A Systematic Literature Review of Keyphrases Extraction Approaches

https://doi.org/10.3991/ijim.v16i16.33081

Lahbib Ajallouda[1(✉)], Fatima Zahra Fagroud[2], Ahmed Zellou[1], El habib Benlahmar[2]
[1]SPM-ENSIAS, Mohammed V University, Rabat, Morocco
[2]LTIM–FSBM, FSBM Hassan II University, Casablanca, Morocco
lahbib_ajallouda@um5.ac.ma

Abstract—The keyphrases of a document are the textual units that characterize its content such as the topics it addresses, its ideas, their field, etc. Thousands of books, articles and web pages are published every day. Manually extracting keyphrases is a tedious task and takes a lot of time. Automatic keyphrases extraction is an area of text mining that aims to identify the most useful and important phrases that give meaning to the content of a document. Keyphrases can be used in many Natural Language Processing (NLP) applications, such as text summarization, text clustering and text classification. This article provides a Systematic Literature Review (SLR) to investigate, analyze, and discuss existing relevant contributions and efforts that use new concepts and tools to improve keyphrase extraction. We have studied the supervised and unsupervised approaches to extracting keyphrases published in the period 2015–2022. We have also identified the steps most commonly used by the different approaches. Additionally, we looked at the criteria that should be evaluated to improve the accuracy of keyphrases extraction. Each selected approach was evaluated for its ability to extract keyphrases. Our findings highlight the importance of keyphrase extraction, and provide researchers and practitioners with information about proposed solutions and their limitations, which contributes to extract keyphrases in a powerful and meaningful way effective.

Keywords—keyphrases extraction, systematic literature review, text mining, natural language processing

1 Introduction

The considerable volume of documents published each year creates a problem to analyze or summarize them. For example, according to [1], nearly 17,000 articles were published in the first quarter of 2020 concerning only COVID-19. To improve the use of this textual data, Keyphrase provides information to understand the content of a text. There are many methods that have provided practical solutions to improve automatic keyphrases extraction, these methods are classified according to [2] into two sets, the first includes unsupervised methods and the second includes supervised methods. These methods have been exploiteded in many NLP applications such as information retrieval, text summarization, text classification, and text clustering. But its performance was not

satisfactory. Some reviews [2], [3] shed light on the challenges faced by these methods, and provide solutions to improve the performance of these methods, but these reviews only included the methods that were published before 2019, while there are many modern methods that appeared and were not included in the reviews, especially the methods that predicts key phrases not mentioned in the document.

Therefore, this paper aims to provide a comprehensive review of the techniques for extracting and predicting keyphrases in the document. The review aims to analyze and discuss the literature on proposed solutions that has been published in recent years. In addition to supervised and unsupervised methods for key phrase extraction, our article reviews methods that predict key phrases not mentioned in the document. Also, we will aim to study the identification of candidate keyphrases and discuss criteria that should be evaluated to improve the accuracy of extraction and prediction of key phrases. For each approach, we examine evaluation metrics, datasets used, extraction accuracy, and a discussion of evaluation findings. Finally, we provide solutions to improve the performance of extraction and generation of keyphrases. We will also suggest promising research directions.

The rest of this paper is structured as follows. Section 2 presents the background and preliminaries. Our research objectives are detailed in Section 3. Section 4 represents the methodology used to carry out this systematic review. Section 5 reports and analyzes the results, while Section 6 discusses and critiques the findings, describes the directions of the research and states the limitations of this review. We conclude our article with Section 7, which also contains future directions for research.

2 Background and preliminaries

Automatic keyphrase extraction (AKE) is a domain of text mining that aims to identify the most useful and important terms that give meaning to document content [4]. This section introduces the steps in the AKE process, and the domains that could benefit from using keyphrase extraction techniques.

2.1 Applications

Automatic keyphrase extraction is used in many domains dealing with textual data, such as text classification [5], document clustering [6], document summarization [7], and search engines [8]. Although some studies have attempted to limit these domains like [9], which limited their use to five domains, due to importance of the information provided by the keyphrases, the AKE can also be exploited in many other domains such as recommender systems [10], web mining [11], bibliometric analysis [12], and sentiment analysis [13].

2.2 Keyphrases extraction process

The keyphrase extraction process goes through a set of steps. Merrouni et al. in [3] defined it, in five main steps as shown in Figure 1, where the text goes through the

preprocessing step, which aims to remove unnecessary textual units. In order to eliminate the noise in the basic text. Many techniques are used, such as tokenization, stop word removal, stemming, and normalization.

According to, [14] and [15], candidate keyphrases are terms that do not contain punctuation or stop words and have morphosyntactic structures "adjective* noun+", for example, ("Big data", "Computer engineering", etc.). Many techniques used to select candidate keyphrases, such as Part-Of-Speech, N-grams [16], and Noun-Phrase-Chunks [17].

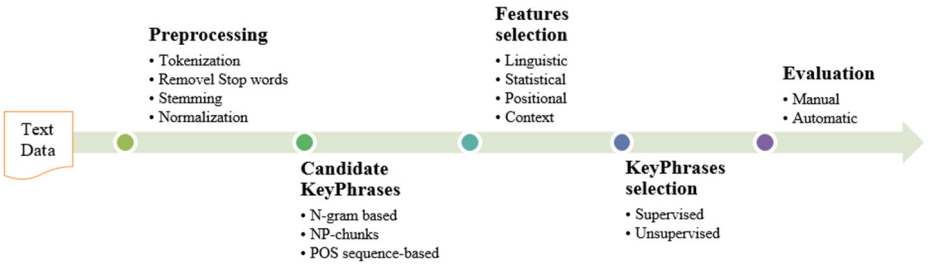

Fig. 1. Keyphrases extraction process

In the third step, each method selects the features of the candidate phrase on which it will rely to determine the keyphrases. According to [18], these features can be classified into statistical, positional, linguistic, and contextual features. The keyphrases are extracted, either via a supervised or unsupervised approach. Supervised approaches mainly teach how to classify candidate keyphrases into "keyphrases" or "non-keyphrases". Unsupervised approaches view this task as a ranking problem. The set of candidate keyphrases is ranked according to a weighting score. The first n candidates are considered keyphrases.

The evaluation process is the last step which aims to know the performance of the approach used to extract the keyphrases. This evaluation is based on dataset available in the literature (scientific articles, and news) and can be carried out manually or automatically, via several metrics, such as precision, recall, F-score, Mean Reciprocal Rank (MRR), and Mean Average Precision (MAP).

3 Research objectives

In recent years, a number of technologies have appeared to improve the automatic processing of text data. This development has greatly improved the performance of the techniques used to extract keyphrases. This highlights the importance of a systematic literature review that provides a comprehensive overview of the recent techniques used to extract keyphrases. Through this SLR, we identify and evaluate the conclusions of the AKE approaches published between 2015 and 2022. The research objectives include studying candidate keyphrase selection techniques, while defining criteria and requirements that must be evaluated to improve the accuracy of keyphrase extraction. Additionally, for each selected article, we will review and discuss evaluation metrics

and results as well as the features and datasets used. Therefore, our findings will provide researchers, and practitioners with information for future investigations for automatic keyphrase extraction.

3.1 Research methodology

To carry out our SLR, which aim to accomplish a specific sequence of detailed steps to gather as much research as possible, several works concerned with the methodology of carrying out systematic reviews of the literature are proposed [19] and [20]. Our study follows the guidelines provided in [21]. We also read some published SLR such as [22] and [23] to get a general idea of how to create SLR.

Our SLR has three main phases. The planning phase which includes the definition of the desired objectives and the predetermination of the research strategy followed. The conduction phase includes the selection of primary studies, the assessment of their quality, as well as the extraction and synthesis of applicable information. The last phase is the results which include an effective interpretation of the results obtained, according to the objectives of the review.

Research Protocol. In this study, we applied a scientific research protocol, comprising several steps. Figure 2 presents the steps followed to perform this protocol.

Fig. 2. Phases of the research protocol

The research questions. We identified a set of research questions (see Table 1) to achieve the main objective of our study, which is to obtain a state of the art on keyphrase extraction techniques, by examining the articles published during the period 2015–2022.

Table 1. The research questions

Research Questions	Motivation and Purpose
RQ1: What techniques can be exploited to identify candidate keyphrases?	Identify the techniques used by keyphrase extraction approaches, to eliminate unnecessary phrases.
RQ2: What techniques can be used to extract keyphrases?	Highlighting the most commonly used algorithms, by keyphrases extraction systems.
RQ3: How to estimate the precision of the proposed approaches?	Identify the techniques and datasets used to validate the solutions
RQ4: What are the most realistic and scalable AKE software?	In order to introduce researchers to this kind of software and to motivate them to develop it or to implement others
RQ5: What obstacles must be overcome to improve the accuracy of keyphrases extraction?	Specify requirements that remarkably affect the efficiency and performance of keyphrase mining systems

Electronic Data Sources. In this study, we used a strategy, based on multiple electronic data sources (EDS), to collect related work. We conducted an online search by five electronic data sources (see Table 2). These EDS include all the journals and conference proceedings of high-quality to automatic keyphrases extraction approaches. We also applied a snowball search strategy by a bibliographic analysis of the selected articles to find more related articles.

Table 2. Electronic data sources adopted in the study

Num	EDS Name	Address
EDS1	ACM Digital Library	https://dl.acm.org
EDS2	DBLP	https://dblp.org
EDS3	IEEE Xplore	https://ieeexplore.ieee.org
EDS4	ScienceDirect	https://www.sciencedirect.com
EDS5	Google Scholar	https://scholar.google.com

Search keywords. We defined the keywords for the research, using specific terms, in order to collect as many relevant articles as possible in the study. The set of keywords that we used to implement an SLR are "keyphrase extraction", "Keyphrase generation", and "keyword extraction". Next, we designed the search strings for each data source to check the title, summary, and keywords, except for Google scholar which only allows titles search.

