Journal of Theoretical and Applied Information Technology
31st October 2021. Vol.99. No 20

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SYSTEMATIC LITERATURE REVIEW OF AUTOMATIC KEYWORD EXTRACTION ALGORITHMS: TEXTRANK AND RAKE

MANAR YAHYA[1], DERAR ELEYAN[2], AMNA ELEYAN[3]

[1,2] Department of Applied Computing, Technical University-Kadoorie, Tulkarem, Palestine

[3]Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M15 6BH, United Kingdom

E-mail: [1]manar.my.yahya@gmail.com, [2]d.eleyan@ptuk.edu.ps, [3]a.eleyan@mmu.ac.uk

ABSTRACT

There are various techniques available in text mining for text summarization which will provide the concise meaningful text from the original text document. Keywords give the summary of the text and help to understand the information described in the text document. Traditional approaches to extract useful Keywords from a text depends on human effort heavily, and because of the difficult manual extraction and consuming much time, the Automatic Keyword Extraction algorithm has been used to extract a keyword efficiently that reduces the scope for human errors and saves time. TextRank and RAKE are approaches based on unsupervised techniques to extract the keywords from the text. This research systematically identifies peer-reviewed literature that discusses to utilize TextRank and RAKE algorithms for automatic keyword extraction from texts and presents a comparison between these keyword extraction approaches. Also, this systematic review presents the latest applications which use keyword extraction approaches.

Keywords: *Text Summarization, Automatic Keyword Extraction, TextRank, RAKE, Unsupervised Approach*

1. INTRODUCTION

Due to the increase of the massive amount of text on the internet, searching the related texts in a specific subject takes more time and more effort, to avoid that, summarizing these texts is used. A summary is a cut short of text without altering the meaning, which includes the important theme as well. Evaluation and review these texts to create a concise text summarization manually are a time expensive and it is a stressful task specially when dealing with long texts [1]. Therefore, to reduce the efforts and save time, with the same performance as achieved in manual way, an automated summarization is required [2].

Also, readers are not interested in reading a long piece of text and they tend to read the important parts of the text, which raised the demand of automatic text summarization. A text summarization is to extract the most informative portions as a compressed version keep the main content of the text such that help to understand the information quickly and save a lot of time by reading the summary instead of the whole text to get the main idea [3][4].

Automatic text summarization is a time-saving process where computers are capable of creating the summaries faster than humans. Also, it can be scaled to different languages depending on a suitable algorithm whereas humans are bounded by the extent of their experience in a particular language. As well, automatic text summarization has a wide usage in different fields. There are common applications of the text summarization such as in the news articles, search engine results, review summarization, scientific articles, emails, improving performance in text analytics, social media platforms [5]. For instance, social media platforms such as Twitter and Facebook can review through thousands of posts for a given topic, understand the overlapping content, and then summarize this content to generate summaries which used to attract users online. Also, social media platforms are using for political purposes where the majority of political campaigns all over the world use social media as their main tool to reach out to their supporters. Text summarization also is used to answer user queries directly in search

results as in search engines which extract the text from ranked and credible websites and generate a summary for this text which is returned as an answer to the query [4][6].

There are two different techniques to summarize the text: Extract and Abstract summarizations, the extraction way of summarization selects the important and related words in sentences and create a meaningful summary of the text whereas the abstraction way rephrases the text to represent it in the short form [7]. The extractive summaries give the most important piece of text, which can provide an idea of the text content and may tool up a certain sentence which can be used for quotation and citation [8]. Extractive text summarization involves the text pre-processing, extraction of words and phrases that are relevant to the topic of research paper and assembling them to produce a meaningful summary. Keyword extraction from the text is through finding the relevant keywords which describe the content of a text and scoring the candidate keywords using supervised or unsupervised techniques, then the "best" ranked are selected as a keyword of the text.

As far as we know, an existing a systematic literature reviews related to keyword extraction algorithms, specially such as our SLR, which covered a TextRank and RAKE algorithms and their applications is very limited. An existing papers are discussing a keyword extraction techniques or comparing between the keyword extraction algorithms individually, for example, the survey paper which was performed by Baruni and Sathiaseelan is the one of recent survey papers in automatic keyword extraction [9]. In this survey paper, the authors present a concise description on automatic keyword extraction techniques such as KEA, TF-IDF, RACK, TextRank with a discussion of their workflow. While another paper contains workflow of TextRank and RAKE algorithms with a comparison of their performance on a corpus of research paper and the authors inferred that RAKE gives the best results compared to TextRank algorithm [10]. Also, Thushara et al. in their paper compare the performance of unsupervised keyword extraction algorithms; Position Rank, TextRank and RAKE by experimenting the algorithms on research documents [2]. They found that Position Rank gives a bit better results and they noted that the TextRank takes a long time when extracting the keywords from large size research document.

On the other hand, our paper seeks to focus on recent existing literature published from 2019 to early of 2021 which concerning the automatic keyword extraction approaches, the TextRank and RAKE (Rapid Automatic Keyword Extraction) algorithms specially, and their applications, in addition to comparing their performance based on previous experiments, where we cover more than one side in keyword extraction topic integrally, as you will see in next sections.

1.1. Research Goals

This research aims to analyze present studies that are focused on automatic keyword extraction algorithms such as TextRank and RAKE, examine their conclusions, and summarize the research efforts. We generated three research questions, which are presented in Table 1, to help in concentrating the research.

Table 1 Research Questions

Research Questions (RQ)
RQ1: What are the most recent Automatic Keyword Extraction applications?
RQ2: How TextRank and RAKE algorithms extract the keywords from text?
RQ3: Which of these approaches (TextRank and RAKE) has a better performance?

