Journal of Theoretical and Applied Information Technology
31st January 2019. Vol.97. No 2

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

AQAS: ARABIC QUESTION ANSWERING SYSTEM BASED ON SVM, SVD, and LSI

[1] MOY'AWIAH A. AL-SHANNAQ, [2] KHALID M.O. NAHAR, [3] KHALDOUN MK.H HALAWANI
[1, 2] Computer Science Department, Faculty of Information Technology & Computer Sciences, Yarmouk University, Irbid-Jordan
[3] Computer Science Department, Faculty of Information Technology, Hebron University, Hebron-Palestine
E-mails: [1] moyawiah.s@yu.edu.jo, [2]khalids@yu.edu.jo, [3] kldoon@hebron.edu

ABSTRACT

In alliance to the emerging internet technologies and services, the field of questions answering was one of the most trending topics. It is being used in multiple applications ranging from search engines to smart and complicated home assistance devices. In this paper, we are proposing an enhanced method and system for question answering that serve Arabic language questions. This system provides accurate paragraph level answers that extract its information out of documents dataset in different fields. The proposed system uses Support Vector Machine (SVM), Single Value Decomposition (SVD), and Latent Semantic Index (LSI) to classify the query in two phases. The method has been tested on a set of queries in different fields (classes) against a documents dataset of size 10,000 documents in 10 classes. The testing shows promising and accurate output for each of the test cases. Average classification accuracy reaches 98% using document classification metrics.

Keywords: *Support Vector Machine (SVM), Latent Semantic Index (LSI), Question Answering System (QAS), Arabic Language, SVD*

1.0 INTRODUCTION

Question answering systems are special systems designed in a way that they are capable of answering questions from different languages. The systems use natural language processing and information retrieval mechanisms to generate the desired answers from the question input. Question answering systems go one step further by analysing the data related to the questions before processing an answer. Arabic question answering systems are some of the most popular systems used today. This is supported by the fact that Arabic is the 6th most popular language in the world today with over 350 million speakers [1]. The popularity of Arabic question answering systems is further supported by the fact that the internet has seen a lot of Arabic content being provided on the web.

Most of the question answering systems work using two major principles; information retrieval and natural language processing. This implies that the system can retrieve the information from the web to generate the required answer [2]. On the other hand, some systems have a knowledge base that is constructed using semantics for the generation of the required answers [3]. Additionally, the process of answering an Arabic question can be segmented into three distinct phases; analysis of the question, passage retrieval, and the extraction of an answer. Question answering systems handle various types of questions based on the structure and nature of the data required for the extraction of an answer [4]. The questions can be factoid which can be answered by a simple word answer or open-ended questions that require the retrieval of numerous amounts of information to generate the required answer.

Many systems have been developed with the aim of improving the question answering paradigm. The systems might vary with the structure and mode of operation, but the aim is to provide a very accurate answer as possible [5]. The purpose of this paper is to propose question answering systems specifically targeting the Arabic language. The proposed system not only retrieve the relative documents, but also refer to target paragraph inside the document that might have the answer.

Past research have concentrated on enhancing modules of the QA pipeline, for example, question preparing, data recovery, data extraction. Ongoing work on printed entailment has appeared on QA

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

results, when utilized for sifting and positioning answer

Albeit unlabelled information investigation through probabilistic graphical models can enhance data extraction, conceiving an apparatus with reasonable generative models for the given common language undertaking is a test. This work assists with such comprehension by means of broad recreations and advances and affirms a speculation clarifying the hypothesis behind the impact of unsupervised pre-preparing for the last discriminant learning errand [6].

The rest of our paper includes: the related work reviewed in section 2, our methodology showed in section 3, the results and discussion are presented in section 4, and we conclude our paper in section 5. Finally, all referred references are stated in the reference section.

2.0 RELATED WORKS

Bassam, Hani, and Lytinen have proposed the adoption of a question answering system referred to as the QARAB [7]. The system intakes, natural queries and articulates them in the Arabic language with the aim of getting short and accurate answers. The primary source of knowledge is an Arabic newspaper text obtained from Al-Raya, Qatar. Its source of knowledge is the traditional Information Retrieval method combined with Natural Language Processing (NLP).

QARAB aim is to recognize text passages that respond to natural language queries [7]. A summary of the task would be as follows: provided with some queries in Arabic, give answers to the questions rooted in these expectations. First, the feedback should not extend through documents. Secondly, it should originate from the Arabic newspaper text obtained from Qatar's Al-Raya journal. QARAB's QA processing has three steps [7]. First, it processes the query. Secondly, it gets the documents possessing responses from the IR system. Lastly, it analyses the documents in a similar manner it processes the inquiries and showing sentences that may possess the answer.

