Artificial Intelligence Review (2021) 54:4731–4770
https://doi.org/10.1007/s10462-021-10010-6

A survey on different dimensions for graphical keyword extraction techniques

Issues and Challenges

Muskan Garg[1]

Published online: 23 April 2021

Abstract
The transmission from offline activities to online activities due to the social disorder evolved from COVID-19 pandemic lockdown has led to increase in the online economic and social activities. In this regard, the Automatic Keyword Extraction (AKE) from textual data has become even more interesting due to its application over different domains of Natural Language Processing (NLP). It is observed that the Graphical Keyword Extraction Techniques (GKET) use Graph of Words (GoW) in literature for analysis in different dimensions. In this article, efforts have been made to study these different dimensions for GKET, namely, the GoW representation, the statistical properties of GoW, the stability of the structure of GoW, the diversity in approaches over GoW for GKET, and the ranking of nodes in GoW. To elucidate these different dimensions, a comprehensive survey of GKET is carried in different domains to make some inferences out of the existing literature. These inferences are used to lay down possible research directions for interdisciplinary studies of network science and NLP. In addition, the experimental results are analysed to compare and contrast the existing GKET over 21 different dataset, to analyse the Word Co-occurrence Networks (WCN) for 15 different languages, and to study the structure of WCN for different genres. In this article, some strong correspondences in different disciplinary approaches are identified for different dimensions, namely, GoW representation: 'Line Graphs' and 'Bigram Words Graphs'; Feature extraction and selection using eigenvalues: 'Random Walk' and 'Spectral Clustering'. Different observations over the need to integrate multiple dimensions has open new research directions in the inter-disciplinary field of network science and NLP, applicable to handle streaming data and language-independent NLP.

Keywords Keyword extraction · Graph of words · Language networks · Word co-occurrence networks

✉ Muskan Garg
muskanphd@gmail.com

[1] Amity University Rajasthan, Jaipur, India

1 Introduction

1.1 Motivation

As per Digital 2020 reports Social (2020), 4.5 billion people are using the internet which is increasing with evolving era of pandemic disease of COVID-19 at a much faster rate than the normal. People are sharing more information and many economic activities are carried online in different countries to reduce the mobility, in-person interactions and to follow lock-down instructions. The change in communication patterns during and after the lock-down, unemployment and other social changes has given rise to the different problems like Psychiatric disorders Guessoum et al. (2020); depression, anxiety, and sleep disturbances due to increased time spent on the internet Gualano et al. (2020). The evaluation of medical reports, legal documents and scientific articles are some other important application domains of the textual information retrieval. It is mandatory to pre-process the documents to identify important words or segments in the text for such application domains. The graphical techniques are independently studied for several different applications, languages, domains, genres and theoretical aspects of the word distribution resulting into a large body of research. The key idea of this article is to connect dots between the network science and the graphical keyword extraction.

1.2 Introduction to keyword extraction

The keyword extraction from textual documents is one of the most promising areas of Natural Language Processing (NLP) and Information Retrieval (IR). It has undergone the years of research and development to bring out useful and actionable insights out of the textual documents.

Keyword extraction is one of the most elementary research for Natural Language Processing (NLP) and Information Retrieval (IR). Keyword extraction is defined as identifying the most relevant words or the set of words which describe the central theme of any document Abilhoa and De Castro (2014), Lahiri et al. (2017), Mihalcea and Tarau (2004). An Automatic Keyword Extraction (AKE) is a process of feeding any document/ documents as an input to a device which shall automatically process the information and shall provide important words or segments which are directly applicable to other NLP problem domains like text summarization Bharti and Babu (2017), Lin (2004), Litvak and Last (2008), topic detection Liu et al. (2010), Bougouin et al. (2013), event detection Garg and Kumar (2018a), and indexing, to name a few. Indirectly, keyword extraction is applicable to industrial applications like automatic question-answering machine, spam detection, rumours detection, false information detection, reputation analysis, sentiment analysis and many more.

Keyword extraction can be formulated on the basis of statistical, graphical, learning based and hybrid techniques. The Graphical Keyword Extraction Techniques (GKET) are observed as more scalable, and computationally less expensive than the other Automatic Keyword Extraction (AKE) techniques.

1.3 Graphical keyword extraction technique: Different dimensions

The patterns and structure of conversations in streaming data changes with time. Thus, recent research advancements are more inclined towards the GKET to study the communication patterns and the connections among words in the Graph of Words (GoW). The GKET are applicable to the ever-changing patterns of words in the GOW evolved from textual documents. The graphical patterns are more scalable and computationally less expensive as compared to other keyword extraction approaches due to the development of efficient Python libraries Tixier et al. (2016) which are used for generating and analyzing GoW.

Recently, much of the research work and developments are observed in multiple dimensions of GoW analysis. The idea behind this article is to study different dimensions in which the complex network of words and/ or information are examined for an application domain of AKE. GKET are comprehensively studied to connect and link these dimensions among each other. To the best of my knowledge, none of the existing survey papers Nasar et al. (2019), Siddiqi and Sharan (2015) have given multiple research dimensions and directions in the field of GKET, as discussed in this article.

Understanding and development of many efficient tools[1]Paranyushkin (2019), Tixier et al. (2016), Zhang et al. (2018), availability of online dataset[2], and implemented methods[3] have made this field even more rich and visible to many academic researchers for working over GKET. The abbreviations used in this article are enumerated in Table 1. The major contributions of this article are based on the five different dimensions for using GoW to extract Keywords which are described in Fig. 1.

The first dimension is the representation of GoW which is observed in many different versions on the basis of the characteristics of its nodes and its connectivity. Although, there are many different types of GoW representations which are possible to represent the word distribution in textual documents, but there are six major GoW representations which are used for GKET, namely, Sliding window based GoW, Context Aware Graph (CAG), Multi-Layer Network (MLN), Heterogeneous Information Network (HIN), Line Graphs and Bi-gram Word Graph. These GoW representations are discussed on the basis of directed/ un-directed graph, weighted/ un-weighted graph, and adjacent/ all-pair adjacency network.

The second dimension is the statistical properties of GoW. After representation of the word distribution of textual data as GoW, different statistical properties are studied. Initially, researchers used to consider the Term Frequency- Inverse Document Frequency (TF-IDF) to rank nodes (words) in the GoW. However, in the last few decades, graph based metrics are used for identifying the characteristics of nodes and edges. Also, other statistical properties like the degree-distribution, the small-world properties, and the scale-free property are studied for NLP in existing literature. However, for GKET, the feature extraction and the feature selection are widely used for different types of permutations and combinations of the existing graphical metrics over GoW.

The third dimension is the stability of the structure of GoW. The stability of the structure of GoW is defined as the extent to which the properties and semantics of the network remains similar for increase or decrease in the size of dataset. It keeps track of the robustness and the scalability of a network to understand some common patterns.

[1] https://github.com/yinruiqing/corpus2graph
[2] https://github.com/LIAAD/KeywordExtractor-Datasets
[3] https://github.com/boudinfl/pke

Table 1 List of Abbreviations

Abbreviation	Definition
ACC	Average Clustering Coefficient
AI	Artificial Intelligence
AKE	Automatic Keyword Extraction
AOF	Average Occurrence Frequency
ASP	Average Shortest Path
ASPL	Average Shortest Path Length
cTPR	context-sensitive Topical PageRank
CAG	Context Aware Graph
CC	Clustering Coefficient
CSTR	Computer Science Technical Reports
COVID	COrona VIrus Disease
GKET	Graphical Keyword Extraction Techniuqes
GoW	Graph of Words
HIN	Heterogeneous Information Network
HITS	Hyperlink-Induced Topic Search
IR	Information Retrieval
KDD	Knowledge Data Discovery
KECNW	Keyword Extraction using Collective Node Weight
LDA	Latent Dirichlet Allocation
MADM	Multiple Attribute Decision Making
MAP	Mean Average Precision
MCFS	Multi-Cluster Feature Selection
MLN	Multi-Layer Networks
MpR	MultipartiteRank
MRR	Mean Reciprocal Rank
NFA	Number of False Alarms
NLP	Natural Language Processing
PCA	Principle Component Analysis
PFA	Principle Feature Analysis
POS	Part of Speech
RT	Re-Tweet
SBKE	Selectivity Based Keyword Extraction
SMART	System for the Mechanical Analysis and Retrieval of Text
TF – IDF	Term Frequency – Inverse Document Frequency
TKG	Twitter Keyword Graph
TPR	Topical PageRank
URL	Uniform Resource Locator
WCN	Word Co-occurrence Network
WWW	World Wide Web

The stability of GoW ensures the possibility to propose network models for GoW which can be further enhanced as language independent and genre independent model. This dimension is observed for social media text, Chinese and English documents. Since,

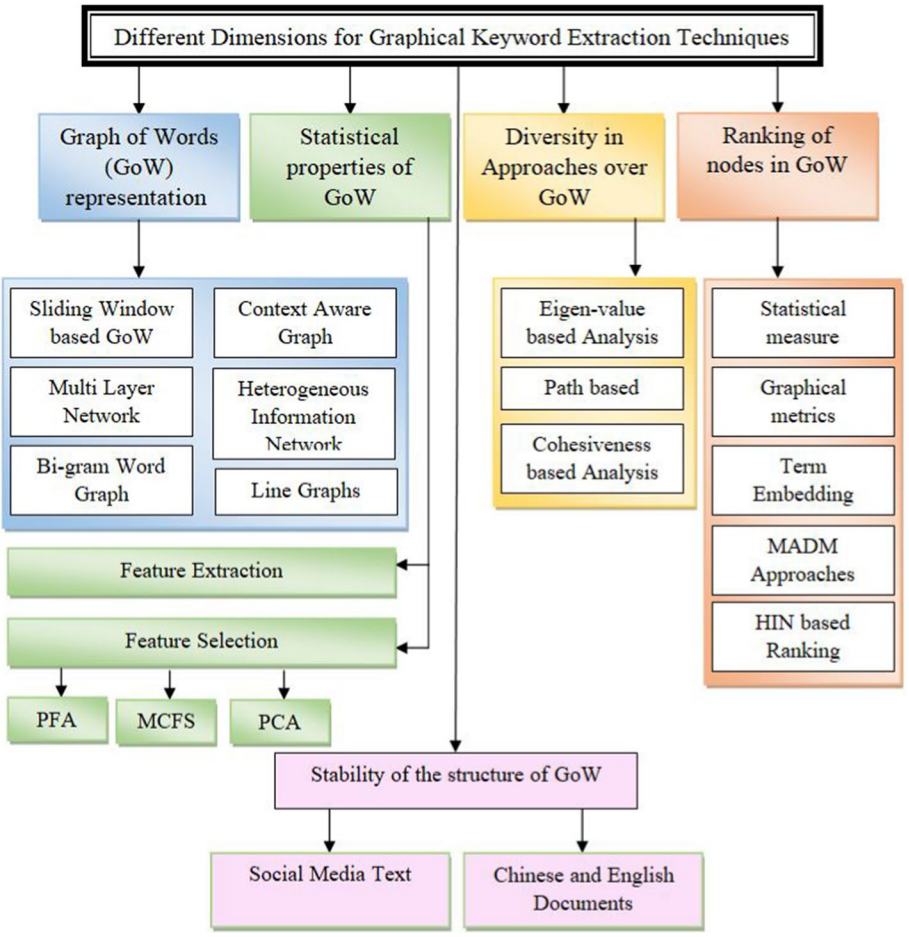

Fig. 1 Outline of the article: Different dimensions for GKET over graph of words

there is not much existing literature in this dimension, it is decribed in this article to recommend this approach as a potential research direction for GKET.

The fourth dimension which is discussed in this article is Diversity in Approaches over GoW for GKET. The GKET are classified into three types, namely, Eigen-value based Analysis using random-walk based approaches; Path-based Analysis using GoW evolved from phrases as path of words; and Cohesiveness based Analysis using network properties of GoW.

The fifth dimension is ranking of nodes in GoW which describes different types of ranking measures used for ranking nodes (words) in GoW. There is not much research work on the node ranking for different types of GoW representations. However, the five types of node ranking for GoW are observed in literature which are the statistical measure, the graphical measure, the term embedding, the Multiple Attribute Decision Making (MADM) approaches, and the HIN based ranking.

