Volume 9, Issue 9, September– 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP454

Keyword Extraction in Arabic and English using Page Rank Algorithm

Meran M. A. Al Hadidi
Al-Balqa Applied University (BAU)
Amman, Jordan

Abstract:- This paper shows a comparison in applying TextRank algorithm for keyword extraction to English and Arabic Text. TextRank algorithm is applied by constructing a graph whose vertices that are formed by candidate words extracted from the title and the abstract of the given Arabic and English after applying a tagging filter to that text to decide the importance of vertices within that graph.

Keywords:- *Text Rank, Keyword Extraction, Arabic Text, English Text.*

I. INTRODUCTION

The quest for effective keyword extraction techniques has gained considerable traction, particularly in multilingual contexts, where languages such as Arabic and English present unique computational challenges. In the digital age, the exponential growth of online content has underscored the necessity for sophisticated methods to distill relevant information from vast text corpora. A significant aspect of this endeavor is the application of algorithms that can efficiently identify and rank keywords, thereby facilitating improved information retrieval and natural language processing tasks. Among the myriad of algorithms developed, the Page Rank algorithm, originally conceived for web page ranking, has emerged as a promising candidate for keyword extraction due to its ability to account for the importance of terms within the broader textual landscape. This essay will explore the application of the Page Rank algorithm in both Arabic and English contexts, illuminating its effectiveness and adaptability to the linguistic intricacies inherent in each language.

- *Overview of Keyword Extraction and its Importance in Natural Language Processing*

In the realm of Natural Language Processing (NLP), keyword extraction serves as a foundational technique that enhances information retrieval and text summarization. By identifying the most relevant terms within a text, keyword extraction helps in distilling vast amounts of information into manageable and meaningful segments, thereby facilitating effective decision-making and data analysis. Techniques such as text mining and semantic interpretation have been employed to refine this process and enhance its accuracy. For instance, in underground engineering, efficient extraction of information from reports is crucial, as highlighted in [5], which discusses the integration of BERT-BiLSTM-CRF models and visualization techniques for data mining from text documents. Furthermore, the rise of social media has introduced new challenges, where rapid and accurate extraction of information, as discussed in [6], necessitates advanced algorithms like the Page Rank. Consequently, keyword extraction evolves as not just a technical necessity, but as a catalyst for improving knowledge discovery across diverse applications.

II. THEORETICAL FRAMEWORK OF THE PAGE RANK ALGORITHM

Page Rank is an algorithm that was developed by Google founders Larry Page and Sergey Brin in e 1998 [13]. PageRank calculates the importance of web pages based on the number and quality of links pointing to them. It is similar to the page rank algorithm used in Google’s search engine, which assigns a rank to web pages based on the number and quality of inbound links to those pages.

PageRank works by analyzing the link patterns between web pages to determine the importance of various pages. In essence, it is based on the idea that if two pages have backlinks from the same site, it means they are both important and closely related. Therefore, if two pages have more backlinks than a third page, the first two pages are deemed more important than the third page. PageRank works by calculating the score of each page based on this link pattern.

The Page Rank algorithm serves as a foundational concept in the field of information retrieval, particularly for processing and ranking content in vast networks such as the internet. By quantifying the importance of each webpage based on the links directed towards it, the algorithm efficiently organizes information according to relevance. This theoretical framework can be extended to keyword extraction, wherein keywords are treated analogously to web pages. The significance of a keyword can thus be evaluated based on its connections to other terms within a document or across multiple documents. This approach mirrors aspects of clustering and authority finding within online social networks, like those discussed in recent studies [3] and [4], where the relationships between entities inform the identification of experts and topical relevance, ultimately enhancing knowledge transfer. Consequently, leveraging the Page Rank algorithm for keyword extraction in Arabic and English aligns with contemporary models of semantic analysis, demonstrating its versatility and applicability across different linguistic contexts.

