Journal of Theoretical and Applied Information Technology
31st January 2025. Vol.103. No.2

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

UTILIZING WORD EMBEDDING'S FOR AUTOMATED QUERY EXPANSION IN ARABIC INFORMATION RETRIEVAL: A BLENDED METHODOLOGY

YASIR HADI FARHAN[1] , BOUMEDYEN SHANNAQ[2] , SAID AL MAQBALI[3] , , OUALID ALI[4], BASEL BANI-ISMAIL[5], MUSTAFA TAREQ ABD[6] , MOHANAAD SHAKIR[7]

[1] Department of Medical Physics, College of Applied Sciences, University of Fallujah, Iraq
[2,4]Management Information System Department, College of Business (CoB), University of Buraimi (UoB), Oman
[3]Information Technology Unit, University of Buraimi (UoB), Oman
[4]Computer Sciences Department, College of Arts & Science, Applied Science University, Manama, Bahrain
[5]Faculty of Information Technology, Majan University College, Al Buraimi, Oman
[6]Department of Computer Science, University of Technology-Iraq, Iraq

E-mail: [1]yasir.hadi@uoanbar.edu.iq ,[2]boumedyen@uob.edu.om

ABSTRACT

Search engines face a critical challenge in addressing query-document vocabulary mismatch during Information Retrieval (IR), when user queries do not match the document content. Automatic Query Expansion (AQE) has been widely used to mitigate this issue by identifying related terms. This research work presents a new hybrid AQE technique which includes DMNs within the BM25 model and two techniques, namely EQE1 & V2Q including DAN. It can be asserted that the hybrid technique is characterized by the optimality of the retrieval performances of the two networks, with reduced query drift. Experimental evaluation reveals that EQE1+(DANs+DMNs) obtained P@10 of 44.20LDilde MAP of 33.10 for the TREC 2001, whereas V2Q+(DANs+DMNs) obtained MAP of 30.40 and P@10 of 39.30. However the proposed method BM25+DMNs achieved the highest average MAP of 38.10 for TREC 2001 surpassing all the methods presented in this study. However, it is suggested that additional improvements employing the enhanced embeddings or fine-tuning of the hybrid solutions be implemented due to the drawbacks in expanding the query and positioning of vectors.
Keywords: *Automatic Query Expansion, Information Retrieval, Word Embedding, Deep Averaging Networks, Deep Median Networks*

1. INTRODUCTION

The first problem facing search engines is that the average query is only 2–3 words long [1–3]. This shortcoming stems from users' inability to articulate their informational needs optimally to allow search engines to meet those needs satisfactorily. Belkin's Anomalous State of Knowledge (ASK) hypothesis [4] has identified this problem due to lack of knowledge among the users. AQE techniques have been developed to overcome this problem with techniques such as expanding the user queries with related terms to enhance the capture of the intention of the user in the current document retrieval processes [5] &[6] . AQE is used in today's major search engine like Google, yahoo and other where auto complete features or suggested words are used [7].

There are two main clans of AQE techniques that is the global and the local. Global techniques use outside sources such as thesauri, of which WordNet is a popular and valuable source for finding related terms to a given word [8]. Local techniques on the other hand utilize relevance feedback, whereby using results from a particular retrieval in order to update the query with the most relevant terms [9, 10]. There was the Pseudo Relevance Feedback (PRF), a local technique that assumes the first set of

documents are relevant and extracts new working terms from them [3].

AQE techniques experience difficulties. International techniques tend to produce ambiguous terms derived from ontological resources such as WordNet and therefore require word disambiguation prior to their use in added keyword queries. PRF, a local technique, might be impacted with noise terms and multiple topics documents within top ranked results may also influence its performance [11]. Hence, while the techniques of AQE enhance precision and recall separately, it is challenging to optimize both for concomitantly [12].

Recent research has been directed towards the elimination of these drawbacks using semantic modeling techniques such as Word embedding (WE). 'WE' methods encode words as points in a dense, low dimensional space that encodes syntax and semantic relationship [13– 15]. There are architectures like Continuous Bag-of-Words (CBOW), and Skip-gram (SG) in the Word2Vec-tools that predicts words based on context. These embeddings allow AQE techniques to find terms semantically related for query expansion [15]. However, traditional WE-based AQE directly retrieves the candidate terms without taking into account the interaction and the correlation between the terms within the source query.

Some of the works that serve as recommendations for enhancing AQE effectiveness are modeling query semantics in the form of collective vocabulary terms. This approach makes it easier to have candidate terms nearly similar to the topics of the primary query. Another method is to use an image classifier for terms, Deep Averaging Networks (DANs) that averages the embedded word vectors to classify related terms comprehensively [24, 25].

However, research in WE has significantly advanced, little innovation of WE application in AIR can be identified. Indeed, Arabic has a rather rich morphology and relatively limited ontological potential, that makes it difficult to solve important NLP problems, such as AQE [16–18]. Previous work in AIR mostly deals with stop warding methods and arranging the documents according to their similar stemmed terms [19–23].