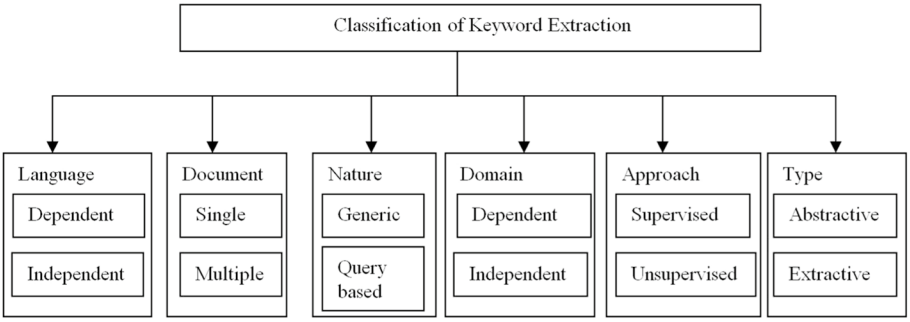

Fig. 2 Classification of different keyword extraction techniques

1.4 Organization of the paper

The article is organized into different Sections. Section 2 describes the background information about classifications of GKET and enlist the dataset used for keyword extraction in existing research. Section 3 explains the pre-processing of textual information for the keyword extraction which is used for 'the representation of GoW', the first dimension for GKET. The GoW is considered as a complex network and its statistical properties are discussed in Sect. 4. Section 5 elucidates the stability of the structure of GoW for different languages and genres. Section 6 describes the diversity in the approaches for GKET by using GoW. Further, the ranking of nodes in GoW is discussed in Sect. 7. In Sect. 8, the experimental results are analysed to compare and contrast the existing GKET over 21 different dataset; to analyse the Word Co-occurrence Networks (WCN) for 15 different languages; and to study the structure of WCN for different genres. Section 9 discuss different applications and gives an outlook of open challenges and future research directions. Finally, Sect. 10 concludes the article.

2 Background information

There are many GKET which are developed in last 30 years. Although, this research area is not newly introduced but the idea of exploring real-time conversations by using GoW is a new and potential research domain which has been minimally explored in literature. The keyword extraction techniques can be supervised and unsupervised. However, the supervised keyword extraction techniques are out of the scope of this article. This Section contains the information about the classification of GKET and enlists the existing dataset for AKE.

2.1 Classification of the graphical keyword extraction techniques

The AKE approach is usually characterised on the basis of many classifications with different perspectives as shown in Fig. 2. The classifications discussed in this article are limited to the GKET.

Language Independence Academic researchers are working on AKE from texual documents of different languages, for instance, Telugu Naidu et al. (2018), Chinese Chen et al. (2020a), Li et al. (2017), English (KPTimes and JPTimes), German Kölbl et al. (2020), Dutch Sterckx et al. (2018), Polish (Pak2018), Spanish Aquino and Lanzarini (2015), Portuguese Marujo et al. (2013), and French Bougouin et al. (2013). The structure of the word distribution in Chinese and English documents is well studied using GoW evolved from these documents and different patterns are observed.

Single Document and Multi-document Keywords can be classified as single document keyword extraction and multi-document keyword extraction based on number of documents it considers for identifying keywords in single iteration.

Generic/Query based The Query based keyword extraction techniques are used when there is demand over the specific type of keywords, for instance, in medical domain, the keywords can be related to various diseases, treatment names, sentiments or emotions about a patient, or how important the treatment is. The generic keyword extraction techniques give keywords usually by considering the nouns and the adjectives Tixier et al. (2016) using Part of Speech (POS) tagging based filtration.

Application Domain/Genre Dependent A keyword extraction technique is said to be an application domain dependent or genre dependent if it is created only for specific type of documents. It is observed that usually application specific or genre specific keyword extraction technique is developed and hence, there are very few techniques which are generalized and are scalable to multiple genres, domains or languages Campos et al. (2020).

Supervised/Unsupervised Techniques Supervised techniques are useful when keywords have to be extracted from static and application specific documents. Supervised keyword extraction techniques are usually genre dependent because training is held in the same domain as that of testing domain. The unsupervised keyword extraction techniques do not require excess and repetitive patterns in data for training and testing. It is directly applied over textual documents and are highly recommended for industrial applications over real-time and dynamic data.

Abstractive or Extractive This classification is new in the area of keyword extraction. This was earlier used for text summarization, namely, abstractive text summarization and extractive text summarization. The Abstractive keyword extraction techniques are those techniques which are not necessarily mentioned in the local document, for instance, CollabRank Wan and Xiao (2008). Alternatively, most of the GKET are extractive approaches which give those words that are present in the corpus Wan and Xiao (2008), Tixier et al. (2016).

2.2 Keyword extraction dataset

In this Section, the existing dataset for AKE are discussed which are enlisted in Tables 2 and 3, for dataset from year 2004 to 2015 and from year 2016 to 2020, respectively.

Every keyword extraction dataset consist of textual documents. Each document can be short text document or long text document. Many variations in the length of a dataset is observed while considering different genres of the text. The results for different genres using the GKET differ on the basis of length of the document, size of the dataset, and language of the dataset. Since the AKE is based on the subjective analysis, academic researchers use the objectivity of the annotated dataset for reliable validation of AKE approaches. Thus, the annotation of an AKE dataset is very important step and annotations are classified as author annotated, experts annotated, crowdsourcing or any of the combination of

Table 2 Existing dataset for keyword extraction from 2003 to 2015

Name	#Doc	General description
CSTR	1800	The computer science technical reports are considered in this dataset which was introduced in 1999. Witten et al. (2005)
Inspec	2000	The abstracts from 1998 to 2002 of Computers and Control, and Information Technology. Hulth (2003)
eBooks	101	The eBook dataset is randomly chosen from all kinds of fields. Huang et al. (2006)
Nguyen2007	211	Corpus of scientific publications annotated for keyphrases Nguyen and Kan (2007)
Wiki20	20	Links all important phrases in a document to Wikipedia articles.
Schutz2008	1231	It consists of research papers selected from PubMed Central and are distributed across 254 different journals. Schutz (2008)
Fao30	30	FAO of the UN, 30 documents. Crowdsourcing by six professional annotators at FAO. Medelyan et al. (2009)
Fao780	780	FAO of the UN, 780 documents. Crowdsourcing by six professional annotators at FAO. Medelyan and Witten (2008)
Krapivin2009	2304	Full CS journal scientific articles in ACM from 2003 to 2005 Krapivin et al. (2009)
Citeulike180	180	The dataset is based on a subset of CiteULike.org containing documents that are indexed with at least three keywords on which at least two users have agreed. Medelyan et al. (2009)
SemEval2010	244	Full scientific articles from ACM, created for SemEval2010 Task 5
SemEval2010	284	ACM Digital Library papers in four ACM 1998 classification distributed systems; information search and retrieval; distributed artificial intelligence (multiagent systems); and social and behavioral sciences (economics). Kim et al. (2010)
MPQA	535	It is based on data of news reports available from 187 various US and foreign news sources from June 2001 to May 2002.
Marujo2012	450	The Portuguese news stories were adapted in English to carry out keyword extraction in English language. Marujo et al. (2013)
500N-KPCrowd-v1.1	500	10 different categories (art and culture; business; crime; fashion; health; politics us; politics world; science; sports; technology) with 50 docs per category
110-PT-BN-KP	110	News from the European Portuguese ALERT Broadcast News database. Marujo et al. (2013)
Wikinews	100	French corpus created from the French version of WikiNews that contains 100 news articles published between May 2012 and December 2012. Bougouin et al. (2013)
KDD	704	The abstracts from the articles of KDD conference papers are collected Caragea et al. (2014)

Table 2 (continued)

Name	#Doc	General description
WWW	1248	The abstracts from the articles of WWW conference papers are collected Caragea et al. (2014)
Blogs	14000	Authors merged all the posts in a blog into one large document, because TF-IDF is usually applied to a set of single documents. No special features or linking were used in this dataset. Park et al. (2014)
CACIC	888	Spanish articles published between 2005 and 2013 in the Argentine Congress of Computer Science Aquino and Lanzarini (2015)
PubMed	500	Full-text papers collected from PubMed Central, which comprises over 26 million citations for biomedical literature Song et al. (2015)

Table 3 Recent dataset for keyword extraction from 2016 to 2020

Name	#Doc	General description
INND	2371	Documents related to Times of India, The Hindu, Hindustan Times, Indian Express Thomas et al. (2016)
SemEval2017	500	Paragraphs from Science Direct journal articles from Computer Science, Material Sciences and Physics. Augenstein et al. (2017)
KP20k	567830	It contains Computer Science articles for comparing abstractive keyword generation Meng et al. (2017)
Emails	212 + 107	Dataset was extracted from the Enron collection (Single email and email threads) and manually classified as either "private" or "corporate". Lahiri et al. (2017)
Pak2018	50	Polish abstracts of journal publications on technical topics collected from Measurement Automation and Monitoring Campos et al. (2020)
Sterckx2018	1200-2000	Three Belgian media companies' data in Dutch language is used to construct total of three datasets, each having a different focus (The first one is public-service broadcaster VRT which is subset of its official news channel website "De Redactie" and its specialized sports section Sporza. The second media company, Sanoma, owns publishing of lifestyle, fashion, and health magazines, which is used to build Lifestyle Magazines test collection. Lastly, Belga, the third media company, has a digital press that is used to construct News articles dataset.)
Twitter 5 dataset	7000	Tweets collected from Donald Trump, Harry Potter, IPL, Uri Attack and American Election. Biswas et al. (2018)
Disaster Tweets	122266	This dataset contains the Tweets regarding Boston Bombing, Hurricane Harvey, and Hurricane Sandy. Chowdhury et al. (2019)
KPTimes, JPTimes	279,923, 10000	The corpus contains editor assigned keywords that were collected by crawling New York Times news website and Japan Times online news portal. Gallina et al. (2019)
German Text	100673	Authors collected 100,673 texts in German language from heise.de. For each document the headline and the text body is available. Kölbl et al. (2020)
Chinese	10,000	The Chinese corpus comes from the NLPIR microblog corpus from June 2012 to July 2012 in Sogou Lab,2. Chen et al. (2020a)

these. Some dataset are annotated before carrying the experiments but few are annotated after results are obtained and are defined as the post-hoc evaluation Lahiri et al. (2017).

The keyword extraction dataset are proposed since 20 years for different domains and different genres. However, the pattern of old keyword extraction dataset vary from those of newly introduced dataset in terms of updated active vocabulary of users/ people around the world. There is need to link and study the connections of words in GoW. Recent studies over the structure and dynamics of GoW may help in determining new insights for real-time conversations or streaming data. Hence, Table 2 shows the existing dataset for keyword extraction till 2015 and Table 3 shows the list of keyword extraction dataset after year 2016. There is need to compare and contrast the existing techniques and study the new ones for the dataset list given in Table 3, especially for informal/ ill-formed text (Micro-blogs and digital conversations).

3 The graph of words representation

In literature, the GoW is generated by using different rule and different approaches in literature. These are classified on the basis of the conceptual graphs and the co-occurrence networks. One of the most prominent approach to represent textual data for GKET is the WCN. The words are tokenized and linked to each other as the path of the nodes where each node represents a term. A window is assigned to every document whose size determines the number of tokens connected in the form of fully-connected (directed/ undirected) network for the tokens of that window. For generating GoW, the pre-processing is performed over textual documents which is discussed in this Section.

3.1 Pre-processing documents

The pre-processing step in NLP is one of the most challenging and interesting research areas. The tokenization of the text gives an individual term which is considered as the candidate keyword. The candidate keywords act as the set of words in a sample out of which the most important keywords are extracted. The pre-processing approaches are different for well formed text and informal text.

3.1.1 Well-formed text

The documents which are prepared using the computer system is considered as official documents which are called well-formed text, for example, the news articles, the scientific articles, and the books to name a few. Well-formed documents are comparatively easier to handle for NLP.

Earlier, the researchers used to remove stop-word list from textual documents Matsuo and Ishizuka (2004). This list was obtained from the SMART Salton (1988) information retrieval system. Further, a new method is developed, YAKE Campos et al. (2020) for internal stop-words removal which has significantly improved the performance of the proposed keyword extraction technique.

The lemmatization and stemming are equally important steps in pre- processing Rose et al. (2010). Stemming is performed using the Porter stemmer Matsuo and Ishizuka (2004), Porter (1980). Porter Stemmer is implemented in R SnowballC Package Tixier et al. (2016). The Part Of Speech (POS) tagging and screening was performed using

OpenNLP R Package Tixier et al. (2016). A tokenization tool was introduced as Treebank-WordTokenizer to tokenize textual documents which is implemented using the python Natural Language ToolKit Sarkar (2019).

3.1.2 Ill-formed/ Informal text

The textual data contains the grammatical error, use of slang, abbreviations and short forms of the frequently used words in daily life.The pre-processing steps helps in identifying the most important tokens which may or may not give redundant information about the text. Some additional research areas in ill-formed text pre-processing are text-normalization, and Hashtag segmentation. The ill-formed text is usually found in short text and Miroblogs.

The usernames and retweet symbol do not contribute anything significant for the problem domain of keyword extraction and may act as noise Biswas et al. (2018). Thus, usernames and retweet symbols are removed. Sometimes other regular expression like URLs, Hashtags are handles separately.