The structure of QARAB contains the IR system and the NLP System. The IR system is a chip off the Salton's vector model. Initially, text obtained from the Al-Raya newspaper is processed to achieve an inverted file system that has answers to the questions provided. The role of the IR system is to retrieve documents with information essential to the question [7].

Omar, Lamia, and Paolo are concerned with the scarcity of QA system for the Arabic language [8]. Thus, they are proposing the Arabic Definition Question Answering System (DefArabicQA). It is rooted in the pattern approach to recognizing accurate and exact data about an organization using the internet. The approach used uses a linguistic analysis without language comprehension capacity. DefArabicQA recognizes candidate definition with the assistance of lexical patterns, the heuristic rules filter candidate definitions and uses the statistical approach to rank them. When using Google as the web source to answer 50 questions, about 54% of the questions are answered. While using Wikipedia and Google as the Web sources to answer the 50 questions, about 64% of the questions are answered. However, several words are excluded in the definition answer since the snippet is truncated.

Sman and Maryam have proposed the Arabic Question-Answering system referred to as the AQuASys (Arabic Question Answering System) [9]. AQuASysis designed to assist users to pose queries in the Arabic language to retrieve accurate answers in the same language. The system responds to queries linked to named entity of any kind; quantity, organization, person, time, location, and many more. Thus, the system inputs a query commencing with how, when, where, what, and who. Nonetheless, posing queries in Arabic interrogative form results in the extension of questioning nouns playing a similar role as the interrogative nouns preferred by the already developed Arabic QA systems. The performance of AQuASys is measured over several questions offered by native Arabic speakers in the testing stage. The architecture of AQuASys, which is composed of sentence filtering, query analysis, and ranking modules determines the accuracy of feedback provided.

Yassine, Abdelouahid, and Paolo [10] are concerned that most facets of the QA systems are language-reliance. Therefore, when building the system, the target language peculiarities should be put into consideration. To this end, they have proposed the Arabic QA system (ArabiQA). ArabiQA comprises of several structures. First is the question assessment module that determines the type of question and the relevant keywords. The passage retrieval model estimates the most accurate answers while the answer extraction module depicts the relevant answers. The test-set for the system has a total precision of 83%, which implies that it is an efficient approach for accurate extraction of answers to the factoid questions. The accuracy of the entire QA system is not recognized as it lags in the implementation stage [10]. It demands further improvements to provide answers to more complex questions than the factoid ones.

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

Bouma et al [11] have proposed the CLEF for the English and Arabic QA tasks. The system is greatly dependent on syntactic information. CLEF has greatly advanced with the inclusion of two innovations. First, Wikipedia was added as its document extension, an internet encyclopaedia presented in varying languages. The XML files are preprocessed to index the document collection for more accurate information retrieval. Lastly, the essential plain text is extracted and parsed. Secondly, the test queries are varied in topics [11]. To a certain topic, queries may presuppose or refer to information from former responses or questions to the presented questions. An anaphora resolution system is created to recognize the anaphoric elements. In addition, it identifies the desirable antecedent in the topic's first question or answers. Lastly, the information retrieval facet has also been improved. The question expansion rooted to blind significant feedback and synonym-lists boosts the IR module's reciprocal rank. It has a query classification module that uses the question class dedicated to the English source queries and automatically translated Arabic questions.

Finally, the internet is a huge source of knowledge. Thus, it becomes complex to identify accurate information. The recent search engines only provide effective answers rather than the exact answer to the query user [11]. To this end, Question Answering (QA) systems provide exact and effective answers to any question asked in the natural or native language. The QA systems mentioned above, such as QARAB, ArabiQA, and AQuASys are effective in providing precise and exact answers in the native Arabic language without limitations to query development rules, precise question language, or the precise knowledge domain.

Finally, and based on previous work, we accept that, for an inquiry presented by a client, the report sets D are recovered by a web crawler dependent on the inquiry extended from the inquiry. Our point is to fabricate a measure to describe comparable ities between a given inquiry and every applicant section/sentence s ∈ D in the recovered records dependent on likenesses of their shrouded subjects. Consequently, we manufactured bayesian probabilistic models on section level instead of report level to unequivocally extricate their concealed subjects. In addition, the way that there is restricted measure of recovered records D per question (∼100 archives) makes it speak to construct probabilistic models on entries instead of documents and characterize semantically sound gatherings in sections as idle ideas.

3.0 METHODOLOGY

In this section, the methodology of our work will be described briefly. The methodology used is decomposed into two phases; the first phase will classify the query into the corresponding class using Support Vector Machine (SVM), while the second phase uses the Latent Semantic Indexing (LSI) to retrieve the relevant documents with the selected paragraphs that has the answer.

3.1 Query Classification

Query Classification is an automatically labelling a query regarding target taxonomy. SVM algorithm learns to distinguish between a set of classes based from a training set that have some labelled examples for the target classes. The SVM represents document as point in high dimensional space, where the documents in each class represent positive examples, while the other documents represent negative examples.

