INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 07, July 2023, pp : 232-240

ENHANCED ARABIC INFORMATION RETRIEVAL FOR INFORMED DECISION-MAKING: EMPOWERING POLITICAL SEARCH

Mansour Al-Helalat[1]

[1]Lecturer, Department of Computer Science, University of Jordan, Amman, Jordan

ABSTRACT

Google searches play a crucial role in providing accurate and relevant information, particularly in the domain of politics. Enhancing Google searches for political matters is essential due to the complex nature of politics and the need for reliable and up-to-date information. This paper proposes an approach to improve information retrieval in Arabic language sources by addressing the challenges associated with the language's complexity. The proposed approach includes several steps. First, tokenization divides the user's query into individual word segments, allowing for further processing based on the selected domain. Unification ensures consistent representation of Arabic letters by addressing the variations caused by diacritics. Stop-words and special characters that offer little semantic value are removed to improve precision. The approach incorporates light stemming using the "Khoja" stemmer, which generates relevant terms without over-generating them. Term generation expands the range of terms by using a mechanism proposed by "Sarf" and selecting the top ten terms with the highest term frequency on Google. Finally, the query is updated on the backend to increase the bag of words for evaluation. To evaluate the effectiveness of the proposed approach, precision is used as the primary metric. Google Search Engine (SE) serves as the benchmark for comparison, considering its efficiency in Arabic language information retrieval. The precision values of applied queries related to the Politician domain are recorded per page, both in their original plain form and after being updated using the proposed approach. The results demonstrate that the proposed approach improves precision compared to the original plain queries. For instance, precision increases from 0.40 to 0.65 for Q1 and from 0.65 to 0.95 for Q3. These findings highlight the effectiveness of the approach in enhancing the retrieval of relevant documents in Arabic language information retrieval systems. The systematic process presented in this research contributes valuable insights for improving the performance of information retrieval systems in the Arabic language. Further research can focus on refining and optimizing the approach, exploring its applicability to other domains, and addressing any remaining challenges to ensure its effectiveness in real-world scenarios.

Keywords: Google Searches, Political Search, Information Retrieval, Arabic Language, Complex Nature of Politics, Reliable Information.

1. INTRODUCTION

Google searches play a crucial role in providing information and insights on various topics, including political matters. As the primary search engine used by millions of people worldwide, Google's ability to deliver accurate and relevant search results is of utmost importance, particularly in the context of political information retrieval. Improving Google searches for political matters holds significant value for several reasons [1]. Firstly, politics is a complex and dynamic field that encompasses a wide range of topics, including government policies, elections, international relations, and public opinion. Access to reliable and up-to-date information is vital for individuals, researchers, journalists, and policymakers to make informed decisions and understand political landscapes. Enhancing Google searches in the political domain can help overcome challenges related to information overload and misinformation. With the vast amount of information available online, users often face difficulties in finding credible and authoritative sources. By improving search algorithms and incorporating domain-specific knowledge, Google can better prioritize trustworthy and accurate political information, thereby reducing the risk of misinformation and promoting informed political discourse [2].

Furthermore, efficient political searches can foster transparency and accountability. Access to comprehensive and diverse political information empowers citizens to hold their representatives and governments accountable. It allows individuals to monitor political developments, track policy changes, and engage in informed discussions and debates [3]. Improved Google searches for political matters also facilitate democratic participation. By providing users with easy access to a wide range of political perspectives, opinions, and news sources, Google can contribute to a more inclusive and democratic public sphere. Users can explore diverse viewpoints, engage with different ideologies, and gain a deeper understanding of complex political issues, fostering a more informed and engaged citizenry. In conclusion, the importance of improving Google searches for political matters cannot be overstated. By enhancing search algorithms, prioritizing reliable sources, and providing users with access to diverse political information, Google can play a vital role in promoting transparency, accountability, and democratic participation. Continued efforts to refine and optimize search capabilities in the political domain are crucial in our increasingly digital and politically interconnected world.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 07, July 2023, pp : 232-240