3.2 Sliding window based graph of words

After pre-processing, GoW are generated using the tokenized version of the textual document. An edge-weight in the network measures the frequency of terms which occurs together in a document depending upon the window size. The underlying assumption is that all the words present in a document have association with the other words. Modulo is a window size outside of which the relationship is not taken into consideration. It is observed that more the window size of sliding window, less will be precision Hulth (2003). Thus, generally the window size is considered to be 3 Rousseau and Vazirgiannis (2015) as shown in Fig. 4. However, if sliding window is set as 2, it becomes word-adjacency graph where two terms are connected only if they are adjacent to each other Garg and Kumar (2018a) as shown in Fig. 3.

Depending on the nature of these interactions, an edge and a graph can be weighted/ unweighted, directed/ undirected and adjacent pairs/ fully connected Abilhoa and De Castro (2014). The directed graph is considered to maintain the lexical sequence and an undirected graph, otherwise; a weighted graph is used for considering the edge weight as co-occurrence frequency; and an adjacent connected graph is used for connected adjacent terms only. However, the sliding window is used for semi-connected or intermediate graphs in case of connections. To decide the static value of sliding window (connections), a parameter is defined. The results of keyword extraction may vary with changing values of this parameter as different features are extracted depend upon these parametric values.

3.3 Context aware graph

The context set by a sentence is often used by the consecutive sentences, imparting continuity in communication. This phenomenon is called entailment, which is a well studied concept in linguistics Duari and Bhatnagar (2019). In this method, the window slides over two consecutive sentences and the candidates co-occurring therein are linked. This eliminates the need of integer-valued window-size parameter, and captures the contextual co-occurrence of words in text.

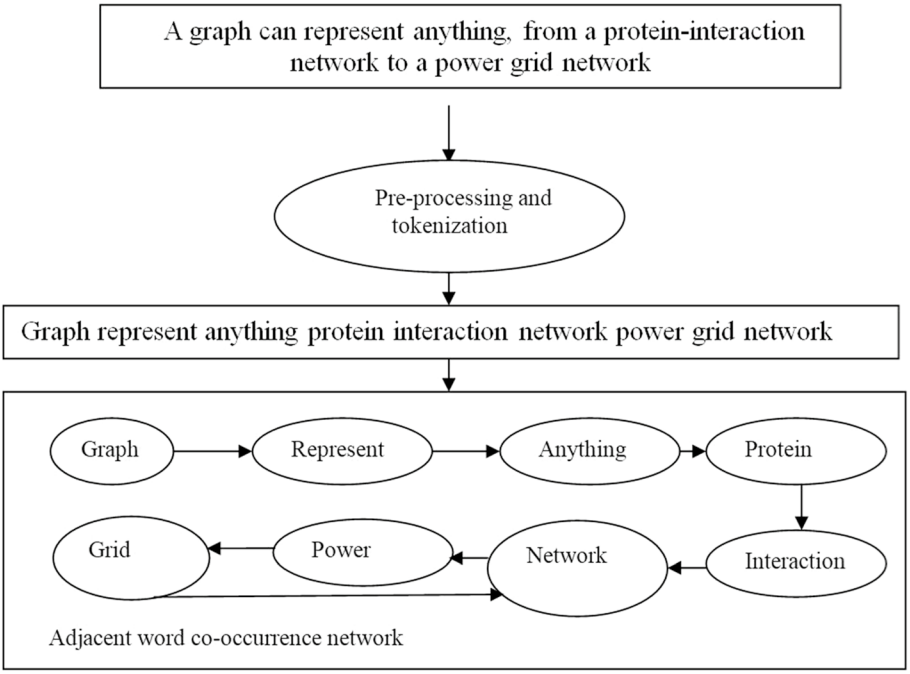

Fig. 3 Generating adjacent word co-occurrence network

3.4 Multi-layer network

Apart from 1-Dimensional homogeneous networks, there are multi-dimensional and heterogeneous types of networks which can be used for document representation. As the name suggests, multiple layers are considered for identifying keywords from the textual documents. Although, these two layers can be given any subjective context, but in recent literature, Multi-Layer Networks (MLN) are studied for NLP using three layers namely, words – syllable – grapheme Martinčić-Ipšić et al. (2016) for Croatian and English languages.

The MLN of words was considered to propose a network model Yang et al. (2018). However, every layer contains homogeneous type of data. In this network, sentence s is linked to all those words which are contained in s. Sentences are connected to each other on the basis of a similarity score like Jaccard coefficient or cosine similarity. the eigenvalues for words in GoW can be synthesized using this arrangement.

3.5 Heterogeneous information network

Another type of graphical representation for textual documents is Heterogeneous Information Network (HIN). In this network, meta-paths are introduced for different types of data and relations among them. These meta-paths are used to find similarities among different entities. In literature, the HIN has not been used for keyword extraction yet but this can be potential research area in near future. The advantage of using HIN is

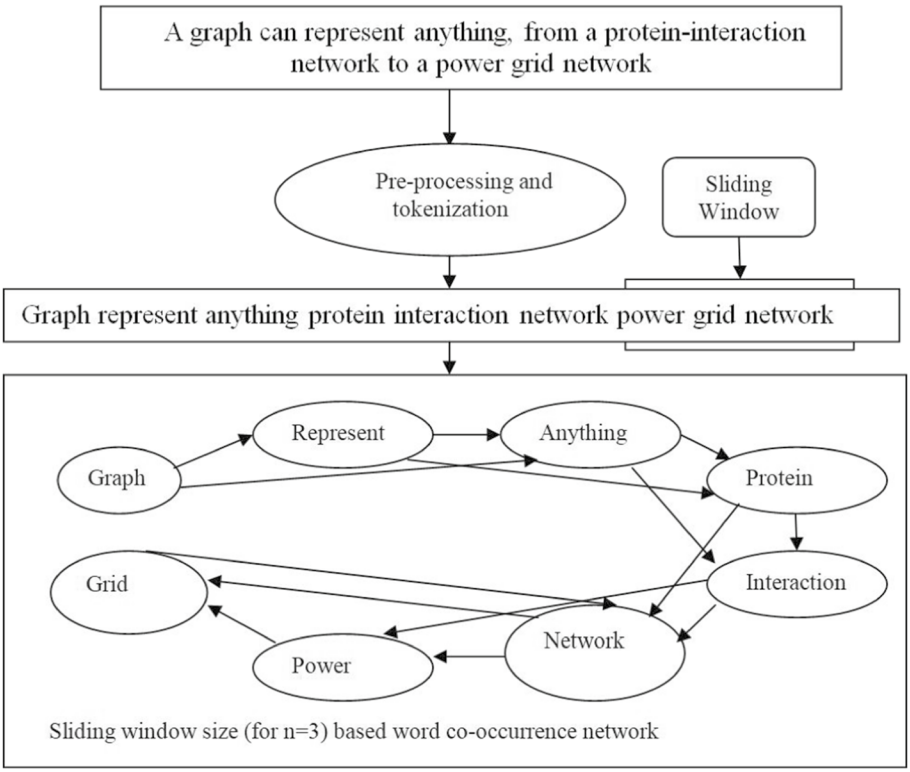

Fig. 4 Generating sliding window size (for n=3) based word co-occurrence network

that constraint-based random walk techniques can be used for finding similarities among words Meilian and Danna (2020).

3.6 Line graphs

The line graphs are those graphs which represents co-occurrence as a node of the graph and occurrence of word as an edge of the graph Harary and Norman (1960). Other terms used for the line graph are the covering graph, the derivative, the edge-to-vertex dual, the conjugate, the representative graph or the edge graph. In the context of complex network theory, the line graph of a random network preserves many of the properties of the network such as the small-world property (the existence of short paths between all pairs of vertices) and the shape of its degree distribution. It is observed that the WCN follow small-world property Liang et al. (2009), Gao et al. (2014), Garg and Kumar (2018a). Thus, line graph can be used for AKE using graphical properties for nodes. The research work in literature has discovered many useful insights by using various characteristics of edge weights in WCN, for instance, TKG Abilhoa and De Castro (2014), BArank Garg and Kumar (2018a) and NErank Bellaachia and Al-Dhelaan (2012). Thus, line graphs Evans and Lambiotte (2010) can also be used for analysis of GoW in future.

3.7 Bi-gram word graph

The GoW having each node as bi-gram connected with another bi-gram having the same later word and former word in tail node and head node, respectively. Such bi-gram Word Graph Rudra et al. (2016) have been used in literature for identifying disaster specific Tweets during Social Media Analysis. This shows that bigram word graph and line graph have strong correspondence among each other. Since there is minimal research work over line graph for GoW, it was not observed in literature while introducing Bi-gram Word Graph. Thus, it will be interesting to link the both GoW representation in different domains.

4 Statistical properties of graph of words

It is observed that the AKE is a classification approach. The textual document is used to identify candidate words. These candidate words are then classified as keyword or non-keyword. In long textual documents, the keywords varies from the range of 20 to 40 but non-keywords are more than tens of thousands. The ratio between these two varies and thus, there exists imbalanced keyword distribution in textual documents. Hence, it becomes difficult to analyse textual data using machine learning algorithms. The statistical studies are performed for GKET to identify a subset of candidate words which shall retain the properties and semantics of the GoW of original text Duari and Bhatnagar (2019). Before we study the impact of statistical properties over GoW, different feature extraction and feature selection techniques are discussed in this Section which are used for GKET in literature.

4.1 Feature extraction

There are many existing studies in literature which shows the collection of graphical features for words Beliga et al. (2015), finds correlation between them Duari and Bhatnagar (2020), and select significant features among all Duari and Bhatnagar (2019). The graphical and statistical features for word distribution in a textual document as represented in GoW are shown in Table 4.

The features of words are extracted from graphical representation of GoW. Different types of the GoW are generated for GKET as discussed in Sect. 3. The significance of these features in GoW are further used to identify the correlation between them using statistical testing Duari and Bhatnagar (2019), Vega-Oliveros et al. (2019). Recently, the reduction in number of features is performed using feature selection technique Vega-Oliveros et al. (2019) in literature.

4.2 Word feature selection

To reduce the dimensionality, the feature selection method is used in literature for selecting only significant features and to make the AKE computationally less expensive.

Table 4 GraphicalfFeature extraction for words in textual document

Features	Inferences for GoW	References
Distance from a central Node	The closer a node is to the most central node more is the probability that it is an important keyword.	Biswas et al. (2018)
Selectivity centrality	A measure of importance of a node by measuring the fraction of vertex strength and vertex degree	Beliga et al. (2016)
Degree	If a word is occurring in association with other words in sliding window from textual document	Beliga et al. (2015), Boudin (xxxx), Lahiri et al. (2014)
TF-IDF	Statistical measure to find important words using term frequency and its inverse document frequency	Matsuo et al. (2001), Lahiri et al. (2014), Meladianos et al. (2017), Wan and Xiao (2008), Campos et al. (2020), Duari and Bhatnagar (2020)
Betweenness centrality	The capacity of information transmission from vertices and showing more inclination towards *path based network* of GoW	Vega-Oliveros et al. (2019), Beliga et al. (2015), Boudin (xxxx)
Eigenvalue centrality	A word, connected to other words in a sliding window is known by the importance of its neighbour.	Duari and Bhatnagar (2020), Lahiri et al. (2014), Boudin (xxxx), Abilhoa and De Castro (2014)
PageRank	The word co-occurrence network where connected words are believed to contribute towards the importance of each other	Tsatsaronis et al. (2010), Duari and Bhatnagar (2019)
Closeness centrality	the lengths of the shortest paths from each word to the rest of the GoW.	Lahiri et al. (2014), Duari and Bhatnagar (2020), Abilhoa and De Castro (2014), Beliga et al. (2015), Boudin (xxxx), Mihalcea and Tarau (2004), Zhou et al. (2013)
K-core decomposition	To identify the cohesive sub-graph of the GoW (Important words are more connected to each other)	Rousseau and Vazirgiannis (2015), Meladianos et al. (2017), Duari and Bhatnagar (2019), Duari and Bhatnagar (2020)
Clustering coefficient	Clustering coefficient measures the presence of triangles (cycles of order three) in the GoW. Higher for sliding window 3	Duari and Bhatnagar (2019), Garg and Kumar (2018a)
Eccentricity	The shortest sequence of edges that connect the shortest path between two terms (supports the feature of a path based network)	Duari and Bhatnagar (2019)
Structural holes	Created on removal of the bridge of clusters for other words	Duari and Bhatnagar (2019)
Position of a node	Position and occurrence of a word in textual document	Hotho et al. (2005), Biswas et al. (2018), Florescu and Caragea (2017), Awan and Beg (2020)
Semantic connectivity	Linking distinct concepts in GoW	Campos et al. (2020)
Casing	The abbreviation and uppercase representation of a word in GoW	Campos et al. (2020)
Information Centrality	All paths that originate with a word.	Xie (2005)

Table 4 (continued)

Features	Inferences for GoW	References
Structural diversity index	Normalized entropy of weight of edges	Lahiri et al. (2014)
Information gain	A measure of expected reduction in entropy on the basis of 'usefulness' of an attribute	Song et al. (2003)
Shannon's entropy	Difference between intrinsic and extrinsic entropy	Yang et al. (2013)
N-Gram	generally maintains the lexical sequence of the text	Rose et al. (2010), Hulth (2003), Pudota et al. (2010)
Jaccard similarity	a statistical similarity measure	Zhou et al. (2013)
Cosine similarity	a statistical similarity measure	Erkan and Radev (2004), Wan and Xiao (2008)

Three major feature selection techniques for AKE are used in literature which are discussed in this Section.