Exclusion and inclusion criteria. This study focuses on articles published during the period 2015–2022. We first analyzed the studies according to titles, years of publication, keywords, and abstracts. To select or exclude any article, we defined in Table 3 the exclusion and inclusion criteria.

Table 3. Inclusion and exclusion criteria

Inclusion Criteria	Exclusion Criteria
IC1 Articles related to research questions (Q1–Q5)	EC1 Articles not dealing with keyphrases extraction
IC2 Journal or conference articles	EC2 Duplicate papers in EDS
IC3 Articles written in English only	EC3 Working papers
IC4 Articles published between 2015 and 2022	EC4 PhD dissertations, tutorials, editorials, magazines.

Quality assessment. During this phase, each paper in the final group went through an evaluation process to measure its quality. For this, we used an evaluation checklist containing six Qualities (see Table 4), we assigned the highest weight to Qualities Q3 and Q5, which respectively deal with the architecture of the proposed solution and comparing the results of the article along with other articles. The qualities, study objectives Q1, related work Q2, evaluation of results Q4, and statement of results Q6 were of low importance.

Table 4. Quality assessment checklist

Num	Qualities	Vote		Weight
		Answer	Score	
Q1	Are the objectives of the study clear in the article?	Yes	1	1
		Partly	0.5	
		No	0	
Q2	Did the study examine related work?	Yes	1	1
		Partly	0.5	
		No	0	
Q3	Did the study clearly identify and discuss the proposed solution?	Propose a new solution to extract keyphrases and describes its architecture	1	2
		The study discusses the proposed solution	0.5	
		The proposed solution is not well defined or discussed	0	
Q4	Was the study evaluated empirically?	Implement the proposed solution and use it in real application	1	1
		The study provided only the implementation	0.5	
		The study did not provide any implementation or results	0	
Q5	Did the study compare the results of the proposed solution with other studies?	Yes	1	2
		Partly	0.5	
		No	0	
Q6	Did the study present a clear statement of findings?	Yes	1	1
		Partly	0.5	
		No	0	

For each article, its quality score is calculated using the formula (1), by considering the score for each question S_i as well as its weight W_i

$$QS = \sum_{i=1}^{6} \frac{(S_i \times W_i)}{8} \times 100 \quad (1)$$

4 Results

This section is devoted to summarizing the data extraction results obtained by applying the research protocol detailed in section 3, with the aim of analyzing the results of each research question. In order to provide a comprehensive review of automatic keyphrases extraction.

4.1 Overview of research articles

The first step allowed us to collect 607 articles (see Figure 3). Next, the article titles were checked for duplicates. This process allowed us to remove 187 articles and only

keep 420 articles. These articles were reviewed according to the exclusion and inclusion criteria described above. Where the number of articles decreased to 159 research articles. After looking at it, we found that only 61 of the 159 articles were relevant to extracting the keyphrase. These articles were supplemented with five more articles after reviewing the reference lists in related articles. In the last step, the six quality assessment criteria in Table 4 were taken into consideration to ensure that the included articles would make a valuable contribution to our SLR. Articles whose score was less than 62% (average of the scores) are eliminated. in the end, we kept 60 articles.

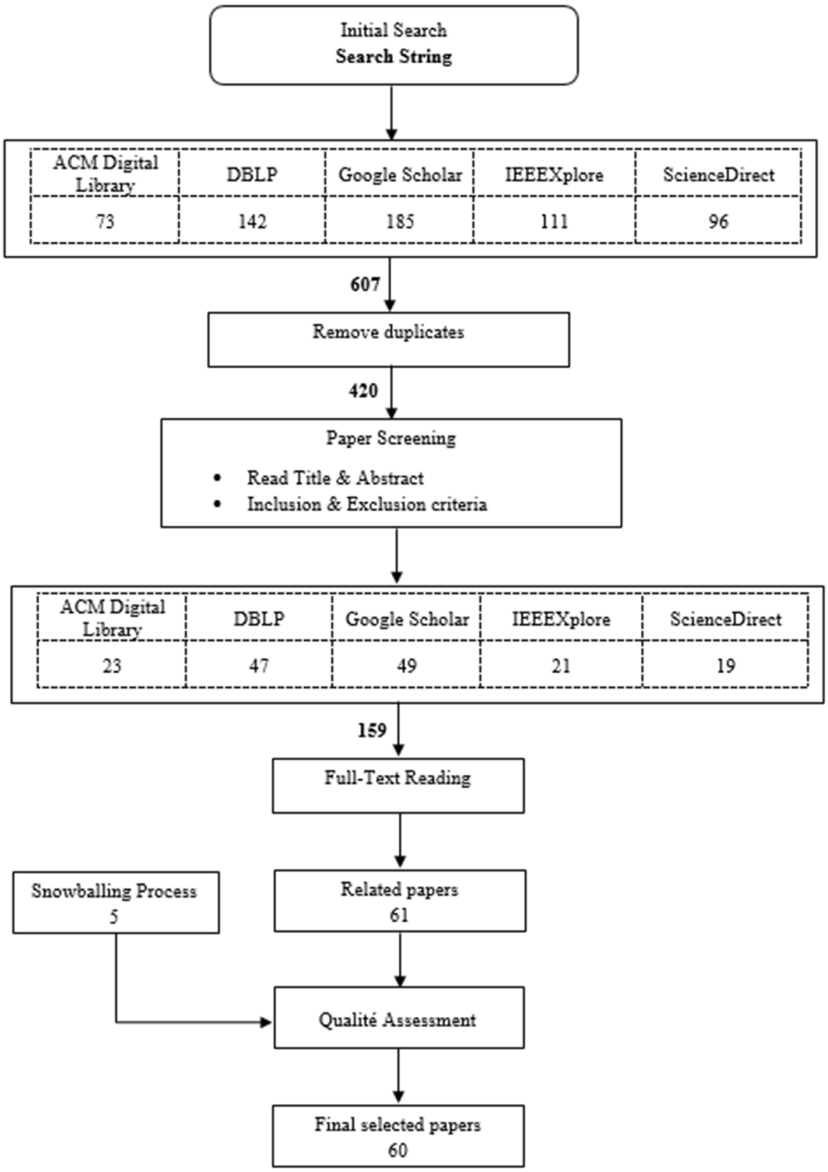

Fig. 3. Overview of final papers selection process

4.2 Classification of selected papers

To classify the selected articles, we have distributed them by EDS, and year of publication. We distributed according to the year of publication (2015–2022), the 159 articles that we obtained after applying the exclusion and inclusion criteria, view Figure 4. The first thing to note is the increase in the number of studies from 2018 to 2019. Indicating the growing interest in developing keyphrase extraction methods. It should also be noted that the results for 2022 are not final. We have also divided the articles before and after the quality assessment phase.

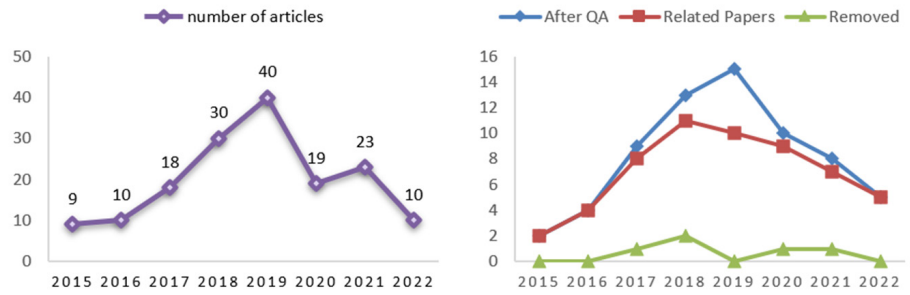

Fig. 4. Distribution of scientific articles according to the year of publication

Our statistics, on data sources, show that Google Scholar and DBLP contain the highest number of relevant articles, see Figure 5, of the 60 articles selected, 55% were published in journals and 45% were presented at conferences.

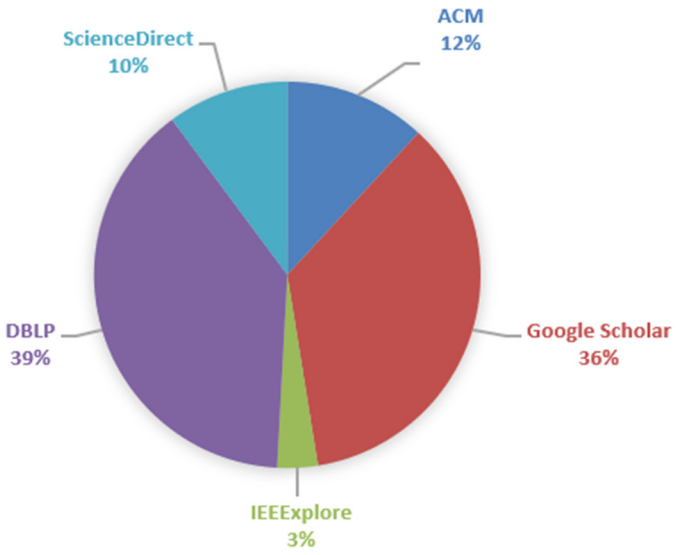

Fig. 5. Distribution of scientific articles according to data sources

4.3 Review and discuss results

In this part, we will discuss the results of this systematic review in order to shed light on the research questions that were raised in the fifth section.