The steps for building the SVM model are as the following:

1. **Pre-processing:**

In this step, the documents are tokenized into words based onto the spaces between the words. After tokenization, a stop word elimination used to remove un-useful words (... ،عن ،الى ،من) based on Arabic stop words list. Then, a light Arabic stemmer used for stemming the set of words in order to remove various word suffixes, and to reduce the number of words.

2. **Feature Extraction:**

In this step, a feature vector is built for each document, where the feature set are all distinct terms in the documents set, and the values are the Term Frequency-Inverse Document Frequency (TF-IDF), which reflect how important a word is for a set of documents in a corpus. Equation (1) shows the TF-IDF calculation.

$$TF_{IDF} = TF * IDF \quad (1)$$

Where TF is the term frequency of the word or term in the document, and the IDF is the word frequency across all documents represent in Equation (2).

$$IDF = log(\frac{n_t}{N}) \quad (2)$$

Where, (n_t) is the number of documents contained the term, and (N) is the number of documents in the referenced corpus.

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

For example, say the term "cat" appears x amount of times in a 10,000,000 million document-sized corpus (i.e. web). Let's assume there are 0.3 million documents that contain the term "cat", then the IDF (i.e. log {DF}) is given by the total number of documents (10,000,000) divided by the number of documents containing the term "cat" (300,000).

IDF (cat) = log (10,000,000/300,000) = 1.52

TF_IDF (cat) = (TF*IDF) cat = 0.12 * 1.52 = 0.182

A good rule of thumb is, the more your content "makes sense" to the user, the more weight it is assigned by the search engine. With words having a high TF*IDF weight in your content, your content will always be among the top search results, so you can:

- stop worrying about using the stop-words,
- successfully hunt words with higher search volumes and lower competition,
- be sure to have words that make your content unique and relevant to the user, etc.

3. SVM Model

The Decision Tree, Radial Bases, Forest Decision Trees, Nearest Neighbour, Fuzzy Classifier, Deep Learning Classifier, and Support Vector Machine are well-known machine learning algorithms used in the literature.

For classification and categorization. The most distinguished one of these algorithms is called SVM which is the best among other learning algorithms according to [12][13].

SVM is a common classifier which separates instances by a hyperplane. In supervised learning, SVM produces an optimal strait line that separates between categorizes, as shown in Figure 4.

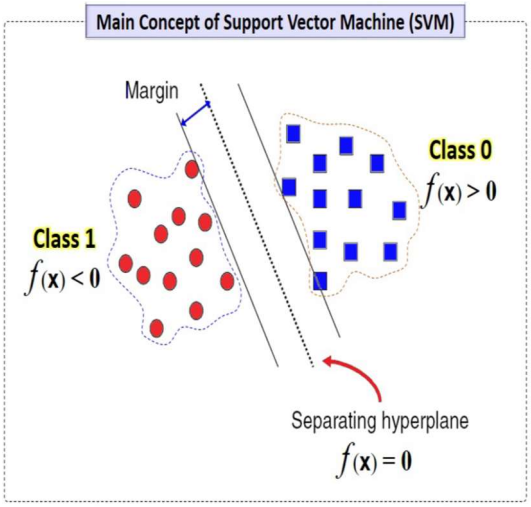

Figure 1 : Support Vector Machine Hyper Plan

In fact, the SVM algorithm finds a line with maximum margins between categorizes, because the optimal separating hyperplane maximizes the margin of the training data [14]. The SVM is a built-in function in many softwares. Sequential Minimal Optimization (SMO) algorithm is SVM implementation of WEKA open source software, which is implemented by [15] [16]. The SMO is one of the most efficient solutions for the SVM algorithm. It is based on solving a series of small quadratic problems, where in each iteration, it uses only two variables, those variables are selected in the working set in order to save time [17].

The SVM as a classifier is considered to be one of the most powerful statistical learning techniques [18][19]. The SVM method successfully addresses and solves different kinds of problems, it depends on three key reasons. The first reason is that the tangible of SVM which was shown by the mathematical and theoretical foundation can be seen as a solid support for classifications and solving problems. The second reason is that SVM was confirmed to be suitable to manage complex data that including high dimensions like text and image data [20] [21]. The third one is that SVM has a potential success in the pattern recognition domain and it confirmed its effectiveness in the image processing field [22] [23]. Consequently, the SVM consists of class classifiers which are mathematically constructed from the summation of a kernel function as stated in Eq. (1) below.