4.2.1 Principle feature analysis (PFA)

The Principal Feature Analysis (PFA) is a wrapper method that wraps the search for the best feature subset around a clustering algorithm. The method computes the eigen-values and the empirical orthogonal functions of the co-variance matrix of original features, construct a new subspace dimension of reduced dimension, and cluster using K-Means Vega-Oliveros et al. (2019). Finally, the corresponding original features are the closest to the mean of each cluster.

4.2.2 Multi-cluster feature selection

The Multi-Cluster Feature Selection (MCFS) is a filter method that essentially reduces the selection to a search problem. The manifold regularization can be best preserved based on the spectral analysis and the L1-regularized least-squares regression problem. The recent studies on the structure of microblog WCN show the significance of spectral analysis of GoW Garg and Kumar (2018a).

4.2.3 Principle component analysis (PCA)

The Principal Component Analysis (PCA) is a data reduction method that employs the correlation matrix of the features. It calculates the eigen-values and principal components (PC) producing a new set of features. Each PC is the linear combination of the original features. The first principal component (PC1) is used as a new feature Vega-Oliveros et al. (2019), Gibert et al. (2011) with the highest variance of information, such that the representation is as faithful as possible to the original data.

In recent years, the statistical study for word distribution has given the $\sigma - index$. The normalized standard deviation of the word's spacing distribution is computed in successive occurrence is called $\sigma - index$. In this feature, higher value indicates higher term relevance. They have proved that self-attraction of words is linked to their relevance to the text considered. This index was further improved by studying on random text and finding other factors like skewness of the word distribution in the text Herrera and Pury (2008). Additionally, they have proposed some other statistical factors forward distribution in textual documents while studying geometrical distribution and entropy of the random text.

Recently, similar statistical studies of words distribution in textual documents are considered to proposed hybrid techniques Duari and Bhatnagar (2019) so that they can balance the classification problem of word into keyword or non-keywords.The mean and standard deviation of word distribution is observed for various keyword extraction dataset. Authors have proposed sCAKE keyword extraction Duari and Bhatnagar (2019), Duari and Bhatnagar (2020) technique using statistical analysis based on these observations. These observations help in identifying the behavior of GoW evolved from textual documents. The statistical analysis of word distribution in different keyword extraction dataset Vanyushkin and Graschenko (2020) is still a potential area of research. The mean and standard deviation of keyword extraction dataset varies from one dataset to another. These statistical analyses have given a new research direction to study GoW statistically.

5 Stability of the structure of the graph of words

The stability of GoW is the extent to which the statistical features and other network science metrics remains stable as the size of dataset gets increased. This factor supports the scalability of network science metrics and statistical features over GoW well. Various observations have been made in literature for different domains, different genre and different languages to test the stability of the GoW. Some of the predominant studies for this dimension are studies over Social Media text, Chinese and English documents which are described in this Section.

5.1 Social media text

The GoW is analysed for different languages like English Gao et al. (2014), Chinese Liang et al. (2009), and Microblogs Garg and Kumar (2018a) to name a few. Recently, in the study of the structure of WCN for Microblogs Garg and Kumar (2018b), the network properties are observed by increasing the number of Microblog in the dataset from 100 to 100k. It is observed that the number of edges increases at higher rate than that of the number of nodes. This shows that when words are used with each other, edges make different connections among them. Although the network is stable but some deviations are observed with increase in size of the Corpus. The clustering coefficient depends upon how well the Microblog WCN is connected. As the Tweet Corpus size gets increased, the number of edges increases at higher rate than the number of nodes which shows highly connected network and thus, the clustering coefficient increases.

5.2 Chinese and English documents

It is observed that as the length of the documents gets decreased, the edge-to-node ratio also gets decreased which suggests that there is a certain range of words which are in the active vocabulary of the author. Thereafter, the connectivity among nodes is increased at faster rate than that of increase in number of nodes. This shall prevail till the context of information remains same. The diameter of GoW gets increased as the size of dataset is reduced. This is evident from the fact that since the number of edges are reduced for smaller dataset, the number of links are less and thus, the diameter is increased.

Similarly, many other inferences lead to the fact that there is a growing need to build the language network based model for deeper analysis of GoW. Although, the static GoW can give useful insights about the textual data, there is huge scope for the real-time data analysis by incorporating such network science based models. The word distribution among different types of text, for instance, poems, articles, Microblogs and others may vary from each other. Thus, new network science models can be generated by understanding the semantics of the GoW evolved from the corresponding genres. Also, in future, this analysis may lay down a foundation for the language-independent statistical analysis for information retrieval by generating a generalized objective function by studying semantics of GoW evolved from text of different languages.

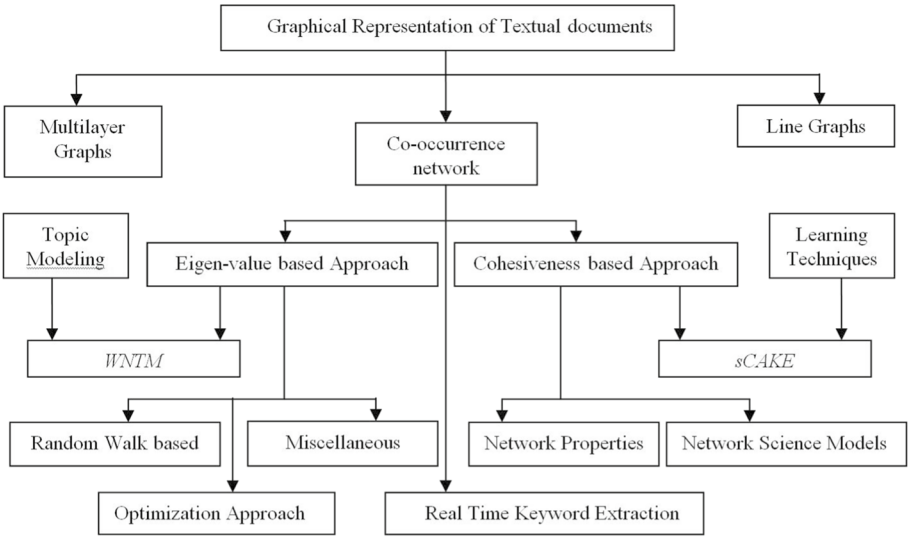

Fig. 5 Co-occurrence Network based approaches for GKET using GoW

6 Diversity in approaches over graph of words for AKE

The GoW is processed to obtain keywords using different approaches which are discussed in this Section. Though there are several ways for GoW representation, it is observed that co-occurrence based networks are the most widely used networks for AKE from textual data in literature. After extensive literature survey, the co-occurrence network based AKE are segregated into three major domains, namely, the Eigen-value based Approaches, the Cohesiveness-based Approaches and the Real-time keyword extraction as observed from Fig. 5. The properties which are studied for such documents helps in making different inferences about data and thereby introducing diverse approaches for AKE using the same GoW representation.

6.1 Eigenvalue based analysis

The Eigen-value determines the importance of a node by using the important score of its neighbouring nodes. Initially, every node is given a node score (which is usually taken as degree of a node in the graph) and an edge is given an edge score (in-case of GoW, it is usually considered as an edge-weight). The scores are updated using PageRank algorithm, iteratively, as the agent moves randomly in graph. This random walk process helps in re-distribution of the importance of node scores. Many academic researchers have developed new node-scores and edge-scores which improved the participation of nodes and edges in GoW.

On the basis of different relations among various nodes (words) in a GoW, the graph based PageRank algorithm is used in the form of TextRank algorithm Mihalcea and Tarau (2004). The dataset which used in these experiments is a collection of 500 abstracts from the Inspec database and manually assigned keywords are used as an annotated dataset.

TextRank algorithm is implemented for varying values of 'co-occurrence window' N from 2 to 10 over both 'directed' and 'undirected' network. It is observed that TextRank gives the best performance over undirected network with the 'co-occurrence window' of 2. TextRank identifies the connections between various entities in a text, and implements the concept of recommendation. A node recommends other related nodes, and the strength of the recommendation is recursively computed based on the importance of the nodes making such recommendation. Thus, it was found to be suitable for keyword extraction from GoW. An extended version of TextRank is recently proposed as the Tag-TextRank Peng et al. (2012), by introducing a tag which calculates the importance of a node in a weighted-term graph. This weighted-term graph calculates the edge weight for a word-co-occurrence pair by a certain tag. This tag is responsible for calculating final score for importance of a term.

In TopicRank Bougouin et al. (2013) keyphrase extraction approach, candidate keyphrases are clustered into topics and used as vertices in a complete graph. A graph-based ranking model is applied to assign a significance score to each topic. Keyphrases are then generated by selecting a candidate from each of the top-ranked topics. This approach is tested and validated over four different datasets of English and French language. In PositionRank Florescu and Caragea (2017), the PageRank scores are evaluated using the position-biased normalized form of Matrix M which is produced using undirected and weighted graph. The position of a word in different documents is considered to identify the importance of a node. Experimental results have shown that the PositionRank outperforms TextRank, SingleRank and ExpandRank.

It is observed that calculation of eigenvalues for GoW can help in dimensionality reduction in spectral clustering Garg and Kumar (2018b) and in identifying tight concentrations in random walk graph Bougouin et al. (2013) .

6.2 Cohesiveness based analysis

The cohesiveness based analysis suggests that the GoW is useful in determining the relation of words among each other. More cohesive the network is, more is the probability of important words to connect with each other. In existing literature, many community detection algorithm, graph decomposition based models (like k-core decomposition Rousseau and Vazirgiannis (2015), k-bridge decomposition, and k-truss decomposition), network cliques Rousseau and Vazirgiannis (2015), and assortativity measures Garg and Kumar (2018b) are used to identify the important keywords in GoW.

Cohesiveness is the property of network where nodes are connected to each other and represents a denser network. Cohesiveness in GoW is shown when words are closer and co-occur along with each other in the predefined sliding window. It is assumed that words which are more connected represent a topic and thus, a set of keywords which describe the essence of the text Meladianos et al. (2017). The cohesive sub-graphs are extracted on the basis of different network science properties such as assortativity, k-truss decomposition, k-core decomposition, network cliques, network motifs and community detection which are described in this Section.

6.2.1 K-Core decomposition

The library to represent graph of words is given as GoWvis Tixier et al. (2016). A fundamental difference when dealing with text with random walk models is that the paramount importance of cohesiveness: keywords not only need to have important connections but

also to form dense substructures with these connections. Therefore, it is observed that keywords are more likely to be found among the influential spreaders of a GoW – as extracted by degeneracy-based methods Tixier et al. (2016). The CoreRank algorithm was proposed using k-core decomposition method to extract keywords from some real-time conversations Meladianos et al. (2017).

6.2.2 K-Truss decomposition

K-truss is a triangle-based extension of k-core decomposition. More precisely, the K-truss sub-graph of G is its largest sub-graph where every edge belongs to at least $K2$ triangles. As compared to k-core, K-truss does not only prune-out nodes based on the number of their direct links, but also based on the number of their shared connections which captures cohesiveness more accurately. As a result, the K-trusses are smaller and denser sub-graphs than the k-cores, and the maximal K-truss of GoW approximates its densest sub-graph Tixier et al. (2016), Duari and Bhatnagar (2019) in a much better way.

6.2.3 K-Bridge decomposition

K-bridge decomposition is the modified form of k-core decomposition. The k-bridge decomposition uses reducing of edge-weight below k in GoW in-place of nodes reduction. The edge-weight is usually fixed as co-occurrence frequency between two words by default. However, there are varying mechanism to assign the edge-weight to edges in a GoW. This edge based decomposition has been used for identifying influential segments Garg and Kumar (2018b) and to introduce BArank Garg and Kumar (2018a), a keyphrase extraction techniques from Microblog word co-occurrence network.

6.2.4 Network Motif

There are some specific patterns of sub-graph or group of nodes which occur repeatedly in GoW. Such patterns are called *Network Motifs*. Recently, some researchers created four dataset and Network Motifs property was used to propose KSM (Keyword extraction for Single document using Motifs) Chen et al. (2019). It is observed that KSM gives better results than TextRank Mihalcea and Tarau (2004), WS-rank Yang et al. (2016) and SalienceRank Teneva and Cheng (2017).