RQ1: "What techniques can be exploited to identify candidate keyphrases?" The choice of keyphrases is always made from a group of candidate keyphrases, hence the importance of knowing the techniques of extracting candidate keyphrases from a text. To address RQ1, we review the techniques that improve the extraction of candidate keyphrases adopted in the studies.

Most of the articles reviewed noted that keyphrases are noun phrases that consist of one or more words and do not include stop words. Therefore, the process of extracting candidate keyphrases must take this into account. The noun phrase appears in many different patterns. In [24], the authors found 56 of these models from a training dataset. The important models are shown in Table 5. In order to define noun phrases, the majority of articles use the Part-Of-Speech (POS) technique. POS converts a phrase into a list of tuples (word, tag). It assigns parts of speech to each word, such as verb, noun, and adjective.

Table 5. The important model for POS Pattern

POS pattern	Description
< *DT*>? < *JJ* > * < *NN*. > +	Begins with an optional determinant DT, followed by zero or more JJ adjectives. followed by one or more NN names.
< *JJ*>? < *NN* > +	Begins with multiple optional adjectives JJ, followed by one or more NN nouns.
(< *JJ*> | < *NN* >)* < *IN* >?(< *JJ*.>| < *NN*.>)* < *NN*.>	Begins with an adjective or noun <JJ | NN> followed by an optional subordinating conjunction IN followed by several adjectives or nouns and ends with noun NN.
< *NN* > + < *JJ* >?	Begins with one or more nouns NN, followed by zero or more optional adjective JJ.
< *PRP* >? < *JJ*> * < *NN* > +	Begins with an optional personal pronouns PRP, followed by zero or more JJ adjectives. followed by one or more NN names.

In order to reduce the number of candidate key phrases, a set of techniques has been exploited. The authors [25] suggested eliminating n-grams that do not have a minimum incidence. The remaining phrases are then classified according to TFIDF to keep the most important. The authors of [26] use the concordance called phraseness, which measures the probability that a sequence of words can be considered as a phrase. Authors [27] propose a neural model of three layers: The embedding layer, the token-level BiLSTM layer and the CRF tagging layer to extract a set of candidates. The authors of [28] did a deep syntactic analysis of the text using the feeling parser and obtained a syntax tree of the text to achieve higher coverage of the text document. The authors of [29] considered the past participle of the verb (VBN) as an adjective and the gerund of the verb (VBG) as a noun. The noun phrase form has been modified to take into account the VBN and VBG tags to extract the candidate phrase. There are also studies suggesting a complete process for extracting candidate key phrases [15].

RQ2 "What techniques can be used to extract keyphrases?" After we have reviewed methods for identifying candidate keyphrases, we now focus on reviewing the techniques used in approaches to extracting keyphrases from documents. In recent years several approaches have been published. These methods employ many techniques. The methods we studied can be classified into four categories, graph-based models, statistical models, embedding models, and deep learning models.

Statistical approaches. The statistical technique is one of the oldest techniques used to extract keyphrases. However, we find that some new methods have been adopted in the AKE process. Among these approaches we find, Giamblanco et al. propose in [30] Key-LUG, an unsupervised approach to extract keyphrases using Newton's law of gravitation. Key-LUG uses a new weighting method that combines both the character length of a word and the frequency of a word in a document. Rabby et al. in [31] provide a tree-based automatic keyphrase extraction technique that uses nominal statistical knowledge. Campos et al. in [32] offer YAKE, an unsupervised approach that uses the statistical text function to extract keyphrases from text. The advantage of YAKE is that it does not use a dataset or dictionaries and it handles various document sizes, and Rabby et al. in [9] proposes TeKET, is a domain-independent technique that uses limited statistical knowledge and does not require any data. To extract keyphrases, it uses KePhEx (Keyphrase Extraction Tree), a new variant of a binary tree. KP-Rank [33] is an AKE approach based on LSA (latent semantic analysis) and clustering techniques and an algorithm based on phrase, paragraph and section frequencies for ranking candidate phrases. Merrouni et al. propose in [34], an unsupervised method that combines linguistic, statistical, semantic, and structural features of a text to identify keyphrases especially in long texts. The method relied on defining the candidate phrases on the parse tree, filtering and part of speech tag approach, which helped to control the computational complexity. Badrul et al. propose in [35] a keyphrase concentrated area (KCA) as a new feature to extract the keyphrase from applying some statistical operations. The proposed method is multilingual and not related to a specific field.

Graph-Based approaches. Representing text in graph form is among the most attractive techniques for researchers, as it has been used in a large number of keyphrase extraction methods (see Figure 6). Most of the traditional methods adopt the relation of co-occurrences between the phrases of the document. Therefore, modern methods have attempted to add the phrase position, syntactic, and semantic relation between phrases in order to improve the extraction of keyphrases. Wen et al. 2016 propose in [36] the use of similarity and co-occurrence between phrases as a new edge weight. For that they used Word2Vec to represent the candidate phrases, and the distance of cos to calculate the similarity between the connected phrases. Florescu et al. consider in [37] that exploiting the phrase locations in the document when calculating the weight can improve the extraction of keyphrases. For this, they proposed a method called PositionRank, which exploits all the positions of the occurrences of the phrase when calculating the score. Figueroa et al. propose in [38] RankUp which applies backpropagation to improve keyphrase extraction algorithms based on graphs. Chen et al. [39] group the co-occurrence relation and the semantic relation, to build a multi-relational graph. Perez et al. propose in [40] an approach that combines lexico-syntactic models and graph-based topic modeling.

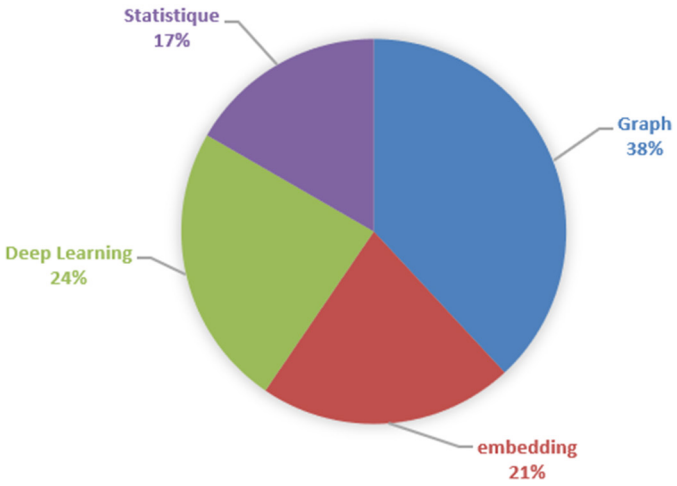

Fig. 6. Distribution of articles according to the technique used

Boudin in [41] represents the document as a complete directed multiparty graph. This representation allows the ranking algorithm to fully exploit the interrelationship between topics and candidates and allows the inclusion of keyphrase selection preferences in the subject. Sun et al. propose [42] DivGraphPointer, an approach that combines the recent approaches based on neural networks and advantages of ranking methods based on graphs. Li et al. [43] suggest adding the idea of topic-based clustering to a graph-based ranking, to embed semantic information. Prasad et al. propose in [44] the Glocal technique (global-local portmanteau) a graph convolution model to incorporate the global importance of the node in the local convolution process for supervised learning on graphs. Tosi et al. propose in [45] C-Rank approach, explores concept links to improve keyphrase extraction. C-Rank constructs a co-occurrence graph with the concepts annotated as vertices. Then, it weighs the vertices using their centrality in the graph.

Dong et al. [46] use several features such as, the total number of sentences in which the target word appears in the sentence, the sum of the inverses of the word's position in the document, and the LDA model to extract the topic information. These features are used when calculating scores by the PageRank algorithm. Yeom et al. propose in [47] a model that exploits modified C-Value to overcome biased co-occurrence frequency and loss of position information. The proposed model uses a corpus of documents as input. The modified C-Value method is applied to recalculate the scores of the keyphrase candidates. Chen et al. propose in [48] an approach based on a three-way decision graph. Using TWDT (three-way decision theory), candidate keyphrases are divided into positive domain, border domain, and negative domain according to graph-based attributes.

Luo et al. propose in [49] a model to classifie candidate phrases according to a structural score and a semantic score. The structural score is calculated using a graphic ranking algorithm. The semantic score is calculated by the similarity between the candidate and all phrases. Yeon et al. propose in [50] an HSN (Hierarchical semantic network) to extract keyphrases using centrality metrics. The identification of hierarchical relationships between keyphrases is motivated by [9]. TOP-Rank [51] is a technique that integrates both topical and positional information. TOP-Rank, based on Positionrank [37]

and applies clustering by topic to similar phrases using cosine similitude and TF-IDF. The highest ranked phrase from each topic is considered the keyphrase. Yang et al. propose in [52] the use of a graph convolutional network (GCN) on document graph to capture core features of text. With the aim of ensuring consistency of the generated keyphrases with the document. Venktesh et al. propose in [53] a method based on representing the text with a graph, where the nodes contain the candidate phrases, while the edge weights represent the semantic association between these nodes. This is done by using embedding techniques to represent the candidate phrases. The proposed method uses a ranking algorithm to select keyphrases.

