International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Comparative Analysis on Automatic Keyphrase Extraction (AKPE) Techniques

Dr. R.Mangai Begum, J.S. Baruni

[1] *Department of Information Technology, St.Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India*

[2] *Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India*

[1]jskmangai@gmail.com,[2]baruni.js@gmail.com

***Abstract*— Extracting Keyphrase from the large corpus manually is a tedious task. To overcome this challenging task Automatic Keyphrase Extraction (AKPE) techniques have been used for Keyphrase extraction. Keyphrase extraction is the task of automatically selecting a small set of phrases that best describe a given free text document. Keyphrase constitutes a succinct conceptual summary of a document, which is very useful in digital information management systems for semantic indexing, faceted search, document clustering, and classification. This paper contains a brief description of automatic Keyphrase extraction techniques which consist of various algorithms and also a comparative study on automotive Keyphrase extraction techniques such as KEA, TF-IFD, RAKE, TextRank.**

***Keywords*— Keyphrase Extraction, Techniques, Comparison, Keyphrase Extraction Algorithm, Term Frequency- Inverse Frequency Document, Rapid Automatic Keyphrase Extraction, TextRank.**

I. INTRODUCTION

Nowadays information is one of the most powerful and important weapons in the modern world. Every moment we are getting an increasingly large amount of data or information from various sources like emails, web pages, electronic documents, etc. [4]. But all the sources cannot fulfill the user's expectation of the readers since it is more difficult to find the appropriate information from a huge amount of document [5]. It is very much difficult for a human being to find out the summary or extracting the main topics from a large body of text for the very fast- growing information [13]. Keyphrase extraction regarded as the crucial method for data analysis. The primary mission of important Keyphrase extraction is to extract a specific group of phrases or words which highlight the main content of the documents. Automatic clustering, automatic filtering, automatic indexing, automatic summarization, information visualization, topic detection, and tracking are the basic data mining applications related to Keyphrase extractions. Automatic Keyphrase extraction plays an important role in many applications of natural

Language Processing (NLP)[15].Automatic keyword extraction provides an efficient and effective way to summarize text from a large corpus[10]. Document Keyphrase has enabled fast and accurate searching for a given document from a large text collection, and have exhibited their potential in improving many natural language processing (NLP) and information retrievals (IR) tasks, such as text summarization, text categorization, opinion mining, and document indexing. The keywords are single words, while Keyphrase are made up of a few words. Keyphrase are single-token or multi-token expressions that provide the essential information of a sentence or document. Many approaches to Keyphrase extraction generally used only the textual content of a target document to extract Keyphrase [11]. Automatic Keyphrase extraction concerns the automatic selection of important and topical phrases from the body of a document.

The purpose of Keyphrase and keyword extraction was done along with the brief description of different feature selection metrics generally used to rank the candidate keywords and key phrases according to their importance in the analyzed text. This paper is dividing into the following sections: Section II explains background work and section III discusses different techniques of Automatic Keyphrase extraction, section IV shows a comparative study on automatic Keyphrase extraction (AKPE), and section V provides the conclusion of this paper.

II. BACKGROUND WORK

Keyphrase extraction concerns the selection of representative and characteristic phrases from a document that express all aspects related to the document's content [2]. Accurate extraction of key phrases is particularly important in the (academic) publishing industry for carrying out several important tasks, such as the recommendations of new articles or books to customers and the analysis of content trends. The versatility of Keyphrase renders Keyphrase extraction a very important document processing task [13] .Keyphrase can be used for semantically indexing a collection of documents either in place of their full-text or in addition to it, enabling semantic and faceted search[9]. Keyphrase extraction can be used for query expansion in the context of pseudo-relevance feedback [8]. Keyphrase extraction can also serve as features for document clustering and classification. Furthermore, the set of extracted Keyphrase can be viewed as an extreme summary of the corresponding document for human inspection, while the individual Keyphrase can guide the

extraction of sentences in automatic document summarization systems. A Keyphrase extraction system classified into two procedures which extracts a list of words/phrases that serve as candidate Keyphrase using some heuristics and determining which of these candidates Keyphrase are correct Keyphrase using supervised or unsupervised approaches [1]. A Keyphrase extraction method involves selecting phrases and sentences from the large corpus and including them in the final summary [9]. Keyphrase Extraction consists of some common techniques such as KEA, TF-IDF, RAKE, and TextRank to preprocess and vectorize free text and also evaluates its effectiveness by running them. KEA, an algorithm used for automatically extracting Keyphrase from the text which identifies candidate Keyphrase using lexical methods [18]. TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents [12]. RAKE takes a simple set of input parameters and automatically extracts keywords in a single pass, making it suitable for a wide range of documents and collections. TextRank succeeds in identifying the most important sentences in a text-based on information exclusively drawn from the text. The sentences with the highest rank are selected for inclusion in the abstract [3].

