ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

Arabic information retrieval system using the neural network model

Issam Al-Hadid[1], Suha Afaneh[2], Hassan Al-Tarawneh[3,] Hiba Al-Malahmeh[4]

Department of CIS, Faculty of IT, Isra University, Amman, Jordan [1]

Department of CS, Faculty of IT, Isra University, Amman, Jordan [2,3]

Department of MIS, Faculty of MS, Isra University, Amman, Jordan [4]

Abstract: Information Retrieval (IR) for Arabic language has gained significant attention and emerged as one of the research topics that has been studied by Arabic and foreign researchers. The goal of this research is to apply the IR using Neural Network (NN) Model on natural Arabic language text documents to solve the problem of retrieving the Arabic information from documents' database. Furthermore, all stored documents must be indexed with keywords classification that describe the exact content of each document, which makes it impossible to retrieve all related documents more computational time to classify and update the stored documents. IR using NN applies to solve the problem of documents indexing, classification and retrieving the related documents using Terms of weight and Normalization. The computational results have been compared with the Vector Space Model (VSM) and showed an improvement of NN training time compared with VSM load document time.

Keywords: Arabic, Information Retrieval, Neural Network, Vector Space.

I. INTRODUCTION

Arabic IR considered as an active research area that satisfy more than 400 million Arabic speakers needs to retrieve text, images and videos on the internet and other applications. The growing number of Arabic users and uploaded documents on the internet has encouraged researchers to develop many different Arabic IR models and systems to enhance the process of retrieving Arabic related documents. this research propose IR using the NN Model technique used to obtain relevant measures between a query and documents, it is based on statistical, linguistic, and knowledge-based approach that has been studied for many years in the hope of achieving human-like performance in IR [1].

Many models existed and used in the IR such as Manual systems (Boolean, Fuzzy) [2, 3], which is based on Boolean logic and set theory, where both documents and quires are described as set of terms, Automatic (Vector space, Latent semantic indexing) [4, 5] where Vector Space represents the documents as vectors of identifiers, and Latent Semantic Indexing is based on the conceptual meaning similarity from the body of the text documents and query, and adaptive (Genetic algorithms, Neural Networks) [6, 7], Genetic Algorithms modified the description of the documents and the query and uses a matching functions to find the relation between it. And NN uses the terms weight and Normalization weight to find the related documents. According to Doszkocs, Riggia and Lin [8], the NN model in IR differ from the other traditional information processing models in many ways but the most important one is the self-processing with intelligent behavior which will lead to an active processing agents (nodes and links). There are many useful properties for the NN such as; Adaptively, Nonlinearity, Contextual Information, Input-Output Mapping, Fault Tolerance.

The aim of this research is to improve the existing Arabic IR techniques used to retrieve text documents, and to solve the problem of indexing and classifying documents' database keywords.

This paper consists of six sections, section one discuses the introduction of the Arabic IR using NN Model, and presents the aim of this work, section two discuses the related work to our research in order to identify the weakness and strength that motivated us to overcome the weakness and increase the strengths of Arabic IR. Section three describes the standard IR Model using NN and its basic terms explanation. Section four proposes algorithm of Arabic IR using NN Model. Section five reports the finding and results of this research, and section six presents the conclusion and discussion of this research.

II. RELATED WORK

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

There are many studies considered the IR in other languages (not Arabic); Ataer and Duygulu [9] presented a method for the retrieval of Ottoman documents based on word matching, his approach applied on the large archives (Ottoman documents), where the documents segmented into word images and then uses a hierarchical matching technique to find the similar instances of the word images. Clarke and Cormack [10] used the Boolean Model; where the queries should be well defined and only the documents matching the query are retrieved, they integrated many features such as phrase matching, truncation and stemming into their model. In Arabic language, several researchers studied IR in many researches, Taghva, Elkhoury and Coombs [11] implemented a root-extraction stemmer for Arabic without a root-dictionary, Larkey, Ballesteros and Connell [12] developed several light stemmers based on