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + d\,, \qquad (3)$$

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

Where $K(x, x_i)$ constructs the summation of the kernel function modelled by [24]. The t_i denotes the ideal outputs when $\sum_{i=1}^{N} \alpha_n t_i = 0$ and $\alpha_i > 0$. From the training set an optimization process in [25], the x_i are the support vectors. The ideal and optimal results can be 1 if the corresponding support vector is in class0, or it can be -1 if the corresponding support vector is in class1. In terms of classification, the class decision happens when the value of f(x), is above a specific threshold or below it. The kernel function K (.,.) is constrained and limited to contain specific properties, which are well-known as the Mercer condition and it can be expressed as in Eq. (2).

$$K(x, y) = b(x)^t . b(y) , \qquad (4)$$

Where $b(x)^t$ implies a mapping according to the input space. Here, the x indicates to be possibly an infinite dimensional space. Finally, the Mercer condition here is responsible to guarantee that the validation of the margin concept, and the optimization of the SVM is limited to definite and boundaries [26].

Specifically, the optimization condition depends on a maximum margin concept, as depicts in Figure 4. Therefore, the system of SVM chooses an appropriate high dimensional space to put and place the best hyperplane that has the maximum margin. As a result, the training of input data points set will be located on the boundaries of the support vectors that are based on Eq. (2). These boundaries are represented by two solid lines, as shown in Figure 4. Modelling of these two boundaries is the main aim of the SVM training process.

In this paper, we carefully decided to use SVM learning algorithm, because SVM generates the hyperplane that classifies the training instances with high speed and more accuracy compared to other traditional clustering methods. However, other traditional clustering methods mainly depend on probability distributions when training data are classified as been demonstrated by [27].

The input query must be represented by the previous steps as the documents in the corpus. The SVM classifier used is the inner product kernel, where the input query classified to one of the classes. Figure 2 represents the steps for the SVM classification method.

SVM classifier model was used to anticipate the entailment scores for query/question (q/a) sets. To portray the closeness between q/a sets we do the following: (I) highlights spoken to by likenesses between semantic parts, e.g., subject, object, action word, or named-substance types found in q/a sets, and (ii) lexical highlights spoken to by lexicon syntactic arrangements, for example, n-gram word covers, or cause and entailment relations found from WordNet. For a given inquiry q, we rank the applicant sentences s dependent on anticipated entailment scores from the classifier.

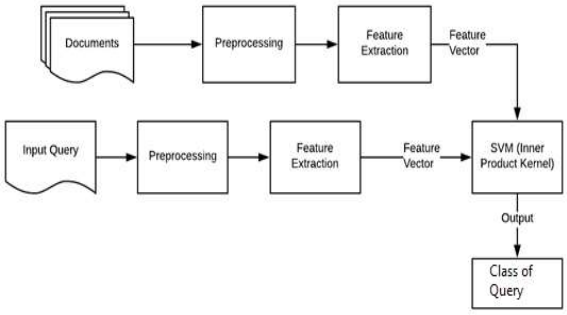

Figure 2 SVM Classifier Steps.

3.2 Latent Semantic Indexing (LSI)

The LSI is a powerful statistical technique information retrieval in textual representation domain. Mainly, the LSI is used lately to solve lexical matching problems and is considered one of the best tools used to maintain the semantic information between the words with high accuracy [28][29][30]. Therefore, LSI is useful in many applications like: author recognition, plagiarism detection, search engines, and text similarity. Typically, LSI projects, runs queries on documents in a space with semantic dimensions well-known as "latent". Latent is statistically adapts the conceptual indices instead of individual words in the process of retrieval [29][31][32]. Consequently, the main contexts of LSI is to examine if a specific word exists or not by applying the similarity process through the documents. Finally, the behaviour of LSI model is similar to human learning when people learn the language of mother tongue to acquire a new vocabulary.

The retrieval process of LSA is basically used the truncated singular value decomposition (SVD) algorithm [33] [31]. SVD works to estimate the structure in word usage in over all the documents, if there are some underlying or latent structure at these words' usage, they are partially obscured or not clearly estimated. Therefore, the retrieval process is done by utilizing well-prepared database for the previous singular values and vectors that are obtained by truncated SVD algorithm. The executable run illustrates more robust indicators of meaning with better performance to these processes of derived vectors against the individual one.

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

The main steps to build the process of LSA model are: i) Preprocess the collected documents by stemming, splitting composite words, and removing the stop words, ii) building the frequency matrix, which is performed by constructing the term document matrix (TDM), iii) applying the weight functions in order to increase the efficiency of the information retrieval process, and to help to allocates weights to the terms based on their occurrences. This weight functions work to replace each element with the product of a Local Weight Function (LWF), and the Global Weight Function (GWF). While the LWF responsible to find the frequency of a specific word within a text, where GWF inspects a term's frequency in over all the documents, iv) decomposing the initial variables, and v) start project queries of LSA for information retrieval.

