Keyword Extraction Using A Graph-Based Approach

Tarique Khan, Danish Kumar, Kamlesh Kumar Soothar

Abstract— Due to the increasing rate of text over the Internet, it is very difficult to retrieve the relevant information regarding the user. To overwhelm with these types of problems more research work has been done in information retrieval and text analytics so far and it is one of the most trending topics for research regarding the keyword extraction. There are many types of data regarding the observations and analysis such as graphical data and others. Data can also be generated by the user, by considering social media, Wikipedia or any other resources. Most of the people generate their data by Twitter (social media, considered as one of the most popular platforms for crawling the short text, because it contains 280 characters per tweet). Keyword extraction is a process where a text is given to the computer and the computer returns a set of keywords that recommended topical words and phrases from the content of documents. It helps the reader to understand the summary or at least the core idea of the document without reading the whole document. As a result, the prospect readers do not waste their valuable times reading the irrelevant documents comprehensively. Generally, by searching the keywords, users could find related posts to an event. In this paper, we have used a graph-based keyword extraction algorithm over four different real datasets collected from Twitter on different terms. After that, we have applied the graph-based algorithm over these datasets and finally get the most important keywords. Word-clouds give us the graphical representation of the importance of keywords by their bold nature.

Index Terms— Keywords, Graph-based model, Wordclouds, Unstructured Data, Twitter.

1 INTRODUCTION

KEYWORD extraction is a task (aka the set of techniques) for extracting required keywords from the text. Keyword extraction technique also used to retrieve the relevant information on the basis of a given query from the huge data. There are various real applications of keyword extraction as Web search, Text summarization, Trending on Twitter and so on. Due to rapidly increasing usage of the Internet, there are more than enough user-generated content websites and applications are created and more are coming daily. Most of the websites and applications are belonging to short text such as Twitter, Quora, and StackOverflow, etc. The user-generated content becomes more and more fragmented and short. To manage a huge number of short user-generated content it became increasingly important for the web application service provider. Its management is based on information retrieval and accuracy. There are many techniques of information retrieval, keyword extraction is one of the most identified technique around the globe in terms of the research area, nowadays. Keyword extraction plays a major role in the different fields such as text retrieval, text clustering, text summarization, and indifferent information processing fields. By the keyword extraction, we can go through the document whether it is relevant or irrelevant, many documents might contain more than enough pages and processing through them will also take much time. In recent time many researchers analyzed keywords extraction from the documents containing short text, we can also call it Micro-blogs.

Micro-blogs have been recently attracting people to express their opinions and socialize with others. There are so many micro-blogs websites but Twitter is one of the most popular micro-blog sites. We have considered analysis for the Twitter social network. Twitter allows a user to tweet (a text message with a max length of 280characters) and tweets can contain 280 of length 280 characters, images, and videos. All users of twitter can see the post without any constraint. In twitter one can follow others without the following confirmation this made twitter a fast social media for news viral. It has more than 336 million users (as per 2019) actively. Smartphones and other web apps are user-friendly, even a person can use social networking services with a bit of knowledge of operating a smart-phone. This lead to the rapid growth of data in social networks. Now the challenging task is how to use these data usefully and extract useful information. The data on the Internet is unstructured, the way to organize the data in the structured form we use graph. Rely on the nature of data to be modeled, the graph could be directed/undirected or weighted/un-weighted. As like there are many complex data such as online social network, in which an entity is represented by a set of features and multiple relations exists between an entity pair.

By the touch in recent research of text analytics area, there have been lots of work done and the rate of work in the future is increasing rapidly. The problem of my thesis work is about to extract the most important keywords from the text, and that text is graph-based. Given a set of Twitter documents that are all related by containing a common search phrase (i.e., a topic), the data is crawled from Twitter (social network) on the basis of some countable tweets from four different events. The tweets are preprocessed by NLTK (Natural Language Tool Kit), in which stemming, punctuation, stop words removal, and URL removal processed. With only consideration of the text, hashtags, and timing of tweets, all the other unnecessary things removed by NLTK. To generate a set of keyword that is relevant to the document and finds out the most relevant keywords by ranking them. Before calculating the score of keywords through the TextRank algorithm, We created an incident matrix for normalization of words. Finally, the top-k keywords are being selected to evaluate the context of the

International Journal of Scientific & Engineering Research, Volume 10, Issue 8, August-2019
ISSN 2229-5518

document.