III. KEYPHRASE EXTRACTION TECHNIQUES

Keyphrase extraction is concerned with automatically extracting a set of representative phrases from a document that concisely summarizes its content. By using Keyphrase extraction, the user can avoid all the hassle of sorting through their data manually to pull out key information [2]. The automatic Keyphrase extraction consists of various algorithms for extracting Keyphrase from the large corpus such as the KEA algorithm, TF-IDF algorithm, RAKE algorithm, and TextRank algorithm.

A. KEA Algorithm

KEA is a platform-independent which is used for automatically extracting Keyphrase from text documents. KEA can also use for free indexing with a controlled vocabulary and automatic tagging [18]. KEA consists of two stages as follow

- Training the document
- Extract candidate phrases.

Both stages choose a set of candidate phrases from their input documents and then calculate the values of certain attributes for each candidate. First input documents for training are filtered to regularize the text and determine initial phrase boundaries. The input stream is split into tokens such as sequences of letters, digits, and internal periods. Then punctuation marks, brackets, and numbers are replaced by phrase boundaries.

After cleaning the input document candidate phrases in the document are limited to a certain maximum length without any stopwords. The next stage in identifying the candidate phrase is the stemming process which removes suffix in all the words.

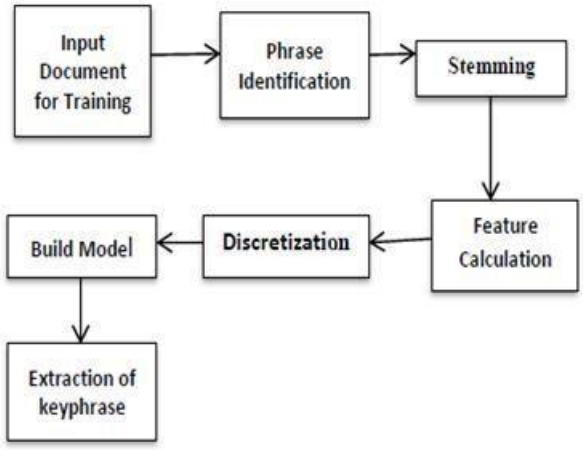

Figure 1. The workflow of KEA Algorithm

When candidate phrases are identified, the two features are calculated for and used in training and extraction. Both features are real numbers so the discretization table for each feature is derived from the training data. KEA uses the Naïve Bayes technique because it is simple and gives the best results. KEA determines candidate phrases and feature values and then applies the model built during training that extracts new phrases from the document. For example, digital libraries, or any depositories of data also use Keyphrase to organize and provide relative access to their data.

B. TF-IDF Algorithm

TF-IDF(Term Frequency Inverse Document Frequency) score finds the words that have the highest ratio of occurring in the current document rather than the frequency occurring in the larger set of documents. The TF-IDF weight [12] is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus [17]. Variations of the TF-IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

TF-IDF for a word in a document is calculated by multiplying two different metrics:

- The term frequency of a word in a document
- The inverse document frequency of the word

The Term Frequency (TF) measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more time in long documents than shorter ones. Thus, the term frequency is often divided by document length as a way of normalization [11]. The Inverse Document Frequency (IDF) for the word acrossa set of documents which means, how common or rare a

word is in the entire document set. The paper length calculates a total number of occurrences of words after separating words ina given abstract using white spaces as a delimiter. When more common a word is indicated then it is closer to 0. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain words [13], and calculating the logarithm. So, if the word is very common and appears in many documents, this number will approach 0. Otherwise, it will approach 1. Multiplying these two numbers results in the TF- IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document.

