w frequency
Betweenness centrality	Centrality calculates the measure of number of paths a word w has for word-to-word connectivity
Closeness centrality	
Eigenvector centrality	
TKG	Calculates the significance of word w in Twitter feeds when word co-occurring graph is generated. Performs best by using all neighbouring edge schemes with edge weighing and inverse co-occurrence frequency and closeness centrality
Clustering coefficient	Measures the extent of similarity among neighbours of word w
Influence	Calculates the difference and maximum value among number of predecessors or successors co-occurring with word w Calculates the average number of predecessors or successors co-occurring with word w
HITS (max (H, A))	
HITS (avg. (H, A))	
TF-IDF	Statistical measure of importance of word w
KeyWorld	Based on the average of shortest path measure between two words. Used small-world phenomenon
PageRank	Measures the impact of neighbouring words co-occurring with words on word w using their votes
LexRank	
TextRank	
Eccentricity	Measures reciprocal of number of co-occurrences of pair of words in given feeds. Zero for isolated node

perform omparably significant results. However, eccentricity and SBKE are least significant measures for keyword extraction from uncertain user-generated text. Among all the basic and derived metrics, strength and HITS metrics have proved to be useful in terms of *F*-measure as observed in Table 2. When *F*-measure is considered for low-valued elements, it is observed that higher is the value of metrics, better it is for keyword extraction and vice versa. However, it has been observed that small values of clustering coefficient for each node give better keyword extraction. This shows that clustering coefficient is non-beneficiary attribute. Moreover, *Tf-Idf*, eccentricity and selectivity-based keyword extraction have proved to be not much significant measure for keyword extraction from Twitter. Inference for different keyword extraction metrics have been shown in Table 3.

Semantics of textual networks has been observed. Different features can be used as important statistical measure for identifying relevant and influential terms. As per observation, degree measure and DegExt experimentation signify similar semantics

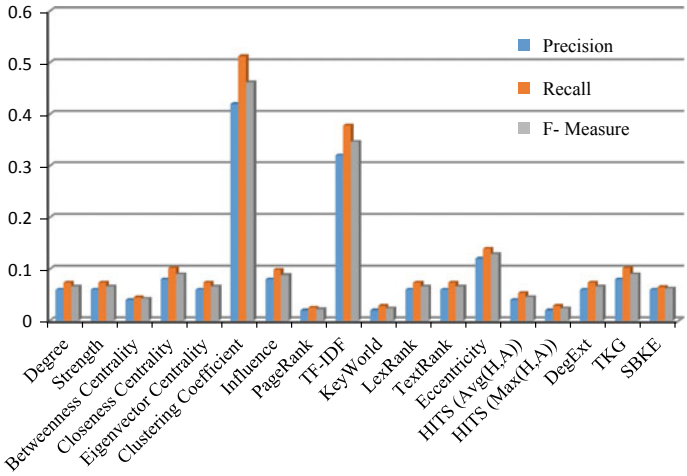

Fig. 1 Graphical representation of performance measures for bottom ten keywords using different keyword extraction metrics

and thus overlap. Metrics with high precision and recall values for top-ranked keywords and low precision and recall values for low-ranked keywords have found to be better than other metrics. On the contrary, clustering coefficient has opposite nature. Although *KeyWorld* metrics marks significantly irrelevant terms, it is computationally expensive and thus is not suitable for large-scale data analysis.

As observed in Table 4, metrics with non-weighted textual graph gives meaningful results. However, strength is a significant measure for textual networks, and thus, edge-based metrics (co-occurrence frequency-based metrics) for weighted textual networks may mark influential words. On the basis of need, adjacent pair and all pair neighbouring models have been used. However, majority of value and centrality-based metrics outperforms using all pair neighbouring model, whereas neighbourhood and its vote-based metrics used adjacent neighbour model for better performance. The inference for this parameter signifies that relation of every word to every other in a document signifies its value as how important the word is for the network and adjacent word pair signifies what impact does influential (important) word have on its neighbour. Similarly, undirected graph is used for neighbourhood-based metrics and directed graph is used for measuring incoming and outgoing links to other words. Also, directed textual network may have better lexical output as sequence of words. Based on this inference, a metric can be developed and it identifies dominant phrase from textual networks which may provide more meaningful results than just keywords.

For each topic, set of different keywords has been obtained. Unique words from this set are represented as ground truth for analysing relevant Twitter feeds. For differently selected metrics, the values have been obtained for each keyword of topic 4. Normalized graph for values of keywords for differently selected metrics has been obtained in Fig. 2. It has been clearly observed that the word '*victim*' has

Table 4 Inference for different keyword extraction metrics for textual networks

Keyword extraction algorithm	*F*-measure		Weighted/ non-weighted metrics	All pair neighbour/ adjacent neighbour	Directed/ undirected
	Top	Bottom			
Strength	0.502074	0.066096	Non-weighted	All pair neighbour	Undirected
Betweenness centrality	0.481900	0.042353	Weighted	All pair neighbour	Undirected
Clustering coefficient	0.043871	0.461710	Non-weighted	All pair neighbour	Undirected
PageRank	0.479843	0.022222	Non-weighted	Adjacent neighbour	Directed
HITS (avg. (H, A))	0.502074	0.045802	Non-weighted	Adjacent neighbour	Directed
HITS (max (H, A))	0.525771	0.023529	Non-weighted	Adjacent neighbour	Directed
TKG	0.502074	0.089725	Non-weighted	All pair neighbour	Undirected

not appeared in text. Also, words '*Richard*' and '*Bowes*' are co-occurring and have same values in most of the metrics, and thus, '*Richard Bowes*' may represent candidate keyword or named entity. Also, the last word '*Hospital*' has high clustering coefficient and strength. However, other values for these keywords are low.