2 Related Works

An increasingly web-based community add the data explosion over the Internet. The volume of increasing unstructured and huge data on the Internet brings people much convenience but makes information processing and information extraction from them very difficult. Keyword extraction is a trending topic in text analytics field and there are many ways of the keyword extraction. The search engine works based on keyword extraction such as we type a word in the search bar and it shows the suggestions further in the search bar, after searching google shows us many links related to that keyword.

Yujun Wen et al.[1] talks about the classifications of the keyword extraction methods. In this paper authors performed classifiers to extaract keywords from news articles. They build a candidate keyword graph model based upon TextRank, by calculation of similarity between words as transition probability of nodes, then by an iterative method calculate the score of words and finally pick the top N keywords as the final result. To extend the work of keyword extraction from graph-based models many other researchers drag this area in different aspects of the graph. Further Jian Cao et al.[2] describes the way to improve graph based keyword extraction, where they proposed a method to compute importance of co-occurrence word in a document and apply it in graph approach to find more representative phrases also they introduce words co-relation degree in document language network to improve performance when extracting average number of keywords in documents.

Further Rafiqul et al. [3] proposed a new improved method for keyword extraction using random walk model by considering position of terms within the document and information gain (IG) of terms corresponds to the whole set of document, they also incorporate mutual information (MI) of terms with help of a random walk model to extract keywords from documents. They created a random walk model by the TextRank before this algorithm there are several types of the random walk have been created if we consider the TextRAnk than we know how previously it founded successfully in a number of applications, including web link analysis, social networks, citation analysis, and more recently in several text processing applications. Their work of post processing phase was similar to the Mihalcea et al. [4].

Jing Zhou et al. [5] describes the work for ranking the result of keywords over the structured data by the proposed approach. Their work is based upon the schema graph-based approach to keyword search which comprises of a candidate network (CN) generation and its evaluation phase. By this idea, the ranking process can also be done to the keywords which result in optimized words from the document.

Mihalcea et al. [6] describe the type of data extracted by them and they converted the text in the form of the graph as structured data. They introduce the TextRank, a graph-based ranking models for graphs extracted from natural language text, they evaluate and investigate the applications of TextRank to two language processing tasks containing unsupervised keywords and sentence extraction and shows the results obtained with TextRank are competitive with state-of-the-art systems developed in these areas.

Larry Page et al. [7] proposed the Google Page Rank to find out the web page's popularity score. The Page Rank formula presented in front of the world in Brisbane at seventh World Wide Web conference (WWW98) by Sergey Brin and Larry Page (founders of Google in 1998).

Sergey Brin et al.[8] they present Google a paradigm for a search engine (large scale) that makes the heavy use of structure in hypertext, the assumption of their proposed work is designed to crawl and indexing to the web optimized and generate much more satisfying search results. They worked in the search engines for the most important and frequent results on search engine and scaling the web pages with the web.

Wengen Li et al. [9] authors proposed work for the combined algorithms PageRank and TextRank for better precision and recall. Firstly they create a matrix (keyword matrix) and then use traditional TextRank for keyword extraction. Slobodan Baliga et al. [10] describes an overview of graph-based keyword extraction methods and approaches, in which many graphs can be considered for the keyword extraction analyze and compare. They suggest future work and boost the development of a new graph-based approach for keyword extraction.