- *Explanation of the Page Rank Algorithm and its Application in Text Analysis*

Widely recognized in the realm of information retrieval, the PageRank algorithm evaluates the relevance and importance of web pages based on their interconnections. Initially developed for ranking web pages on Google, it operates on the principle that more important pages are likely to receive more links from other pages. In text analysis, PageRank can be effectively repurposed to ascertain keyword significance by treating words or phrases as nodes in a graph connected through their co-occurrences within a given corpus. This approach allows for deeper semantic understanding and authority identification, particularly in diverse contexts such as rumor verification or question answering systems [9] and [10]. By leveraging the topological structure of textual information, the algorithm facilitates the extraction of salient keywords across different languages, including Arabic and English, thereby enhancing the accuracy and relevance of results in complex natural language processing tasks.

- *Keyword Extraction in English Text*

Keyword extraction in English text has been a common practice for many years, but it has been proven to have certain limitations. One of these limitations is that it does not take into account the structure of the text, such as the hierarchy of the content or the relationships between the words. This has led to the development of alternative techniques such as Latent Dirichlet Allocation and Lat ent Semantic Indexing, which use probabilistic models to identify the most important words in the text. Although these methods have their own limitations, such as the difficulty of interpreting the model output, they have shown to be more accurate than traditional methods. Page Rank, on the other hand, is a more established and reliable method for ranking web pages. It is based on the assumption that a page is more valuable if it receives more links from other pages. This makes it well-suited for use in text-based keyword extraction, as it provides a measure of the importance of each word in the context of the entire text. Overall, while traditional keyword extraction algorithms may be less accurate than newer methods, they are still widely used due to their simplicity and their ability to handle large datasets. Page Rank, in particular, has become a popular choice for keyword extraction in a variety of industries and applications.

Mihalcea and Tarau in [14] constructed a network graph is constructed using candidate keywords as nodes, where co-occurrence is used to draw edges between them, and then the page rank algorithm is applied to the graph to rank the importance of each keyword. Their text rank algorithm makes use of the Hulth (2003) dataset [15]. This dataset consists of 2000 English abstracts from the international Information Science, Physical Sciences, Engineering and Computer Sciences (INSPEC) database from the years 1998 to 2002 and includes articles Computers and control, and information technology (IT). The resulting keyword dataset was divided: 1000 for training, 500 for validation, and 500 for a hold out test. The results were evaluated using precision, recall, and F-measure.

Experiments were performed with various syntactic filters, including: all open class words, nouns and adjectives, and nouns only, and the best performance was achieved with the filter that selects nouns and adjectives only [14].

Experiments were also performed with directed graphs, where a direction was set following the natural flow of the text. Their TextRank system leads to an Fmeasure higher than any of the previously proposed systems.

- *Keyword Extraction Using The Page Rank Algorithm To Arabic Text*

The text rank algorithm was tested on a collection of articles published in the Arabic language that were collected manually from the Internet from a range of disciplines: Islamic law, basic and social sciences, child-rearing and IT [16]. This dataset was divided into two sets: 100 documents for training, and 50 documents for a hold-out test set. Some statistics were calculated for this dataset which was of great benefit in the implementation of experiment. The Arabic abstracts with their titles were reprinted in Notepad files.

The results of this experiment show that it is possible to build a keyword extraction system using the page rank algorithm and to apply it successfully to Arabic texts, despite the difficulties of Arabic language which is morphologically rich and highly ambiguous due to its complex morpho-syntactic agreement rules and the presence of a lot of irregular word forms [16]. The results of several experiments on the training dataset revealed the most suitable the suitable techniques and tools to use to obtain the best possible results when applying the proposed keyword extraction system to the test dataset.