2. LITERATURE REVIEW

Existing research in the field of Arabic information retrieval has primarily focused on addressing the complexities of Arabic morphology. Numerous stemming methods have been proposed to enhance natural language processing (NLP) applications for Arabic text. These methods can be broadly classified into two categories: root-based approaches and stem-based approaches. Ghwanmeh et al. [4] introduced a novel and efficient approach to text classification that aims to enhance the process. The approach utilizes an algorithm to automatically classify documents based on their content. The key concept behind the algorithm is the selection of feature words from each document that effectively capture the main ideas expressed within. The algorithm generates a list of the primary subjects identified in the document, providing valuable insights into its content. Furthermore, the paper explores the impact of Arabic text classification on Information Retrieval. It investigates how the classification process can enhance the retrieval of relevant information from Arabic language sources. By examining the effects of text classification in this context, the paper contributes to a deeper understanding of the role and significance of classification techniques in Information Retrieval tasks. Bessou et al. [5] proposed a method for indexing and retrieving Arabic texts using natural language processing techniques. Our approach leverages the concept of templates in word stemming to replace words with their corresponding stems. This technique has demonstrated its effectiveness in improving information retrieval by reducing silence during the retrieval phase and returning highly relevant results. To evaluate the performance of our proposed algorithm, ESAIR (Enhanced Stemmer for Arabic Information Retrieval), we conducted a series of experiments. The results obtained from these experiments indicate that our algorithm achieves a high accuracy rate, up to 96%, in extracting the exact root of Arabic words. This high accuracy in root extraction significantly contributes to the improvement of information retrieval in Arabic texts. By applying our method, researchers and practitioners can benefit from enhanced indexing and retrieval capabilities for Arabic texts. The use of template-based word stemming offers a reliable and accurate approach for handling Arabic language intricacies and optimizing the retrieval process. The findings of our experiments validate the effectiveness of our approach and highlight its potential for various applications in the field of Arabic information retrieval. Mustafa et al. [6] developed a system framework that enables users to retrieve Arabic information based on queries written in slang language. This framework utilizes a context-free grammar approach to facilitate the conversion between slang and classical Arabic. Through this framework, we aim to bridge the gap between colloquial slang and formal Arabic, enabling effective information retrieval in both contexts. To validate the effectiveness of the framework, we will focus on applying it to the colloquial slang used in specific regions of North Jordan, namely Irbid, Ajloun, Jerash, Mafraq, and AlRamtha city. This region-specific analysis will allow us to fine-tune the framework and adapt it to the unique linguistic characteristics and expressions found in these areas. Additionally, we will create a specialized file for handling non-Arabic words commonly used in the slang language. This will ensure the accurate processing and understanding of queries that involve mixed languages or include foreign words. Furthermore, stop words, which are commonly used but carry little semantic value, will also be addressed in our framework. By excluding stop words from the retrieval process, we can improve the precision and relevance of the retrieved information.

3. METHODOLOGY

In the realm of information retrieval systems, handling the complexity of the Arabic language structure poses unique challenges. To overcome these challenges and enhance the performance of such systems, a specific approach can be implemented. This approach involves mapping a given term to its different morphological forms on a backend implementation, taking into account the specificity of the domain in question.

To provide a comprehensive understanding of the experiment process, let’s delve into each step to gain a deeper insight into the approach:

- **Step 1: Tokenization**
  In this initial step, the user's query is segmented into individual word segments. Depending on the selected domain, further processing is carried out on the words that fall under that domain. General operations are performed, including the removal of stop-words and specific characters, which contribute to refining the query.
- **Step 2: Unification**
  The Arabic language encompasses a vast array of diacritics that can be perplexing for machines to decipher. To mitigate this complexity, a unification process is applied, bringing all letters to a specific form. For example, variations like (إ, ا, آ, أ) are unified to (ا). This unification ensures consistent representation across all Arabic letters.
- **Step 3: Removing Stop-Words and Special Characters**
  Certain words, known as stop-words, are recurrent in many documents and offer little semantic value. Examples include words like "في" (in) and "على" (on). To prevent the retrieval of irrelevant documents, these stop-words are eliminated. Furthermore, special characters that do not contribute to semantic understanding are also removed.