This work also proposes a novel hybrid AQE approach using Deep Median Networks (DMNs) [24] and DANs [25] with the help of sentence embedding. This method enriches AIR for Arabic documents by rectifying the problem of term match and the general retrieval performance. The proposed hybrid approach shows that it is possible to address specific difficulties of Arabic NLP by integrating advanced semantic modeling. This study uses sentiment analysis with deep learning models, including XLNet and BERT, on a novel Rotten Tomatoes dataset to summarize movie reviews effectively [26] . This study improving data analytics integration for strategic decision-making and organizational competitiveness. Such work could leveraging word embedding to enhance search engines [27].

This study detecting semantic errors with 95.6% accuracy in NLP text correction, and the results could uses word embedding and weighted federated learning to enhance search engines by [28].

This study leverages word embedding to enhance search engine accuracy, proposing a unified Database Security Meta-model (DBSM) for improved security understanding and solutions. The outcomes could be leverages word embedding to enhance search engine accuracy [29].leveraging word embedding for search engines could be benefit several application as investigated in the following studies , This study, systematically reviews information systems' impact on public sector performance, highlighting skill development, ROI, and strategic decision-making[30].

This study uses bibliometric analysis and word embedding to explore digital transformation trends in banking, highlighting FinTech, innovation, and technology adoption from 2009 to 2023[31].

This study uses bibliometric analysis to explore knowledge management in healthcare, highlighting trends, key themes, leading countries, and collaboration patterns from 1996 to 2023[32].

2. RELATED WORKS

Situations like the following often occur: Moreover, there are difficulties that include Vocabulary mismatch commonly [33] & [3]. To these, researchers have offered a number of solutions, of which Automatic Query Expansion (AQE) is one of the best known. AQE techniques alter the user queries by including useful terms making the IR systems more accurate and precise [34]. Out of all methods in AQE, Word Embedding (WE) has received considerable attention [13-15, 35].

2.1 Word Embedding in NLP and IR

WE stands for the language modelling and feature learning methodology that is a part of the Natural Language Processing paradigm which initiates semantic interpretation of the text. WE, for example, is a semantic vector space model in which words are real-numbered vectors in a given corpus. Two representation types are commonly used: Local representation based on the occurrence of a word and distributed representation, where the words in

similar context are nearer in the WE space [36] &[37]. For instance, the Arabic terms "ذكر" (male) and "انثى" (female) are contextually similar, and their vectors appear closer in the WE space.

2.2 Applications of Word Embedding in AQE

Researchers have explored various applications of WE in AQE to improve IR performance:

2.2.1 Local Embedding-based AQE

[15]proposed a novel AQE technique using local embeddings derived from retrieved documents. Using the Kullback–Leibler divergence language model, they re-ranked documents based on local embeddings. Experimental results showed that local embeddings outperformed global techniques in document retrieval.

2.2.2 High-quality Vector Representations

[14]compared AQE techniques by employing high-quality vector representations based on deep learning. Words like "driver" and "taxi" were represented by similar vectors, and cosine similarity measured the semantic distance between terms. Tests on four CLEF collections with skip-gram architecture demonstrated statistically significant improvements over baseline models like Mutual Information (MI) and Pseudo-Relevance Feedback (PRF)[38, 39].

2.2.3 Relevance Feedback Models

[13]proposed a relevance feedback (RF) model incorporating semantic relationships via embedded word vectors. By using Kernel Density Estimation (KDE) for weighted co-occurrence calculation, they established a framework for leveraging term compositionality in AQE. Experiments on TREC test collections showed the KDE-based RF approach outperformed traditional feedback models.

2.2.4 Word2Vec-basedAQE

[40]used Word2Vec to train Skip-gram and CBOW models on the TREC Washington Post Corpus. They implemented query reweighting strategies and measured term similarity using Euclidean distance. Results revealed that reweighting improved retrieval efficiency compared to assigning equal weights to expanded query terms[41].

2.2.5 Arabic Text Retrieval Using AQE

[42]integrated WE semantic similarities into various IR models, including Language Models (LM), Log-Logistic Distribution Model (LGD), Smoothed Power-Law Model (SPL), and Okapi BM25. Their approach addressed vocabulary mismatches using three neural WE processes—Skip-gram, CBOW, and GloVe. Strategies such as selecting similar words from retrieved texts or the complete corpus demonstrated better performance than baseline WE-based models[43-46].

2.3 Advances in Contextual AQE Strategies

While most AQE methods assume all query terms contribute positively, neglecting query context often leads to inefficiencies. Researchers have proposed context-aware solutions:

2.3.1 Query-Guided AQE (V2Q)

[12] introduced V2Q, which filters unnecessary terms from queries, thus avoiding detrimental expansion. By representing words using Word2Vec, they tested two Approaches: Query-Driven Association (Q2V) and Prospect-Driven Association (V2Q).V2Q outperformed Q2V by focusing only on terms with high semantic similarity to the query[32].

2.3.2 Embedding-based Query Expansion (EQE1)

[47]proposed EQE1, which leverages WE vectors to add semantically similar terms to queries. This technique assumes that query terms are conditionally independent, emphasizing semantic alignment between expanded and original terms. Experiments using TREC collections demonstrated that embedding-based approaches surpassed baseline methods like Maximum Likelihood Estimation (MLE) and heuristic-based query expansion (VEXP) [48] &[49].