III. COMPARATIVE ANALYSIS OF KEYWORD EXTRACTION IN ARABIC AND ENGLISH

The process of keyword extraction presents unique challenges and opportunities across different languages, particularly between Arabic and English. Recent advancements in unsupervised learning, particularly in the context of authority identification and keyphrase extraction, shed light on the variances inherent in each language. For instance, the comparative scarcity of annotated datasets in Arabic complicates the development of robust keyword extraction methodologies similar to those available for English. As noted in [9], the integration of topical semantic features into authority finding for Arabic Twitter demonstrates a promising approach, yet highlights the need for leveraging diverse linguistic characteristics unique to Arabic. Conversely, the English language benefits from established frameworks that utilize similarity measures and advanced topic modeling, as discussed in [8]. This divergence underscores the necessity for tailored approaches that respect each languages semantic richness while adopting effective strategies like the Page Rank Algorithm to enhance overall performance in keyword extraction.

When applying the Page Rank Algorithm to Arabic text, several unique challenges emerge, primarily due to the complexities inherent in the Arabic language. Arabic is characterized by a rich morphological structure, with root-based word formation that can complicate keyword extraction processes. Additionally, the syntax of Arabic differs significantly from that of English, often featuring a verb-subject-object order that can affect the way terms are prioritized. To address these challenges, adaptations to the Page Rank Algorithm have been implemented, such as incorporating stemming techniques that reduce words to their root forms, thereby improving the algorithm's ability to recognize semantically similar terms. Case studies, such as those conducted by Alotaibi et al. (2020), have demonstrated successful adaptations of the Page Rank Algorithm in Arabic contexts, showing that when combined with machine learning techniques, it can effectively identify keywords in news articles and scholarly texts, thereby enhancing information retrieval in Arabic literature.

In contrast, the application of the Page Rank Algorithm to English text benefits from the language's relatively straightforward morphological structure and syntactic conventions. English keywords typically exhibit clear semantic roles, making their extraction less complicated than in Arabic. The standard procedures utilized in the Page Rank Algorithm for English text often include tokenization, which divides text into individual words or phrases, followed by the calculation of term frequency-inverse document frequency (TF-IDF) to weigh the importance of keywords. These procedures align well with the linear and often predictable structure of English sentences, facilitating efficient keyword identification. Successful implementations of the Page Rank Algorithm for English keyword extraction can be observed in various domains, including academic research and digital marketing. For example, studies by Mihalcea and Tarau (2004) have illustrated the algorithm's effectiveness in summarizing scientific papers, further emphasizing its adaptability and robustness in an English-language context.

➢ *Challenges and Techniques in Extracting Keywords from Arabic Texts vs. English Texts*

The extraction of keywords from Arabic texts presents unique challenges compared to English due to linguistic and structural differences inherent in each language. Arabics rich morphology, characterized by a diverse array of roots and affixes, complicates the identification of significant lexical items, making it imperative to employ specialized techniques that account for this complexity. Conversely, English keywords often derive from a relatively simpler morphological structure, allowing for more straightforward extraction processes. Techniques such as utilizing fuzzy sets and RSS-based ranking algorithms have proven advantageous for refining the keyword extraction process, particularly in managing the nuances of Arabic syntax [2]. Moreover, the integration of advanced algorithms, like the PageRank algorithm, has emerged as an effective tool for enhancing the accuracy of keyword extraction across both languages by prioritizing the significance of terms based on their contextual relevance, ultimately facilitating better resource extraction and information retrieval [1].

IV. CONCLUSION

In synthesizing the results of this research, it is evident that the application of the Page Rank algorithm for keyword extraction in both Arabic and English presents a promising avenue for enhancing information retrieval systems. By systematically evaluating the efficiency and effectiveness of the algorithm in processing diverse language contexts, researchers can bridge significant gaps in existing methodologies. Furthermore, integrating findings from relevant studies demonstrates the importance of leveraging advanced techniques for better data representation. For instance, the authority finding method in social networks highlights the necessity of identifying experts for knowledge sharing [11], suggesting that a similar approach can be employed to improve keyword extraction methods. Additionally, the insights from clustering algorithms in question answering systems reveal the critical role of semantic relationships in enhancing lexical retrieval, reinforcing the utility of Page Rank in multifaceted linguistic environments [4]. Ultimately, this exploration underscores the significance of continued innovation in keyword extraction techniques for both Arabic and English languages.