1.2. Contributions

This SLR offers the following to help anyone interested in keyword extraction algorithms continue their study:

We determine 24 primary studies on automatic keyword extraction methods such as TextRank, RAKE, and their applications from 2019 to early 2021. This list of studies could be used by others to further their own research in this area.

We chose 14 primary studies that met the quality criteria we established. These studies can prove to be useful benchmarks for comparing to other similar studies.

We present the workflow of TextRank and RAKE algorithms and compare between their performance for keyword extraction from text.

To encourage future work in this area, we make representations and provide guidelines.

The architecture of this paper is breakdown as follows. The methodology to select the primary studies for analysis is described in Section 2. The conclusions of the analysis of all the primary studies chosen are presented in Section 3. The findings relevant to the earlier outlined research questions are discussed in Section 4. A study conclusion and recommendations for further research were in Section 5.

2. RESEARCH METHODOLOGY

We conducted the SLR in accordance with the guidelines established by Kitchenham and Charters [6] in order to realize the goal of answering the questions of the study. To allow for a full examination of the SLR, we attempted to progress through the planning, conducting, and reporting phases of the review in iterations.

2.1. Primary Studies Selection

Passing keywords to the searches of the particular magazine or search engine was used to highlight primary studies for selection. The keywords were chosen to foster the emergence of research that would help answer the research questions. The Boolean operators that were be utilized are AND and OR. The search queries were as follows:

"Automatic Keyword Extraction" AND ("RAKE" OR " Rapid Automatic Keyword Extraction") OR ("TextRank" OR "Text-Rank")

"Automatic Keyword Extraction" AND "text summarization" AND ("TextRank" OR "RAKE")

Google Scholar was the platform which was used to conduct the search. Depending on the Google Scholar platform, searches were performed against the title, keywords, and abstract. We did the searches from March 10th to March 21st, 2021, and processed all studies that had been published up to that point. We used the inclusion/exclusion criteria, in Section 2.2, to filter the results from searches, allowing us to produce a set of results that met the inclusion criteria.

2.2. Inclusion and Exclusion Criteria

This SLR includes the studies that focus on automatic keyword extraction algorithms, particularly TextRank and RAKE, and their applications. Papers could also offer an experiment of TextRank and RAKE performance. Papers must be written in English and been peer-reviewed product. Table 2 displays the major inclusion and exclusion criteria.

2.3. Selection Results

From the initial keyword searches in the Google Scholar database, a total of 258 studies were identified. After removed duplicated studies and non-open access papers, the total number of studies was decreased to 115. The number of papers left to read after running the research through the inclusion/exclusion criteria was 41. The 41 papers were read in sections (abstract, introduction, and conclusion), and the inclusion/exclusion criteria were applied again, reducing to 24 papers. The number of papers to be included in this SLR has been determined to be 24.

Table 2 Inclusion and exclusion criteria for primary studies

Inclusion criteria	Exclusion criteria
The paper must contain information related to automatic keyword extraction, TextRank or RAKE algorithms and their applications.	Non-English language papers.
The paper experiences the performance of TextRank and RAKE algorithms on different data sets and compares between them.	Papers which compare between the performance of TextRank and another algorithm or compare between the performance of RAKE and another algorithm.
The paper must be a peer-reviewed product published in a conference proceeding or journal indexed on Scopus or SCImago.	

2.4. Quality Assessment

Kitchenham and Charters [11] provided guidelines for assessing the quality of primary research that were determined. This allowed for an evaluation of the papers' relevance to the research goals, with consideration for any indicators of study bias and the validity of experimental results. Four randomly selected papers were subjected to the quality assessment method in order to evaluate their effectiveness, based on the approach employed by Hosseini et al. [12]. The following are the phases of the process:

Phase 1: Automatic Keyword Extraction and their approaches. The paper mainly focused on the automatic keyword extraction topic and it presented enough detail for how these approaches were executing.

Phase 2: Automatic Keyword Extraction applications. There enough detail present in the study for how the way of employing an automatic keyword extraction approach in an application, which will assist in answering research question RQ1.

Phase 3: TextRank algorithm. The papers present a mechanism of the TextRank algorithm in an effort to assist in answering RQ2.

Phase 4: RAKE algorithm. The papers present a mechanism of the RAKE algorithm in an effort to assist in answering RQ2.

Phase 5: Performance of TextRank and RAKE. The papers compare between the performance of TextRank and RAKE and find which has the best performance, which will assist in answering research question RQ3.

Phase 6: Context is important. In respect to the research objectives, enough context must be presented.

We discovered that 10 studies did not meet two or more of the checklist categories after using this checklist for quality evaluation on primary studies that were identified, thus we eliminated those papers from the SLR, as shown in Table 3.

Table 3 Excluded Studies

Checklist Criteria Phase	Excluded Studies
Phase 1: Automatic Keyword Extraction approaches	[S5] [S11] [S12]
Phase 2: Automatic Keyword Extraction applications	[S4] [S12]
Phase 3: TextRank algorithm	[S3] [S4] [S5] [S11] [S13] [S17] [S22]
Phase 4: RAKE algorithm	[S1] [S3] [S13] [S15] [S17] [S22]
Phase 5: Performance of TextRank and RAKE	[S1] [S3] [S4] [S5] [S11] [S12] [S13] [S15] [S22]
Phase 6: Context	[S1] [S3] [S4] [S12] [S17] [S22]

2.5. Data Extraction

The data extraction procedure was tested on an initial four studies and then it expands to include all studies that passed the quality assessment process. For each study, we extracted the data from it, categorized and put it in the spreadsheet to collect all extracted data. The data pursued the following categories:

Context Data: Information regarding the goal of the study.

Qualitative Data: Author's findings and conclusions

Quantitative data: An observed data that has been gathered via experimentation and research.