Embedding approaches. In recent years, various embedding techniques have been proposed [54], Table 6 shows them. This development has encouraged researchers to develop keyphrase extraction methods based on these techniques. Zeng et al. propose in [55] an algorithm that uses the word embedding technique. It is semi-supervised, integrates word frequency, the effects of word co-occurrence, and the semantic relationship between words. The words in the document are grouped according to the vector distance between the words. Bennani et al. [56] consider the candidate keyphrases most similar to the vector representing the document as keyphrases. For this, they propose EmbedRank, an approach based on embedding phrases and documents for the extraction of keyphrases, as opposed to embedding standard individual words. Papagiannopoulou et al. propose in [57] an unsupervised method of extracting keyphrases, based on the calculation of an average vector of the words of the title and the abstract, of a document called the reference vector. The candidate keyphrases are classified according to their cosine similarity with the reference vector.

Mahata et al. propose in [58] Key2vec, an approach using the title and the summary, as excerpts from the topic. Each phrase in the topic snippet is represented by a phrase embedding model. The final theme vector is obtained by adding the vectors of the theme extract. Key2vec calculates the cosine distance between the theme vector and each candidate. The ranking of keyphrases extracted from them using PageRank weighted by theme similarity. Toleu et al. propose in [59] KeyVector, an unsupervised approach based on the calculation of three classification scores: global semantic score, calculation of the semantic relationship between the document and the candidate phrases, and weighted Topics, calculated by the semantic relationship between the topic and the documents and the topic's internal score, is the ranking of keyphrases within each topic. Each candidate keyphrase is ranked according to its values for the three scores.

Table 6. Different models of word and phrase embedding

Model	Vector Dimension
Word2vec [60]	200
Doc2Vec [61]	300
Glove [62]	300
Sent2vec [63]	700
Infertsent [64]	4096
Elmo [65]	1024
BERT [66]	700
USE [67]	512

Fan et al. propose in [68] incorporate local context of the word graph, topical information expressed in the document, and co-occurrence between words, which are important for keyphrase extraction. A new PageRank-based ranking model is designed to extract keyphrases by taking advantage of these features. Sun et al. [69] use SIF a phrase embedding model [70] to extract the relationship between phrase embeddings and the topic of the document. Next, they combined the ELMo autoregressive [65] with SIF to calculate phrase embeddings and document embedding. Cosine similarity is used to calculate the distance between candidate phrases and the topic. Rafiei et al. propose in [29] GLEAKE (Global and Local Embedding Automatic Keyphrase Extraction), an unsupervised method for AKE using a combination of local indices and semantic information from candidate phrases. GLEAKE is based on a local word embedding model to assign a syntax vector to each candidate keyphrase and the document. Ajallouda et al. confirme in [71], that calculating the similarity between candidate phrases and a document is not performed in long documents, so they suggested dividing the document into parts and then calculating the average similarity of the candidate phrases with the parts of the document that are probable to contain keyphrases. Wang et al. propose in [72], to rely on Bert embedding technic to extract keyphrases from the document, by use the whitening operation and reduce dimensionality, well as word frequency information.

Deep Learning approaches. Deep learning (DL) approaches for keyphrase extraction typically apply the process AKE in an encoder-decoder framework, which first encodes documents from input by vector representation, then generate keyphrases with decoders. In recent years several DL approaches have been proposed. Jonathan et al. propose in [73] an approach incorporate the DBN (Deep Belief Networks) as a classifier, uses factual sentiment as a new feature of the keyphrase. Helmy et al. [74] consider that recent AKE deep learning approaches are not applicable to documents in Arabic. For this, they propose a DL model developed on the basis of the LSTM and uses AraVec for word embedding, to extract keyphrases from Arabic documents. Alzaidy et al. [75] discusse keyphrase extraction as a sequence labeling problem. For this, they propose a model AKE, which combines CRF and Bi-LSTM, to extract keyphrases from scientific documents. This model inserts a Bi-LSTM layer between the output and input layers in order to exploit dependencies in the text. Patel et al. [76] Build a complex labeling model use the Bi-LSTM-CRF network, which incorporates long distance information about an input sequence as well about the output sequence. Sahrawat et al. [77] formulate AKE as a sequence tagging using a BiLSTM-CRF, where phrases from the input text are represented at the using deep embedding. They propose to use deep contextual integration models (BERT, SciBERT, and ELMo) instead of the use of fixed word embedding models (word2vec, Glove and FastText).

Xiong et al. propose in [78], Beyond Language Understanding Keyphrase Extraction (BLING-KPE), an approach addresses the challenges of AKE in documents from varying domains and content qualities. BLING-KPE uses a convolutional transformer architecture to model language properties in web documents. The BLING-KPE process has two main components, the embedding of hybrid words where each word is represented by its ELMo embedding, position integration and visual features. The convolutional transformer to model n-grams and their interactions. BLING-KPE first composes the hybrid word embeddings into n-gram embeddings using CNN. The final score of an n-gram is calculated by an anticipation layer on the transformer. Zhu et al. propose in [79] a neural network-based approach, which uses bidirectional long-short

memory (BLSTM) and conditional random field (CRF), for the extraction of scientific keyphrases. Word representation is done by concatenation of word embedding, POS embedding, and dependency embedding. Then the Bi-LSTM layer takes the word representation as input and generates more complex functionality for the input phrase. Finally, CRF is added to predict the sequence of labels for the phrase. Zhou et al. propose in [80] an approach that uses the memory array [81] to capture the long-range contextual information hidden in the textual data. they use CRF model to capture dependencies between adjacent words in a sequence of text and determine if a candidate phrase is a keyphrase. Huanqin et al. propose in [82] to exploit the keyphrases mentioned in the document, in order to generate keyphrases not mentioned in the document by using the mask-predict method.

RQ 3 "How to estimate the precision of the proposed approaches?" After analyzing the proposed methods for extracting keyphrases, we review the proposed measures in order to evaluate these methods, well as the datasets used.

In order to know the precision of keyphrase extraction for any approach, an evaluation process must be performed. The approach is applied to a set of data and the extracted keyphrases are compared via evaluation metrics to a set of manually assigned keyphrases. There are many datasets that have been exploited in the evaluation process by the methods studied. In Figure 7 we present ten datasets that were used by most of articles studied.

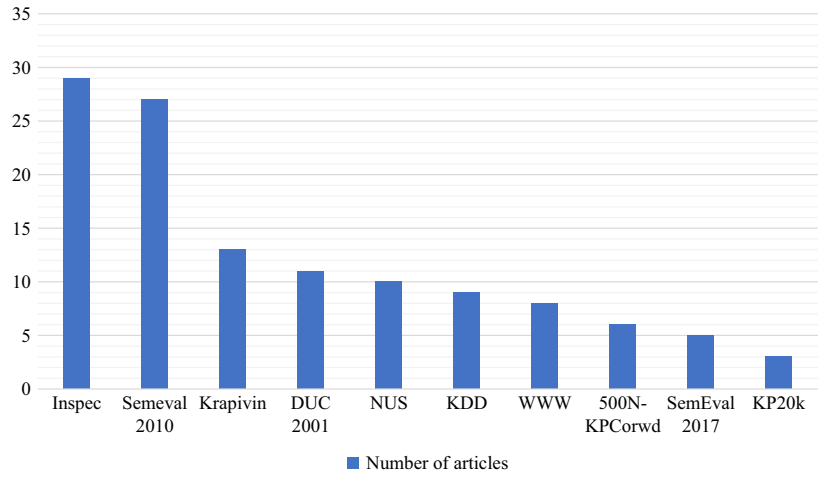

Fig. 7. Comparison of datasets based on the number of articles in which they were used

Inspec, Semeval2010 and Krapivin are the most used datasets. It is not surprising that all three datasets are widely used. Because its scientific publications have a professional expression and a clear semantics compared to other datasets. Each paper used in these groups has its own keyphrases assigned by the authors which make the evaluation process somewhat accurate. However, the only problem that datasets, is the dependence of most of them on articles in the English language, and therefore the methods that extract keyphrases from articles in another language such as Arabic, for example, are difficult to evaluate despite some attempts such as WikiAll [83] composed of 100 documents collected on Arabic Wikipedia. Each document has its own keyphrases written by its authors.

Datasets. In order to evaluate and develop their approaches. The authors need textual sources. That is to say scientific publications, news documents, and abstracts of articles. Table 7 presents these sources.

Table 7. Datasets used for AKE evaluation

Type	Dataset	Language	Docs	KP/Doc
Full-text Papers	NUS [84]	English	211	10
	Krapivin [85]		2300	6
	PubMed [86]		1300	5
	Citeulike-180 [87]		180	5
	Semeval2010 [88]		282	15
Paper Abstracts	Inspec [89]	English	2000	10
	KDD [90]		755	4
	WWW [90]		1300	5
	KP20k [91]		550K	4
	KPTimes [96]		260K	5
News	DUC-2001 [92]	English	308	10
	WikiAll [83]	Arabic	100	–
	110-PT-BN-KP [93]	Portuguese	110	28
	500N-KPCorwd [94]	English	500	46
	Wikinews [95]	French	100	10

Metrics. Most authors trust three metrics of precision, recall and F1-score, due to their accuracy and ease of use. Some works such as [32] which was also based on mean mean precision (MAP), and [68] that used mean reciprocal rank (MRR). Figure 8 shows the percentages of use of these metrics in the articles we studied.