2.3.3 Deep Averaging Networks (DANs)

[25]developed a new AQE method, DANs, which used the average vector of original query terms to determine candidates for expansion. By considering the overall query context, DANs avoided query drift—a common issue with term-specific expansions[50].

2.3.4 Limitations and Future Directions

While WE-based AQE methods have significantly improved IR systems, challenges remain. For instance, individual term-based expansions often neglect query-level context, leading to suboptimal retrieval. However, semantic relationships are sometimes very basic with little attention to the distinctions between the meanings. While methods like V2Q and DANs have resolved these problems to some extent, the incorporation of contextual and semantic information remains somewhat separate from the actual queries. The future research should expand to dynamic weighting mechanisms and implementation of statistical neural hybrid models for improvement of the expansions plans.

Implementation of WE in AQE ensures efficient translation to bridge vocabulary gaps and optimizes query reformulation making IR reinvent itself. Studies of engines such as local embeddings of

objects with high-quality vector representation or models based on a set of relevance feedback have revealed high usability of the developed approaches in different datasets and conditions. However, more context-sensitive, and semantically sophisticated expansion methods still constitute an important research topic. Further developments in this area will contribute to the increased efficiency of IR systems in terms of satisfying various users' requirements.

3. METHOD

3.1 Word2Vec and Hybrid Approach

Word2Vec is a deep learning toolkit presented by [39] which looks at a text corpus as input and provides word vectors as output. There matches a vocabulary by training data and analyses vector interpretations of words [51].

These are as follows: Continuous Bag of Words (CBOW) and Skip-Gram (SG). The idea is to give a middle word in cases of CBOW model, or a context in cases of SG model, by simply taking a distributed representation of the surrounding words or a representation of an input word.

There is a proposed approach for getting the candidate vectors of DANs, which are AQE techniques [25]. Nevertheless, the location of query term vectors matters in DANs, for which DMNs are employed to build candidate expansion sets [24]. They increase the performance of AQE by finding the median vector of the original query term vectors as well as using it to derive expansion vectors. This research therefore proposes the use of a combined DAN and DMNs to improve on the Arabic Information Retrieval (AIR). The AQE technique works in synergy with probabilistic models such as BM25, EQE1, and V2Q in advancing performance. In other respects, DMNs use sentence embedding to get text vector representation and optimization in datasets contain syntactic variability like Arabic text. Researchers evaluated the DMNs approach on AQE by integrating it into three IR models: BM25, EQE1, and V2Q. Corpus used by them was the Arabic TREC 2001/2002 news articles dataset of the year 1994-2000 .The study undertaken were with 25, 50 and 75 queries over subsets. The Arabic TREC 2001/2002 is the available benchmark for Arabic IR research till date [3].

Table 1. Statistics for Arabic T REC Collections

Collections	TREC 2001	TREC 2002	TREC 2001/2002
Number of queries	25	50	75
The average number of words/queries	4.88	3.28	4.08
Number of documents	383,872		
Number of tokens	76 million		
Number of unique words	666,094		
Size (compressed)	209 MB		
Size (uncompressed)	869 MB		

3.2 Basic Query Expansion based on DMNs (BM25+DMNs)

Derived from the probabilistic framework, the Okapi BM25 model is superior to the rest of the algorithms in term weighting based on Term Frequency, Inverse Document Frequency, and Document Length. Equations 1 and 2 provide the definitions of its scoring functions

$$\sum_{i \in Q} log \frac{\frac{(r_i + 0.5)}{R - r_i + 0.5}}{\frac{(n_i - r_i + 0.5)}{(N - n_i - R + r_i + 0.5)}} \cdot \frac{(k_1 + 1)f}{K + f_i} \quad (1)$$

$$K = k_1 \left((1 - b) + b \cdot \frac{dl}{avdl} \right) \quad (2)$$

Parameters k1 , k1, and K are given ,where k1 = 1.2 , k2=0 to 1000 and b=0.75 is used , qf indicates the frequency of query term , dl refers to document length and avdl refers to average document length of the collection. Equation 3 describes cosine similarity and, Equation 4 depicts another model of Word2Vec known as CBOW which computes target words from surrounding context, where |C| represent corpus size and c is the size of a dynamic window.

$$Cosine(A, B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2} \times \sqrt{\sum_{i=1}^{n} {B_i}^2}} \quad (3)$$

$$\frac{1}{|C|}\sum_{t=1}^{|C|}\log[\log[P(wt\backslash wt-c,\ \ldots,\ wt-1,\ wt+1,\ ..,\ wt+c)]] \quad (4)$$

$$m(\overrightarrow{X_k}) = \vec{X}\left[\frac{n}{2}\right] + \vec{X}\left[\frac{n+1}{2}\right] \quad (5)$$

Equation number five is used to find the median vector of the original query term vectors, where Xk represents the arranged set of vectors corresponding to a specific query, where n denotes the vector dimension, typically set to 300.