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

heuristics and a statistical stemmer based on co-occurrence for Arabic retrieval, Then they compared the retrieval effectiveness of their stemmers and of a morphological analyzer. Chen and Gey [13] proposed an approach to the cross language retrieval which was to translate the English topics into Arabic using online English-Arabic machine translation systems, and they reported on the construction of an Arabic stop list and two Arabic stemmers, and the experiments on Arabic monolingual retrieval, English-to-Arabic cross-language retrieval. Hmeidi , Kanaan and Evens [14] have put together a corpus and designed and built an automatic IR system from scratch to handle Arabic data. They have implemented both automatic and manual indexing techniques for this corpus. A long series of experiments using measures of recall and precision has demonstrated that automatic indexing is at least as effective as manual indexing and more effective in some cases. They have also compared the retrieval results using words as index terms versus stems and roots, and conclude that root indexing is more effective than word indexing. Also, IBM [15] proposed a new method to IR, they built two systems for cross-language experiments with English queries and Arabic documents. One system approached translation and retrieval as entirely separate tasks: they used a machine translation system to translate the Arabic documents into English, and then did the retrieval with a standard English IR system, the other system incorporated the parameters of a machine translation model directly into IR scoring formula; by accessing an innovative Arabic morphological analyzer. Abu salem, Al-Omari, and Evens [16] investigated how to improve the performance of an Arabic IR System (Arabic-IRS) by imposing the retrieval method over individual words of a query depending on the importance of the WORD, the STEM, or the ROOT of the query terms in the database. An extended version of the Arabic-IRS system is designed, implemented, and evaluated to reduce the number of irrelevant documents retrieved. Wedyan, Alhadidi, and Alrabea [17] studied the retrieval effectiveness achieved when he applied a successful automatic query expansion method in Arabic text retrieval based on an automatic thesaurus. Kanaan, et al. [18] implemented a system to deal with Arabic IR and then they examined some of the common relevance feedback strategies that have been shown to be effective in other languages; their proposed system is based on Boolean model and VSM. Many previous work applied NN with IR to retrieve information on other languages such as; Doszkocs, Reggia, and Lin [8] used models that represent information as a network of weighted, interconnected nodes. He [19] investigated the applications of some NN models in IR systems. Mokriš and Skovajsová [1] used the NN model to describe IR system which retrieves information from the text documents in natural language and comes from the IR system using statistical. Kim and Raghavan [20] developed a NN Model, where the rule weights can be adjusted by the users' relevance feedback. Using NN with IR improved the performance of retrieving information. However, there are leaking of previous work using IR systems with NN for Arabic language.

III. Neural Network Model in IR

IR using the NN Model for Arabic language is a technique which is used to obtain relevant measures between a query and documents. The model is consists of three layers; Query Terms Layer, Documents Terms Layer and Documents Layer. The Model uses the normalized weight to compute the degree of similarity between the documents' terms and the queries' terms. The degree of similarity (similarity measures) is a function that computes degree of similarity between documents and query, the Cosine similarity measure will be used in the NN model which measures cosine of the degree between document-query vectors.

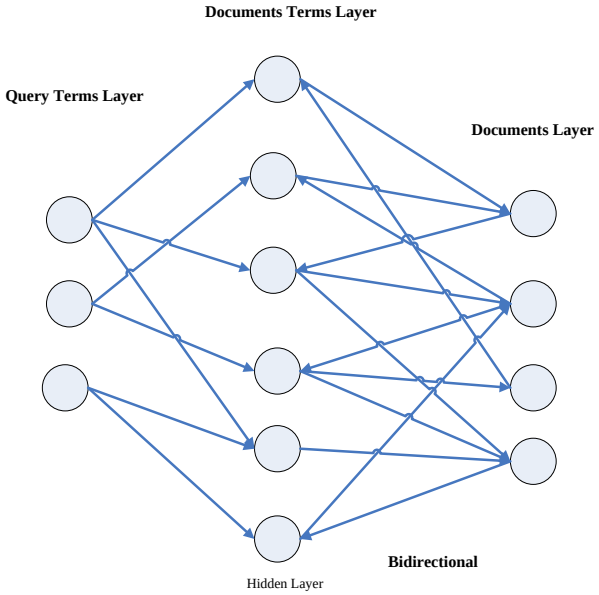

Fig. 1 NN Model in IR

IV. Proposed Model

The application that we have implemented using the VB.NET (2005) can uploaded with unlimited documents number using the dynamic allocation, it is tested on a collection of 224 Arabic documents. Figure 2 illustrates the process of Arabic IR using NN Model.