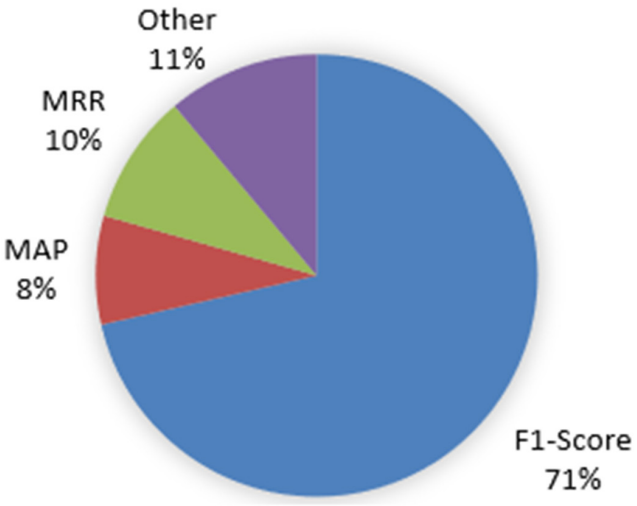

Fig. 8. The proportions of using evaluation metrics

RQ 4 "What are the most realistic and exploitable AKE systems?" The growing demand for the use of keyphrases in NLP fields has prompted researchers to develop efficient systems of extracting keyphrases. In this section, we present these systems with the aim of enabling researchers to identify and exploit them in order to develop them or produce new, more efficient systems.

In recent years, several AKE systems have been developed, some of them apply a single AKE approach, like KEA [4], which is a keyphrase extraction system, which works at the technique [97]. KEA is developed by the JAVA language. It is available under the GNU General Public License. Pytextrank is a python implementation, works like TextRank, with some modifications, notably the graph also contains verbs but are not selected as keyphrases. Also uses lemmatization instead of stemming.

RAKE (Rapid Automatic Keyword Extraction) implemented in Python, works by technique [98]. RAKE selects phrases that are at least five characters long, phrases that make up at most three words and appear in the text at least four times. TopicCoRank, also implemented by Python [99], builds two graphs, the first represents the document and the second represents the domain, which allows extracting keyphrases that belonged to its domain. The source for TopicCoRank, is also available under GitHub. Seq2Seq is one of the few systems that rely on neural networks to extract keyphrases. It is also implemented in Python [91]. Finally, YAKE is a lightweight system, implemented by python which relies on the statistical approach [100] to select keyphrases. This system does not need an external corpus. It's also available under GitHub.

Table 8. Automatic keyphrase extraction systems

Implementation Language	Software	Approach	Type	Language
Python	Pytextrank	[101]	Unsupervised	Multilanguage
	RAKE	[98]	Unsupervised	Multilanguage
	TopicCoRank	[99]	Unsupervised	En/Fr
	Seq2Seq	[91]	Supervised	En
	YAKE	[100]	Unsupervised	Multilanguage
	PKE	[37], [41], [83], [92], [95], [102], [103], [104]	Supervised/ Unsupervised	Multilanguage
Java	KEA	[97]	Supervised	Multilanguage
	Maui	[87]	Supervised	Multilanguage
	CiteTextRank	[92], [101]	Unsupervised	En
	Sequential Labeling	[87], [102], [105], [106]	Supervised	En
C++	KE package	[92], [101]	Unsupervised	En

There are other systems that work according to a combined technique, we mention them, CiteTextRank is a system implemented in Java and uses several techniques including, TfIdf, TextRank, SingleRank and ExpandRank. Also, PKE is a system implemented in Python and works by eight techniques [37], [41], [83], [92], [95], [102], [103], and [104] (c.f; Table 8). KE package is a system implemented in C++,

it works by the techniques that CiteTextRank, but it only processes English texts. The last system is Sequential Labeling, implemented in Java and works by the techniques [87], [102], [105], and [106] (c.f; Table 8).

Indeed, English is the default language for all these systems, but users can use other languages except TopicCoRank which uses English and French, while Seq2seq, CiteTextRank, Sequential Labeling, and KE package only use English. Python and Java remain among the preferred programming languages for developing keyphrase extraction systems. Figure 9 shows the proportion of systems developed in each pro-gramming language.

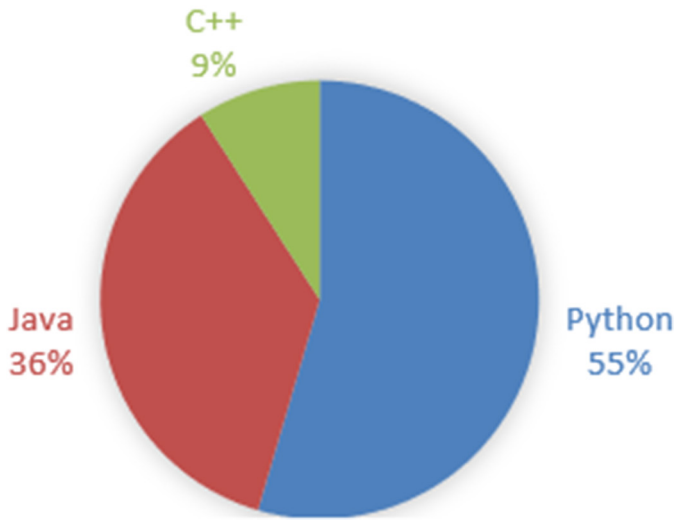

Fig. 9. Percentage of use programming languages in AKE systems

RQ 5 "What obstacles must be overcome to improve the accuracy of keyphrases extraction?" The different methods that we studied tried to improve the performance of extracting phrases from the document. These performances vary according to the language and the document domain, the length and the number of keyphrases extracted, their type, the values of the hyperparameters as well as the availability of data learning used by supervised methods.

The evaluations carried out by the authors of these methods show that no method can extract keyphrases very efficiently. Considering these results, various challenges are revealed at the different stages of the keyphrase extraction: The preprocessing, the functionalities used and the identification of candidate keyphrases. At the preprocessing stage, most methods remove stopwords and use stemming and normalization techniques. These operations are always linked to the writing language of the document. Additionally, the use of individual phrases may refer to a different meaning of the text context, which negatively affects the performance of the keyphrase extraction method. When defining candidate keyphrases, most methods consider these phrases either as a single word or as a multi-word noun phrase and thus exclude any phrases that differ from this structure. In addition, despite the exploitation of resources, such as Wikipedia

and external terminology databases. Most of these methods neglect the relationship between synonyms and phrase abbreviations, resulting in keyphrases that are different in writing but maybe linguistically equivalent. Indeed, future approaches should take into account the understanding of the text.

Most of the current AKE methods select only the keyphrases mentioned in the text, while more than 50% of the keyphrases of a document are not mentioned, which is exemplified by the datasets that are used to evaluate the performance of the AKE. Table 9 shows the percentage of the presence of the keyphrases in 5 most datasets used for AKE evaluation. Thus, care should be taken to generate keyphrases that are not mentioned in the document.

Table 9. Percentage of present and absent keyphrases in datasets

Dataset	Present KP	Absent KP
Inspec	73,58%	26,42%
Krapivin	55,67%	44,33%
NUS	54,64%	45,36%
SemEval2010	44,37%	55,63%
KP20K	62,77%	37,23%

The difficulties that reduce the effectiveness of keyphrase extraction vary from one method to another. The performance of graph-based approaches is affected by changing the frequency of the co-occurrence window. Therefore, the use of semantic information when creating the graph could be an alternative solution to word windows. In addition, most of them create graphs of words and not of phrases, also the score of a phrase which consists of more than one word based on sum or an average score of the words which compose it. And this reflects negatively on the results of these methods, especially for long documents. For statistical methods, the characteristics they adopt only discover the importance of each phrase of the document on the basis of its repetition and coexistence with other phrases, while the semantic relationship between words is neglected. This leads to ignoring rare keyphrases and focusing on repeated words in the document, especially in short texts. In addition, the values of the hyperparameters used also affect the performance of these approaches. The performance of embedding approaches is affected by the quality of the phrase embeddings. In addition, the vector representation of multi-word phrases always depends on the representation of their words, and some specific phrases are difficult to represent, such as biomedical terms, the same for abbreviations.

Indeed, Deep Learning approaches depend on the learning dataset. They require a large dataset to be more efficient. Despite the fact that deep learning reduces the use of features, they require considerable computational time for training. This explains why they are not used by AKE systems. Also, most of these methods formulate keyphrase extraction as a sequence tagging task, not a ranking task. So maybe an important phrase won't be extracted in the first place. Additionally, these methods do not exploit document subjects during the extraction process, which could increase the efficiency of these methods.

5 Discussion

In this section, we will generally discuss the problems encountered in the approaches proposed in the articles studied. In addition, we will present the limits of this SLR, and we suggest recommendations that should be taken into account in the future.

5.1 Discussion of problems

Despite the effort made in recent years by researchers to develop keyphrase extraction. However, through our study of 60 articles published between 2015 and 2022, it is clear that these methods still have not reached the desired level due to certain problems that limit their effectiveness. Most of these methods deal with scientific documents and are therefore difficult to use in other types of texts. Texts lose information during the preprocessing process due to stopword removal, stemming, and normalization. The process of extracting candidate phrases based on choosing which nominal phrases are repeated in the text, which sometimes leads to the exclusion of phrases that may be important, such as biomedical texts in which the repetition of important phrases is low. In addition, most of these techniques do not work well for keyphrases containing synonyms or abbreviations.

As for the use of features, its implementation depends on the approach used. We find that some authors prefer the use of statistical features, while others prefer the use of grammatical, morphological or semantic features. Others have used external features based on certain resources, such as Wikipedia and terminology databases, despite their computational cost. However, most approaches do not use all of the features of the phrase. The above-mentioned issues negatively affect the keyphrase extraction step, which is also faced by other issues that limit its performance. One of the problems that arises during the keyphrase extraction process is that some phrases are included with others. Sometimes a few extracted phrases are different in writing but similar in meaning. Most of the methods still haven't solved the problem of removing keyphrases that rarely appear in the document. Note that most of these methods have been tested in scientific articles so that the textual structure affects the performance of these methods. There are some shortcomings in the evaluation of keyphrase extraction approaches, as the test is limited to comparing the proposed method with traditional approaches by applying it to limited types of data. In addition, the evaluation results change depending on the number of keyphrases extracted and the length and nature of the text. Our study also shows that no method is applied to all types of data to guarantee its effectiveness. Often the proposed method is designed according to the target application and the dataset to be processed.