Google News and Similar Document Analysis Technique: Expansion of the Query using DANs and DMNs

To identify the central vector among the initial query term vectors, Equation (5) is applied:

m(vk) = Median of the vectors Xk where Xk={x1,x2…xin} and in = 300 m(vk) = Median of the vectors Xk where Xk={x1,x2…x in } and in = 300 . This work proposed a new method combining both query expansion with DANs and Median Networks query expansion giving it the name of DMN-DANs. First incorporated into the BM25 probabilistic model, it is referred to as BM25+(DANs+DMNs), comprising several steps:

1. In this case calculate the average vector avg(vk)avg(vk) of original query terms based on DANs.
2. Calculate the median vector m(vk)m(vk) of DMNs.
3. Identify the most similar vectors in the Word Embedding (WE) corpus to both avg(vk)avg(*vk*) and m(vk)*m*(*vk*), creating two sets W={w1,w2,…,wk}*W*={*w*1,*w*2,…,*wk*} and V={v1,v2,…,vk}*V*={*v*1,*v*2,…,*vk*}.
4. Calculate similarities of vectors in W*W* and V*V* to avg(vk)avg(*vk*) and m(vk)*m*(*vk*), selecting vectors with similarity ≥0.7≥0.7, a threshold supported by [33, 46].
5. Retrieve words corresponding to the selected vectors using the WE model, incorporate them into the original query, and retrieve documents.

The hybrid nature lies in leveraging both **DANs** and **DMNs** candidate sets, improving **Automatic**

3.3 Hybrid Expansion with EQE1 and V2Q

EQE1, proposed by Zamani and Croft [40], identifies candidate vectors similar to all query terms. To reduce complexity, this study compares candidate vectors against avg(vk)avg(*vk*) and m(vk)*m*(*vk*), selecting those with ≥0.7≥0.7 similarity. After adding the corresponding words to the query, new documents are retrieved.

V2Q+(DANs+DMNs) incorporates prospect guidance with hybrid expansion. Candidate vectors are selected using similarity measures against avg(vk)avg(*vk*) and m(vk)*m*(*vk*), producing sets W*W*, V*V*, and D*D*. Words from vectors meeting the threshold are added to the query. This hybrid method optimizes query expansion, utilizing the strengths of both techniques.

4. EXPERIMENTS

4.1 Experimental Setup:

The hybrid AQE technique was tested using the **TREC 2001/2002 Arabic newswire dataset**, widely used in Arabic text retrieval [22, 43, 52]. This dataset includes 2,117 documents spanning **May 1994–December 2000**. The **Word2Vec process** trained the WE corpus using **383,872 Arabic articles** from Agence France Presse, totaling over 1 GB after UTF-8 encoding. Three query collections were evaluated: **TREC 2001** (25 queries), **TREC 2002** (50 queries), and their combination (**TREC 2001/2002**) with 75 queries.

For stemming, the **Farasa stemmer** [22, 25, 46] was employed due to its effectiveness in Arabic. Retrieval was limited to the top 100 documents.

4.2 Evaluation Baselines

Five baselines were compared:

1. **BM25**: A probabilistic model without expansion.
2. **DANs** [25]: Using averaged vectors for expansion.
3. **DMNs** [24]: Employing median vectors.
4. **EQE1** [47]: Embedding-based query expansion.
5. **V2Q** [12]: Prospect-guided query expansion.

Experiments were conducted using the **Whoosh search engine** [52], with semantic similarity measured via cosine similarity.

4.3 Evaluation Metrics

The study evaluates retrieval effectiveness using the following metrics:

1. **Mean Average Precision (MAP)**: Applied to the top 100 documents, reflecting overall retrieval precision.
2. **P@10**: Precision at the top 10 retrieved documents.
3. **Recall-Precision Curves**: Visualizing performance across standard recall levels.

The equations for these metrics are shown below:

$$Recall = \frac{|Rel| \cap |Ret|}{|Rel|} \quad (6)$$

$$Precision = \frac{|Rel| \cap |Ret|}{|Ret|} \quad (7)$$

$$MAP = \frac{1}{N}\sum_{i=1}^{n} AP_i \quad (8)$$

Where:

Ret is the total number of retrieved documents,

Rel is the total number of relevant documents in the dataset, and

$AP = \sum_{i=1}^{n} Precision_i \cdot \Delta Recall_i$.

Where $\boldsymbol{i}$ is the rank in the sequence of retrieved documents, $\boldsymbol{n}$ is the number of retrieved documents, $\boldsymbol{Precision_i}$ is the precision at cut-off $\boldsymbol{i}$ in the list, and $\Delta\boldsymbol{Recall_i}$ is the change in recall from items $\boldsymbol{i-1}$ to $\boldsymbol{i}$.

5. RESULTS AND DISCUSSION

The following tables present precision at 10 (P@10) and MAP values for the proposed AQE methods and baseline approaches, evaluated on TREC 2001 (25 queries), TREC 2002 (50 queries), and TREC 2001/2002 (75 queries) collections. The experiments used the same dataset corpus, considering the top 100 retrieved documents.