Figure 1 represents the PRISMA flow diagram which displays the number of papers selected at each stage of the process as well as the papers depletion rate, from searching for the initial keywords from the Google Scholar platform to the final number of primary studies picked.

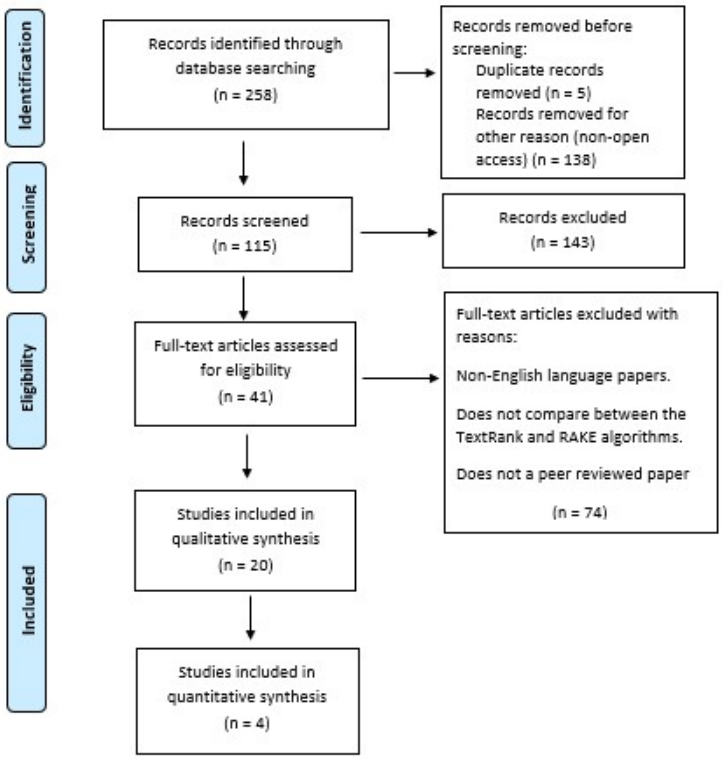

Figure 1 PRISMA Flow Diagram

2.6. Data Analysis

To answer the research questions, we collected the data depending on the qualitative and quantitative data categories and analyzed the selected primary studies.

2.6.1. Publications over time

There is a growing up in the number of research studies that focusing on automatic keyword extraction approaches and applications as shown in the Figure 2, which depicts the number of primary studies published between 2019 and early 2021.

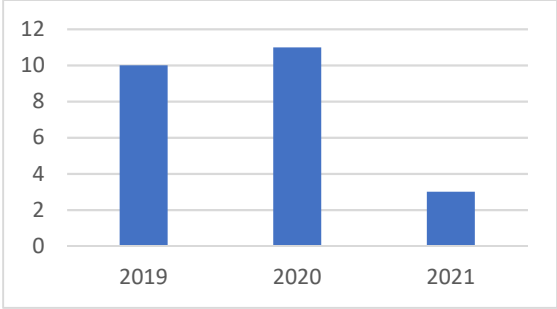

Figure 2 Number of Primary Studies Published Over Time

2.6.2. Counts of important keyword

An analysis of keywords was performed across all 24 papers in order to connect common themes throughout the selected primary studies. The total number of times distinct words appeared across all primary studies are shown in Table 4. The most frequent keyword in our dataset, excluding the authors selected keywords, i.e.," keywords", "extraction" and "TextRank", is" classification" as shown in the table, indicating that there is an interest in employing keyword extraction approaches in classification applications.

Table 4 Counts of Primary Studies Keywords

Keywords	Count
text	792
extraction	473
keyword	410
TextRank	379
keyphrase	311
classification	302
performance	221
RAKE	162
approach	144
automatic	139
unsupervised	119
NLP	116
summarization	86

3. FINDINGS

We read the primary study papers and extracted the relevant qualitative and quantitative data from each primary study and then summarized it as shown in Table 5. We selected the primary studies that focus on automatic keyword extraction approaches as TextRank and RAKE and their mechanisms.

The percentage breakdown of themes for the 24 primary studies which met quality assessment and included in the data analysis is shown in Figure 3.

The themes identified in the primary studies highlight that 33% of all studies into automatic keyword extraction were interested with keyword extraction applications in different fields. As well that 33% of primary studies focused on the TextRank algorithm work mechanism while the work mechanism of the RAKE algorithm was discussed on 19% of these studies. The comparison between TextRank and Rake algorithms is the least theme was discussed with 11%.

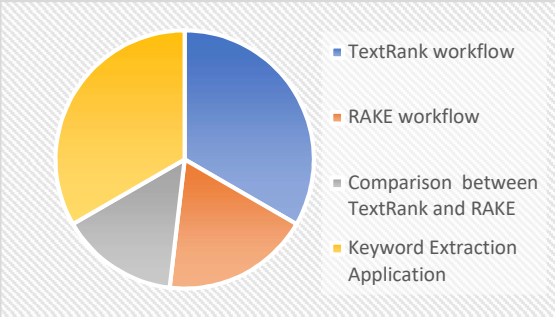

Figure 3 Chart of Primary Study Themes

Table 5 Primary Studies Key Findings and Themes

Primary Study	Qualitative and Quantitative Data Extracted	Theme
S2	This paper is a comparative study of unsupervised keyword extraction algorithms. The authors compare the performance of the Position Rank algorithm which take into consideration the position of the word and its frequency in a document, with TextRank and RAKE by experimenting the algorithms on search documents. In their study, they found that the TextRank takes a long time when extracting the keywords from large size research document.	TextRank Algorithm, RAKE Algorithm, Comparison between TextRank and RAKE
S6	The authors introduce a methodological framework to identify the future work sentences in scientific papers. And to understand the concepts presented in identified future work sentences, they extracted the keywords from it using the RAKE algorithm.	RAKE Algorithm, Keyword Extraction application