- **Step 4: Stemming**
  Differentiating this approach from other Arabic morphological systems, a light stemming process is employed. Specifically, the terms within the selected domain undergo a light stemming process using the "Khoja" light stemmer. This choice is motivated by the fact that the root of Arabic terms can generate an overwhelming number of derived terms. Light stemming ensures the generation of a manageable set of relevant terms.
- **Step 5: Term Generation**
  After stemming the specific terms, a mechanism proposed by "Sarf" is applied. This mechanism generates all possible forms of a term and selects the top ten terms with the highest term frequency when searched on Google. This expansion of the term range facilitates a more comprehensive retrieval process.
- **Step 6: Query Update**
  Following the completion of the previous steps, the original query is updated on the backend of the application. This update increases the bag of words and prepares the query for evaluation in the subsequent section.

  By following this systematic approach, the proposed method aims to address the intricate structure of the Arabic language in information retrieval systems. Through the mapping of morphological forms, unification of letters, removal of stop-words and special characters, light stemming, and term generation, the approach strives to enhance the retrieval of relevant information from Arabic language sources. Table 1shows the methodology in the python implementation.

Table 1: Methodology in the python implementation

```
# Step 1: Tokenization
def tokenize_query(query):
word_segments = query.split() # Split query into individual word segments
# Perform additional processing based on selected domain
processed_segments = process_segments(word_segments)
return processed_segments
# Step 2: Unification
def unify_letters(query):
unified_query = query.replace("ٱ" ,"أ").replace("ٱ" ,"إ").replace("ٱ" ,"ٱ")
# Apply unification to other Arabic letters as needed
return unified_query
# Step 3: Removing Stop-Words and Special Characters
def remove_stop_words(query):
stop_words = ["على" ,"في"] # Define a list of stop-words
processed_query = ' '.join([word for word in query.split() if word not in stop_words])
# Remove special characters using regular expressions if needed
processed_query = remove_special_characters(processed_query)
return processed_query
# Step 4: Stemming
def perform_stemming(query):
stemmed_terms = []
for term in query.split():
if term in selected_domain_terms:
stemmed_term = perform_light_stemming(term)
stemmed_terms.append(stemmed_term)
processed_query = ' '.join(stemmed_terms)
return processed_query
# Step 5: Term Generation
def generate_terms(query):
expanded_terms = []
for term in query.split():
```

```
expanded_terms.extend(generate_all_possible_forms(term))
top_ten_terms = select_top_terms(expanded_terms)
processed_query = ' '.join(top_ten_terms)
return processed_query
# Step 6: Query Update
def update_query(original_query):
updated_query = process_backend_update(original_query)
return updated_query
# Example usage
user_query = "السياسة العالمية في الحروب"
processed_query = tokenize_query(user_query)
processed_query = unify_letters(processed_query)
processed_query = remove_stop_words(processed_query)
processed_query = perform_stemming(processed_query)
processed_query = generate_terms(processed_query)
updated_query = update_query(processed_query)
```

4. RESULTS

The primary metric utilized in this research is precision, considering the unknown recall for the given queries. Google Search Engine (SE) is selected as the benchmark for comparison due to its efficient results in Arabic language information retrieval (IR) systems. The precision metric is calculated using the formula:

$$\text{Precision} = \frac{\text{Number of Relevant Documents}}{\text{Number of Relevant Documents}}$$

To evaluate the precision, each query is entered into the Google SE, and the cumulative precision per page is recorded. For instance, if a query yields 99 out of 100 relevant documents on the ten pages, it scores 99%. Table 1 showcases the applied queries specifically related to the politician domain. Table 2 presents the precision values for each query, both in its original plain form and after being updated using the proposed approach. Figures 1 to 10 showcase the precision values for each query.