LSI grows from the problem of how to discover relevant documents from search words. The fundamental difficulty grows when we compare words to find relevant documents because what we want to do is comparing the meanings or concepts beyond the words. LSI attempts to solve this problem by mapping queries into a large document and doing the comparison in this space.

3.3 Singular Value Decomposition (SVD)

SVD is considered an ancient numerical analysis technique that was discovered long ago to serve many applications [34]. Beltrami and Jordan in the 1870's [35] used SVD for real square matrices. After that, and in 1902 Autonne's used it for complex matrices [36] [37]. However, SVD was improved by Eckart and Young Later, in 1939, to include rectangular matrices as cited [36]. Recently the SVD becomes one of the most important numerical techniques used in image processing applications, such as image watermarking, image hiding, image compression and noise reduction [34] [36]. SVD could be applied on any medium. But mostly, it is famous in image processing. The widespread use of SVD is due to its important features and characteristics, which are the following [34] [36] [38]:

1. Good stability of the singular values S A of any image, i.e., no significant changes to the SVs of images will occur upon the addition of small perturbations.

2. The singular values (S) of an image specify its algebraic properties, representing an image's luminance, whereas the singular vectors (U and V) represent the geometry properties of an image.

3. Singular values are in descending order, and many of them have small values compared to the first singular value. Updating or ignoring these small singular values at the reconstruction stage leads to a slight and negligible effect on an image's quality.

SVD can be applied on square or rectangle matrices. After the classification step, A Term Document Matrix (TDM), and Singular Value Decomposition (SVD) are built for each class. The TDM is the basic step for building the term vector, which is the basic step for LSI. TDM contains the distinct terms, and the documents in the class, where the entries are TF-IDF. The SVD is the decomposing of a matrix into a product of three matrices. The decomposition exposes all the best properties and feature of the matrices. The SVD calculated for each class, Equation (3) present the SVD Formula.

$$A = USV^t \quad (5)$$

• U: is an m × k matrix, the columns of the matrix are the eigenvectors of the AA^t matrix; left eigenvectors.

• S: is a k×k matrix, the diagonal elements are the singular values of A; so, all non-diagonal elements are zero by definition.

• V: is an n × k matrix, the columns are the eigenvectors of theA^t A matrix; right eigenvectors.

The output of the SVD calculation will be the set of term vectors, and document vectors. The term vector reflects the relevancy between the term and all the terms, while the document vector reflects the relevancy a certain document to all other documents.

The LSI uses the cosine similarity to find the relation between each term vectors of the query and the document vectors. Equation (4) shows the cosine similarity (c) formula.

$$C = \frac{\vec{A}.\vec{B}}{\left|\left|\vec{A}\right|\right|\left|\left|\vec{B}\right|\right|} = \frac{\sum_{i=0}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}} \quad (6)$$

4.0 RESULTS AND DISCUSSION

In this section, the proposed question answering system will be evaluated in terms of recall, precision, and F-measure. The performance metrics calculation based on true positive, true negative, false positive, and false negative. Each of them is briefly described in

Table 1 below.

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

Table 1: Metrics Description

The Term	Its Meaning
TP	Documents that belong to the target topic and it have been classified correctly.
TN	Documents that do not belong to the target topic and it have been classified correctly.
FP	Documents that belong to the target topic, but they have been classified to the wrong topic.
FN	Documents that do not belong to the target topic and it have been classified to the wrong topic.

For the classification part, the precision, recall, and F-measure calculated according to Equations (7) to (9) [39].

$$Presicion = \frac{TP}{TP+F} \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

$$F-measure = 2\ \frac{Precision\ .Recall}{Precision+Rec} \quad (9)$$

For the second phase (information retrieval part), the precision is calculated according to Equation (10) [40].

$$Presicion = \frac{\{RvD\ \} \cap \{RtD\}}{\{\mathrm{RtD}\}} \quad (10)$$

Were RvD is the Relevant Documents and RtD is the Retrieved Documents. The documents dataset used in question answering system is described in Table 1. The dataset size is 10000 documents that contains 10 classes of Arabic documents, each class contains a variety number of documents.

Table 2: Arabic Document Dataset.

Class	Number of documents
Economy	1600
History	1550
Women and Family	1400
Islamic and religion	1200
Sport	1000
Health	1000
Astronomy	557
Law	944
Children's Stories	726
Food and Recipes	1223
Total	**10000**

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

4.1 Experimental Results

In this section, we discussed the conducted experiments. Twenty different queries were tested, the queries were selected randomly from the documents. In order to calculate the precision, recall, and f-measure for each class, two different queries from each class were selected. Figure 3 shows the precision, recall, and F-measure results for classification steps, as the results show that all the classes have approximately 98% a precision, 97% recall, and 98% f-measure. The method was tested against twenty. The precision, recall, and f-measure ratio may be less, if the system tested over a larger number of queries but, the tested cases were very complicated, and the results were very promising.