S7	The authors built a new automatic keyword extraction method utilizing the unsupervised automatic keyword extractors TextRank, RAKE and TAKE. Their ensemble method follows these steps, firstly, the method receives a list of candidate keywords from each automatic keyword extractor, filters these candidate keywords by removing any candidate keyword that selected by a single extractor and consists of a single word, then combines and recalculates the scores of candidate keywords. Finally, extracts keywords by applying dynamic threshold functions. The dynamic threshold functions used are calculation the overall mean and any candidate keyword has scores higher than the mean is extracted as a keyword for the document, and the other function used the overall median instead of the mean. They evaluated the performance of their ensemble method and the other automatic keyword extraction methods (TextRank, RAKE and TAKE) by applying these methods on a data set contains 2000 abstracts in English for journal papers from Computer Science and Information Technology from the Inspec database. They divided this data set into 3 sets: Set 1 which contains 1000 abstracts, and each of the Set 2 and Set3 contains 500. Then they found that their ensemble method achieved better overall performance.	Keyword Extraction approach, TextRank Algorithm, RAKE Algorithm, Comparison between TextRank and RAKE
S8	This paper proposes a framework to extract and analyze the personal interests and preferences of microblog users. In this framework, the authors used the TextRank method to extract the user interest candidate terms, but they improved it by importing TF-IDF into it as a factor. They did this step because that the TextRank extracts the keywords depending on the relation between terms without taking the term frequency in its consideration while the TF-IDF considers the term frequency in keyword extraction. By experiments, they found that the performance of the improved TextRank method better than TextRank and TF-IDF.	TextRank Algorithm, Keyword Extraction Application
S9	This paper presents a comparative survey for keyword extraction methods and evaluates the performance of these methods by examining these methods on different datasets with varying forms, sizes and types. In this paper, the authors examined the keyword extraction approaches on those datasets: Amazon, SemEval dataset, Stack Exchange and TMDB dataset. From Amazon, they selected the first 100 reviews from Automotive category under Amazon Product data. These selected reviews are unstructured, have spelling and grammar errors, and contain slang in their content. And each review has an average of 50 words. From SemEval, which consists of 280 formal scientific articles collected from the ACM Digital Library, they selected the first 100 articles and these articles are structured with specified heading and body sections. Stack Exchange is a set of question-answer groups on different topics and fields, and each group related to a specific topic. For this experiment, the authors selected 1000 documents randomly. Also, they selected a subset of 1000 random brief abstract of movies from TMDB dataset. By comparing the results of TextRank and RAKE methods, they found that the TextRank method beats RAKE on Amazon and SemEval datasets while RAKE outperforms than TextRank for the TMDB dataset and they achieved a close performance for the Stack Exchange dataset.	Comparison between TextRank and RAKE
S10	The paper presents a method to solve the problem of semantic duplication in the Intelligent Test Paper Generation Method. As a part of this method, the authors in this paper employed the TextRank algorithm to extract the keywords of test questions, more details about the TextRank algorithm will explain in the next section.	TextRank Algorithm, Keyword Extraction Application
S14	This paper proposed a method for online travel review classification which can help to extract the tourist's opinions about their travel and their future destinations from their comments. Opinion's analysis process included the keywords extraction step. To determine keyword extraction algorithm which can be used in this method with better performance, the authors implemented three keyword extraction algorithms namely TF-IDF, LDA and TextRank on four datasets from the TripAdvisor website and they evaluated the results. Thus, they found that TextRank is the most suitable way to extract the text keywords from online travel reviews. TextRank had the better performance because the keywords were extracted from short text where the reviews became shorter after preprocessing step.	TextRank Algorithm, Keyword Extraction Application

S16	This paper presents the workflow of TextRank and RAKE algorithms. The authors implemented TextRank and RAKE in Python=3.7 and to evaluate the performance of these algorithms, they choose literature abstracts from Arxiv NLP papers randomly. After extracting the keywords from this abstract, they compared the results and they found that RAKE extracted the keywords more efficiently than TextRank and given the best result.	TextRank Algorithm, RAKE Algorithm, Comparison between TextRank and RAKE
S18	The authors of this paper proposed an unsupervised approach to extract the most important information which describes the topic from the documents written in the Urdu language. A proposed method was TOP-Rank and its extracts the keywords from the documents with takes the position of keywords in document in consideration. And the authors found that their approach extracts the topic from Urdu language document more efficiently than existing approaches.	Keyword Extraction Approach
S19	In this paper, the authors introduced a method for Vietnamese texts classification in an effective way. Their method collects a Vietnamese text from Vietnamese news websites, extracts the keywords from these texts and classifies these texts depends on news categories. The authors in their method used the TextRank algorithm to extract the most significant keywords.	TextRank Algorithm, Keyword Extraction application
S20	This paper focuses on employment of natural language application in policy documents where the authors introduce a meta-algorithmic modelling framework based on natural language processing techniques: information extraction, automatic summarization, and automatic keyword extraction, for processing internal bank policies. In the automatic summarization process, to generate an individual summary for each document, the authors selected the TextRank algorithm which works on a single document at a time to extract the keywords.	Keyword Extraction Application, TextRank Algorithm
S21	The authors developed an approach to automate the literature review process from collecting articles up to evaluating. They used a rapid automatic keyword extraction algorithm in text mining processes in their approach.	Keyword Extraction Application, RAKE Algorithm
S23	This paper presents an approach to extract the patent keywords from patent texts. The authors used the TextRank algorithm for extracting patent keywords with an improved way to calculate the node rank value.	TextRank Algorithm, Keywords Extraction Application
S24	This paper proposed a method to analyze the movie reviews using k-mean and TextRank to extract the topic of review. The authors used k-mean to classify the reviews in categories and extract the keywords from each category using TextRank where the extracted keywords explain the topic.	TextRank Algorithm, Keyword Extraction Application

4. DISCUSSION

Searching for initial keyword highlighted that the interest of keyword extraction topic is growing up continuously where there are significant number of papers relevant to keyword extraction topic. A subset of the primary studies chosen are experimental proposals for applications are used a keyword extraction as a part of its strategy, while another part of the primary studies chosen are focused on TextRank and RAKE algorithms mechanism and evaluated their performance by implementing them on different datasets and comparing the results.