6.3 Real-time keyword extraction

Real-time conversation takes place in social media data. It becomes difficult to manually identify keywords and hence, trending topics which are discussed on internet platforms due to huge amount of data. Recently, CoreRank algorithm was proposed using k-core decomposition method to extract keywords from real-time conversations Meladianos et al. (2017). CoreRank has outperformed the baseline algorithm of PageRank. Although, EmbedRank is also suitable for the real-time processing of large amounts of Web data Bennani-Smires et al. (2018).

Apart from these three classifications, another *path-based* approach is observed in literature. However, it was introduced in 2010 to extract keyphrases and since then, no considerable progress has been made in this direction.

6.4 Path based analysis

When adjacent-pairs of words are used to generate a graph, a path based network is obtained. The path based network shows the relation between the incoming degree and the out-going degree within a directed graph. The path based GoW is considered to identify keyphrases from textual documents which is out of the scope of this article. An important research work was carried out to propose Opinosis Ganesan et al. (2010) to identify extractive keyphrases from text using path based graph of words. Also, it has been used for identifying influential segments from Microblog WCN Garg and Kumar (2018b) .

7 Ranking of nodes in graph of words

The idea behind ranking of nodes in GoW is to identify the extent to which the words are relevant or important in the given document. The ranking techniques were used to rely on heuristics before the TextRank Mihalcea and Tarau (2004) algorithm was proposed in existing literature. The PageRank based approaches like TextRank, PositionRank Florescu and Caragea (2017), and TopicRank Bougouin et al. (2013) are proposed to rank the most relevant terms in top of the list. However, Multi Attribute Decision Making (MADM) methods used Analytical Hierarchical Process to rank keywords Ramay et al. (2018) and keyphrases Garg and Kumar (2018b) in the document using different graphical properties. The ranking of nodes in GoW is evaluated using the performance evaluation measures for ranking of keywords, namely, Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR). Different ranking approaches which are used in literature over GoW are enumerated in this section.

7.1 Statistical measures based ranking

The microblogs of social media data are ill-formed and user-generated. In the beginning of last decade, a Topical PageRank (TPR) and a context-sensitive Topical PageRank method (cTPR) were proposed as the first step of keyword ranking Zhao et al. (2011). TPR uses the latent topic distribution inferred by Latent Dirichlet Allocation (LDA) to perform the ranking of noun phrases extracted from documents. The ranking procedure consists of running PageRank K-times, where K is the number of topics used in the LDA model. SalienceRank Teneva and Cheng (2017), a modification of TPR only needs to run PageRank once and extracts comparable or better keyphrases on benchmark datasets.

7.2 Graphical metrics based ranking

Microblog may contain uncertain information which is difficult to process. Many graph based features are used for keyword extraction and ranking. The degree-centrality measure is used to propose Twitter Keyword Graph (TKG) for identifying and ranking important keywords Abilhoa and De Castro (2014) which is tested and validated over different types of WCN. Later, this initiative was outperformed by another keyword extraction technique, namely, Selectivity Based Keyword Extraction (SBKE) Beliga et al. (2016). In SBKE, selectivity based centrality measure is proposed. Further, various other graphical features were used to propose Keyword extraction using collective node weight (KECNW) Biswas et al. (2018). The graphical keyword extraction and ranking has been studied for various

languages and documents Liang et al. (2009), Gao et al. (2014), and short text Garg and Kumar (2018a).

7.3 Term embedding based ranking

In contrast, Term Ranker Khan et al. (2016) mitigates this issue by learning vector representations of term candidates (i.e. term embedding) and utilising it to capture similarities along with relation strength between terms for adding edges between nodes, and hence, a well connected graph was build. This graph was further used for ranking terms and obtain top ranked terms as keywords.

7.4 MADM Approaches

In recent times, as academic researchers started working on the structure of WCN of different textual documents Liang et al. (2009), Gao et al. (2014), the graphical features are used for generating an objective function to provide an optimized solution by ranking nodes. A keyword Ramay et al. (2018) and keyphrase ranking Garg and Kumar (2018a) techniques are proposed by using the Multiple Attributes Decision Making (MADM) approach, namely, Analytical Hierarchical Process (AHP).

7.5 HIN based ranking

The HIN is used for ranking and clustering entities. The advantage of using HIN is that some constraint-based random walk techniques can be used for finding similarities among words Meilian and Danna (2020). The HIN has not been used in many application domains of NLP yet but it has huge potential as there are different dimensions of textual information which can be used as features to find the similarity between words and sentences.

8 Experiments and evaluation

This Section of experimental analysis is introduced to show the snippets of the possible directions for various research directions for AKE using GKET. It is observed that in existing literature, there is no direct comparison between AKE from the ill-formed data and the well-formed data. Since, the semantics of GoW evolved from different type of data is different, the domain/ genre/ language of the textual document plays pivotal role in identifying suitable AKE technique. In these case studies, efforts have been made to compare the results over different dataset for ill-formed and well-formed dataset.

8.1 Case Study 1: Comparison of eigen-value based GKET for various textual dataset

The eigen-value based GKET are implemented over 21 dataset, namely, TextRank, SingleRank, TopicRank, TopicalPageRank (TPR), and MultipartitePageRank (MpR) as shown

Table 5 AKE using Eigen-value based GKET for various textual dataset

Dataset	TextRank	SingleRank	TopicRank	TPR	MpR
110-PT-BN-KP	0.207	0.275	0.256	0.186	0.179
500N-KPCrowd-v1.1	0.111	0.157	0.172	0.158	0.172
Inspec	0.098	0.378	0.289	0.361	0.307
Krapivin2009	0.121	0.097	0.138	0.104	0.146
Nguyen2007	0.167	0.158	0.173	0.148	0.190
PubMed	0.071	0.039	0.085	0.052	0.089
Schutz2008	0.118	0.086	0.258	0.123	0.238
WWW	0.059	0.097	0.067	0.101	0.077
KDD	0.050	0.085	0.055	0.089	0.065
SemEval2010	0.149	0.129	0.195	0.125	0.199
SemEval2017	0.125	0.449	0.332	0.443	0.335
CACIC	0.067	0.087	0.149	0.009	0.137
CiteULike180	0.112	0.066	0.156	0.072	0.178
FAO30	0.077	0.066	0.154	0.107	0.149
FAO780	0.083	0.085	0.137	0.108	0.145
Pak2018	0.041	0.022	0.022	0.043	0.019
Theses100	0.058	0.060	0.114	0.083	0.117
WICC	0.082	0.133	0.146	0.027	0.136
Wiki20	0.074	0.038	0.106	0.059	0.091
WikiNews	0.175	0.248	0.218	0.193	0.170
Microblogs	0.084	0.185	0.524	0.193	0.578

in Table 5. The experimental results obtained from the eigen-value based approaches for AKE from well-formed data is taken from existing research over AKE Campos et al. (2020). The results are obtained as F-measure for k@10 which means for top 10 keywords extracted. To compare the AKE for well-formed data with that of ill-formed data, the Microblog are collected over 10 different topics to generate a toy dataset.

The UTweet10 refers to the Un-balanced set of data for Tweets over 10 topics. UTweet10 dataset is extracted between 02 December 2019 and 04 December 2019 which contains 10 topics having 1000 Tweets for each topic. Out of these 10k Tweets, unique Tweets are extracted for each instance which does not contain any repetitive information. This gives unbalanced dataset which contains different number of Tweets for each topic with an average of 246 Tweets. The dataset is extracted using Tweepy API and Python 2.7 version from Twitter. The collection of Tweet IDs and ground truth key-phrases is made available online[4]. The agreement study was carried for collecting the ground truth for each topic by using *Fleiss Kappa statistics*Randolph (2005). The overall agreement was calculated over 2 categories (agree, disagree) separately which gives 82.8% and 78.1% for the agreement study of UTweet10 dataset. Thus, the comparative analysis is made for AKE from both well-formed and ill-formed data using eigen-value based GKET.

As observed from the given Table 5, the dataset which are used in this study are available in public domain[5]. To compare and constrast the implementation of traditional

[4] https://github.com/drmuskangarg/UTweet10.

[5] https://github.com/LIAAD/KeywordExtractor-Datasets.

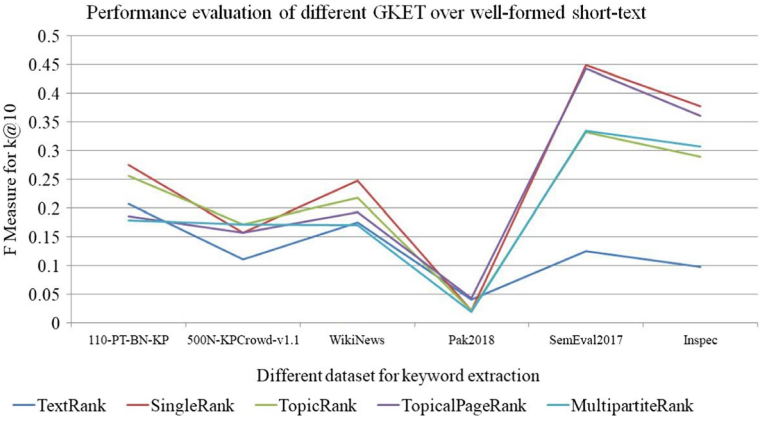

Fig. 6 Performance evaluation of different GKET over well-formed short-text

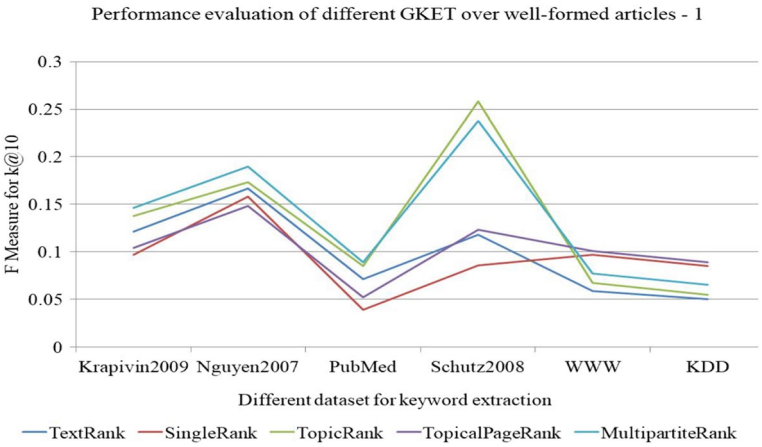

Fig. 7 Performance evaluation of different GKET over well-formed articles - 1

eigen-value based measures for both formal and informal text over the existing dataset, the dataset are classified into various groups. The first group contains all those dataset which contains well-formed textual information of shorter text (abstract, news reports, Paragraph). The first set of dataset is represented in Fig. 6. The second and third group contains all those dataset which contains well-formed long-text (Research Papers) and are represented in Figs. 7 and 8, respectively. The fourth group is assembled with the size of the dataset varying from microblogs (UTweet10) and books (thesis100). The dataset which are taken in between the two dataset in fourth group, to compare and contrast the implementation over varying size of corpus, are those which shows high variation for different techniques and outperforms the other dataset of their group. The performance evaluation for the fourth dataset is shown in Fig. 9.