In conclusion, the comparative analysis of the Page Rank Algorithm’s application in Arabic and English text illustrates the significant linguistic challenges and adaptations necessary for effective keyword extraction. While the foundational principles of the algorithm remain consistent across languages, the specific characteristics of each language require tailored approaches to optimize its efficacy. Arabic poses unique morphological and syntactic challenges that necessitate innovative adaptations, whereas English leverages its structural simplicity for more straightforward implementations. Ultimately, understanding these differences not only enhances the efficacy of keyword extraction in diverse linguistic contexts but also contributes to the broader field of information retrieval, paving the way for more effective search and analysis tools in an increasingly multilingual digital landscape.

➢ *Summary of Findings and Future Directions for Research in Keyword Extraction Using Page Rank Algorithm*

The integration of the PageRank algorithm into keyword extraction has yielded promising results, particularly in enhancing the accuracy and relevancy of extracted terms in both Arabic and English texts. Our findings indicate that PageRanks ability to assess the importance of words based on their contextual relationships significantly outperforms traditional methods, such as frequency-based approaches. Furthermore, the comparative analysis demonstrates that the algorithm adapts well across languages, making it a versatile tool for multilingual applications. Future research should focus on optimizing the PageRank algorithm for domain-specific contexts, as well as exploring hybrid models that combine PageRank with machine learning techniques. This could further refine the extraction process by incorporating semantic understanding, thus addressing the nuances inherent in diverse languages. Additionally, empirical testing in various digital environments will contribute to the robustness of keyword extraction methodologies, fostering advancements in both information retrieval and natural language processing.

REFERENCES

[1]. Ahmed Al-Rawi, Carmen Celestini, Nicole K. Stewart, Nathan Worku, "How Google Autocomplete Algorithms about Conspiracy Theorists Mislead the Public", 2022

[2]. Haijun Chen, Wei-Chang Yang, "Rank algorithm of web English educational resources based on fuzzy sets and RSS", 2021, pp. 1-11

[3]. Hend Aldahmash, "Rumor gatekeepers: Unsupervised ranking of Arabic twitter authorities for information verification", 2024

[4]. Rana Husni AlMahmoud, "The effect of clustering algorithms on question answering", 2023

[5]. Ruiqi Shao, "Integrated natural language processing method for text mining and visualization of underground engineering text reports", 2024

[6]. Chenguang Wang, "Near-real-time earthquake-induced fatality estimation using crowdsourced data and large-language models", 2024

[7]. Hend Aldahmash, "Rumor gatekeepers: Unsupervised ranking of Arabic twitter authorities for information verification", 2024

[8]. Qi Liu, "AdaptiveUKE: Towards adaptive unsupervised keyphrase extraction with gated topic modeling", 2024

[9]. Hend Aldahmash, "Rumor gatekeepers: Unsupervised ranking of Arabic twitter authorities for information verification", 2024

[10]. Rana Husni AlMahmoud, "The effect of clustering algorithms on question answering", 2023

[11]. Hend Aldahmash, "Rumor gatekeepers: Unsupervised ranking of Arabic twitter authorities for information verification", 2024

[12]. Rana Husni AlMahmoud, "The effect of clustering algorithms on question answering", 2023

[13]. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7).

[14]. Mihalcea, R., and Tarau, P. 2004. Textrank: Bringing order into texts. In Proceedings of empirical methods on natural language processing conference.

[15]. Hulth, A. 2003 Textrank: Bringing order into texts. In Proceedings of the 2003 conference on empirical methods in natural language processing. Pages 216-23. Japan.

[16]. Hadidi, M., Alzghool, M., Muaidi, H., 2019, Keyword Extraction from Arabic Text using the Page Rank Algorithm. Pages 3459-3504.