RQ1: What are the most recent Automatic Keyword Extraction applications?
A keyword extraction technique was used in several fields, where the latest studies indicated that there are many applications that uses the keyword extraction in its approach. These applications were varied in several fields as follows:

- Education and Scientific Research- keyword extraction was used in applications to avoid the redundancy of questions in electronic test papers, to understand the trends of scientific research and topics which interested the researchers, and to automate the literature review process [S10] [S6] [S21] [S23].
- Tourism- to know the tourist opinion about his travel and his future destinations, the keyword extraction was used to classify the online travel reviews [S14].
- Personal Interest and preferences- it can be extracted from the user social media accounts by analyzing their posts and shares, [S8] is an application to extract the personal interests for microblog users.
- classification applications- used a keyword extraction as a part of its strategy, such as Vietnamese texts classification [S19], internal bank policies classification [S20], movie reviews analysis [S24].

RQ2: How does TextRank and RAKE algorithms extract the keywords from the text?

TextRank is a graph-based keyword extraction algorithm used to extract the keywords from a single document. It is an unsupervised approach which does not need training and it is based on the PageRank algorithm used on the Google search engine. TextRank divides the text into several units (e.g., words, sentences), represents it as a directed graph and utilizes the adjacent relationship between words for ranking keywords in the texts.

The keyword is extracted using TextRank as follows:

- Dividing a given text into sentences and tagging parts of speech.
- For each sentence segmentation and part-of-speech tagging, a stop words are filtered out, where common nouns, verbs, adjectives, gerunds and punctuations are considered as stop words. Only words belonging to the specified part-of-speech are retained as candidate keywords.
- Building the candidate keyword graph G = (V, E), where V is the node set containing the candidate keywords, and E is the set of edges in the graph, and each edge represents the co-occurrence relationship between two nodes.
- Calculating the weight of each node in a graph using follows formula, for a node l_i, the weight of this node is:

$$S_{TextRank}(l_i) = (1-d) + d \times \sum_{l_j \in in(l_i)} \frac{w_{ji}}{\sum_{l_k \in out(l_j)} w_{jk}} S_{TextRank}(l_j)$$

Equation 1 Weight of Node

Where w_{ij} indicates the weight of edge from node l_j to node l_i, and d is damping coefficient, which is generally set to be 0.85.

The following example explains how to calculate the weight of node, if we have a sentence contains the candidates keywords (word1, word2, word3, word4) and the candidate keywords graph model for this sentence was represented as follows in Figure 4.

Initially, the weight of all nodes is equaled, and it equals 0.15 in this example where the initial weight of node l_i = (1-0.85) + 0.85 * 0. The weight of word1 node calculating by the above formula as follows:

S(word1) = (1-0.85) + 0.85 * (0.15/1 + 0.15/2) = 0.34, where the word1 have two in-edge edges from (word2, word3), word2 has one out-edge and word3 have two out-edge. And so on for other nodes.

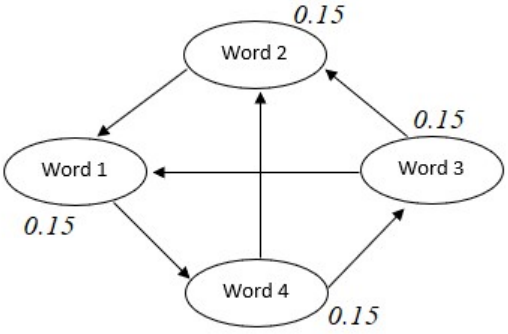

Figure 4 Candidate Keywords Graph

- Sorting the node weights in reverse order to get the most important K words which are obtained as candidate keywords.
- Marking the obtained candidate keywords in the original text and combining the adjacent ones into multi-word keywords.

This algorithm discussed in [S2] [S7] [S8] [S10] [S14] [S16] [S19] [S20] [S23] [S24].

And Figure 5 presents a sample graph build for keyword extraction from an Inspec abstract using the

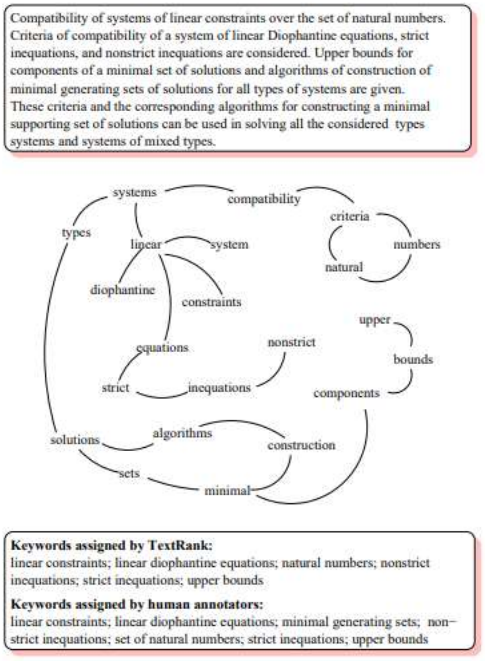

Figure 5 Sample Graph Build for Keyword Extraction from an Inspec Abstract Using TextRank

TextRank algorithm [13].