Table 2: The tested queries

Number	Query
Q1	عزيز العدالة الاجتماعية وتقليل الفجوة
Q2	الشفافية والمساءلة في العمل السياسي
Q3	تعزيز حقوق الإنسان والحريات الأساسية
Q4	نظام السعر المحدد في الاردن
Q5	فرص العمل والاستقرار الاقتصادي
Q6	التعليم الجيد والصحة العامة
Q7	التعاون وتحقيق المصالح العامة
Q8	تعزيز الديمقراطية ومشاركة المواطنين في صنع القرارات السياسية
Q9	مكافحة الفساد وتعزيز النزاهة في الحكم
Q10	تعزيز الأمن الوطني ومكافحة الإرهاب والتطرف

Through this evaluation process, the precision of the information retrieval system is assessed, providing insights into the effectiveness of the proposed approach in enhancing the retrieval of relevant documents for the given queries.

Table 3: The tested queries: precision per page

Query	page 1	page 2	page 3	page 4	page 5
Q1	0.40	0.40	0.30	0.30	0.20
Q1*	0.65	0.60	0.35	0.35	0.30

Q2	0.50	0.40	0.20	0.20	0.10
Q2*	0.85	0.55	0.35	0.35	0.30
Q3	0.65	0.65	0.40	0.30	0.30
Q3*	0.95	0.80	0.60	0.60	0.50
Q4	0.50	0.25	0.25	0.20	0.20
Q4*	0.80	0.60	0.50	0.50	0.50
Q5	0.60	0.50	0.50	0.40	0.30
Q5*	1	0.90	0.70	0.70	0.60
Q6	0.50	0.45	0.40	0.40	0.40
Q6*	0.70	0.65	0.60	0.60	0.50
Q7	0.60	0.45	0.30	0.20	0.20
Q7*	0.95	0.60	0.50	0.45	0.45
Q8	0.70	0.70	0.50	0.40	0.30
Q8*	1	0.85	0.70	0.60	0.60
Q9	0.60	0.50	0.45	0.20	0.10
Q9*	0.90	0.70	0.60	0.50	0.50
Q10	0.70	0.55	0.40	0.40	0.20
Q10*	1	1	0.90	0.80	0.70