According to [11] keyword extraction roughly divided into different approaches such as Statistical Approaches, Machine Learning Approaches, Linguistic Approaches, and Other Approaches. Further authors discuss the keyword extraction on graph-based data, moreover, they classified the graph types. Most of the analysis is done on the co-occurrence graph because it is easy to compute and constructed with the two words simultaneously. Ohsawa et al. [11] proposed an algorithm by automatic indexing by co-occurrence graphs constructed from metaphors called KeyGraph. This algorithm is based on the segmentation of a graph representing the co-occurrence between terms in a document. KeyGraph proved to be a content-sensitive, domain-independent device of indexing. Xialong Wang et al. [12] describes the sentiment analysis in Twitter, our data is crawled from Twitter too. Twitter is a popular platform used over the globe where massive messages with their real feelings posted everyday freely. Jinghua Wang et al. [13] worked on keyword extraction using PageRank. They used two algorithms WordNet and PageRank to propose their work with the help of WordNet they represented a rough, undirected, weighted and semantic graph. They weighted the graph with the relationship of synsets, then they apply the PageRank algorithm on that rough graph to prune the graph and again applied the PageRank algorithm to a pruned graph for the keyword extraction. Marina Litvak et al.[14] proposed work for graph-based keyword extraction, firstly they compared two novel approaches supervised and unsupervised for identifying the

International Journal of Scientific & Engineering Research, Volume 10, Issue 8, August-2019
ISSN 2229-5518

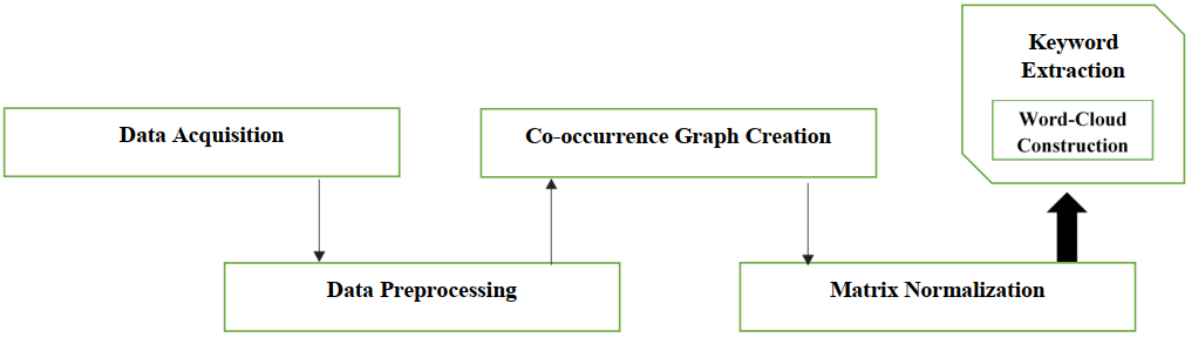

Figure 1: A schematic representation of the proposed approach

keywords, and then they train the classification algorithms on a summarized collection of documents. They execute the HITS algorithm on document graphs under the assumptions that the top-ranked nodes should represent the document keywords.

3 Proposed Approach

A schematic description of the proposed approach is shown using figure 1.The detailed description of various modules of the proposed approach is defined in following sections:

3.1 Data Acquisition

At first, we have to crawl data from twitter. There are two ways through which twitter data can be retrieved, they are Standard search API and Enterprise search API. But we have used Standard search API as it is free to use. If we specify a query, Standard search API retrieves a set of relevant tweets posted within the last seven days matching to that query. We have crawled a set of tweets for different events.

3.2 Data Preprocessing

There are a number of stopwords, punctuations that must be filtered. To get a usable tweet, a pre-processing is done and stop words, punctuation, URL, etc. are being removed. Hashtag needs to be separated because it can be used to form the relationship between the nodes.

3.3 Co-occurrence Graph Creation

After pre-processing of tweets we have saved the tweets into a text file individually. In the text file from the first tweet to the last tweet every word will make a co-occurrence graph by considering his next word as a neighbor. It exploits simple neighbor relation; two words are linked if they are adjacent in the sentence. With this technique, we are converting our text into a graph and we will get graphical data as our input.