C. RAKE Algorithm

Rapid Automatic Keyword Extraction (RAKE) is a well-known Keyphrase extraction method that uses a list of stop words and phrase delimiters to detect the most relevant words or phrases in a piece of text[15]. RAKE's simplicity and efficiency enable its use in many applications where keywords can be leveraged. Based on the variety and volume of existing collections and the rate at which documents are created and collected, RAKE provides advantages and frees computing resources for other analytic methods. Rapid Automatic Keyphrase Extraction algorithm is a domain-independent Keyphrase that tries to Determine Keyphrase.

In a body analyzing the frequency of word and its co-occurrence other words in the text. The first thing the method does is splitting the text into a list of words and removes stop words from that list. This returns a list of what is known as content words [19]. The RAKE algorithm finds strings of words that do not include phrase delimiters or stop words. This produces the list of candidate keywords or phrases. Then a co-occurrence graph is built to identify the frequency that words are associated together in those phrases. A score is calculated for each phrase that is the sum of the individual word's scores from the co-occurrence graph. An individual word score is calculated as the degree of a word divided by its frequency, which weights towards longer phrases. Adjoining keywords are included if they occur more than twice in the document and score high enough. An adjoining keyword is two keyword phrases with a stop word between them. The top T keywords are then extracted from the content, where T is 1/3rd of the number of words in the graph [20].

D. TextRank Algorithm

TextRank is a graph-based ranking model for text processing in an unsupervised framework [14]which can be used in order to find the most relevant sentences in the text and also to find keywords. The TextRank algorithm is inspired by [8] PageRank algorithm which was used by Google to rank the websites. The workflow for the TextRank algorithm as shown below in figure 2

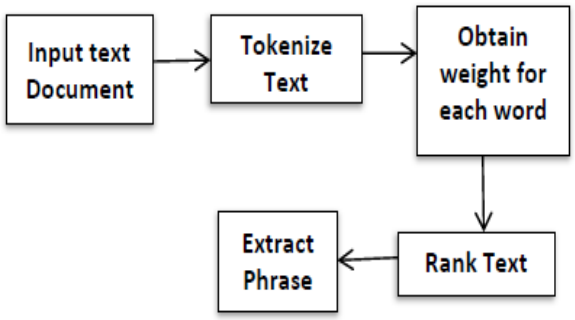

Figure 2. The workflow of TextRank Algorithm

TextRank algorithm [16] creates a graph of the words and relationships between them from a document, then identifies the most important vertices of the graph (words) based on importance scores calculated recursively from the entire graph.

Candidates are extracted from the text through the sentence and then word parsing to produce a list of words to be evaluated. The words are annotated with part of speech tags to better differentiate syntactic use. Each word is then added to the graph and relationships are added between the word and others in a sliding window around the word. A ranking algorithm is run on each vertex for several iterations, updating all of the word scores based on the related word scores, until the scores stabilize the research paper [7].

The words are sorted and the top N is kept (N is typically 1/3rd of the words). A post-processing step loops back through the initial candidate list and identifies words that appear next to one another and merges the two entries from the scored results into a single multi-word entry [20].

IV. COMPARISON OF AUTOMATIC KEYPHRASE EXTRACTION TECHNIQUES

AKPE Techniques	Approaches and Models	Strength	Applied Areas
KEA	- Supervised approach - Naive Bayes Model	- The dependent feature which is knownas high readability feature - High-speed	- Generate Keyphrase
TF-IDF	- Supervised approach - Simple statistical Model	- No need for a labeled corpus and also gives some basic metrics for extraction. - Easy to compute the similarity between documents	- Widely applied for a large corpus

RAKE	- Unsupervised approach - Simple statistical Model	-Easy to implement with low complexityand also there is no need for labeled corpus -Uses long Keyphrase extraction for best performance	- Only extracting Keyphrase from texts documents
TextRank	- Unsupervised approach - Graph-Based Model	-Language independentand also provide credibility assessment enhancement. -The algorithm can do many implementationsat the same time and no need for a huge Corpus for training.	- Applied for small scale text Keyphrase extraction task.