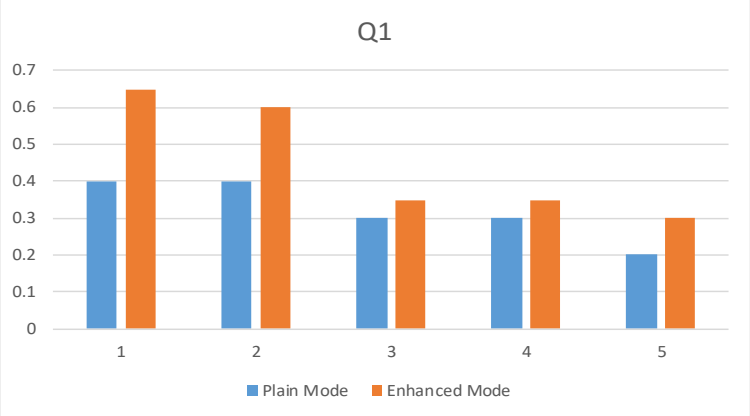

Figure 1: Query1 precision per page

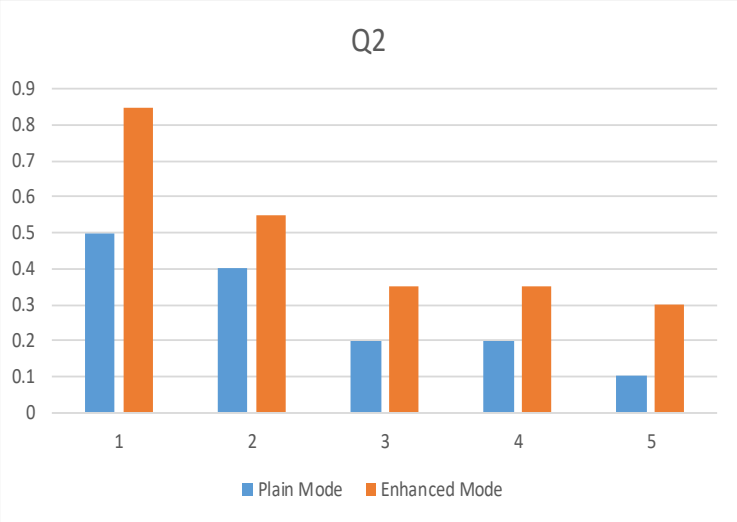

Figure 2: Query2 precision per page

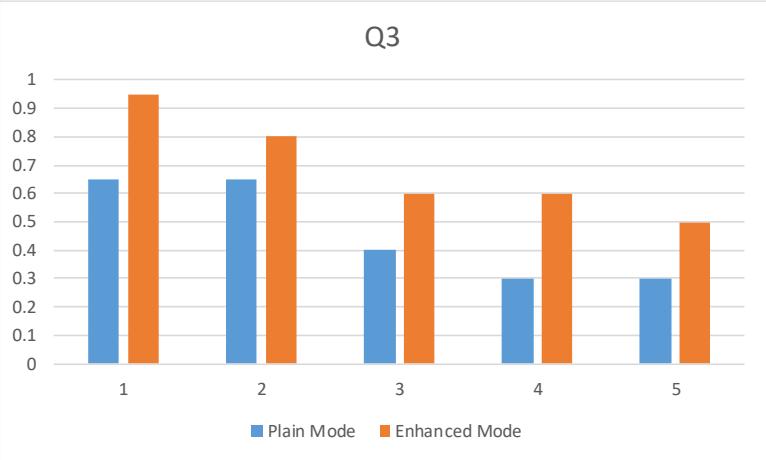

Figure 3: Query3 precision per page

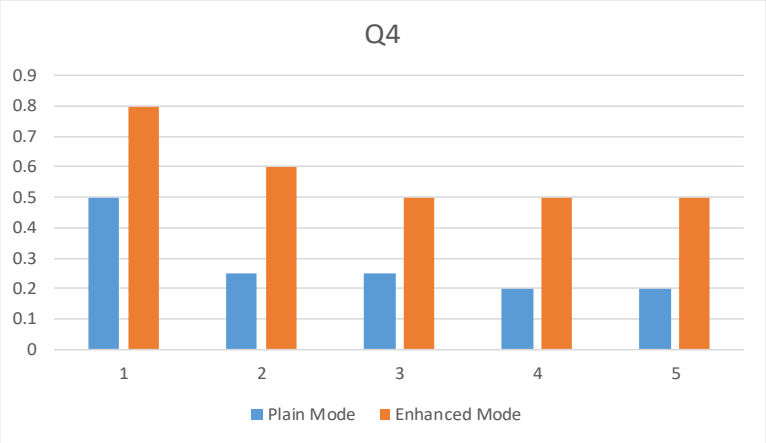

Figure 4: Query4 precision per page

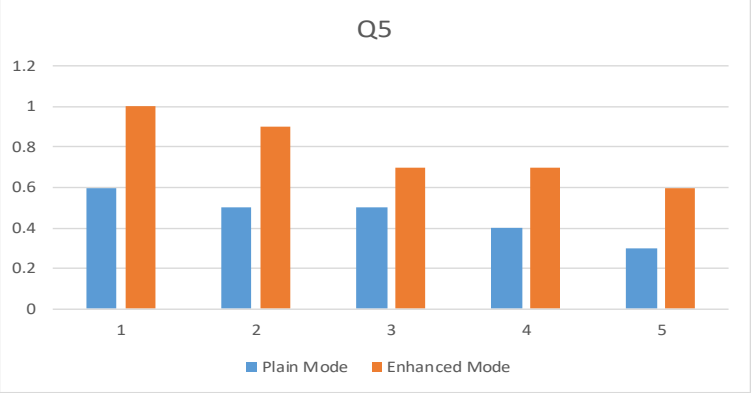

Figure 5: Query5 precision per page

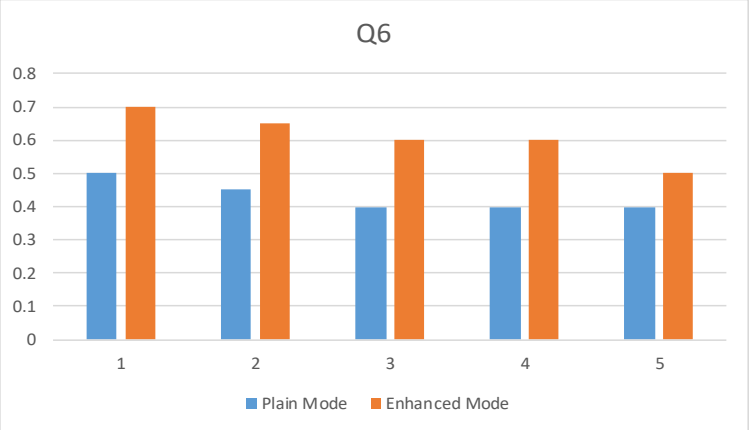

Figure 6: Query6 precision per page

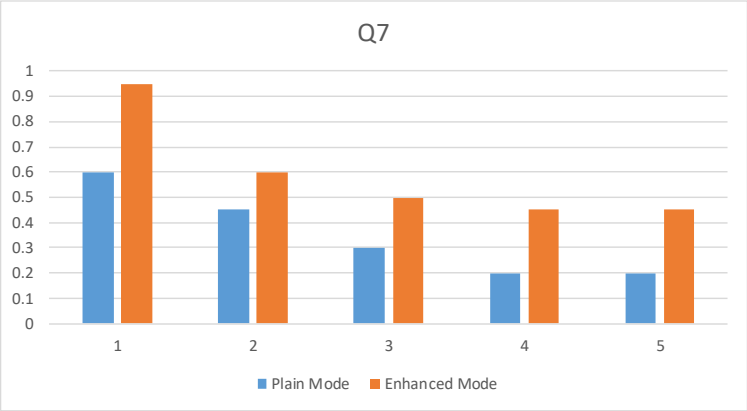

Figure 7: Query7 precision per page

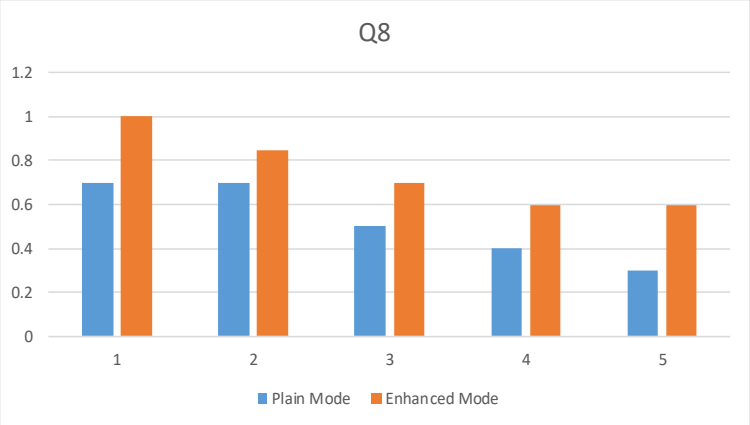

Figure 8: Query9 precision per page

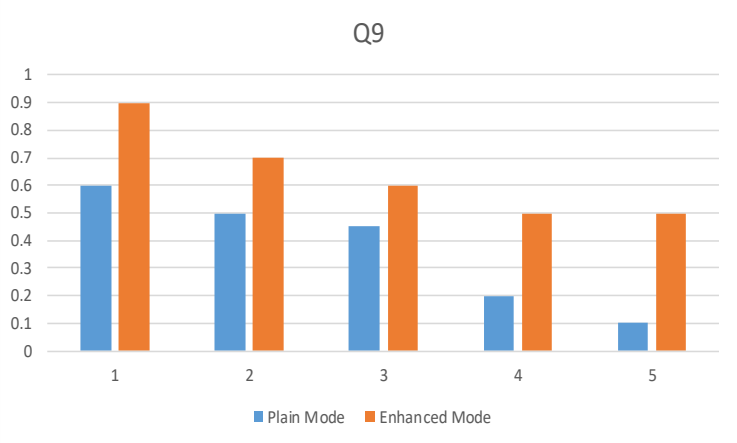

Figure 9: Query9 precision per page

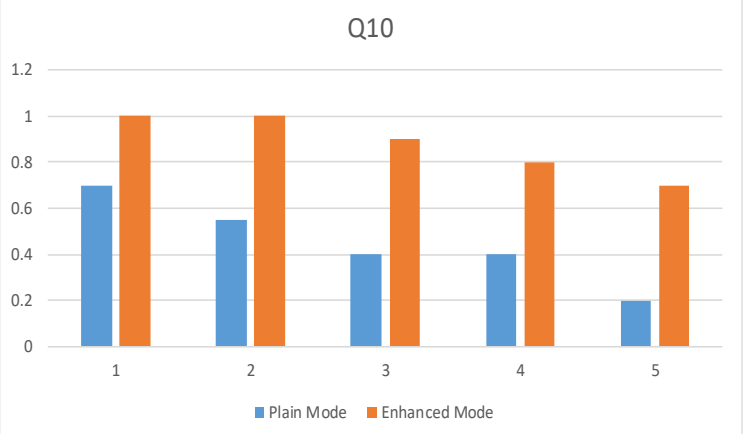

Figure 10: Query10 precision per page

5. DISCUSSION

The proposed approach aims to address the challenges associated with the complexity of the Arabic language in information retrieval systems. By following a systematic process, including tokenization, unification, removing stop-words and special characters, stemming, term generation, and query update, the approach seeks to enhance the retrieval of relevant information from Arabic language sources. Tokenization is the