The following steps describe the algorithm of Arabic IR process using NN Model:

1- Load the documents:
Select the documents collection's folder to be loaded into the System.
2- Stemming all the terms in the documents :
Reduce variant words' forms to their stem (root).
3- Build the inverted files (arrays):
fill the stemmed words into document array.
4- Find the documents terms weight:
find the number of occurrence of the term in the document [21].
5- Find the document term normalized weight:
normalize the words' frequency count to measure the words occurrence relatively to the size of the document [22]
(Steps from 2-5 is the training the NN cells)

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

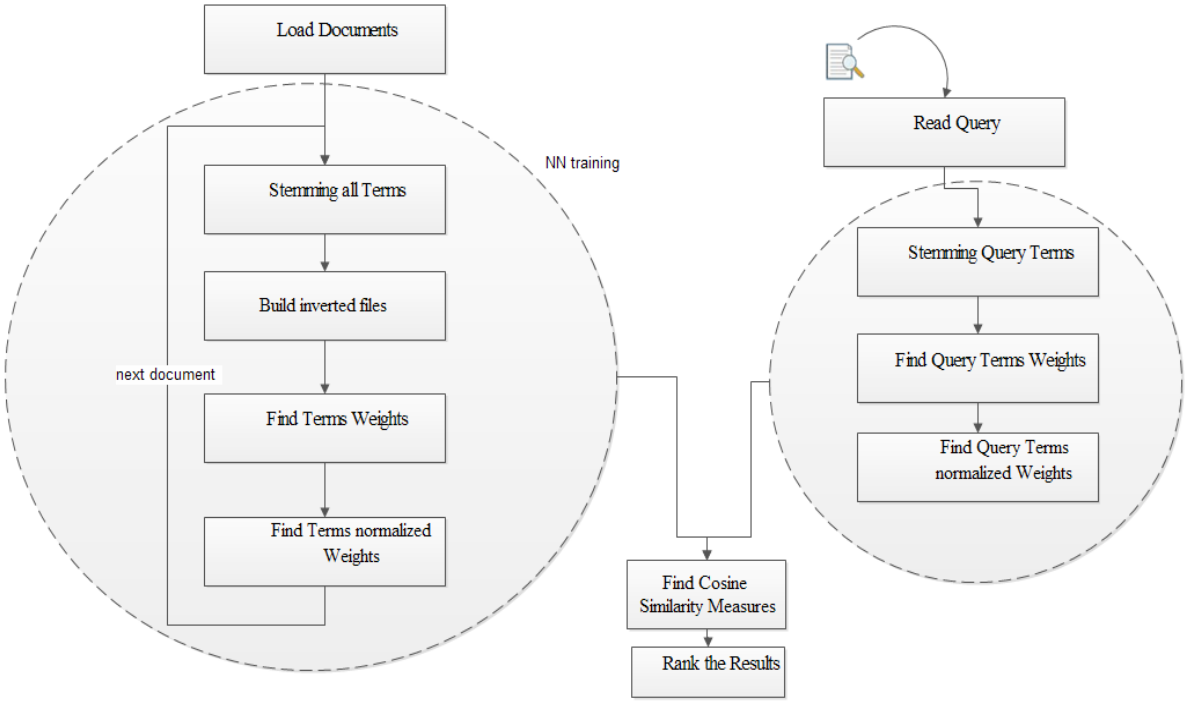

Fig. 2 Process of Arabic IR System using NN Model

6- Read the query:
after the NN training process, system reads the user's query.
7- Stemming the query terms:
reduce query words' forms to their stem(root).
8- Find the query term weight:
Find the number of occurrence of the term in the query.
9- Find the query term normalized weight
Normalize the words' frequency count to measure the words occurrence relatively to the size of the query [22]
10- Find the cosine similarity measures:

$$\mathbf{CosSim\ (d,q)} = \frac{\sum \mathbf{d_{ik}} \cdot \mathbf{q_k}}{\sum \mathbf{d_{ik}^2} \cdot \sum \mathbf{q_k^2}} \qquad \ldots\ldots (1)$$

where:

- d: document
- q: query
- d_{ik}: normalized weight of the k-th element of the document i-th term
- q_k: normalized weight of the k-th element of the query

11- Rank the results
Evaluate the results and sort it according to the query relative documents in the collection.