5.2 Recommendations

The discussion of the problems encountered in the different approaches of AKE gave us an idea of the directions that researchers should work in the future in order to improve the process of AKE. Inevitably, if the interest in improving preprocessing is increased, and the POS and N-Gram modules are merged in the process of extracting

candidate keyphrases, we may have more suitable candidate keyphrases. The keyphrase extraction process is mainly applied to candidate keyphrases by exploiting the features of the latter, which are mainly either statistical, linguistic, structural, or semantic. These features, especially semantics, can be exploited to overcome redundancy issues. The features of the phrase should not be treated independently of the text. Rather, the sentence should be treated through its meaning in the text and not in its general sense. Moreover, despite their computational costs, the use of dictionaries and external corpora will help AKE methods to highlight keyphrases that are rare to appear in the text. Improving the training process can improve the performance of supervised methods, especially if one has a complete and available body of data. Embedding keyphrases and contextual information into deep learning models is showing encouraging results. Therefore, researchers should develop models of deep learning, with a view to using them, especially in large documents. And because these approaches have more ability than others to generate keyphrases that are not mentioned in the document.

Unsupervised methods remain preferred by researchers, 77% of the articles studied in our SLR propose unsupervised approaches. As they are easy to operate in various areas of NLP, they also do not require any prior domain knowledge or data training. But its performance, especially on large documents, remains poor. The exploitation of certain characteristics of the document, such as knowledge of its domain and its topic, can help these approaches to partially overcome the scarcity problem. Therefore, research should focus on this direction to determine keyphrases regardless of their presence or absence in the document. Also, the use of embedding techniques has allowed these methods to process sentences according to their concept in the text, these techniques also remain among the research directions, especially in long texts.

5.3 Limitations of this SLR

Although five digital libraries are used as research resources, which gave us access to a large collection of related scientific articles. However, our study is not exhaustive and does not cover all work related to keyphrase extraction, due to the methodology adopted in the study, which is based on articles published in English. Therefore, all studies published in other languages are ignored. In addition, our study focused on work published from January 2015 to May 2022. Work published after May 2022 was not studied and may be included in future work. In addition, due to the time factor, the evaluation of the proposed methods not having been verified, we limited ourselves to the results published by the authors.

6 Conclusion and future work

In this SLR, we have adopted a comprehensive scientific methodology to understand the research direction related to the extraction of keyphrases from the text, to study the problems of their extraction from the solutions proposed in 60 research articles, published between 2015 and 2022. This work included a series of steps related to a broad search strategy from the choice of keyphrases for the search, through the adoption of a set of inclusion and exclusion criteria. Our study included five research questions

that allowed us to know, the techniques used to remove unnecessary phrases, the algorithms, and techniques used in AKE approaches, the most important evaluation metrics, and datasets to verify the performance of the proposed solutions. Identify the most important AKE systems and their implementation languages, then identify the elements that significantly affect the efficiency and performance of AKE systems. Therefore, this study is useful not only to guide researchers and practitioners but can also be used as a support for the development of new, more precise AKE systems. In future work, we aim to develop the steps of the keyphrase extraction process, in particular the preprocessing and selection of candidate keyphrases. We will find a new method of transforming the text into a graph that does not depend on the window of the words. We will create a dataset composed of documents written in the Arabic language in order to evaluate AKE approaches in Arabic texts, and We will propose an AKE approach that not only extracts key phrases mentioned in the document, but can also generate key phrases that are not mentioned in the document.

7 References

[1] A. Aristovnik, D. Ravšelj, and L. Umek. "A bibliometric analysis of COVID-19 across science and social science research landscape." Sustainability 12.21 (2020): 9132. https://doi.org/10.3390/su12219132

[2] E. Papagiannopoulou, and G. Tsoumakas. "A review of keyphrase extraction." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10.2 (2020): e1339. https://doi.org/10.1002/widm.1339

[3] Z. A. Merrouni, B. Frikh, and Brahim Ouhbi. "Automatic keyphrase extraction: A survey and trends." Journal of Intelligent Information Systems 54.2 (2020): 391–424. https://doi.org/10.1007/s10844-019-00558-9

[4] Z. Nasar, S. W. Jaffry, and M. K. Malik. "Textual keyword extraction and summarization: State-of-the-art." Information Processing & Management 56.6 (2019): 102088. https://doi.org/10.1016/j.ipm.2019.102088

[5] T. Sabri, O. El Beggar, and M. Kissi. "Comparative study of Arabic text classification using feature vectorization methods." Procedia Computer Science 198 (2022): 269–275. https://doi.org/10.1016/j.procs.2021.12.239

[6] L. Abualigah, et al. "Efficient text document clustering approach using multi-search Arithmetic Optimization Algorithm." Knowledge-Based Systems 248 (2022): 108833. https://doi.org/10.1016/j.knosys.2022.108833

[7] R. Srivastava, P. Singh, K. P. S. Rana, & V. Kumar. "A topic modeled unsupervised approach to single document extractive text summarization." Knowledge-Based Systems 246 (2022): 108636. https://doi.org/10.1016/j.knosys.2022.108636

[8] A. Erdmann, A. Ramón, and J. M. Ponzoa. "Search engine optimization: The long-term strategy of keyword choice." Journal of Business Research 144 (2022): 650–662. https://doi.org/10.1016/j.jbusres.2022.01.065

[9] G. Rabby, S. Azad, M. Mahmud, K. Z. Zamli, & M. Rahman. "Teket: A tree-based unsupervised keyphrase extraction technique." Cognitive Computation 12.4 (2020): 811–833. https://doi.org/10.1007/s12559-019-09706-3

[10] D. Pramod, and B. Prafulla. "Conversational recommender systems techniques, tools, acceptance, and adoption: A state of the art review." *Expert Systems with Applications* (2022): 117539. https://doi.org/10.1016/j.eswa.2022.117539

[11] S. Kumar, and R. Kumar. "A study on different aspects of web mining and research issues." *IOP Conference Series: Materials Science and Engineering*. Vol. 1022. No. 1. IOP Publishing, 2021. https://doi.org/10.1088/1757-899X/1022/1/012018

[12] J. Ali, A. Jusoh, N. Idris, A. F. Abbas, & A. H. Alsharif. "Everything is going electronic, so do services and service quality: bibliometric analysis of E-services and E-service quality." *International Journal of Interactive Mobile Technologies* 15.18 (2021): 149. https://doi.org/10.3991/ijim.v15i18.24519

[13] U. Jha, L. Tyagi, D. Kansal, S. Chakraborty, & A. Singhal "A review of sentiment analysis techniques using soft computing approaches." *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2021. https://doi.org/10.1109/Confluence51648.2021.9377031

[14] M. Haddoud, A. Mokhtari, T. Lecroq, & S. Abdeddaïm. "Accurate keyphrase extraction from scientific papers by mining linguistic information." *CLBib@ ISSI*. 2015.

[15] L. Ajallouda, O. Hourrane, A. Zellou, & E. H. Benlahmar. "Toward a new process for candidate key-phrases extraction." *International Conference on Digital Technologies and Applications*. Springer, Cham, 2022. https://doi.org/10.1007/978-3-031-02447-4_48

[16] C. Lioma, and C. K. van Rijsbergen. "Part of speech n-grams and information retrieval." *Revue française de linguistique appliquée* 13.1 (2008): 9–22. https://doi.org/10.3917/rfla.131.0009

[17] A. Handler, M. Denny, H. Wallach, & B. O'Connor. "Bag of what? simple noun phrase extraction for text analysis." *Proceedings of the First Workshop on NLP and Computational Social Science*. 2016. https://doi.org/10.18653/v1/W16-5615

[18] N. Firoozeh, A. Nazarenko, F. Alizon, & B. Daille, "Keyword extraction: Issues and methods." *Natural Language Engineering* 26.3 (2020): 259–291. https://doi.org/10.1017/S1351324919000457

[19] J. Biolchini, P. G. Mian, A. C. C. Natali, & G. H. Travassos. "Systematic review in software engineering." *System engineering and computer science department COPPE/UFRJ, Technical Report ES* 679.05 (2005): 45.

[20] J. Renaud, V. Martin, and P. Dagenais. Les normes de production des revues systématiques: Guide méthodologique. Institut national d'excellence en santé et en services sociaux (INESSS). (2013).

[21] B. Kitchenham, Barbara, and P. Brereton. "A systematic review of systematic review process research in software engineering." *Information and software technology* 55.12 (2013): 2049–2075. https://doi.org/10.1016/j.infsof.2013.07.010

[22] I. Mustapha, N. Khan, M. I. Qureshi, A. A. Harasis, & N. T. Van. "Impact of Industry 4.0 on healthcare: A systematic literature review (SLR) from the last decade." *International Journal of Interactive Mobile Technologies* 15.18 (2021). https://doi.org/10.3991/ijim.v15i18.25531

[23] S. S. A. Shah Kazmi, M. Hassan, S. A. Khawaj, & S. F. Padlee. "The use of AR technology to overcome online shopping phobia." *International Journal of Interactive Mobile Technologies* 15.5 (2021). https://doi.org/10.3991/ijim.v15i05.21043

[24] S. Popova, Svetlana, and V. Danilova. "Keyphrase extraction abstracts instead of full papers." *2014 25th International Workshop on Database and Expert Systems Applications*. IEEE, 2014. https://doi.org/10.1109/DEXA.2014.57

[25] S. Danesh, T. Sumner, and J. H. Martin. "Sgrank: Combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction." *Proceedings of the fourth joint conference on lexical and computational semantics*. 2015. https://doi.org/10.18653/v1/S15-1013

[26] H. Jia, and E. Saule. "Addressing overgeneration error: An effective and effcient approach to keyphrase extraction from scientific papers." *BIRNDL@ SIGIR*. 2018.