initial step, where the user's query is divided into individual word segments. This allows for further processing and refinement of the query based on the selected domain. Unification is then applied to ensure consistent representation of Arabic letters, considering the various diacritics present in the language. Removing stop-words and special characters is crucial to improve the precision of the retrieval process.

Commonly occurring words with little semantic value, such as "في" (in) and "على" (on), are eliminated. Additionally, special characters that do not contribute to semantic understanding are also removed. Stemming, a distinctive feature of this approach, employs the "Khoja" light stemmer to generate relevant terms within the selected domain. Light stemming is chosen to avoid over-generation of terms; as Arabic language roots can generate numerous derived terms. The term generation step utilizes a mechanism proposed by "Sarf" to generate all possible forms of a stemmed term.

The top ten terms with the highest term frequency when searched on Google are selected. This expands the range of terms and facilitates a more comprehensive retrieval process. Finally, the query is updated on the backend of the application, increasing the bag of words for evaluation in subsequent sections. The systematic approach presented in this research aims to improve the performance of information retrieval systems for the Arabic language.

To evaluate the effectiveness of the proposed approach, precision is used as the primary metric. Google Search Engine (SE) is chosen as the benchmark for comparison due to its efficiency in Arabic language information retrieval systems. Precision is calculated by dividing the number of relevant documents by the total number of retrieved documents. The precision values for each query are recorded per page on Google SE, allowing for the assessment of the approach's performance. Table 1 presents the applied queries related to the Politician domain, while Table 2 displays the precision values for each query in its original plain form and after being updated using the proposed approach. The results demonstrate that the proposed approach generally improves precision compared to the original plain queries. For instance, the precision for Q1 increases from 0.40 to 0.65, indicating a substantial improvement. Similarly, Q3 shows an increase in precision from 0.65 to 0.95 after applying the proposed approach.