V. Experiments and Results

For this comparison, the study uses the precision, Recall in order to measures and evaluate the IR Performance, where The Precision is applied to evaluate the correlation of the documents collection query; by measuring the system ability of retrieving only the top- Ranked documents that are mostly relative to the query [23]. Figure 3 shows the relation between recall and precision.

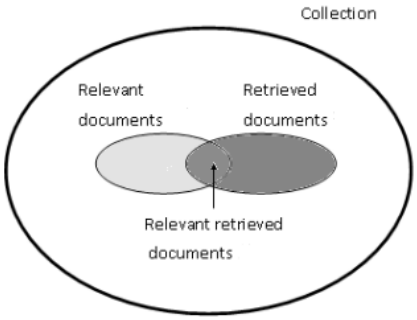

Fig. 3 Precision and recall [24]

The precision can be calculated as the following formula:

$$\text{Precision} = \alpha / \gamma \quad .....(2)$$

Where :

- α is the amount of related documents retrieved.
- γ is the amount of documents retrieved.

On the other hand; Recall is the ability of the search to find all of relevant document items in the collection; it measures the ability of the system to retrieve all documents that are relevant to the query [13].
The recall can be calculated as the following formul

$$\text{Recall} = \beta / \nu \quad .....(3)$$

Where :

- β is the amount of relevant documents retrieved .
- ν is the amount of relevant documents in the collection.

Using the same query in both systems we found the following;
Query = استخدام الحاسب الآلي
Number of retrieved documents (both systems) = 224
The experiments have been carried out to analyse the computational time and search capability of NN for IR Arabic documents .For this comparison, this study used the suggested NN model and the VSM, see Table.1 and figure.4

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 3, Issue 12, December 2014

TABLE I
COMPARISON BETWEEN NN MODEL AND VSM RESULTS

Run time for NN System	Run time using the Vector Space System
Training Time = 38- 39 min Search Time = 2 sec	Load Documents Time = 48-49 min Search Time = 2 sec

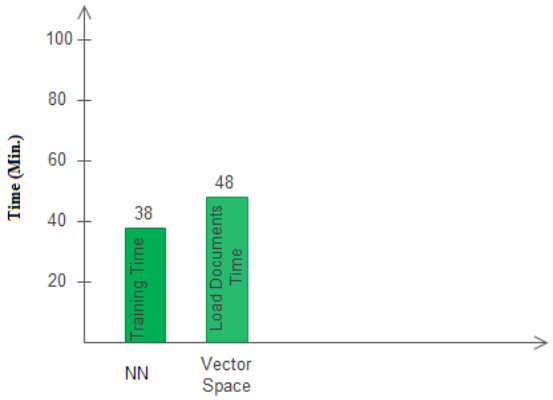

Fig. 1 NN and VSM computational time

Table 1 and Figure 3 showed the average computational time training for the NN Model and VSM loading document time, furthermore, showed the search time using both models. However, the comparison results proved that the Figure running time using NN Model outperformed the VSM by around 10 minutes. Meanwhile, the search time results are the same using both Models which is two Seconds). In table 2 , we compare NN model and VSM for both Recall and Precision.

TABLE II
EXPERIMENTAL RESULTS OF NN MODEL AND VSM OVER 224 RETRIEVED DOCUMENTS

Model	Recall	Precision
NN	0.736842	0.291667
VSM	0.736842	0.291667

Table 2 showed that NN and VSM produced the same results , furthermore, figures 4 and 5 shows the performance of NN in for both Recall and precision results.