3.4 Matrix Normalization

After constructing the graph, we have to extract the keywords for the calculate accuracy of extracted keywords. we need to solve the graph data first and to solve this, we created the matrix because the matrix is the solution of the graph. After the creation of the matrix, we normalized the matrix to make a stochastic matrix (divide the sum of the row to the individual elements of a row in a matrix).

3.5 Word-Cloud Construction

Finally, we will construct word-cloud. Word-cloud is the representation of the extracted keywords with good precision. It is generally a graphical visualization of the topmost extracted keyword.

4 Experimental Setup and Results

This section provides a detailed description of the experiments performed, evaluation metrics, and corresponding results. It also provides a brief description of the datasets used for the evaluation.

4.1 Dataset

In order to perform the evaluation process, the data was gathered from the Twitter social network, Twitter is microblogging service, which permits its user to post tweets, a status message which can have maximum 280 characters which usually use to carry personal views, news information, events information and information related to the different topics. As it is so common for social network users to type the sentence is not proper structure and they do not give importance to the grammar to write the proper sentence.the most important to them is to spread their views rather than writing Grammarly correct sentence and correct typing, and also social network users are now adopted in reading and understanding of such unstructured sentence and Grammarly corrected sentence in social networks. Because of this, all the time we get tweets from the twitters app, it will contain punctuations, stop words, hashtags, abbreviations, and slang language. This creates a bottleneck for the processing of such data. So, for this reason, we consider the natural language preprocessing for our tweet dataset to remove those unnecessary part of tweets which either can affect our result and degrade the performance or does not help the performance improvement and also add more to improve the feature extraction. After this phase, we got preprocessed tweet dataset. hereafter we called it corpus which contains preprocessed tweets without having an unnecessary part as like original tweets, and then we did different feature extraction from the corpus which after we use for the generation of the social graph which has the form of an

International Journal of Scientific & Engineering Research, Volume 10, Issue 8, August-2019
ISSN 2229-5518