RAKE is an unsupervised, domain and language independent algorithm for keyword extraction from document. RAKE extracts the keywords as follows:

- Locating candidate keywords by removing all stop words (a, the, of, in, on, at, etc.). Any word appears between two stop-list words and/or punctuation marks are marked as candidate keywords.
- Building the score-weight matrix by calculating:
  Word frequency: which represents the count of times the word is repeated in the document.
  Word degree: which is the degree of co-occurrence of each word in the document.
  Ratio of degree to frequency.

Figure 6 and Figure 7 explain how to build the score-weight matrix, if we want to extract the keywords from the sentence (rapid automatic keyword extraction is an unsupervised algorithm for keyword extraction from document), the candidate keywords are: rapid, automatic, keyword, extraction, unsupervised, algorithm, document. The co-occurrence graph of candidate keywords in the sentence as follows in Figure 6.

And the score-weight matrix for candidate keywords which calculated from the candidate keywords co-occurrence graph as follows in Figure 7.

- Selecting the top-scored candidates as a keyword of document.

This algorithm discussed in [S2] [S7] [S16] [S32]. And the figures, Figure 8 to Figure 11, represent an example for keyword extraction from Inspec abstract using the RAKE algorithm [14].

RQ3: Which of these approaches (TextRank and RAKE) has a better performance?

A number of primary studies compared the TextRank and RAKE performance on different datasets. Performance of TextRank and RAKE algorithms evaluated using standard and statistical evaluation matrices namely: the precision; which represents the ratio between correct extracted keywords and all extracted keywords, the recall; which represents the percentage between correct extracted keywords and manually assigned keywords and F-measure; is the harmonic mean of precision and recall.

According to the experiment in [S7] which applied on data set contains 2000 abstracts in English for journal papers from Computer Science and Information Technology from the Inspec database. Where 2000 abstracts were divided into 3 sets: Set 1 which contains 1000 abstracts, and each of Set 2 and

	rapid	automatic	keyword	extraction	unsupervised	algorithm	document
rapid	1	1	0	0	0	0	0
automatic	1	1	1	0	0	0	0
keyword	0	1	2	2	0	0	0
extraction	0	0	2	2	0	0	0
unsupervised	0	0	0	0	1	1	0
algorithm	0	0	0	0	1	1	0
document	0	0	0	0	0	0	1

Figure 6 Co-occurrence Graph of Candidate Keywords

	rapid	automatic	keyword	extraction	unsupervised	algorithm	document
Word degree	2	3	5	4	2	2	1
Word frequency	1	1	2	2	1	1	1
degree(w) / frequency (w)	2	1.5	2.5	2	2	2	1

Figure 7 Score-Weight Matrix

Compatibility of systems of linear constraints over the set of natural numbers

Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types of systems and systems of mixed types.

Manually assigned keywords:
linear constraints, set of natural numbers, linear Diophantine equations, strict inequations, nonstrict inequations, upper bounds, minimal generating sets

Figure 8 A Sample Abstract from The Inspec Test Set and Its Manually Assigned Keywords

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

Figure 9 Candidate Keywords Parsed from The Sample Abstract

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

Figure 10 The Score-Weight Matrix for Candidate Keywords

minimal generating sets (8.7), linear diophantine equations (8.5), minimal supporting set (7.7), minimal set (4.7), linear constraints (4.5), natural numbers (4), strict inequations (4), nonstrict inequations (4), upper bounds (4), corresponding algorithms (3.5), set (2), algorithms (1.5), compatibility (1), systems (1), criteria (1), system (1), components (1),constructing (1), solving (1)

Figure 11 Candidate Keywords and Their Calculated Score

Set3 contains 500 abstracts. And TextRank and RAKE algorithms were implemented on these sets. The result shows that TextRank has a better performance compared to RAKE in the three different data sets, where precision of TextRank on Set 1 was 31.2% while a precision of RAKE was 26.0%, which is mean that a TextRank extracted a number of correct keywords more than RAKE. And recall of TextRank was 43.1% while recall of RAKE was 42.2%, which is mean that TextRank extracted a number of correct keywords matched to keywords extracted manually more than RAKE and so on for Set2 and Set3. Table 6 [S7] shows these results.

In [S9], TextRank and RAKE were examined on four different datasets with varying forms, sizes and types, namely: Amazon, SemEval dataset, Stack Exchange and TMDB dataset. From Amazon, the authors selected the first 100 reviews from Automotive category under Amazon Product data and these selected reviews are unstructured, have spelling and grammar errors, and contain slang in their content. And each review has an average of 50 words. From SemEval, which consists of 280 formal scientific articles collected from the ACM Digital Library, they selected the first 100 articles and these articles are structured with specified heading and

body sections. From Stack Exchange which is a set of question-answer groups on different topics and fields, and each group related to a specific topic,1000 documents were selected randomly. Also, a subset of 1000 random brief abstract of movies was selected from TMDB dataset. According to the results as shown in Table 7 [S9], The TextRank beats RAKE on the SemEval dataset, because of the corpus is a set of scientific documents, the topics probably are focused. And TextRank works best than RAKE for recall and F-measure evaluation matrices on the Amazon dataset due to the reviews are focused on automotive products. And TextRank and RAKE achieve close performance for the Stack Exchange dataset. While RAKE outperforms than TextRank for the TMDB dataset where the movie's summaries are centered in a specific genre.

Likewise, the experiment which presented in [S16] that implemented the TextRank and RAKE algorithms on literature abstracts from Arxiv NLP papers randomly selected, was found that RAKE extracted the keywords more efficiently than TextRank and given the best result. While the [S2] noted that TextRank consumes more time to extract the keywords from a large size research document.