The hybrid AQE method (DANs+DMNs) was applied to the BM25 framework, while two expanded approaches, EQE1 and V2Q, were enhanced using (DANs+DMNs) to identify candidate expansion terms. These modified methods, EQE1+(DANs+DMNs) and V2Q+(DANs+DMNs), were hypothesized to recommend better expansion terms and improve AIR systems' retrieval performance.

Comparisons were made against BM25 (no expansion), BM25+DANs, BM25+DMNs, EQE1, EQE1+DANs, EQE1+DMNs, V2Q, V2Q+DANs, and V2Q+DMNs. Table 2 summarizes the results in terms of MAP and P@10.

Table 2. MAP and P@10 for all the models

Techniques	Collections					
	TREC 2001		TREC 2002		TREC 2001/2002	
	MAP	**P@10**	**MAP**	**P@10**	**MAP**	**P@10**
BM25	31.30	42.10	28.70	35.80	29.50	37.90
BM25+DANs	19.40	26.20	28.50	39.60	25.40	35.10
BM25+DMNs	38.10	38.10	33.20	33.20	34.90	34.90
BM25+(DANs+DMNs)	20.00	27.30	17.10	19.30	18.10	22.00
EQE1	30.70	42.90	25.70	33.50	26.90	30.40
EQE1+DANs	32.60	43.40	34.30	43.50	33.70	43.50
EQE1+DMNs	37.20	37.20	37.00	37.00	37.10	37.10
EQE1+(DANs+DMNs)	33.10	44.20	28.60	36.50	30.10	39.10
V2Q	27.00	35.20	26.20	33.30	26.50	33.90
V2Q+DANs	30.20	38.90	32.90	43.90	32.00	42.30
V2Q+DMNs	36.00	36.00	35.60	35.60	35.70	35.70
V2Q+(DANs+DMNs)	30.40	39.30	26.30	33.80	27.70	35.70

The experimental outcomes in Table 2 reveal that the hybrid (DANs+DMNs) technique integrated with the probabilistic BM25 model outperforms the BM25+DANs approach in MAP and P@10 for TREC 2001 but falls short against other baseline methods across all TREC collections. The hybrid EQE1+(DANs+DMNs) technique demonstrates superior performance over the baseline EQE1 in MAP and P@10 for TREC 2001, TREC 2002, and TREC 2001/2002 collections. It also outperforms EQE1+DANs in MAP and P@10 but only for TREC 2001. Similarly, the hybrid V2Q+(DANs+DMNs) surpasses its baseline V2Q in

MAP and P@10 across all TREC collections and exceeds V2Q+DANs for TREC 2001 only.

Before considering whether the outcomes meet the standard of performance laid down in literature, it is important to think about what that means and what the benchmarks established in comparable studies need. For information retrieval processes investigating TREC datasets, it is believed that values that surpass 35 on the MAP and approximate or surpass 40 on the P@10 are capable of penetrating the competition.

For example:

Baseline BM25: MAP values like 31.30 (TREC 2001) and 28.70's (TREC 2002) are quite normal and represent the baseline scores that many experiments use. Nevertheless, the BM25 + DANs + DMNs, with the MAP of 20.00 are lower, hence the drawbacks in vector based expansion strategies need improvement.

EQE1 and V2Q with Extensions: With EQE1+(DANs+DMNs), a MAP of 33.10 (TREC 2001) and P@10 of 44.20 showing it is better than EQE alone 30.70 MAP, 42.90 P@10 relative to the experimental results demonstrate certain enhancements in some cases. Still these figures are small if compared with state of the art hybrid models based on BERT or neural retrieval.

DMNs: MAPs achieved by BM25+DMNs are consistently high, 38.10 on the average on TREC 2001 and hence in sync with perfect performing models in analogous retrieval exercises.

In total, some methods (BM25+DMNs and others) work as or better than literature values, while BM25+(DANs+DMNs) do worse compared to benchmarks. For competitive results in future work, different more sophisticated embeddings such as contextualized models can be applied or the hybrid approaches applied in the study can be further optimized for more effective QA in terms of query expansion and the positioning of the vectors.

6. CONCLUSION AND FUTURE WORKS

To enhance AIR performance, this study proposed a hybrid technique combining DANs and DMNs, leveraging their strengths to generate candidate expansion terms. Word2Vec was utilized for word embeddings, and comparisons were made with BM25, EQE1, V2Q, and their extensions. The hybrid method integrates the median vector from DMNs and the average vector from DANs to address vocabulary mismatch and improve retrieval performance.

The findings demonstrate that the proposed hybrid technique often outperforms baselines like BM25, V2Q, and EQE1 in MAP and P@10 but is limited by vector positioning in the embedding space, similar to DANs. This suggests that while the hybrid approach captures semantic similarity effectively, it remains affected by spatial constraints in the vector space.

Future research should develop techniques that are less influenced by vector positioning, focusing instead on enhancing the semantic representation of query terms. This could enable the retrieval of more relevant expansion terms, further improving the effectiveness of AIR systems.