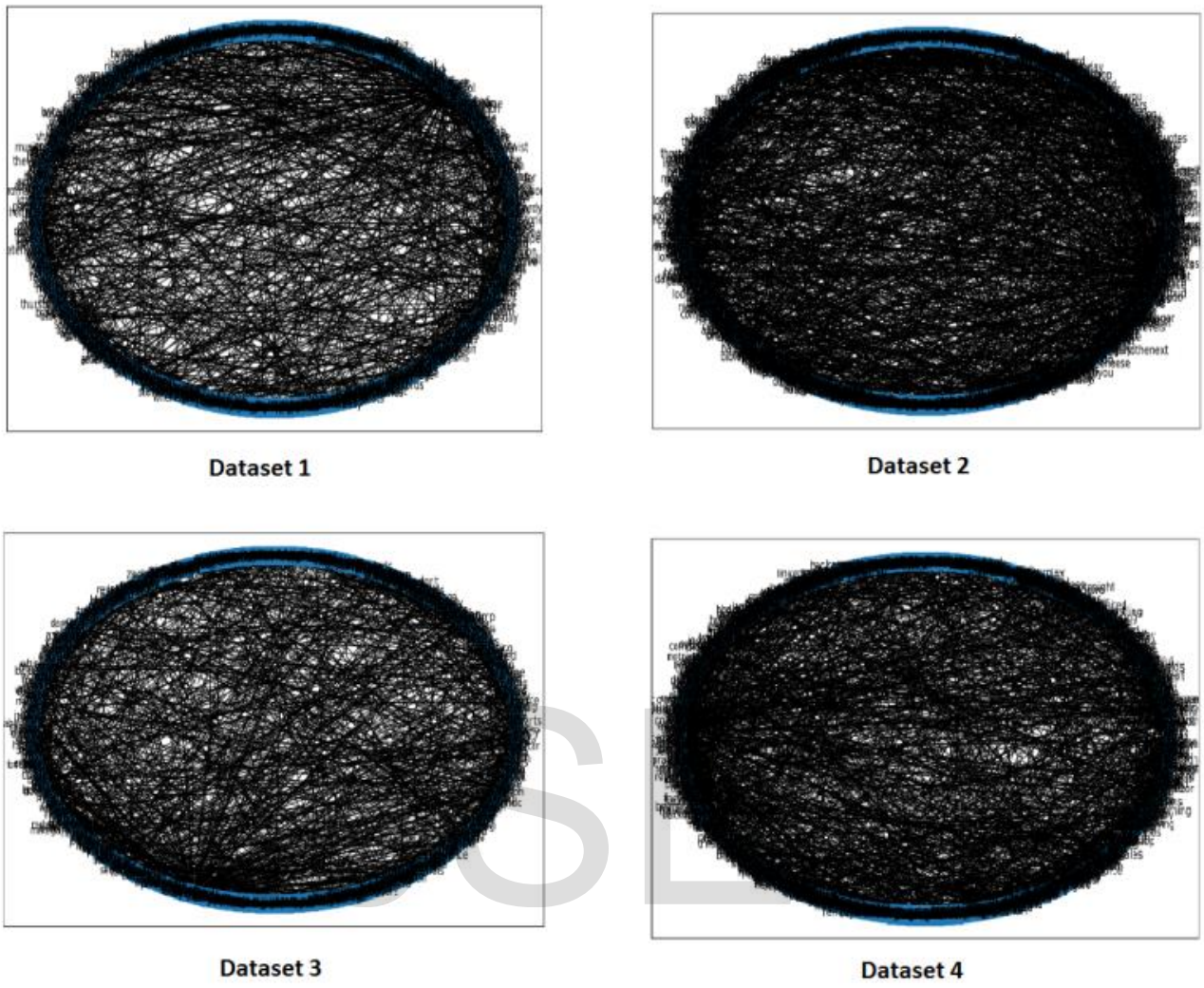

Figure 2: Dataset Representation

attributed graph. After these processes, we use to visualize the result of each step, as almost all our results are in the form of a graph. Hence we use networks python library to plot the graph for our resulted edges pair. Dataset description is given in table 1.

Table 1: Tweets related to various events

Dataset	Events	#tweets
Dataset 1	#AmericanIdol	100
Dataset 2	#Dame	100
Dataset 3	#GoodFriday	100
Dataset 4	#RedSkins	100

The graph will be constructed like dense graph (figure 1) because there are so many words in a corpus.

4.2 Experimental Setup

To experience that how we get the desirable results from our proposed methods and to check the efficacy of the proposed methods, we applied the methodology for the keyword extraction on graph-based approach on the real-world dataset.

4.3 Evaluation Metrics

To evaluate the performance, we have used the standard evaluation metrics called precision which describes the accurately classified keywords over entire keywords as outlined using the given equation. Here, the accurately obtained keywords by the algorithm are interpreted as true positive, TP and the extracted keyword which is not associated with the topics are expressed as false positive, FP.

$$\text{Precision} = TP / (TP + FP)$$

4.4 Evaluation Results

From the evaluation results, it can be said that top-20 and top-40 keywords will be more relevant to the topic. From dataset1, highest precision is 75 for k=20 and k=40 which means top 20 and top 40 keywords will give us the relevant context. For dataset2, k = 20, k = 40, k= 80 and k = 100 are giving same precision of 70%. In dataset3 and dataset4, the highest

International Journal of Scientific & Engineering Research, Volume 10, Issue 8, August-2019
ISSN 2229-5518

Figure 3: Word clouds of the top-60 keywords extracted from each of the four datasets

precision is 75% and this precision is obtained by the top 20 keywords. The precision results of each dataset are illustrated on table 2.

Table 2: Performance evaluation results in term of precision for keyword extraction

Dataset	k=20	k=40	k=60	k=80	k=100
Dataset 1	75%	75%	73.33%	72.50%	71%
Dataset 2	70%	70%	68.33%	70%	70%
Dataset 3	75%	67.50%	68.33%	70%	68%
Dataset 4	75%	65%	65%	66.25%	67%

Word clouds are based on the frequency of individual words found in the available text after stop words removal. The most frequented words in the corpus are visualized by the different techniques, the word cloud is one of them. It is common and simple and mostly presented on the Internet and websites also with explanation and without explanation. A commonly cited issue with word clouds is that they can hinder understanding since they lack information about the relationship between words. We also represented the keywords of my different dataset on the word cloud.