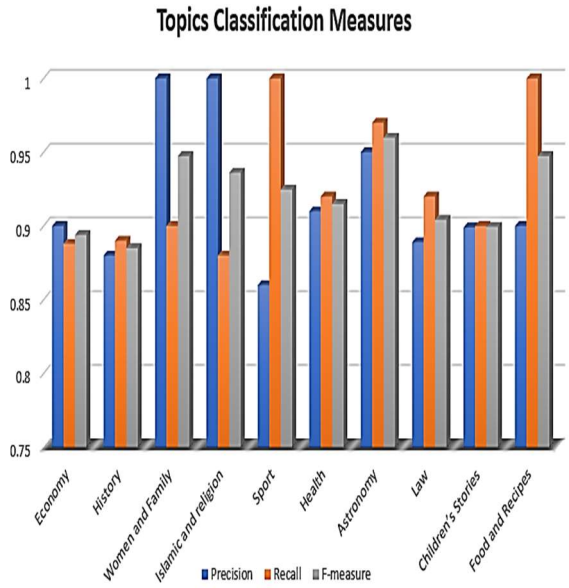

Figure 3: Precision, Recall, and F-measure Results

Figure 4 shows the precision results for the LSI phase, this reflect how relevancy the retrieved documents is according to the queries. The results show that the precision for various queries ranges from 60% to 100%. The minimum precision value was for query 17, while the average precision for all queries was 88%.

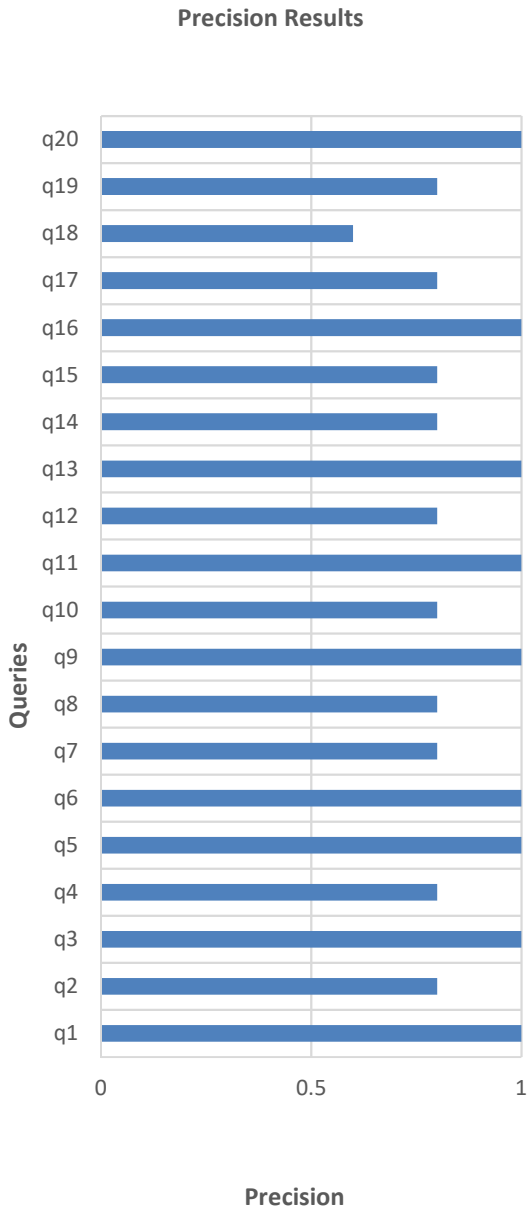

Figure 4: Precision Results for LSI.

Figure 5 shows the Graphical User Interface (GUI) of the proposed system. In the Ask tab, the query must be inserted in the ask tab. The first row of the results shows the class of the queries, and the other rows show the retrieved documents according to the classified class sorted as their relevancy to the query.