ACKNOWLEDGMENT

We sincerely thank the University of Buraimi for their generous financial support in funding this research article. Their contribution has been instrumental in the successful completion of this work.

REFERENCES

[1] H. K. Azad and A. Deepak, "Query Expansion Techniques For Information Retrieval: A Survey," Information Processing & Management, vol. 56, no. 5, pp. 1698-1735, 2019.

[2] B. Shannaq, "Enhancing Human-Computer Interaction: An Interactive and Automotive Web Application - Digital Associative Tool for Improving Formulating Search Queries," in Advances in Information and Communication, vol. 921, K. Arai, Ed., in Lecture Notes in Networks and Systems, vol. 921. , Cham: Springer Nature Switzerland, 2024, pp. 511–523. doi: 10.1007/978-3-031-54053-0_35.

[3] H. Farhan Yasir, M. Noah Shahrul Azman, and M. Mohd, "Survey of Automatic Query Expansion for Arabic Text Retrieval," (in En), Journal of Information Science Theory and Practice, vol. 8, no. 4, pp. 67-86, 12/30 2020.

[4] Boumedyen Shannaq*, "Investigating the Distribution of Arabic and English Keywords and Their Progress Over Different Text File Formats," American Journal of Computing Research Repository, vol. 1, no. 1, Art. no. 1, 2013, doi: 10.12691/ajcrr-1-1-1.

[5] M. A. Raza, R. Mokhtar, N. Ahmad, M. Pasha, and U. Pasha, "A Taxonomy And Survey Of

Semantic Approaches For Query Expansion," IEEE Access, vol. 7, pp. 17823-17833, 2019.
[6] M. Esposito, E. Damiano, A. Minutolo, G. De Pietro, and H. Fujita, "Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering," Information Sciences, vol. 514, pp. 88-105, 2020.
[7] F. Cai and M. De Rijke, "A Survey Of Query Auto Completion In Information Retrieval," Foundations And Trends® In Information Retrieval, vol. 10, no. 4, pp. 273-363, 2016.
[8] D. Pal, M. Mitra, and K. Datta, "Improving Query Expansion Using Wordnet," Journal of the Association for Information Science and Technology, vol. 65, no. 12, pp. 2469-2478, 2014.
[9] B. Shannaq, R. Adebiaye, T. Owusu, and A. Al-Zeidi, "An intelligent online human-computer interaction tool for adapting educational content to diverse learning capabilities across Arab cultures: Challenges and strategies," J. Infras. Policy. Dev., vol. 8, no. 9, p. 7172, Sep. 2024, doi: 10.24294/jipd.v8i9.7172.
[10] K. P. K. RA, N. A. Andreeva, K. V. Frolov, and B. Shannaq, "The info logical approach to develop edutainment systems," St. Petersburg institute for Informatics and Automation of Russian RAS, Academy of Sciences, vol. 199178, Accessed: Jul. 01, 2024. [Online]. Available: https://scholar.google.com/scholar?cluster=5299013294889662098&hl=en&oi=scholarr.
[11] S Boumedyen Shannaq* and F. AlAzzawi, "On the Development of Novel Arabic Documents Classifier," International Journal of Advanced Science and Technology, vol. 29, no. 3, Art. no. 3, 2020.
[12] F. C. Fernández-Reyes, J. Hermosillo-Valadez, and M. Montes-y-Gómez, "A Prospect-Guided Global Query Expansion Strategy Using Word Embeddings," Information Processing & Management, vol. 54, no. 1, pp. 1-13, 2018.
[13] D. Roy, D. Paul, M. Mitra, and U. Garain, "Using Word Embeddings for Automatic Query Expansion," CoRR, vol. abs/1606.07608, / 2016.
[14] M. ALMasri, C. Berrut, and J.-P. Chevallet, "A Comparison Of Deep Learning Based Query Expansion With Pseudo-Relevance Feedback And Mutual Information," in European Conference On Information Retrieval, Italy, 2016, pp. 709-715: Springer.
[15] F. Diaz, B. Mitra, and N. Craswell, "Query Expansion with Locally-Trained Word Embeddings," CoRR, vol. abs/1605.07891, / 2016.
[16] K. Alsmearat, M. Al-Ayyoub, and R. Al-Shalabi, "An Extensive Study Of The Bag-Of-Words Approach For Gender Identification Of Arabic Articles," in 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), Aqaba, Jordan, 2014, pp. 601-608: IEEE.
[17] B. Shannaq, "Improving security in intelligent systems: how effective are machine learning models with tf- idf vectorization for password-based user classification," vol. 102, nov. 2024.
[18] G. Mohsen, M. Al-Ayyoub, I. Hmeidi, and A. Al-Aiad, "On The Automatic Construction Of An Arabic Thesaurus," in 2018 9th International Conference On Information And Communication Systems (ICICS), Amman, Jordan, 2018, pp. 243-247: IEEE.
[19] B. Shannaq, O. Ali, S. A. Maqbali, and A. Al-Zeidi, "Advancing user classification models: A comparative analysis of machine learning approaches to enhance faculty password policies at the University of Buraimi," J. Infras. Policy. Dev., vol. 8, no. 13, p. 9311, Nov. 2024, doi: 10.24294/jipd9311.
[20] B. Shannaq, D. K. Muniyanayaka, O. Ali, B. Bani-Ismail, and S. Al Maqbali, "Exploring the role of machine learning models in risk assessment models for developed organizations' management decision policies," J. Infras. Policy. Dev., vol. 8, no. 13, p. 9364, Nov. 2024, doi: 10.24294/jipd9364.
[21] M. Mustafa, H. AbdAlla, and H. Suleman, "Current Approaches In Arabic Ir: A Survey," in International Conference on Asian Digital Libraries, Hyderabad, India, 2008, pp. 406-407: Springer.
[22] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, "Farasa: A Fast And Furious Segmenter For Arabic," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, California, 2016, pp. 11-16: Association for Computational Linguistics.
[23] S. B. Guirat, I. Bounhas, and Y. Slimani, "Combining Indexing Units For Arabic Information Retrieval," International Journal of Software Innovation (IJSI), vol. 4, no. 4, pp. 1-14, 2016.
[24] F. Yasir Hadi, S. Mohanaad, T. Mustafa Abd, and S. Boumedyen, "Incorporating Deep