Zero value of betweenness centrality indicates that the word has either no in-degree or no out-degree. Also, it can be observed that linking from one word to another, for instance '*London*' to '*Riots*' and '*Riots*' to '*dies*', is dropping and rising in high range of values, and thus, weight edges as strength play pivotal role in identifying co-occurring important words which represent influential nodes in textual networks. As observed from Fig. 2, all measures except strength and clustering coefficient give highest values for '*Richard Bowes*'. The inference for lower value of strength indicates that for each node, number of occurrences has been considered by strength, and thus, most frequently occurring words are given higher values, for instance '*very*' which is not significant.

5 Conclusion and Future Work

The user-generated data on social media is analysed using textual networks. To evaluate the performance of basic and derived metrics for textual networks, we have used Twitter data. In this analysis, unsupervised context-independent graph-based keyword extraction techniques have been implemented and discussed. It is observed that out of 17 identified network-based metrics, 14 metrics proved to be beneficiary and one as non-beneficiary attribute. The two metrics fluctuate with

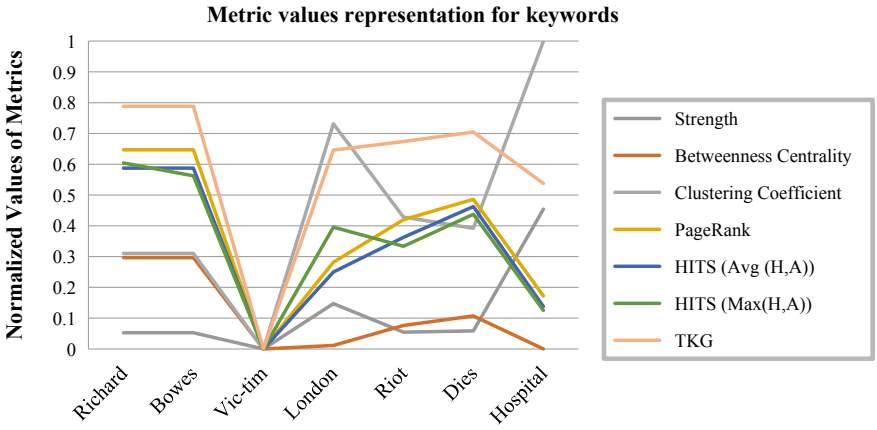

Fig. 2 Normalized metric values for differently selected keyword extraction metrics

change in data set. Differently selected metrics have been analysed, and semantics of textual networks have been discussed. It is observed that in order to maintain lexical sequence of occurrence of words, directed adjacent and weighted graph should be constructed. Majority of network metrics outperforms traditional *Tf-Idf* statistical method. Using this analysis, many useful insights can be obtained for different real-world applications including text mining, topic tracking and detection, event detection and opinion mining. In future, edge-based network metrics can be studied for word co-occurrence directed graph for identifying co-occurring keyphrases.

References

1. Allan J (ed) (2002) Topic detection and tracking: event-based information organization. Kluwer Academic Publisher
2. Amancio DR (2015) Probing the topological properties of complex networks modeling short written texts. PLoS ONE 10(2):e0118394
3. Matsuo Y, Ohsawa Y, Ishizuka M (2001) Keyworld: extracting keywords from document s small world. In: International conference on discovery science, Nov 2001. Springer, Berlin, Heidelberg, pp 271–281
4. Ohsawa Y, Benson NE, Yachida M (1998) KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor. In: Proceedings. IEEE international forum on research and technology advances in digital libraries, 1998, ADL 98, Apr 1998. IEEE, pp 12–18
5. Erkan G, Radev DR (2004) LexRank: graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22:457–479
6. Mihalcea R, Tarau P (2004) TextRank: bringing order into texts. Association for Computational Linguistics, Barcelona, Spain, pp 104–401

7. Palshikar GK (2007) Keyword extraction from a single document using centrality measures. In: International conference on pattern recognition and machine intelligence, Dec 2007. Springer, Berlin, Heidelberg, pp 503–510
8. Litvak M, Last M (2008) Graph-based keyword extraction for single-document summarization. In: Proceedings of the workshop on multi-source multilingual information extraction and summarization, Aug 2008. Association for Computational Linguistics, pp 17–24
9. Sayyadi H, Hurst M, Maykov A (2009) Event detection and tracking in social streams. In: ICWSM, May 2009
10. Boudin F (2013) A comparison of centrality measures for graph-based keyphrase extraction. In: International joint conference on natural language processing (IJCNLP), Oct 2013, pp 834–838
11. Abilhoa WD, de Castro LN (2014) A keyword extraction method from twitter messages represented as graphs. Appl Math Comput 240:308–325
12. Beliga, S., Meštrović, A., & Martinčić-Ipšić, S. (2016). Selectivity-based keyword extraction method. International Journal on Semantic Web and Information Systems (IJSWIS), 12(3), 1-26.
13. Matsuo Y, Ishizuka M (2004) Keyword extraction from a single document using word co-occurrence statistical information. Int J Artif Intell Tools 13(01):157–169
14. Grineva M, Grinev M, Lizorkin D (2009) Extracting key terms from noisy and multitheme documents. In: Proceedings of the 18th international conference on world wide web, Apr 2009. ACM, pp 661–670
15. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2 (1):1–8
16. Lahiri S, Choudhury SR, Caragea C (2014) Keyword and keyphrase extraction using centrality measures on collocation networks. arXiv preprint arXiv:1401.6571
17. Zhou, X., & Chen, L. (2014). Event detection over twitter social media streams. The VLDB journal, 23(3), 381-400.
18. Petrović S, Osborne M, Lavrenko V (2010) Streaming first story detection with application to twitter. In: Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics, June 2010. Association for Computational Linguistics, pp 181–189