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

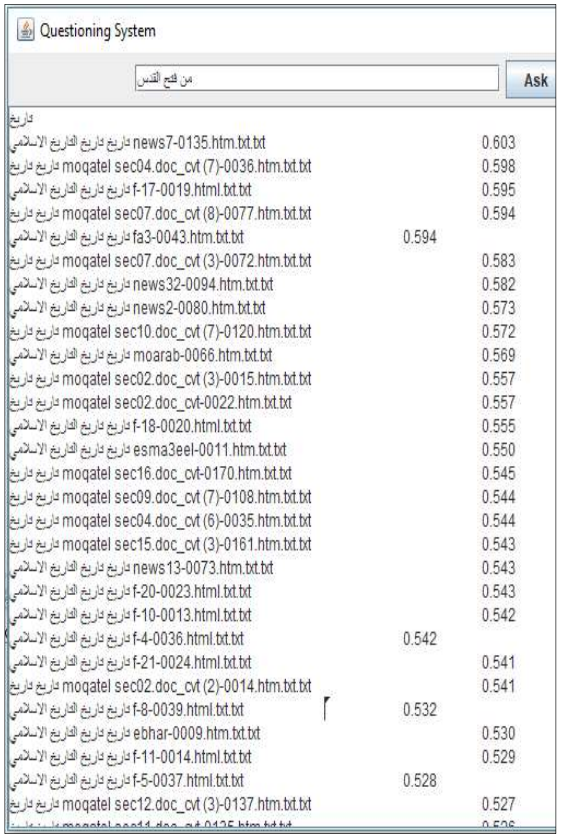

Figure 5: GUI of the proposed system

The second column in the GUI of Figure 5 represents the rank value of the documents. Lastly, we pay the reader attention to that, the accuracy of this work is subjected to some limitations such as; accuracy of the corpus, comprehensiveness of the corpus, and availability of nosey data inside the corpus.

For evaluating and comparing our paper results we refer to the baseline Pattern: it is a direct utilized directed classifier display exhibited in [41] as our standard QA demonstrate. Their datasets, gave in http://www.eecs.berkeley.edu/~asli/asliPublish.html, are q/a sets from TREC errand.

5.0 CONCLUSION

In this paper, a question answering system was proposed. The proposed system uses the LSI techniques to find the relation between the queries, and the documents. The LSI does not only find the match between the terms, it discovers the hidden relations between the terms according to complex mathematical model SVD. The SVD has an intensive computation that requires large working memory, to solve this problem, the query first classified by SVM classifier, then the LSI calculation will be only on the query class. The LSI results were promising. The method has been tested on a set of queries in different fields (classes) against Arabic documents dataset of size 10,000 with 10 classes. The testing results shows promising and accurate output for each of the test cases. Average classification accuracy reaches 98% measured by document classification metrics.

REFERENCES

[1] S. Ray and K. Shaalan, "A Review and Future Perspectives of Arabic Question Answering Systems," *IEEE Trans. Knowl. Data Eng.*, no. 12, pp. 3169–3190, 2016.

[2] A. Albarghothi, F. Khater, and K. Shaalan, "Arabic Question Answering Using Ontology," *Procedia Comput. Sci.*, vol. 117, pp. 183–191, 2017.

[3] M. Shaheen and A. M. Ezzeldin, "Arabic Question Answering: Systems, Resources, Tools, and Future Trends," *Arab. J. Sci. Eng.*, vol. 39, no. 6, pp. 4541–4564, 2014.

[4] H. Al-Chalabi, S. Ray, and K. Shaalan, "Semantic based query expansion for Arabic question answering systems," *Proc. - 1st Int. Conf. Arab. Comput. Linguist. Adv. Arab. Comput. Linguist. ACLing 2015*, pp. 127–132, 2016.

[5] H. Maluf, A. Chalabi, and S. Ray, "Question Processing for Arabic Question Answering System في نظام سؤال جواب للغة العربية Question Processing for Arabic معالجة السؤال .Question Answering System," 2015

[6] G. Tur and M. Park, "LDA Based Similarity Modeling for Question Answering," in *Proceedings ofthe NAACL HLT 2010 Workshop on Semantic Search*, 2010, no. June, pp. 1–9.

[7] B. Hammo and S. Lytinen, "QARAB: A Question Answering System to Support the Arabic Language," *ACL2002Computational Approaches to Semit. Lang.*, p. 11, 2002.

[8] O. Trigui, L. H. Belguith, and P. Rosso, "DefArabicQA : Arabic Definition Question Answering System," *Work. Lang. Resour. Hum. Lang. Technol. Semit. Lang. 7th Lr.*, 2010.

[9] S. Bekhti and M. Al-Harbi, "AQuASys: A Question-Answering System For Arabic," *WSEAS Int. Conf. Proceedings. ...*, pp. 130–139, 2013.

[10] Y. Benajiba, P. Rosso, and A. Lyhyaoui, "Implementation of the ArabiQA Question Answering System's components," *Proc. Work. Arab. Nat. ...*, no. 329, pp. 1–5, 2007.

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

[11] G. Bouma, G. Kloosterman, J. Mur, G. Van Noord, L. Van Der Plas, and J. Tiedemann, "Question answering with joost at CLEF 2007," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5152 LNCS, pp. 257–260, 2008.

[12] J. Huang, J. Lu, and C. X. Ling, "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy," in *Third IEEE International Conference on Data Mining*, pp. 553–556.

[13] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. their Appl.*, vol. 13, no. 4, pp. 18–28, Jul. 1998.

[14] A. Statnikov, C. F. Aliferis, D. P. Hardin, and I. Guyon, *A Gentle Introduction to Support Vector Machines in Biomedicine: Volume 2: Case Studies and Benchmarks*. World Scientific, 2013.