Median Networks for Arabic Document Retrieval Using Word Embeddings-Based Query Expansion," JOURNAL OF INFORMATION SCIENCE THEORY AND PRACTICE, vol. 12, no. 3, pp. 36-48, 2024.

[25] Y. H. Farhan, S. A. Mohd Noah, M. Mohd, and J. Atwan, "Word-embedding-based query expansion: Incorporating Deep Averaging Networks in Arabic document retrieval," Journal of Information Science, vol. 49, no. 5, pp. 1168-1186, 2021.

[26] Department of Computer and Software Technology, University of Swat, Swat, Pakistan, S. S. Khan, Department of Computer Science, IQRA National University, Swat, Pakistan, Y. Alharbi, and College of Computer Science and Engineering, University of Hail, Hail, Saudi Arabia, "Sentiment analysis of movie review classifications using deep learning approaches," Int. j. adv. appl. sci., vol. 11, no. 8, pp. 146–157, Aug. 2024, doi: 10.21833/ijaas.2024.08.016.

[27] College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia and M. Ramzan, "The significance and application of data analytics models for strategic management," Int. j. adv. appl. sci., vol. 11, no. 1, pp. 87–94, Jan. 2024, doi: 10.21833/ijaas.2024.01.010.

[28] School of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan et al., "Optimizing semantic error detection through weighted federated machine learning: A comprehensive approach," Int. j. adv. appl. sci., vol. 11, no. 1, pp. 150–160, Jan. 2024, doi: 10.21833/ijaas.2024.01.018.

[29] Department of Computer Sciences, Faculty of Computing and Information Technology, Northern Border University, Rafha, Saudi Arabia and A. Alshammari, "Structuring and organizing database security domain from big data perspective using meta-modeling approach," Int. j. adv. appl. sci., vol. 11, no. 2, pp. 180–194, Feb. 2024, doi: 10.21833/ijaas.2024.02.019.

[30] Centre for Research in Development, Social and Environment (SEEDS), Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia, M. K. M. Kasim, A. Awang, Centre for Research in Development, Social and Environment (SEEDS), Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia, M. Jaafar, and Centre for Research in Development, Social and Environment (SEEDS), Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia, "A systematic literature review on the effect of information systems on the performance of government officials," Int. j. adv. appl. sci., vol. 11, no. 3, pp. 46–54, Mar. 2024, doi: 10.21833/ijaas.2024.03.006.

[31] Department of Commerce, School of Social Sciences and Languages, Vellore Institute of Technology, Vellore, India, B. Lavanya, A. Dunstan Rajkumar, and Department of Commerce, School of Social Sciences and Languages, Vellore Institute of Technology, Vellore, India, "Paradigm shift in the digital transformation of the banking sector: A bibliometric analysis," Int. j. adv. appl. sci., vol. 11, no. 3, pp. 115–126, Mar. 2024, doi: 10.21833/ijaas.2024.03.013.

[32] Faculty of Economics, Commerce and Management Sciences, University of Oum El Bouaghi, Oum El Bouaghi, Algeria, F. Yahiaoui, K. Chergui, Faculty of Economics, Commerce and Management Sciences, University of Oum El Bouaghi, Oum El Bouaghi, Algeria, R. Aichouche, and Faculty of Economics, Commerce and Management Sciences, University of Oum El Bouaghi, Oum El Bouaghi, Algeria, "Mapping the growth and trends of knowledge management in health organizations: A bibliometric exploration," Int. j. adv. appl. sci., vol. 11, no. 3, pp. 158–174, Mar. 2024, doi: 10.21833/ijaas.2024.03.017.

[33] Boumedyen Shannaq, "Novel Algorithm for Differentiating Authorized Users from Fraudsters by Analyzing Mobile Keypad Input Patterns during Password Updates," TEM Journal, vol. 13, no. 4.