5 Conclusion

Keywords provide a dense representation of the content of a document. Graph-based methods for keyword extraction are inherently unsupervised and have fundamental aim to build a network of words and then rank the nodes. There are a lot of approaches to extract the keywords. But for microblogging dataset, only a few methods work because of the small length of the texts (tweets). The famous TF-IDF does not work over this dataset because it does not consider the grammatical relations. So, considering the grammatical coherence and cohesion, we have selected a graph-based approach. In this paper, a detailed description of existing approaches for keyword extraction is considered; the review of related work on supervised and unsupervised methods with a special focus on the graph-based methods. At first, we have crawled tweets and created 4 different datasets. A preprocessing is done to make the dataset fit for our model. Furthermore, we have created a co-occurrence graph from this dataset. We applied the graph-based keyword extraction algorithm over this co-occurrence graph that will give the score for each keyword. Finally, we ranked the keywords according to its importance value. It will also calculate the precision of the extracted keywords. If we want to visualize those extracted keywords then we can also visualize those important keywords from the dataset through the word cloud. Even though word cloud representation is very common nowadays, there are so many representations are available on the Internet. It means to create graph-based data, extract keywords from that dataset calculate the scores of keywords from data and calculate the precision of extracted keywords are the basic assumptions of this proposed work. There is a vast field of text analytics and many advance types of research are coming in front of us in today's time so we can extend this work regarding many fields. In the future, we can apply our model over the dataset crawled from the communication portal. This will give the current trend and it will also reduce the effort. Just observing these keywords, we will understand the overall context of the dataset.

References

[1] Y. Wen, H. Yuan, and P. Zhang, "Research on keyword extraction based on word2vec weighted textrank," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2016, pp.2109–2113.

[2] J. Cao, Z. Jiang, M. Huang, and K. Wang, "A way to improve graphbased keyword extraction," in *2015 IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2015, pp. 166–170.

[3] M. R. Islam and M. R. Islam, "An improved keyword extraction method using graph based random walk model," in *2008 11th International Conference on Computer and Information Technology*. IEEE, 2008, pp.225–229.

[4] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2004, pp. 170–173.

[5] J. Zhou, X. Yu, Y. Liu, and Z. Yu, "Ranking keyword search results with query logs," in *2014 IEEE International Congress on Big Data*. IEEE, 2014, pp. 770–771.

[6] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural*

International Journal of Scientific & Engineering Research, Volume 10, Issue 8, August-2019
ISSN 2229-5518

language processing, 2004, pp. 404–411.

[7] W. S. X. Tao, "An introduction to pagerank algorithm theory [j]," *Library and Information Service*, vol. 2, 2003.

[8] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[9] W. Li and J. Zhao, "Textrank algorithm by exploiting wikipedia for short text keywords extraction," in *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)*. IEEE, 2016, pp. 683–686.

[10] S. Beliga, A. Mestrovi ˇ c, and S. Martin ´ ci ˇ c-Ip ´ si ˇ c, "An overview of graph-based keyword extraction methods and approaches," *Journal of information and organizational sciences*, vol. 39, no. 1, pp. 1–20, 2015.

[11] Y. Ohsawa, N. E. Benson, and M. Yachida, "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-ADL'98-*. IEEE, 1998, pp. 12–18.

[12] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1031–1040.

[13] J. Wang, J. Liu, and C. Wang, "Keyword extraction based on pagerank," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2007, pp. 857–864.

[14] M. Litvak and M. Last, "Graph-based keyword extraction for singledocument summarization," in *Proceedings of the workshop on Multisource Multilingual Information Extraction and Summarization*. Association for Computational Linguistics, 2008, pp. 17–24.