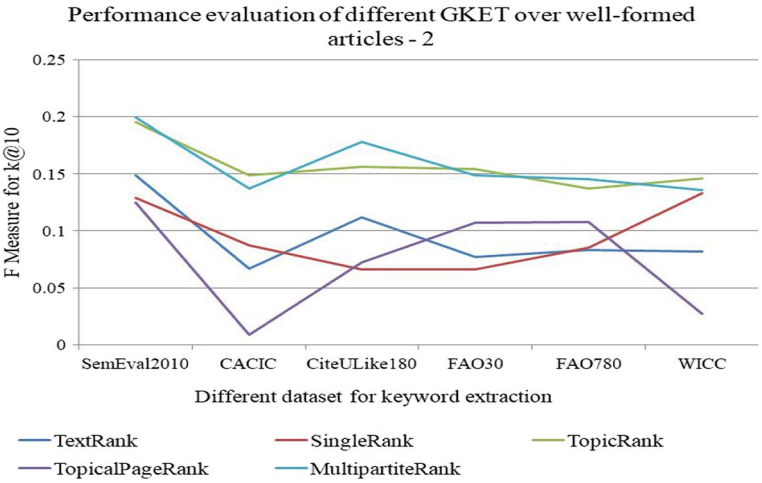

Fig. 8 Performance evaluation of different GKET over well-formed articles - 2

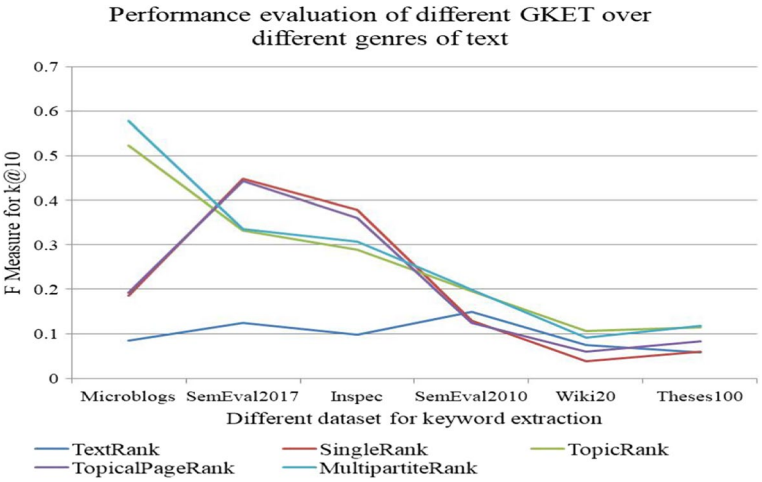

Fig. 9 Performance evaluation of different GKET over different genres of text

Observations for Group 1 As observed from Fig. 6, as per the characteristics of the dataset, the dataset 110-PT-BN-KP and 500N-KPCrowd-v1.1 are the news dataset and of miscellaneous domain which contains 110 and 500 documents, respectively. Further, it is observed that SemEval2017 and Inspec dataset have information about 493 miscellaneous paragraphs and 2000 abstracts of computer science articles, respectively. It is observed that TextRank shows more deviation for SemEval2017 and Inspec dataset than the other dataset. However, all the existing techniques shows the least performance over Pak2018 dataset which can be due to 50 miscellaneous abstract (insufficient data) or due to different language (Polish). The best results are shown for SemEval2017 which can be due to well connected words in paragraphs of 593 miscellanous articles. This shows that the pattern of word distribution in abstract varies than that of normal paragraph

Table 6 AKE using Eigen-value based GKET for various existing dataset

Language	Average Degree	ASPL	CC
Microblog	2.166	3.144	0.076
Belarusian	4.819	3.797	0.100
Bulgarian	5.690	3.354	0.186
Chinese	8.684	2.944	0.283
Croatian	5.353	3.479	0.151
Czech	4.945	3.627	0.199
English	9.043	2.964	0.299
Macedonian	6.206	3.225	0.220
Polish	4.983	3.628	0.118
Russian	4.504	3.891	0.091
Serbian	5.348	3.485	0.147
Slovak	5.166	3.592	0.128
Slovenian	5.367	3.406	0.164
Ukrainian	4.865	3.814	0.096
Upper-Sorbian	5.347	3.550	0.131

writing because the abstract cover wider aspects in minimum possible words in which words shows minimal connections among each other.

Observations for Group 2 As observed from Fig. 7, the dataset for which the implementation of existing techniques varies to a large extent is Schutz2008. Although, all the dataset have similar nature of research papers related to computer science domain in English language, the dataset Schutz2008 documents contains higher number of gold keys per documents which come out to be about 44% on an average. On the contrary, the number of gold keys per document in other dataset is less than 12%. This observation supports this argument that the GKET which rely more on the term-frequency gives different results than those which are less dependent on term-frequency parameter. The GKET which belong to the former nature are TopicRank and MultipartiteRank and those which rely over the latter part are TextRank, SingleRank, and TopicalPageRank. Also, the procedural details of TopicRank and MultipartiteRank are similar to each other.

Observations for Group 3 This group of dataset is similar to that of the previous group which contains research papers in the corpus for keyword extraction as shown in Fig. 8. It is observed that the variation in the results of existing techniques for each dataset is more than Group 2. The results of F measure for k@10 as shown for both the dataset of Food and Agriculture Organization (FAO) (FAO30 and FAO780) are similar to each other and there is no significant difference in the results of existing GKET. The results for both FAO dataset are different from that of other dataset because the domain which is used in this dataset is Agricultural domain and thus, the pattern of word distribution in articles related to agriculture may vary from those of computer science/ miscellaneous domains.

In addition to this, TopicalPageRank shows least results for CACIC and WICC dataset which contains documents in Spanish text. Thus, the variation in language have severe effect over results. The overall performance of MultipartiteRank and TopicRank is significantly better than the other eigen-value based GKET.

Observations for Group 4 The Group 4 is assembly of many different dataset whose size varies from Microblog to book. It is observed from Fig. 9 that TopicRank and MultipartiteRank performs better than other existing techniques. Since the semantics of Microblog

Table 7 Different features for WCN of different genres

Features	Microblogs	Essays	Novels	Science articles	News reports
Length	183466	1142	5224	961	748
#Nodes	34925	440	1314	426	343
#Edges	116477	826	3199	734	581
Average degree	2.166	3.26	4.69	3.32	3.35
$ASPL/ASPL_r$	1.258/3.366	3.61/4.70	3.29/4.60	3.92/5.04	3.91/4.83
$CC\%/CC_r\%$	0.493/0.00039	10.61/0.97	17.62/0.45	8.15/0.87	7.16/1.02

WCN evolved from social media data is different than the normal text, all the traditional GKET may not give promising results. Also, much deviations are observed for SemEval2017 and Inspec dataset. The performance of GKET descreases as the size of document gets increased and the word distribution gets more complex. To study this complexity, it is important to study the structure and dynamics of GoW for different languages and different genres which can give promising results in future.

8.2 Case Study 2: Comparative discussion over WCN for different languages

The variations in the networks of words for different languages are observed on the basis of average degree, average shortest path length (ASPL), and clustering coefficient (CC). 14 word co-occurrence networks were constructed and studied based on parallel texts of 12 Slavic languages and 2 non-Slavic languages, respectively Liu and Cong (2013). To compare and contrast the behaviour of ill-formed text with that of the well-formed text, the structure of Microblog WCN Garg and Kumar (2018a) is compared with the structure of WCN evolved from 14 different languages as shown in Table 6.

It is observed that the ASPL remains in the range of 3 to 4 for GoW and is not much deviant for different languages. However, the clustering coefficient of Microblog WCN and English WCN is approximately in the ratio of 1:4. This variation in observation for average degree, ASPL and CC in WCN different languages are on the basis of different features which are enlisted as

- Language: The type of connectivity in every language is different because the words and grammar which are used in a language are based on the script which it follows. Every script may have different pattern of connectivity and different essence.
- Vocabulary: There are variations in the number of basic units (Alphabets) and the number of dialects in which the words in active vocabulary may vary for each language. For instance, it is observed that only 33% of the total English dictionary is used by people in the world. This limitation on the number of words which are used in active vocabulary is important so that people can retain and connect with each other for better understanding.
- Case Variations: There are some special features in different languages which can act as open challenge to tackle automatically. For instance, the words in Hungarian language have 18 to 35 cases which means that the meaning of the same word can be identified in different ways depending upon the context. This special feature cannot be handled by automatic NLP.

– Word Order: Most languages depend on word order to convey meaning, at least to a certain extent. Warlpiri is a language of indigenous Australians, spoken by about 3,000 people. Warlpiri is a rare example of a free word order language, where the order words are placed which depends on what the speaker believes deserves the greater emphasis, or that makes logical sense depending on the flow of the conversation.

Thus, the key idea behind these observations is that the language independent research work should be carried for processing WCN. However, the varying features of languages which have been enlisted in this Section is not the exhaustive list and can be explored further.

8.3 Case Study 3: Comparison of the structure of WCN for different Genres

The network properties for the structure of the WCN is observed for different genres including Microblogs, essays, novels, science articles, and news reports in Table 7. All these genres are observed for the English language for full network even if it is disconnected. These graphical properties are used to examine the behaviour of WCN for different genres. This approach is followed to study the first chief characteristic feature of WCN for different languages and different genres. The GoW representation used for this comparison is the WCN with un-directed and unweighted edges in word-adjacency network.

The microblog which are used for this comparison are taken from the *First Story Detection* dataset Petrović et al. (2010) to examine the WCN evolved from random incoming tweets. The number of Microblog which are taken for analysis of the structure of WCN are 25000. The structure of the Microblog WCN is carried Garg and Kumar (2018a) to propose a keyword extraction technique, BArank. Other genres of English language, namely, essays, novels, popular science articles, and news reports are taken from another article which has carries comprehensive studies over English and Chinese language Liang et al. (2009). Only English language based genres are used in this article for direct comparison of different genres. The major observations are - As the length of the WCN is decreased the ratio of CC of the WCN to that of CC of random network, as measured using Erdos-Renyi model is decreasing. This observations justifies the fact that the number of edges keeps on decreasing as we decrease the dataset size. This factor is scalable for multiple genres and to the extent of change in the context of textual documents.

– Different patterns are observed for the connectivity of words among each other. Thus, the average degree of nodes in WCN differs depending upon the genre.
– It is observed that ASPL remains similar to that of ASPL for random network. Both ASPL and CC features varies with random network in the same proportion irrespective of genre. Thus, it is observed that WCN follows the small-world property.
– It can be observed that irrespective of the genre, the edge to node ratio decreases with decrease in size of the dataset due to reduced number of links in smaller dataset.

Other inferences can be made by using assortativity, hierarchical organization and spectral distribution over GoW Garg and Kumar (2018b).

9 Discussion

As per analysis, it is observed that the academic researchers are working in different dimensions over GoW for AKE from textual documents. The behaviour GoW due to words distribution is examined for multiple short-text documents (online microblog) which may vary during communication on internet. Initially, the graph will act like random graph because random words and links between them are being used. The scale-free property is examined for distribution of words when the number of nodes (words) in corpus are between 10^3 and 10^5, and beyond 10^5 as observed from Microblog WCN Garg and Kumar (2018a). It is observed that the rate of increase in nodes is reduced as compared to the rate of increase in edges as the WCN grows bigger and denser with increase in size of data. The small-world property of WCN indicates that the vocabulary used over internet by users is limited. The small-world property for GoW is much required so that it gets easy for users to communicate and understand each other.

This Section is further divided into three parts. The first part examines the comparative results to make inferences about different languages and for different genres. The structure of the WCN is examined to study the pattern of connection among words. The second part enlists the application domains which enables a reader to understand the major areas in which the keyword extraction techniques can be used. The second part gives the information about possible challenges and new research directions.

9.1 Applications

The GKET are applied to various different applications over ill-formed and well-formed text. Some of the most commonly analysed GKET are based on the AKE from Microblogs (Multi-document) and from articles (single document). There are plenty of applications which uses the keyword extraction for information retrieval. There are different application areas which are directly and indirectly associated with keyword extraction from textual documents which are discussed in this Section.

9.1.1 Email

Emails constitute an important genre of online communication. Many of the users often face the daunting task of sifting through increasingly large number of emails on a daily basis. Keywords extracted from emails can help in combating such information overload by allowing a systematic exploration of the topics contained in emails Laclavík and Maynard (2009). The researchers have introduced a new dataset for email keyword extraction and have focused on unsupervised and supervised approaches for keyword extraction from email data Lahiri et al. (2014).

9.1.2 Literature articles

There is vast literature which contains books, scientific articles, newspaper articles to name a few. Any article containing thousands of words needs to be summarized automatically for processing the large number of articles. The initial stage of text summarization is to obtain substantial keywords from textual documents. Several random walk based keyword extraction techniques are proposed for the news articles, for instance, CollabRank Wan and Xiao (2008) and for scientific/ Literature articles Wan and Xiao (2008), Rose et al. (2010).

9.1.3 Short-text/ Microblogs

Massive data is generated over social networking websites by users as first-hand information. Unlike the traditional documents, such text is extremely short and informal. Analysis of such text is used for many applications such as advertising, search, and content filtering Zhao et al. (2017). The keyword extraction technique was proposed for Facebook data and it is observed that there are no keywords for 25% of social snippets Chowdhury et al. (2019).

The shortest documents for AKE are the Microblogs such as Tweets (limited to 280 characters as per the recently updated policy). The AKE from Twitter data earlier used the traditional approaches which were proposed for AKE from SMS (Service Message Service). Recently, GKET are adapted for AKE from Microblogs in English Abilhoa and De Castro (2014) and Croatian Beliga et al. (2016) language due to the unstructured and the user-generated ill-formed text. An unsupervised graph-based keyword extraction method called Keywords from Collective Weights (KCW) which is based on the node-edge rank centrality with node weight depending on the graphical features of words Bordoloi and Biswas (2018). The PageRank algorithm was used to propose NErank Bellaachia and Al-Dhelaan (2012) for identifying keywords from Twitter data which uses node score and edge score.

9.2 Outlook: Future scope and challenges

There is a huge research gap to identify the importance of GoW, find latent patterns which are directly applicable to NLP problem domains, and introduce a generic model for GoW. The GoW is a complex network which can be studied for static and dynamic behaviour to understand the connection among words in streaming data. This understanding can help in identifying the dense sub-graphs using approximation algorithms and cohesiveness based GKET can be used for AKE from real-time data. The similar open research challenges and the future scope in this domain is discussed in this Section.

9.2.1 Text-Normalization

Text Normalization is the process of converting the ill-formed text like short-text from Microblogs into the well-formed text. This is a booming area of research which is much required for rumour detection and false information detection. The academic researchers have generated the dictionary to convert the informal text into well-formed information, but everyday new words and hashtag are created on the basis of new events and discussions over social media. Thus, text normalization is one of the most challenging area of research for social media data.