[27] Q. Liu, D. Kawahara, and S. Li. "Scientific Keyphrase extraction: extracting candidates with semi-supervised data augmentation." *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, Cham, 2018. 183–194. https://doi.org/10.1007/978-3-030-01716-3_16

[28] M. Barreiro-Guerrero, A. Simón-Cuevas, Y. Pérez-Guadarrama, F. P. Romero, & J. A. Olivas. "Applying OWA operator in the semantic processing for automatic keyphrase extraction." *Iberoamerican Congress on Pattern Recognition*. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-33904-3_6

[29] J. R. Asl, and J. M. Banda. "GLEAKE: Global and local embedding automatic keyphrase extraction." *arXiv preprint arXiv:2005.09740* (2020).

[30] N. Giamblanco, and P. Siddavaatam. "Keyword and keyphrase extraction using newton's law of universal gravitation." *2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE)*. IEEE, 2017 https://doi.org/10.1109/CCECE.2017.7946724

[31] G. Rabby, S. Azad, M. Mahmud, K. Z. Zamli, & M. M. Rahman. "A flexible keyphrase extraction technique for academic literature." *Procedia Computer Science* 135 (2018): 553–563. https://doi.org/10.1016/j.procs.2018.08.208

[32] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, & A. Jatowt, A. «YAKE! Keyword extraction from single documents using multiple local features." *Information Sciences* 509 (2020): 257–289. https://doi.org/10.1016/j.ins.2019.09.013

[33] M. Aman, S. J. Abdulkadir, I. A. Aziz, H. Alhussian, & I. Ullah. "KP-Rank: a semantic-based unsupervised approach for keyphrase extraction from text data." *Multimedia Tools and Applications* 80.8 (2021): 12469–12506. https://doi.org/10.1007/s11042-020-10215-x

[34] Z. A. Merrouni, F. Bouchra, and O. Brahim. "HAKE: An unsupervised approach to automatic keyphrase extraction for multiple domains." *Cognitive Computation* (2022): 1–23. https://doi.org/10.1007/s12559-021-09979-7

[35] M. B. A. Miah, S. Awang, M. S. Azad, & M. M. Rahman. "Keyphrases concentrated area identification from academic articles as feature of keyphrase extraction: A new unsupervised approach." *International Journal of Advanced Computer Science and Applications* 13.1 (2022). https://doi.org/10.14569/IJACSA.2022.0130192

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

[37] C. Florescu, and C. Caragea. "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017. https://doi.org/10.18653/v1/P17-1102

[38] G. Figueroa, P. Chen, and Y. Chen. "RankUp: Enhancing graph-based keyphrase extraction methods with error-feedback propagation." *Computer Speech & Language* 47 (2018): 112–131. https://doi.org/10.1016/j.csl.2017.07.004

[39] W. Chen, Z. Liu, W. Shi, & J. X. Yu. "Keyphrase extraction based on optimized random walks on multiple word relations." *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, Cham, 2018. https://doi.org/10.1007/978-3-319-96893-3_27

[40] Y. Perez-Guadarrama, A. Simón-Cuevas, W. Hojas-Mazo, J. A. Olivas, & F. P. Romero. "A fuzzy approach to improve an unsupervised automatic keyphrase extraction process." *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2018. https://doi.org/10.1109/FUZZ-IEEE.2018.8491487

[41] F. Boudin. "Unsupervised keyphrase extraction with multipartite graphs." *arXiv preprint arXiv:1803.08721* (2018). https://doi.org/10.18653/v1/N18-2105

[42] Z. Sun, J. Tang, P. Du, Z. H. Deng, & J. Y. Nie. "Divgraphpointer: A graph pointer network for extracting diverse keyphrases." *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019. https://doi.org/10.1145/3331184.3331219

[43] T. F. Li, L. Hu, J. F. Chu, H. T. Li, & L. Chi. "An unsupervised approach for keyphrase extraction using within-collection resources," *IEEE Access* 7 (2019): 126088–126097 https://doi.org/10.1109/ACCESS.2019.2938213

[44] A. Prasad, and M. Y. Kan. "Glocal: Incorporating global information in local convolution for keyphrase extraction." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019.

[45] M. Tosi, D. Lucca, and J. Cesar dos Reis. "C-rank: a concept linking approach to unsupervised keyphrase extraction." *Research Conference on Metadata and Semantics Research*. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-36599-8_21

[46] H. Dong, J. Wan, and Z. Wan. "Keyphrase Extraction Based on Multi-Feature." *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. IEEE, 2019. https://doi.org/10.1109/MLBDBI48998.2019.00047

[47] H. Yeom, Y. Ko, and J. Seo. "Unsupervised-learning-based keyphrase extraction from a single document by the effective combination of the graph-based model and the modified C-value method." *Computer Speech & Language* 58 (2019): 304–318. https://doi.org/10.1016/j.csl.2019.04.008

[48] T. Chen, D. Miao, and Y. Zhang. "A graph-based keyphrase extraction model with three-way decision." *International Joint Conference on Rough Sets*. Springer, Cham, 2020. https://doi.org/10.1007/978-3-030-52705-1_8

[49] L. Luo, L. Zhang, and H. Peng. "An unsupervised keyphrase extraction model by incorporating structural and semantic information." *Progress in Artificial Intelligence* 9.1 (2020): 77–83. https://doi.org/10.1007/s13748-019-00200-3

[50] Y. Yeon Sung, and S. B. Kim. "Topical keyphrase extraction with hierarchical semantic networks." *Decision Support Systems* 128 (2020): 113163. https://doi.org/10.1016/j.dss.2019.113163

[51] M. N. Awan, and M. O. Beg. "Top-rank: a topicalpostionrank for extraction and classification of keyphrases in text." *Computer Speech & Language* 65 (2021): 101116. https://doi.org/10.1016/j.csl.2020.101116

[52] P. Yang, Y. Ge, Y. Yao, & Y. Yang. "GCN-based document representation for keyphrase generation enhanced by maximizing mutual information." *Knowledge-Based Systems* 243 (2022): 108488. https://doi.org/10.1016/j.knosys.2022.108488

[53] V. Venktesh, M. Mohania, and V. Goyal. "Topic aware contextualized embeddings for high quality phrase extraction." *European Conference on Information Retrieval*. Springer, Cham, 2022. https://doi.org/10.1007/978-3-030-99736-6_31

[54] L. Ajallouda, K. Najmani, A. Zellou, & E. L. Benlahmar. "Doc2Vec, SBERT, InferSent, and USE Which embedding technique for noun phrases?" *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*. IEEE, 2022. https://doi.org/10.1109/IRASET52964.2022.9738300

[55] P. Zeng, P., Q. Tan, Y. Yan, Q. Xie, J. Xu, & W. Cao. "Automatic keyword extraction using word embedding and clustering." *2017 International Conference on Computer Systems, Electronics and Control (ICCSEC)*. IEEE, 2017. https://doi.org/10.1109/ICCSEC.2017.8447033

[56] K. Bennani-Smires, C. Musat, A. Hossmann, M. Baeriswyl, & M. Jaggi. "Simple unsupervised keyphrase extraction using sentence embeddings." *arXiv preprint arXiv:1801.04470* (2018). https://doi.org/10.18653/v1/K18-1022

[57] E. Papagiannopoulou, and G. Tsoumakas. "Unsupervised keyphrase extraction based on outlier detection." *arXiv preprint arXiv: 1808.03712* (2018).
[58] D. Mahata, J. Kuriakose, R. Shah, & R. Zimmermann. "Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 2018. https://doi.org/10.18653/v1/N18-2100
[59] A. Toleu, G. Tolegen, and R. Mussabayev. "Keyvector: Unsupervised keyphrase extraction using weighted topic via semantic relatedness." *Computación y Sistemas* 23.3 (2019): 861–869. https://doi.org/10.13053/cys-23-3-3264
[60] T. Mikolov, K. Chen, G. Corrado, & J. Dean. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
[61] Q. Le, and T. Mikolov. "Distributed representations of sentences and documents." *International conference on machine learning*. PMLR, 2014.
[62] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. https://doi.org/10.3115/v1/D14-1162
[63] M. Pagliardini, P. Gupta, and M. Jaggi. "Unsupervised learning of sentence embeddings using compositional n-gram features." *arXiv preprint arXiv:1703.02507* (2017). https://doi.org/10.18653/v1/N18-1049
[64] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, & A. Bordes. "Supervised learning of universal sentence representations from natural language inference data." *arXiv preprint arXiv:1705.02364* (2017). https://doi.org/10.18653/v1/D17-1070
[65] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, & L. Zettlemoyer. Deep contextualized word representations. arXiv 2018. arXiv preprint arXiv:1802.05365, 12. https://doi.org/10.18653/v1/N18-1202
[66] J. Devlin, M. W. Chang, K. Lee, & K. Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
[67] D. Cer, et al. "Universal sentence encoder." *arXiv preprint arXiv:1803.11175* (2018).
[68] W. Fan, H. Liu, S. Wang, Y. Zhang, & Y. Chang. "Extracting keyphrases from research papers using word embeddings." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Cham, 2019. https://doi.org/10.1007/978-3-030-16142-2_5
[69] Y. Sun, H. Qiu, Y. Zheng, Z. Wang, & C. Zhang. "SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model." *IEEE Access* 8 (2020): 10896–10906. https://doi.org/10.1109/ACCESS.2020.2965087
[70] S. Arora, Y. Liang, and T. Ma. "A simple but tough-to-beat baseline for sentence embeddings." *International conference on learning representations*. 2017.
[71] L. Ajallouda, F. Z. Fagroud, A. Zellou, & E. H. Benlahmar. KP-USE: An Unsupervised Approach for Key-Phrases Extraction from Documents. (IJACSA) International Journal of Advanced Computer Science and Applications, 13(4). (2022). https://doi.org/10.14569/IJACSA.2022.0130433
[72] H. Wang, and J. Li. "Unsupervised Keyphrase Extraction from Single Document Based on Bert." *2022 International Seminar on Computer Science and Engineering Technology (SCSET)*. IEEE, 2022. https://doi.org/10.1109/SCSET55041.2022.00068
[73] F. C. Jonathan, and O. Karnalim. "Semi-supervised keyphrase extraction on scientific article using fact-based sentiment." *Telkomnika* 16.4 (2018): 1771–1778. https://doi.org/10.12928/telkomnika.v16i4.5473
[74] M. Helmy, R. M. Vigneshram, G. Serra, & C. Tasso. "Applying deep learning for Arabic keyphrase extraction." *Procedia Computer Science* 142 (2018): 254–261. https://doi.org/10.1016/j.procs.2018.10.486