[15] J. Platt and others, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. large margin Classif.*, vol. 10, no. 3, pp. 61–74, 1999.

[16] Z.-Q. Zeng, H.-B. Yu, H.-R. Xu, Y.-Q. Xie, and J. Gao, "Fast training Support Vector Machines using parallel sequential minimal optimization," in *Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on*, 2008, vol. 1, pp. 997–1001.

[17] J. Wang, A. Lu, and X. Jiang, "An Improved SMO Algorithm for Credit Risk Evaluation*," *Australas. Data Min. Conf.*, pp. 169–176, 2015.

[18] J. Huang, J. Lu, and C. X. Ling, "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003, pp. 553–556.

[19] L. Wang, *Support vector machines: theory and applications*, vol. 177. Springer Science & Business Media, 2005.

[20] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[21] E. Zarrouk, Y. Ben Ayed, and F. Gargouri, "Hybrid continuous speech recognition systems by HMM, MLP and SVM: a comparative study," *Int. J. Speech Technol.*, vol. 17, no. 3, pp. 223–233, 2014.

[22] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*, 2001, pp. 107–118.

[23] Y. Benayed, D. Fohr, J. P. Haton, and G. Chollet, "Confidence measures for keyword spotting using support vector machines," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003, vol. 1, pp. I--I.

[24] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. their Appl.*, vol. 13, no. 4, pp. 18–28, 1998.

[25] R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *J. Mach. Learn. Res.*, vol. 1, no. Feb, pp. 143–160, 2001.

[26] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[27] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, no. 2, pp. 210–229, 2006.

[28] K. Nahar, H. Al-Muhtaseb, W. Al-Khatib, M. Elshafei, and M. Alghamdi, "Arabic phonemes transcription using data driven approach," *Int. Arab J. Inf. Technol.*, vol. 12, no. 3, pp. 237–245, 2015.

[29] K. M. O. Nahar, M. Elshafei, W. G. Al-khatib, H. Al-muhtaseb, and M. M. Alghamdi, "Statistical Analysis of Arabic Phonemes for Continuous Arabic Speech Recognition," *Int. J. Comput. Inf. Technol.*, vol. 01, no. 02, pp. 49–61, 2012.

[30] K. Nahar, M. Abu Shquier, W. Al-Khatib, H. Al-Muhtaseb, and M. Elshafei, "Arabic phonemes recognition using hybrid LVQ/HMM model for continuous speech recognition.," *Int. J. Speech Technol.*, vol. 19, no. 3, pp. 495–508, 2016.

[31] P. Wiemer-Hastings, K. Wiemer-Hastings, and A. Graesser, "Latent semantic analysis," in *Proceedings of the 16th international joint conference on Artificial intelligence*, 2004, pp. 1–14.

[32] K. M. . Nahar, "Off-line Arabic Hand-Writing Recognition Using Artificial Neural Network With Genetics Algorithm," *Int. Arab J. Inf. Technol.*, vol. 6, no. 6, p. XXX, 2018.

[33] D. M. Blei, "Probabilistic topic models,"

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

Commun. ACM, vol. 55, no. 4, pp. 77–84, 2012.

[34] H. Zhang, C. Wang, and X. Zhou, "A robust image watermarking scheme based on SVD in the spatial domain," *Futur. Internet*, vol. 9, no. 3, p. 45, 2017.

[35] K. M. O. Nahar, N. Alhindawi, O. M. Al-hazaimeh, R. A. E. D. M. Al-, and A. M. Al-akhras, "NLP AND IR BASED SOLUTION FOR CONFIRMING CLASSIFICATION OF RESEARCH PAPERS," vol. 96, no. 16, pp. 5269–5279, 2018.

[36] K. Loukhaoukha, "Image watermarking algorithm based on multiobjective ant colony optimization and singular value decomposition in wavelet domain," *J. Optim.*, vol. 2013, 2013.

[37] S. Gan, Y. Chen, S. Zu, S. Qu, and W. Zhong, "Structure-oriented singular value decomposition for random noise attenuation of seismic data," *J. Geophys. Eng.*, vol. 12, no. 2, pp. 262–272, 2015.

[38] N. M. Makbol and B. E. Khoo, "A new robust and secure digital image watermarking scheme based on the integer wavelet transform and singular value decomposition," *Digit. Signal Process. A Rev. J.*, vol. 33, pp. 134–147, 2014.

[39] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009.

[40] S. Haiduc, V. Arnaoudova, A. Marcus, and G. Antoniol, "The use of text retrieval and natural language processing in software engineering," *Proc. 38th Int. Conf. Softw. Eng. Companion - ICSE '16*, pp. 898–899, 2016.

[41] M. Thint, "A Graph-based Semi-Supervised Learning for Question-Answering at Berkeley," no. August, pp. 719–727, 2009.