5. CONCLUSION AND FUTURE WORK

This SLR has covered out the most recent research on keyword extraction, as well as TextRank and RAKE algorithms. Where this research highlighted applications that uses keyword extraction approaches, and focused on TextRank and RAKE methods. Also, this paper introduced a comparison between TextRank and RAKE to extract keywords from a single document. In our opinion, we think that the TextRank algorithm is better than RAKE algorithm, especially for short sentence because maybe the sentence doesn't contain any duplicated word, where we can employ this algorithm to extract keywords from any text, so that is not limited to extract a keyword from a document, but it can be extract a keyword, for example, from SMS messages, chats, posts on social media sites, news summary, which can help to discover a trending topic.

According to this paper, the TextRank algorithm extracts the most significant keywords from the text while the RAKE extracts the keywords depending on word co-occurrences in text. As well, we noted that extraction of keywords from a document with a small number of words using TextRank given the best results. But there is no general situation with a specific feature for documents (structured /not structured, correct/incorrect grammar, focused on the specific topic or not) to determine where we can use TextRank or RAKE. We need to search more and more, so, we plan to implement TextRank and RAKE algorithms on different data sets with specific features, and compare between the results to find a recommendation that gives a good result for any document. Also, since there are many approaches for extracting the keyword from the text documents, we will apply many comparative experiments between keyword extraction approaches (such as TF-IDF, LDA, PositionRank, TAKE, YAKE, RAKE, TextRank) to determine the best approach.

As a possible future direction, and because of the lack of researches which focus on extraction keyword from Arabic text, we plan to implement TextRank, RAKE and other keyword extraction approaches on Arabic datasets.

In this SLR, we presented the workflow of TextRank and RAKE algorithms in details and we compared between their performance for keyword extraction from text according to the previous experiments, which makes this SLR as a guideline for support any further work related to keyword extraction.

ACKNOWLEDGMENT

The authors wish to thank Palestine Technical University – Kadoorie (PTUK) for supporting this research work as part of PTUK research fund.

Table 6 Results of TextRank and RAKE Comparison [S7]

	Set 1		*Set 2*		*Set 3*	
	RAKE	TextRank	RAKE	TextRank	RAKE	TextRank
Precision	26.0%	31.2%	23.9%	31.0%	24.5%	28.9%
Recall	42.2%	43.1%	42.4%	45.3%	42.0%	43.3%
F-measure	32.1%	36.2%	30.6%	36.8%	30.9%	34.7%

Table 7 Results of TextRank and RAKE Comparison [S9]

	SemEval		Amazon Reviews		Stack Exchange		TMDB	
	RAKE	TextRank	RAKE	TextRank	RAKE	TextRank	RAKE	TextRank
Precision	18.1%	21.2%	35.2%	30.5%	16.9%	15.5%	23.1%	17.3%
Recall	16.7%	20.0%	23.8%	27.6%	14.7%	14.2%	21.1%	16.5%
F-measure	17.4%	20.6%	26.7%	29.0%	15.7%	14.8%	22.1%	16.9%

REFERENCES

[1] Rani, U. and Bidhan, K., "Comparative Assessment of Extractive Summarization: TextRank, TF-IDF and LDA", Journal of Scientific Research, Vol. 65, No. 1, 2021, pp. 304-311.

[2] Thushara, M. G., Mownika, T. and Mangamuru, R., "A Comparative Study on different Keyword Extraction Algorithms", 3rd International Conference on Computing Methodologies and Communication. Surya Engineering College, Erode, India, March 2019.

[3] Gao, S., Chen, X., Ren, Z., Zhao, D. and Yan, R., "From Standard Summarization to New Tasks and Beyond: Summarization with Manifold Information", International Joint Conference on Artificial Intelligence, 2020.

[4] Ghodratnama, S., Beheshti, A., Zakershahrak, M., and Sobhanmanesh, F., "Extractive document summarization based on dynamic feature space mapping", IEEE Access, Vol. 8, 2020, pp. 139084-139095.

[5] Sohail, A., Aslam, U., Tariq, H.I. and Jayabalan, M., "Methodologies and techniques for text summarization: A Survey", Journal of Critical Reviews, Vol. 7, No. 11, 2020, pp. 781-785.

[6] Dehru, V., Tiwari, P.K., Aggarwal, G., Joshi, B. and Kartik, P., "Text Summarization Techniques and Applications", IOP Conference Series: Materials Science and Engineering, International Conference on Applied Scientific Computational Intelligence using Data Science, Jaipur, India, 2021.

[7] Kumari, N. and Singh, P., "Automated Hindi Text Summarization Using TF-IDF and TextRank", Journal of Critical Reviews, Vol. 7, No. 17, 2020, pp. 2547-2555.

[8] Bhargavaa, R. and Sharmaa, Y., "Deep Extractive Text Summarization", Procedia Computer Science, Vol. 167, 2020, pp. 138–146.

[9] Baruni, J. and Sathiaseelan, J., "Comparative Analysis on Automatic Keyphrase Extraction (AKPE) Techniques", International Journal of Scientific Research in Computer Science Applications and Management Studies, Vol. 9, No. 4, 2020.

[10] Baruni, J. and Sathiaseelan, J., "Keyphrase Extraction from Document Using RAKE and TextRank Algorithms", International Journal of Computer Science and Mobile Computing, Vol.9, No. 9, 2020, pp.83-93.

[11] Kitchenham, B. and Charters, S., "Guidelines for performing Systematic Literature Reviews in Software Engineering", Engineering, Vol. 2, 2007, pp. 1051.

[12] Hosseini, S., Turhan, B. and Gunarathna, D., "A Systematic Literature Review and Meta-Analysis on Cross Project Defect Prediction", IEEE Transactions on Software Engineering, 2017.