[75] R. Alzaidy, C. Caragea, and C. L. Giles. "Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents." *The world wide web conference*. 2019. https://doi.org/10.1145/3308558.3313642

[76] K. Patel, and C Caragea. "Exploring word embeddings in crf-based keyphrase extraction from research papers." *Proceedings of the 10th International Conference on Knowledge Capture*. 2019. https://doi.org/10.1145/3360901.3364447

[77] D. Sahrawat, et al. "Keyphrase extraction from scholarly articles as sequence labeling using contextualized embeddings." *arXiv preprint arXiv:1910.08840* (2019).

[78] L. Xiong, C. Hu, C. Xiong, D. Campos, & A. Overwijk. "Open domain web keyphrase extraction beyond language modeling." *arXiv preprint arXiv:1911.02671* (2019). https://doi.org/10.18653/v1/D19-1521

[79] X. Zhu, C. Lyu, D. Ji, H. Liao, & F. Li. "Deep neural model with self-training for scientific keyphrase extraction." *Plos one* 15.5 (2020): e0232547. https://doi.org/10.1371/journal.pone.0232547

[80] T. Zhou, Y. Zhang, and H. Zhu. "Multi-level memory network with crfs for keyphrase extraction." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Cham, 2020. https://doi.org/10.1007/978-3-030-47426-3_56

[81] Y. Uzun. "Keyword extraction using naive bayes." *Bilkent University, Department of Computer Science, Turkey,* 2005. www.cs.bilkent.edu.tr/~guvenir/courses/CS550/Workshop/Yasin_Uzun.pdf

[82] H. Wu, B. Ma, W. Liu, T. Chen, & Nie, D. "Fast and constrained absent keyphrase generation by prompt-based learning." (2022). https://doi.org/10.1609/aaai.v36i10.21402

[83] S. R. El-Beltagy, & A. Rafea. "KP-Miner: A keyphrase extraction system for English and Arabic documents." *Information Systems* 34.1 (2009): 132–144. https://doi.org/10.1016/j.is.2008.05.002

[84] T. D. Nguyen, and M. Y. Kan. "Keyphrase extraction in scientific publications." *International conference on Asian digital libraries*. Springer, Berlin, Heidelberg, 2007.

[85] M. Krapivin, A. Autaeu, and M. Marchese. "Large dataset for keyphrases extraction." (2009).

[86] A. T. Schutz. "Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods." *M. App. Sc Thesis* (2008).

[87] O. Medelyan, E. Frank, and L. H. Witten. "Human-competitive tagging using automatic keyphrase extraction." *Association for Computational Linguistics*, 2009. https://doi.org/10.3115/1699648.1699678

[88] S. N. Kim, O. Medelyan, M. Y. Kan, T. Baldwin, & L. P. Pingar. "SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific."

[89] A. Hulth. "Improved automatic keyword extraction given more linguistic knowledge." *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 2003. https://doi.org/10.3115/1119355.1119383

[90] S. D. Gollapalli, and C. Caragea. "Extracting keyphrases from research papers using citation networks." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 28. No. 1. 2014. https://doi.org/10.1609/aaai.v28i1.8946

[91] R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky, & Y. Chi. "Deep keyphrase generation." *arXiv preprint arXiv:1704.06879* (2017). https://doi.org/10.18653/v1/P17-1054

[92] X. Wan, and J. Xiao. "Single document keyphrase extraction using neighborhood knowledge." *AAAI*. Vol. 8. 2008.

[93] L. Marujo, M. Viveiros, and J. Paulo da S. Neto. "Keyphrase cloud generation of broadcast news." *arXiv preprint arXiv:1306.4606* (2013).

[94] L. Marujo, A. Gershman, J. Carbonell, R. Frederking, & J. P. Neto. "Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization." *arXiv preprint arXiv:1306.4886* (2013).
[95] A. Bougouin, F. Boudin, and B. Daille. "Topicrank: Graph-based topic ranking for keyphrase extraction." *International joint conference on natural language processing (IJCNLP)*. 2013.
[96] Y. Gallina, F. Boudin, and B. Daille. "KPTimes: A large-scale dataset for keyphrase generation on news documents." *arXiv preprint arXiv:1911.12559* (2019). https://doi.org/10.18653/v1/W19-8617
[97] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, C. G. NevillManning. Domain-specific keyphrase extraction. Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, pp. 668–673. San Francisco, CA, USA. (1999).
[98] S. Rose, D. Engel, N. Cramer, & W. Cowley. "Automatic keyword extraction from individual documents." *Text Mining: Applications and Theory* 1 (2010): 1–20. https://doi.org/10.1002/9780470689646.ch1
[99] A. Bougouin, F. Boudin, and B. Daille. "Keyphrase annotation with graph co-ranking." *arXiv preprint arXiv:1611.02007* (2016).
[100] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, & A. Jatowt. "Yake! collection-independent automatic keyword extractor." *European Conference on Information Retrieval*. Springer, Cham, 2018. https://doi.org/10.1007/978-3-319-76941-7_80
[101] R. Mihalcea, and TP. Tarau. "Textrank: Bringing order into text." *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.
[102] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, & C. G. Nevill-Manning. "Kea: Practical automated keyphrase extraction." *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*. IGI global, 2005. 129–152. https://doi.org/10.4018/978-1-59140-441-5.ch008
[103] T. D. Nguyen, and M. T. Luong. "WINGNUS: Keyphrase extraction utilizing document logical structure." *Proceedings of the 5th international workshop on semantic evaluation*. 2010.
[104] L. Sterckx, T. Demeester, J. Deleu, & C. Develder. "Topical word importance for fast keyphrase extraction." *Proceedings of the 24th International Conference on World Wide Web*. 2015. https://doi.org/10.1145/2740908.2742730
[105] C. Caragea, F. Bulgarov, A. Godea, & S. D. Gollapalli. "Citation-enhanced keyphrase extraction from research papers: A supervised approach." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. https://doi.org/10.3115/v1/D14-1150
[106] S. Gollapalli, D. X. Li, and P. Yang. "Incorporating expert knowledge into keyphrase extraction." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 31. No. 1. 2017. https://doi.org/10.1609/aaai.v31i1.10986

8 Authors

Lahbib Ajallouda is PhD Student at Computer Science and Systems Analysis School (ENSIAS), Mohamed V University, Rabat, Morocco. His research interests are primarily in the area of internet of things, search engines, cloud computing, and machine learning, where he is the author/co-author of over 6 research publications. He can be contacted at email: lahbib_ajallouda@um5.ac.ma.

Fatima Zahra Fagroud is PhD Student at Faculty of Sciences Ben M'sick, Hassan II University of Casablanca, Morocco. Her research interests are primarily in the area of internet of things, search engines, cloud computing, machine learning, where she is the author/co-author of over 14 research publications. She can be contacted at email: fagroudfatimazahra0512@gmail.com.

Ahmed Zellou received his Ph.D. in Applied Sciences at the Mohammedia School of Engineers, Mohammed V University, Rabat, Morocco 2008. He is currently a coordinator of the IWIM Web Engineering & Mobile Computing branch at ENSIAS Mohamed V university in Rabat, Morocco. His research interests include Parallel Computing, Information Systems (Business Informatics), and Distributed Computing, where he is the author/co-author of over 72 research publications. He can be contacted by email: ahmed.zellou@um5.ac.ma.

EL Habib Benlahmar received his PhD, computer science at Computer Science and Systems Analysis School (ENSIAS), University Mohamed V, Rabat, Morocco. He is currently a coordinator of the master data science & Big Data at FSBM Hassane II University Casablanca Morocco. His research interests include Educational Technology, Software Engineering, Information Systems (Business Informatics), and Human-computer Interaction, where he is the author/co-author of over 165 research publications. He can be contacted by email: h.benlahmer@gmail.com.

Article submitted 2022-06-07. Resubmitted 2022-07-06. Final acceptance 2022-07-06. Final version published as submitted by the authors.