VI. CONCLUSION

The Automatic Keyphrase Extraction (AKPE) technique helps to sense important words within a corpus, which topics are being discussed and it automates the workflows like tagging incoming survey responses or responding to urgent user queries, allows saving a lot of time. It also provides actionable insights that used to make a better conclusion. In this survey paper, the various approaches and models of Automatic Keyphrase extraction such as KEA, TF-IDF, RAKE, TextRank was briefly described and also comparative study on each AKPE technique has been discussed. In the future, we can implement AKPE techniques to extract Keyphrase from the large corpus and also which technique gives the best result to find exact Keyphrase.

REFERENCES

[1] Gollum Rabby, Saiful Azad1, Mufti Mahmud · Kamal Z. Zamli1, Mohammed MostafizurRahman "TeKET: a Tree-Based Unsupervised Keyphrase Extraction Technique, Cognitive Computational", Published online" March 2020.
[2] Alzaidy. R., Caragea, C., Giles, C.L.: Bi-LSTM- CRF sequence labeling for keyphrase extraction from scholarly documents. In: Proceedings of the World Wide Web Conference, pp. 2551–2557. ACM, 2019.
[3] DebanjanMahata, John Kuriakose, Rajiv Ratn Shah, and Roger Zimmermann "Key2Vec: Automatic Ranked Keyphrase Extraction from Scientific Articles using Phrase Embeddings".
[4] Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem, and KhaledShaalan, "Using Text Mining Techniques for Extracting Information from Research Articles", Chapter in Studies in Computational Intelligence , DOI: 10.1007/978-3-319-67056-0_18 January 2018.
[5] M. Uma Maheswari, Dr. J. G. R. Sathiaseelan. "Text Mining: Survey on Techniques and Applications", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, Volume 6 Issue 6, June 2017.
[6] Jinzhang Zhou, "Keyword extraction method based on word vector and TextRank. Application Research of Computers", 36, 5, 2019.
[7] Suhan pan, Zhiqiang Li, Juan Dai, "An improved TextRank keywords extraction algorithm", ACM TURC '19: Proceedings of the ACM Turing Celebration Conference – China, May 2019.
[8] Florescu, C., Caragea, and C., "PositionRank: an unsupervised approach to keyphrase extraction from scholarly documents", In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 1105–1115, 2017.
[9] Howard, Jeremy, & Ruder, Sebastian, "Universal language model fine-tuning for text classification". arXiv preprint arXiv:1801.06146, 2018.
[10] Sifatullah Siddiqi, AditiSharan, "Keyword and Keyphrase Extraction Techniques: A Literature Review", International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 2, January 2015.
[11] Meng, Rui, Yuan, Xingdi, Wang, Tong, Brusilovsky, Peter, Trischler, Adam, & He, Daqing. "Does Order Matter? An Empirical Study on Generating Multiple Keyphrases as a Sequence",arXiv preprint arXiv:1909.03590, 2019.
[12] zAlzaidy, R., Caragea, C., Giles, C.L.: "Bi- LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents". In: Proceedings of The World Wide Web Conference, pp. 2551–2557. ACM 2019.
[13] Beltagy, I., Cohan, A., Lo, K.: "Scibert: pretrained contextualized embeddings for scientific text", 2019.
[14] S. Anjali, Nair M. Meera, M.G. Thushara, " A Graph-based Approach for Keyword Extraction from Documents, Second International Conference on Advanced Computational and Communication Paradigms", (ICACCP) 2019.
[15] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. "Language models are unsupervised multitask learners. OpenAI Blog", 2019.
[16] Yan Ying, Tan Qingping, Xie Qinzheng, Zeng Ping and Li Panpan, "A graph-based approach of automatic keyphrase extraction", Procedia Computer Science, vol. 107, pp. 248-255, 2017.
[17] Gollapalli, S.D., & Caragea, C. "Extracting keyphrases from research papers using citation networks", 2014.
[18] MG Thushara, SA Sreeremya, and S Smitha, "Kea-based document tagging for project recommendation and analysis" in Recent Findings in Intelligent Computing Techniques, Springer, pp. 285- 295, 2018.
[19] Tayfun Pay, Stephen Lucci, James L. Cox, "An Ensemble of Automatic Keyword Extractors: TextRank, RAKE and TAKE", Computación y Sistemas, Vol. 23, No. 3, 2019.
[20] Isabella Gagliardi and Maria Teresa Artese, " Semantic Unsupervised Automatic Keyphrases Extraction by Integrating Word Embedding with Clustering Methods", June 2020.