[13] Mihalcea, R. and Tarau, P., "Textrank: Bringing order into text", Conference on Empirical Methods in Natural Language Processing, 2004, pp. 404-411.

[14] Rose, S., Engel, D., Cramer, N., and Cowley, W., "Automatic keyword extraction from individual documents", Text Mining: Applications and Theory, 2010, pp. 1-20.

PRIMARY STUDIES

[S1] Liu, A., Du, X. and Wang, N., "Unstructured text resource access control attribute mining technology based on convolutional neural network", IEEE Access, Vol. 7, 2019, pp. 43031-43041.

[S2] Thushara, M. G., Mownika, T. and Mangamuru, R., "A Comparative Study on different Keyword Extraction Algorithms", 3rd International Conference on Computing Methodologies and Communication. Surya Engineering College, Erode, India, March 2019.

[S3] Li, Q., Li, S., Zhang, S., Hu, J. and Hu, J., "A review of text corpus-based tourism big data mining", Applied Sciences, Vol. 9, No. 3300, 2019.

[S4] Blanck, S., Niekler, A. and Kaulisch, M., "Augmenting a Research Information System with automatically acquired category and keyword information", International Society for Scientometrics and Informetrics, Rome, September 2019.

[S5] Lee, J., Lee, T. and In, H.P., "Automatic stop word generation for mining software artifact using topic model with pointwise mutual information", The Institute of Electronics, Information and Communication Engineers Transactions on Information and Systems, Vol. 102, No. 9, 2019, pp. 1761-1772.

[S6] Li, K. and Yan, E., "Using a keyword extraction pipeline to understand concepts in future work sections of research papers", 17th International Conference on Scientometrics & Informetrics, Rome, Italy, September 2019.

[S7] Pay, T., Lucci, S. and Cox, J.L., "An ensemble of automatic keyword extractors: TextRank, RAKE and TAKE", ComputaciÃ n y Sistemas, Vol. 23, No. 3, 2019, pp. 703-710.

[S8] Niu, R. and Shen, B., "Microblog user interest mining based on improved TextRank model", Journal of Computers, Vol. 30, No. 1, 2019, pp. 42-51.

[S9] Kumbhar, A., Savargaonkar, M., Nalwaya, A., Bian, C. and Abouelenien, M., "Keyword Extraction Performance Analysis", IEEE Conference on Multimedia Information Processing and Retrieval, IEEE Computer Society, 2019, pp.550-553.

[S10] Wang, H. and Yang, W., "An intelligent test paper generation method to solve semantic similarity problem"', IOP Conference Series: Earth and Environmental Science, IOP Publishing Ltd, Vol. 252, No. 5., 2019

[S11] Alexandridis, G., Voutos, Y., Mylonas, P. and Caridakis, G., "A geolocation analytics-driven ontology for short-term leases: inferring current sharing economy trends", Algorithms, Vol. 13, No. 59, 2020.

[S12] Gagliardi, I. and Artese, M.T., "Semantic unsupervised automatic keyphrases extraction by integrating word embedding with clustering methods", Multimodal Technologies and Interaction, Vol. 4, No. 30, 2020.

[S13] Albalawi, R., Yeap, T.H. and Benyoucef, M., "Using topic modeling methods for short-text data: a comparative analysis", Frontiers in Artificial Intelligence, Vol. 3, No. 42, 2020.

[S14] Chen, W., Xu, Z., Zheng, X., Yu, Q. and Luo, Y., "Research on sentiment classification of online travel review text", Applied Sciences, Vol. 10, No. 5275, 2020.

[S15] Zhang, M., Li, X., Yue, S. and Yang, L., "An empirical study of TextRank for keyword extraction", IEEE Access, Vol. 8, 2020, pp. 178849-178858.

[S16] Baruni, J. and Sathiaseelan, J., "Keyphrase Extraction from Document Using RAKE and TextRank Algorithms", International Journal of Computer Science and Mobile Computing, Vol. 9, No. 9, 2020, pp. 83-93.

[S17] Tigunova, A., Yates, A., Mirza, P. and Weikum, G., "CHARM: Inferring Personal Attributes from Conversations", Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, November 16–20, 2020, pp.5391–5404.

[S18] Amin, A., Rana, T.A., Mian, N.A., Iqbal, M.W., Khalid, A., Alyas, T. and Tubishat, M., "TOP-Rank: A novel unsupervised approach for topic prediction using keyphrase extraction for Urdu documents", IEEE Access, Vol. 8, 2020, pp. 212675-212686.

[S19] Huynh, H.T., Duong-Trung, N., Truong, D.Q. and Huynh, H.X., "Vietnamese text classification with TextRank and Jaccard similarity coefficient", Technology and Engineering Systems Journal, Vol. 5, No. 6, 2020, pp. 363-369.

[S20] Spruit, M. and Ferati, D., "Text mining business policy documents: applied data science in finance", International Journal of Business Intelligence Research, Vol. 11, No. 2, 2020.

[S21] Tauchert, C., Bender, M., Mesbah, N. and Buxmann, P., "Towards an integrative approach for automated literature reviews using machine learning", Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020, pp.762-771.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[S22] Kumar, A., Seth, S., Gupta, S. and Maini, S., "Sentic computing for aspect-based opinion summarization using multi-head attention with feature pooled pointer generator network", Cognitive Computation, 2021.

[S23] Huang, Z. and Xie, Z., "A patent keywords extraction method using TextRank model with prior public knowledge", Complex and Intelligent Systems, 2021.

[S24] Liu, Y., Liu, B., Yu, J. and Yu, Z., "Multi-Angle Movie Reviews Analysis based on multi model", International Conference on Computer Big Data and Artificial Intelligence. Changsha, China, 24-25 October 2020, IOP Publishing Ltd, 2021.