These findings highlight the effectiveness of the approach in enhancing the retrieval of relevant documents in Arabic language information retrieval systems. The systematic process presented in this research provides valuable insights for improving the performance of such systems and optimizing the retrieval of information in the Arabic language. Future research can further refine and optimize the approach, explore its applicability to other domains, and address any remaining challenges to ensure its effectiveness in real-world scenarios.

6. CONCLUSION

In conclusion, this paper has presented an enhanced approach to Arabic information retrieval, specifically focusing on political search. The importance of accurate and relevant information in the domain of politics, particularly through Google searches, has been emphasized. The proposed approach addresses the challenges associated with the complexity of the Arabic language, offering a systematic process for improving information retrieval in Arabic language sources.

By employing tokenization, unification, removal of stop-words and special characters, light stemming using the "Khoja" stemmer, and term generation based on the "Sarf" mechanism, the approach enhances the retrieval of relevant documents.

The evaluation of the approach using precision as the primary metric, with Google Search Engine as the benchmark, demonstrates its effectiveness in improving precision compared to plain queries. The findings highlight the potential of the proposed approach to empower political search in Arabic language information retrieval systems.

The systematic process and insights provided in this research contribute to the advancement of information retrieval systems in the Arabic language. Future research can focus on further refining and optimizing the approach, exploring its applicability to other domains, and addressing any remaining challenges to ensure its effectiveness in real-world scenarios. Overall, the enhanced Arabic information retrieval approach presented in this paper holds promise for empowering political search and improving the accessibility of accurate and reliable political information in the Arabic language.

ACKNOWLEDGMENT

I would like to extend my sincere gratitude to the esteemed reviewers of this paper for their invaluable feedback and support. Their insightful comments and suggestions have greatly contributed to the improvement of this work. I would

also like to express my appreciation to the International Journal of Progressive Research in Engineering Management and Science for providing the platform to share this research. Furthermore, I would like to express my deepest gratitude to my family and friends for their unwavering support and encouragement throughout this endeavor. Their constant belief in me has been a source of inspiration and motivation.

7. REFERENCES

[1] Seung-Pyo Jun, Hyoung Sun Yoo, San Choi, Ten years of research change using Google Trends: From the perspective of big data utilizations and applications, Technological Forecasting and Social Change, 2018, vol. 130, pp. 69-87,

[2] Muhammed T, S., Mathew, S.K. The disaster of misinformation: a review of research in social media. Int J Data Sci Anal 13, 271–285 (2022). https://doi.org/10.1007/s41060-022-00311-6

[3] Reilly, Shauna & Richey, Sean & Taylor, J.. (2012). Using Google Search Data for State Politics Research An Empirical Validity Test Using Roll-Off Data. State Politics & Policy Quarterly. 12. 146-159.

[4] Ghwanmeh, Sameh & Kanan, Ghassan & Al-Shalabi, Riyad & Ababneh, Ahmad. (2007). Enhanced Arabic Information Retrieval System based on Arabic Text Classification. 461-465. 10.1109/IIT.2007.4430469.

[5] Sadik Bessou and Mohamed Touahria, An Accuracy-Enhanced Stemming Algorithm for Arabic Information Retrieval, Neural Network World, 2014, vol. 24, pp. 117-128

[6] Mustafa Abdel-Kareem Ababneh & Ghassan Kanaan & Ayat Amin Al-Jarrah, 2019. "Enhanced Arabic Information Retrieval by Using Arabic Slang Language," Modern Applied Science, Canadian Center of Science and Education, vol. 13(6), pp. 1-24, June.