9.2.2 Dataset for keyword extraction

New dataset and/ or annotated dataset are required for keyword extraction due to evolving and changing structure of documents. The data is being developed in well-structured form which changes the semantic nature of GoW which is used to study distribution of words. Although, much data is available for keyword extraction from well-formed

documents, there are very few dataset available for keyword extraction from ill-formed data. Even if the dataset is available, there is no specific annotated data/ ground truth information available for keyword extraction. Only few dataset are validated using different statistical tests and agreement studies.

9.2.3 Variation in values of parameters

One of the most important factors considered for PageRank, TextRank and other random walk based algorithms is damping factor. In page linking, the PageRank theory holds that an imaginary surfer who is randomly clicking on links will eventually stop clicking. The probability, at any step, that the person will continue is a damping factor d. In WCN, damping factor is considered as the probability of moving ahead or to move to adjacent word which is usually set as 0.85. The damping factor is used for iteratively distributed algorithm. The structure and behaviour of WCN varies from web page linking in the context of different structures for page linking and WCN as core-periphery structure using follower-followee relationship and path based network, respectively. The follower-followee relationship is suitable for network of users in the social media platforms but not for WCN evolved from content or textual data (like Tweets). Although it is important to study the distribution of words in textual documents, its behaviour varies from that of the HITS algorithm. Thus, it is important to study the effect of damping factor for PageRank over WCN Liu et al. (2018), Chen et al. (2020a).

9.2.4 Real-time keyword extraction approaches

Preferential attachment model is used for predicting the attachment of incoming information with existing network. This model is based on the probability of any incoming node which is randomly being attached to a network. However, in GoW of microblogs, the incoming short-text (Tweet) is considered and pre-processing is performed to identify tokenized words. These words are arranged in a specific arrangement and this arrangement is used to plant words as nodes (both new and existing) and edges as per arrangement of terms in adjacent pattern. A mathematical model can be build to handle these incoming nodes and edges and also, traditional or new mechanism can be proposed for removing old nodes and edges. This network model can be used to identify the nodes of sub-graph with increasing density which may represent a short-story or a new topic.

9.2.5 Network models for words distribution

The behaviour of words distribution in GoW evolved from multiple short text documents may vary during communication on internet. Initially, the graph will act like random graph because random words and links between them are being used, then scale-free property develops for number of nodes in between 10^3 and 10^5 as observed from Microblog WCN Garg and Kumar (2018a) and beyond number of nodes 10^5, the small world property develops in GoW. This saturation is observed because rate of increase in nodes is reduced as compared to rate of increase in edges. The small-world property indicates that there are limited number of words which are in active vocabulary of a user. The vocabulary used over internet by users is limited and this saturation is required so that it gets easy for users to communicate and understand each other.

9.2.6 Finding n-graph from path based networks

The GoW is created using the arrangement of words in a Tweet. This network is known as path based network. These path of short-text documents generate the word adjacency networks which may give sequence of nearly similar values for edge weight. This similarity is expected in the GoW due to repetition of some words or phrases by multiple users, for instance, "Lockdown due to COVID-19" was a common phrase in Tweets during COVID-19 pandemic situation. A new network model can be generated to identify the influential segments or the keyphrases from GoW so that n-grams can be extracted where value of n is dynamic. The keywords can be found by determining the statistically significant difference between the incoming and the outgoing degrees along with the proximity of values towards each other. These words can be grouped together on the basis of the closeness of values of edge weights.

9.2.7 Comprehensive study over keyword extraction from short-text

The major challenge in this area is that none of existing techniques has released the dataset along with annotations in public domain for keyword extraction. The existing GKET for Twitter dataset should be compared and validated over the common dataset. The structure and dynamics of word-distribution in Tweets changes with change in policies of Twitter social media platform. Thus, it is important to compare and contrast the results over same dataset because the data is ever-changing and dynamic which is missing in literature.

9.2.8 Approximation algorithms

As *random walk based* and *decomposition based* keyword extraction approaches use much iteration, the approximation algorithms should be used or proposed to reduce the time complexity. These approximation algorithms may prove to be beneficial for various industrial applications (real-time keyword extraction) as they needs to be computationally less expensive. Also, this approach of approximation algorithms can be useful for real-time applications.

10 Conclusion

An extensive literature survey is carried out for the Automatic Keyword Extraction (AKE) techniques. It is observed that the GoW can be generated in many different ways which are discussed in this article. These GoW representations can be analysed in many ways using different types of graphical approaches. There are many traditional approaches which classified on the basis of direction which it follows, namely, random-walk based, using graphical metrics/ network science properties, real-time AKE over dynamic GoW evolved from streaming data. These approaches are not well-versed with their application over different representations of GoW yet.

An important aspect which is studied in this article is to examine the stability of the structure of GoW for static data from existing literature. However, the stability of dynamic GoW due to its scalablity and robustness is yet to be explored. Also, the statistical properties for GoW are considered in recent literature and these properties are considered as

potential features for examining the importance of nodes. On the basis of these feature extraction and feature selection process, the words are ranked in GoW using statistical measure, graphical measure, term embedding, MADM approaches and HIN based ranking. Although, there are many existing node ranking techniques for a complex network, however, the semantics of GoW is different for different representation of GoW. Thus, there exists the need to examine the pattern of distribution of words in a document. This research work has paved the path to connect all these dimensions for AKE from textual documents. Also, this articles classifies AKE techniques for graphical keyword extraction as evident from existing literature.

In addition, this article reflect on comprehensive elements which are handles in different domains for various dimensions. One of the GoW representation is a line graph and similar representation is referred as newly induced bigram word graph for identifying disasters from GoW evolved from Microblogs. Another observation is that the calculation of eigenvalues for GoW can help in dimensionality reduction in spectral clustering Garg and Kumar (2018b) and in identifying tight concentrations in random walk graph Bougouin et al. (2013). Analysis of the study of semantics of various language networks may result into generalized objective functions which can help to propose a language-independent keyword extraction technique.

Another important contribution of this article is by laying the foundation to compare and contrast the structure and dynamics of the WCN for both well-formed languages and ill-formed languages, and multiple genres. Moreover, interesting insights have been obtained by studying existing eigen-value based GKET over 21 different dataset. It is observed that MultipartiteRank and TopicRank gives better results for ill-formed text. For well-formed text, the SingleRank gives equally comparable performance in addition. However, for Spanish text, TopicalPageRank (TPR) does not give good results but for few dataset like SemEval2017, Inspec and KDD, TPR along with SingleRank has given outstanding performance. This article also contributes the information about background for AKE from textual documents, different classifications of AKE techniques, the application domains in which AKE can be used directly or indirectly to solve a problem, and open challenges along with future research directions in the field of connecting different research dimensions for AKE for GKET.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abilhoa WD, De Castro LN (2014) A keyword extraction method from twitter messages represented as graphs. Appl Mathe Comput 240:308–325. https://doi.org/10.1016/j.amc.2014.04.090

Aquino GO, Lanzarini LC (2015) Keyword identification in Spanish documents using neural networks. Journal of Computer Science & Technology, 15

Augenstein I, Das M, Riedel S, Vikraman L, McCallum A (2017) Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. https://doi.org/10.18653/v1/S17-2091

Awan MN, Beg MO (2020) TOP-Rank: A topicalpostionrank for extraction and classification of keyphrases in text. Comput Speech Lang 65:101116. https://doi.org/10.1016/j.csl.2020.101116

Beliga S, Meštrović A, Martinčić-Ipšić S (2016) Selectivity-based keyword extraction method. Int J Semantic Web Inf Syst (IJSWIS) 12(3):1–26. https://doi.org/10.4018/IJSWIS.2016070101

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
Bellaachia A, Al-Dhelaan M (2012) Ne-rank: A novel graph-based keyphrase extraction in twitter. In 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 1:372-379. https://doi.org/10.1109/WI-IAT.2012.82
Bennani-Smires K, Musat C, Hossmann A, Baeriswyl M, Jaggi M (2018) Simple unsupervised keyphrase extraction using sentence embeddings. In Proceedings of the 22nd Conference on Computational Natural Language Learning (Association for Computational Linguistics), pp. 221–229. https://doi.org/10.18653/v1/K18-1022
Bharti SK, Babu KS (2017) Automatic keyword extraction for text summarization: A survey
Biswas SK, Bordoloi M, Shreya J (2018) A graph based keyword extraction model using collective node weight. Expert Syst Appl 97:51–59. https://doi.org/10.1016/j.eswa.2017.12.025
Bordoloi M, Biswas SK (2018) Keyword extraction from micro-blogs using collective weight. Soc Netw Anal Min 8(1):58. https://doi.org/10.1007/s13278-018-0536-8
Boudin F (October) A comparison of centrality measures for graph-based keyphrase extraction. In Proceedings of the sixth international joint conference on natural language processing, pp. 834-838
Boudin F (2018) Unsupervised keyphrase extraction with multipartite graphs. https://doi.org/10.18653/v1/N18-2105
Bougouin A, Boudin F, Daille B (2013) Topicrank: Graph-based topic ranking for keyphrase extraction. In Proceedings of the 6th International Joint Conference on Natural Language Processing, pp. 543–551
Campos R, Mangaravite V, Pasquali A, Jorge A, Nunes C, Jatowt A (2020) YAKE! Keyword extraction from single documents using multiple local features. Inf Sci 509:257–289. https://doi.org/10.1016/j.ins.2019.09.013
Caragea C, Bulgarov F, Godea A, Gollapalli SD (2014). Citation-enhanced keyphrase extraction from research papers: A supervised approach. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1435-1446. https://doi.org/10.3115/v1/D14-1150
Chen Y, Wang J, Li P, Guo P (2019) Single document keyword extraction via quantifying higher-order structural features of word co-occurrence graph. Comput Speech Lang 57:98–107. https://doi.org/10.1016/j.csl.2019.01.007
Chen J, Hou H, Gao J (2020a) Inside importance factors of graph-based keyword extraction on Chinese short text. ACM Trans Asian Low-Res Lang Inf Process (TALLIP) 19(5):1–15. https://doi.org/10.1145/3388971
Duari S, Bhatnagar V (2020) Complex network based supervised keyword extractor. Expert Syst Appl 140:112876. https://doi.org/10.1016/j.eswa.2019.112876
Duari S, Bhatnagar V (2019) sCAKE: semantic connectivity aware keyword extraction. Inf Sci 477:100–117. https://doi.org/10.1016/j.ins.2018.10.034
Erkan G, Radev DR (2004) Lexrank: Graph-based lexical centrality as salience in text summarization. J Artif Intell Res 22:457–479. https://doi.org/10.1613/jair.1523
Evans TS, Lambiotte R (2010) Line graphs of weighted networks for overlapping communities. Euro Phys J B 77(2):265–272. https://doi.org/10.1140/epjb/e2010-00261-8
Florescu C, Caragea C (2017) Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1105-1115. https://doi.org/10.18653/v1/P17-1102
Gallina Y, Boudin F, Daille B. (2019) KPTimes: A Large-Scale Dataset for Keyphrase Generation on News Documents
Ganesan K, Zhai C, Han J (2010) Opinosis: A graph based approach to abstractive summarization of highly redundant opinions
Gao Y, Liang W, Shi Y, Huang Q (2014) Comparison of directed and weighted co-occurrence networks of six languages. Phys A 393:579–589. https://doi.org/10.1016/j.physa.2013.08.075
Garg M, Kumar M (2018) The structure of word co-occurrence network for microblogs. Phys A 512:698–720. https://doi.org/10.1016/j.physa.2018.08.002
Garg M, Kumar M (2018) Identifying influential segments from word co-occurrence networks using AHP. Cogn Syst Res 47:28–41. https://doi.org/10.1016/j.cogsys.2017.07.003
Gibert J, Valveny E, Bunke H (2011) Dimensionality reduction for graph of words embedding. In International Workshop on Graph-Based Representations in Pattern Recognition, pp. 22-31. https://doi.org/10.1007/978-3-642-20844-7_3
Gollapalli SD, Caragea C (2014) Extracting Keyphrases from Research Papers Using Citation Networks. AAAI 14:1629–1635. https://doi.org/10.1.1.686.7325

Gualano MR, Lo Moro G, Voglino G, Bert F, Siliquini R (2020) Effects of COVID-19 lockdown on mental health and sleep disturbances in Italy. Int J Environ Res Pub Health 17(13):4779. https://doi.org/10.3390/ijerph17134779

Guessoum SB, Lachal J, Radjack R, Carretier E, Minassian S, Benoit L, Moro MR (2020) Adolescent psychiatric disorders during the covid-19 pandemic and lockdown. Psychiatry research:113264. https://doi.org/10.1016/j.psychres.2020.113264