[34] A. Abbache, F. Meziane, G. Belalem, and F. Z. Belkredim, "Arabic Query Expansion Using Wordnet And Association Rules," in Information Retrieval And Management: Concepts, Methodologies, Tools, And Applications: IGI Global, 2016, pp. 1239-1254.

[35] A. El Mahdaouy, S. O. El Alaoui, and E. Gaussier, "Word-Embedding-Based Pseudo-Relevance Feedback For Arabic Information Retrieval," Journal of Information Science, p. 0165551518792210, 2018.

[36] H. K. Kim, H. Kim, and S. Cho, "Bag-Of-Concepts: Comprehending Document Representation Through Clustering Words In Distributed Representation," Neurocomputing, vol. 266, pp. 336-352, 2017.

[37] B. Shannaq and A. Al-Zeidi, “Intelligent Information System: Leveraging AI and Machine Learning for University Course Registration and Academic Performance Enhancement in Educational Systems,” in Achieving Sustainable Business Through AI, Technology Education and Computer Science, vol. 159, A. Hamdan, Ed., in Studies in Big Data, vol. 159. , Cham: Springer Nature Switzerland, 2024, pp. 51–65. doi: 10.1007/978-3-031-71213-5_5.

[38] C. D. Manning, P. Raghavan, and H. Schütze, Introduction To Information Retrieval. Cambridge University Press Cambridge, England: Cambridge university press, 2008.

[39] B. Aklouche, I. Bounhas, and Y. Slimani, "Query Expansion Based on NLP and Word Embeddings," in TREC, 2018.

[40] B. Shannaq, “Unveiling the Nexus: Exploring TAM Components Influencing Professors’ Satisfaction With Smartphone Integration in Lectures: A Case Study From Oman,” TEM Journal, pp. 2365–2375, Aug. 2024, doi: 10.18421/TEM133-63.

[41] A. El Mahdaouy, S. O. El Alaoui, and E. Gaussier, "Improving Arabic Information Retrieval Using Word Embedding Similarities," International Journal of Speech Technology, vol. 21, no. 1, pp. 121-136, 2018.

[42] H. Fang and C. Zhai, "Semantic Term Matching In Axiomatic Approaches To Information Retrieval," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle Washington USA, 2006, pp. 115-122: ACM.

[43] G. Zuccon, B. Koopman, P. Bruza, and L. Azzopardi, "Integrating And Evaluating Neural Word Embeddings In Information Retrieval," in Proceedings Of The 20th Australasian Document Computing Symposium, 2015, pp. 1-8.

[44] D. Ganguly, D. Roy, M. Mitra, and G. J. Jones, "Word Embedding Based Generalized Language Model For Information Retrieval," in Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, Santiago, Chile, 2015, pp. 795-798.

[45] B. Shannaq and M. Shakir, “Enhancing Security with Multi-Factor User Behavior Identification Via Longest Common Subsequence Analysis,” Informatica, vol. 48, no. 16, pp. 73–82, 2024, doi: https://doi.org/10.31449/inf.v48i19.6529.

[46] V. Lavrenko and W. B. Croft, "Relevance-Based Language Models," in ACM SIGIR Forum, New York,, 2017, vol. 51, no. 2, pp. 260-267: ACM.

[47] R. Crimp and A. Trotman, "Refining Query Expansion Terms using Query Context," in Proceedings of the 23rd Australasian Document Computing Symposium, Melbourne, Australia, 2018, p. 12: Association for Computing Machinery (ACM).

[47] Y. H. Farhan, M. Mohd, and S. A. M. Noah, "Survey of Automatic Query Expansion for Arabic Text Retrieval," JISTaP, vol. 8, no. 4, pp. 67-86, 2020.

[48] Y. H. Farhan, S. A. M. Noah, M. Mohd, and J. Atwan, "Word Embeddings-Based Pseudo Relevance Feedback Using Deep Averaging Networks for Arabic Document Retrieval," JISTaP, vol. 9, no. 2, pp. 1-17, 2021.

[49] S B. Shannaq, M. A. Talab, M. Shakir, M. T. Sheker, and A. M. Farhan, “Machine learning model for managing the insider attacks in big data,” presented At The The Second International Conference On Emerging Technology Trends In Internet Of Things And Computing, Ramadi, Iraq, 2023, p. 020013. doi: 10.1063/5.0188358.

[50] B. Shannaq*, Dr. I. R. A. Shamsi, and Dr. F. J. I. AlAzzawi, “Innovative Algorithm for Managing the Number of Clusters,” IJRTE, vol. 8, no. 5, Art. no. 5, Jan. 2020, doi: 10.35940/ijrte.E4875.018520..

[51] B. Shannaq, I. Al Shamsi, and S. Abdul Majeed, “Management Information System for Predicting Quantity Martial’s,” TEM Journal, vol. 8, pp. 1143–1149, Dec. 2019, doi: 10.18421/TEM84-06.

[52] S. Mukherjee and N. Kumar, "Duplicate Question Management and Answer Verification System," in 2019 IEEE Tenth International Conference on Technology for Education (T4E), Bhubaneswar, India, 2019, pp. 266-267: IEEE.