Harary F, Norman RZ (1960) Some properties of line digraphs. Rendiconti del Circolo Matematico di Palermo 9(2):161–168. https://doi.org/10.1007/BF02854581

Herrera JP, Pury PA (2008) Statistical keyword detection in literary corpora. Eur Phys J B 63(1):135–146. https://doi.org/10.1140/epjb/e2008-00206-x

Hotho A, Nürnberger A, Paaß G (2005) A brief survey of text mining. Ldv Forum 20(1):19–62. https://doi.org/10.1.1.447.4161

Huang C, Tian Y, Zhou Z, Ling CX, Huang T (2006) Keyphrase extraction using semantic networks structure analysis. In Sixth International Conference on Data Mining (ICDM'06):275-284. https://doi.org/10.1109/ICDM.2006.92

Hulth A (2003) Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. 216-223. https://doi.org/10.3115/1119355.1119383

Khan MT, Ma Y, Kim JJ (2016) Term Ranker: A Graph-Based Re-Ranking Approach. In FLAIRS Conference, pp. 310-315

Kim SN, Medelyan O, Kan MY, Baldwin T (2010). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 21-26

Kölbl M, Kyogoku Y, Philipp JN, Richter M, Rietdorf C, Yousef T (2020). Keyword Extraction in German: Information-theory vs. Deep Learning. In ICAART, pp. 459-464. https://doi.org/10.5220/0009374704590464

Krapivin M, Autaeu A, Marchese M (2009) Large dataset for keyphrases extraction. University of Trento

Laclavík M, Maynard D (2009) Motivating intelligent e-mail in business: An investigation into current trends for e-mail processing and communication research. In 2009 IEEE Confer-ence on Commerce and Enterprise Computing: 476-482. https://doi.org/10.1109/CEC.2009.47

Lahiri S, Choudhury SR, Caragea C (2014) Keyword and keyphrase extraction using centrality measures on collocation networks

Lahiri S, Mihalcea R, Lai PH (2017) Keyword extraction from emails. Nat Lang Eng 23(2):295–317. https://doi.org/10.1017/S1351324916000231

Li SQ, Du SM, Xing XZ (2017) A keyword extraction method for Chinese scientific abstracts. In Proceedings of the 2017 International Conference on Wireless Communications, Networking and Applications, pp. 133-137. https://doi.org/10.1145/3180496.3180620

Liang W, Shi Y, Chi KT, Liu J, Wang Y, Cui X (2009) Comparison of co-occurrence networks of the Chinese and English languages. Phys A 388(23):4901–4909. https://doi.org/10.1016/j.physa.2009.07.047

Lin CY (2004) Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pp. 74-81

Litvak M, Last M (2008) Graph-based keyword extraction for single-document summarization. In Coling 2008: Proceedings of the workshop Multisource Multilingual Information Extraction and Summarization, pp. 17-24

Liu Z, Huang W, Zheng Y, Sun M (2010) Automatic keyphrase extraction via topic decomposition. In Proceedings of the 2010 conference on empirical methods in natural language processing, pp 366-376

Liu H, Cong J (2013) Language clustering with word co-occurrence networks based on parallel texts. Chin Sci Bull 58(10):1139–1144. https://doi.org/10.1007/s11434-013-5711-8

Liu T, Qian Y, Chen X, Sun X (2018) Damping Effect on PageRank Distribution. In 2018 IEEE High Performance extreme Computing Conference (HPEC), pp. 1-11. https://doi.org/10.1109/HPEC.2018.8547555

Marujo L, Gershman A, Carbonell J, Frederking R, Neto JP (2013). Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization

Marujo L, Viveiros M, Neto JPDS (2013) Keyphrase cloud generation of broadcast news

Martinčić-Ipšić S, Margan D, Meštrović A (2016) Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. Phys A 457:117–128. https://doi.org/10.1016/j.physa.2016.03.082

Matsuo Y, Ohsawa Y, Ishizuka M (2001) Keyworld: Extracting keywords from document s small world. In International conference on discovery science: pp. 271-281. https://doi.org/10.1007/3-540-45650-3_24

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. https://doi.org/10.1142/S0218213004001466
Medelyan O, Frank E, Witten IH (2009) Human-competitive tagging using automatic keyphrase extraction. In Proceedings of the 2009 conference on empirical methods in natural language processing, pp. 1318-1327
Medelyan O, Witten IH (2008) Domain-independent automatic keyphrase indexing with small training sets. J Am Soc Inform Sci Technol 59(7):1026–1040. https://doi.org/10.1002/asi.20790
Meilian LU, Danna YE (2020) HIN_DRL: A random walk based dynamic network representation learning method for heterogeneous information networks. Expert Syst Appl 158:113427. https://doi.org/10.1016/j.eswa.2020.113427
Meladianos P, Tixier A, Nikolentzos I, Vazirgiannis M (2017) Real-time keyword extraction from conversations. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics 2, pp. 462-467. https://doi.org/10.18653/v1/E17-2074
Meng R, Zhao S, Han S, He D, Brusilovsky P, Chi Y (2017) Deep keyphrase generation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 582–92
Mihalcea R, Tarau P (2004) Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404-411
Nasar Z, Jaffry SW, Malik MK (2019) Textual keyword extraction and summarization: State-of-the-art. Inf Process Manage 56(6):102088. https://doi.org/10.1016/j.ipm.2019.102088
Naidu R, Bharti SK, Babu KS, Mohapatra RK (2018) Text summarization with automatic keyword extraction in telugu e-newspapers. In Smart Computing and Informatics, pp. 555-564. https://doi.org/10.1007/978-981-10-5544-7_54
Nguyen TD, Kan MY (2007) Keyphrase extraction in scientific publications. In International conference on Asian digital libraries, pp. 317-326. https://doi.org/10.1007/978-3-540-77094-7_41
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-ADL'98, pp. 12-18. https://doi.org/10.1109/ADL.1998.670375
Paranyushkin D (2019) InfraNodus: Generating insight using text network analysis. In The World Wide Web Conference, pp. 3584-3589. https://doi.org/10.1145/3308558.3314123
Park J, Kim J, Lee JH (2014) Keyword extraction for blogs based on content richness. J Inf Sci 40(1):38–49. https://doi.org/10.1177/0165551513508877
Peng L, Bin W, Zhiwei S, Yachao C, Hengxun L (2012) Tag-TextRank: a webpage keyword extraction method based on tags. J Comput Res Develop 49(11):2344
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, pp. 181-189
Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130. https://doi.org/10.1.1.848.7219
Pudota N, Dattolo A, Baruzzo A, Tasso C (2010) A new domain independent keyphrase extraction system. In Italian Research Conference on Digital Libraries, pp. 67-78. https://doi.org/10.1007/978-3-642-15850-6_8
Ramay WY, Cheng-Yin X, Illahi I (2018) Keyword extraction from social media via AHP. Human Syst Manage 37(4):463–468. https://doi.org/10.3233/HSM-180344
Randolph JJ (2005) Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa, Online submission
Chowdhury JR, Caragea C, Caragea D (2019) Keyphrase extraction from disaster-related tweets. In The world wide web conference, pp. 1555-1566. https://doi.org/10.1145/3308558.3313696
Rose S, Engel D, Cramer N, Cowley W (2010) Automatic keyword extraction from individual documents. Text Min: Appl Theory 1:1–20. https://doi.org/10.1002/9780470689646.ch1
Rousseau F, Vazirgiannis M (2015) Main core retention on graph-of-words for single-document keyword extraction. In European Conference on Information Retrieval, pp. 382-393. https://doi.org/10.1007/978-3-319-16354-3_42
Rudra K, Banerjee S, Ganguly N, Goyal P, Imran M, Mitra P (2016) Summarizing situational tweets in crisis scenario. In Proceedings of the 27th ACM Conference on Hypertext and Social Media, pp. 137-147. https://doi.org/10.1145/2914586.2914600
Salton G (1988) Automatic Text Processing. Addison Wesley
Sarkar D (2019) Processing and understanding text. In Text Analytics with Python, pp. 115-199. https://doi.org/10.1007/978-1-4842-4354-1_3
Schutz AT (2008) Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods." M. App. Sc Thesis. https://doi.org/10.1.1.394.5372

Siddiqi S, Sharan A (2015) Keyword and keyphrase extraction techniques: a literature review. Int J Comput Appl. https://doi.org/10.5120/19161-0607
Social WA (2020) Digital 2020 Global Digital Overview. Erişim Tarihi 18(03)
Song M, Song IY, Hu X (2003) KPSpotter: a flexible information gain-based keyphrase extraction system. In Proceedings of the 5th ACM international workshop on Web information and data management, pp. 50-53. https://doi.org/10.1145/956699.956710
Song M, Kim EHJ, Kim HJ (2015) Exploring author name disambiguation on PubMed-scale. J Inform 9(4):924–941. https://doi.org/10.1016/j.joi.2015.08.004
Sterckx L, Demeester T, Deleu J, Develder C (2018) Creation and evaluation of large keyphrase extraction collections with multiple opinions. Lang Res Eval 52(2):503–532. https://doi.org/10.1007/s10579-017-9395-6
Teneva N, Cheng W (2017) Salience rank: Efficient keyphrase extraction with topic modeling. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 530-535. https://doi.org/10.18653/v1/P17-2084
Tixier A, Skianis K, Vazirgiannis M (2016) Gowvis: a web application for graph-of-words-based text visualization and summarization. In Proceedings of ACL-2016 System Demonstrations, pp. 151-156. https://doi.org/10.18653/v1/P16-4026
Tixier A, Malliaros F, Vazirgiannis M (2016) A graph degeneracy-based approach to keyword extraction. In Proceedings of the 2016 conference on empirical methods in natural language processing, pp. 1860-1870. https://doi.org/10.18653/v1/D16-1191
Thomas JR, Bharti SK, Babu KS (2016) Automatic keyword extraction for text summarization in e-newspapers. In Proceedings of the international conference on informatics and analytics, pp. 1-8. https://doi.org/10.1145/2980258.2980442
Tsatsaronis G, Varlamis I, Nørvåg K (2010) SemanticRank: ranking keywords and sentences using semantic graphs. In Proceedings of the 23rd international conference on computational linguistics (Coling 2010), pp. 1074-1082
Vanyushkin A, Graschenko L (2020) Analysis of Text Collections for the Purposes of Keyword Extraction Task. J Inf Org Sci 44(1):171–184. https://doi.org/10.31341/jios.44.1.8
Wan X, Xiao J (2008) Single document keyphrase extraction using neighborhood knowledge. AAAI 8:855–860
Wan X, Xiao J (2008) CollabRank: towards a collaborative approach to single-document keyphrase extraction. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 969-976. https://doi.org/10.3115/1599081.1599203
Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG (2005) Kea: Practical automated keyphrase extraction. In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, pp. 129-152. https://doi.org/10.4018/978-1-59140-441-5.ch008
Xie Z (2005) Centrality measures in text mining: prediction of noun phrases that appear in abstracts. In Proceedings of the ACL student research workshop, pp. 103-108
Vega-Oliveros DA, Gomes PS, Milios EE, Berton L (2019) A multi-centrality index for graph-based keyword extraction. Inf Process Manage 56(6):102063. https://doi.org/10.1016/j.ipm.2019.102063
Yang Z, Lei J, Fan K, Lai Y (2013) Keyword extraction by entropy difference between the intrinsic and extrinsic mode. Phys A 392(19):4523–4531. https://doi.org/10.1016/j.physa.2013.05.052
Yang F, Zhu YS, Ma YJ (2016) WS-rank: Bringing sentences into graph for keyword extraction. In Asia-Pacific Web Conference, pp. 474-477. https://doi.org/10.1007/978-3-319-45817-5_49
Yang L, Li K, Huang H (2018) A new network model for extracting text key-words. Scientometrics 116(1):339–361. https://doi.org/10.1007/s11192-018-2743-5
Zhang Z, Zweigenbaum P, Yin R (2018) Efficient generation and processing of word co-occurrence networks using corpus2graph. In Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing, pp. 7-11. https://doi.org/10.18653/v1/W18-1702
Zhao WX, Jiang J, He J, Song Y, Achanauparp P, Lim EP, Li X (2011) Topical keyphrase extraction from twitter. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, pp. 379-388
Zhao D, Du N, Chang Z, Li Y (2017) Keyword extraction for social media short text. In 2017 14th Web Information Systems and Applications Conference (WISA), pp. 251-256. https://doi.org/10.1109/WISA.2017.12
Zhou Z, Zou X, Lv X, Hu J (2013) Research on weighted complex network based keywords extraction. In Workshop on Chinese Lexical Semantics, pp. 442-452. https://doi.org/10.1007/978-3-642-45185-0_47

Muskan Garg is working as an Assistant Professor at Amity University Rajasthan, India. She is active in research areas of network science, natural language processing, and social media analysis. She has published her research work in peer reviewed journals and has 6 web of science indexed researcher papers along with a book chapter and a conference paper in COMSNET 2020.