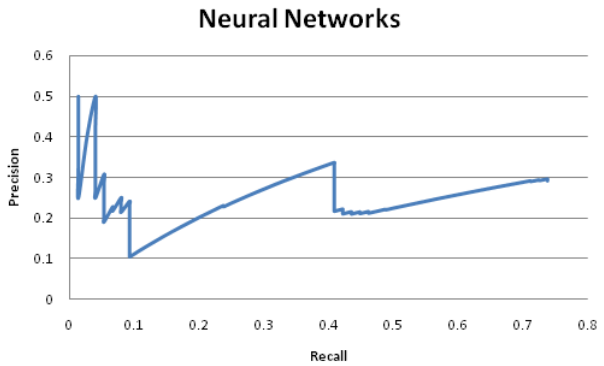

Fig. 5 Recall and Precision- Arabic IR system using the NN Model

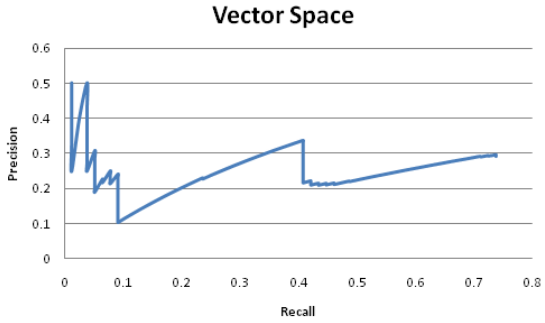

Fig. 6 Recall and Precision- Arabic IR system using the VSM

VI. CONCLUSION

This work has proposed a NN system to retrieve the Arabic documents, by minimize the consuming time, where the standard VSM which leads to more consuming time. The computational results showed that the NN system can enhance the consuming time of retrieving the Arabic documents by compare the results with standard VS model. However, both systems retrieved the same documents using the same query, where the Recall and the Precision for both systems are exactly the same. Therefore, NN outperformed VS system in the running time that needs to retrieve the Arabic documents.
The limitation of this study is the NN training time when new documents are appended to the current collection. In future work this study can be applied on other Arabic collection or huge number of Arabic documents in order to enhance the results of the Recall and Precision.

ACKNOWLEDGMENT

The authors thank Prof. Ghassan Kanaan for the assistant and support provided by him.

REFERENCES

[1] I. Mokriš, and L. Skovajsová, "Development of Neural Network Information Retrieval System from Text Documents", Acta Electrotechnica et Informatica,, Vol. 5 (3), 2005.
[2] O. Cordón, F. Moya, and C. Zarco, "A new evolutionary algorithm combining simulated annealing and genetic programming for relevance feedback in fuzzy information retrieval systems," Soft Computing, vol. 6, no. 5, pp. 308-319, 2002.
[3] E. Herrera and Viedma, "Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach," Journal of the American Society for Information Science and Technology, vol. 52, no. 6, pp. 460-475, 2001.
[4] A. Hotho, A. Nürnberger, and G. Paaß, "A Brief Survey of Text Mining", Ldv Forum, Vol(20), pp. 19-62.
[5] M. Lan, C.-L. Tan, H.-B. Low et al., "A comprehensive comparative study on term weighting schemes for text categorization with support vector machines", Special interest tracks and posters of the 14th international conference on World Wide Web, ACM, 2005, pp. 1032-1033.
[6] D. A. Grossman, Information retrieval: Algorithms and heuristics: Springer, 2004.
[7] W. Maitah, M. Al-Rababaa, and G. Kannan, "Improving The Effectiveness Of Information Retrieval System Using Adaptive Genetic Algorithm," International Journal of Computer Science & Information Technology, vol. 5, no. 5, 2013.
[8] T. E. Doszkocs, J. Reggia, and X. Lin, "Connectionist models and information retrieval," Annual review of information science and technology, vol. 25, pp. 209-262, 1990.

ISSN (Online) : 2278-1021
ISSN (Print) : 2319-5940

[9] E. Ataer, and P. Duygulu, "Retrieval of Ottoman documents", Proceedings of the 8th ACM international workshop on Multimedia information retrieval, ACM, 2006, pp. 155-162.
[10] C. L. Clarke, and G. V. Cormack, "Shortest-substring retrieval and ranking," ACM Transactions on Information Systems (TOIS), vol. 18, no. 1, pp. 44-78, 2000.
[11] K. Taghva, R. Elkhoury, and J. S. Coombs, "Arabic Stemming Without A Root Dictionary", ITCC (1), 2005, pp. 152-157.
[12] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis", Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2002, pp. 275-282.
[13] A. Chen, and F. C. Gey, "Building an Arabic Stemmer for Information Retrieval", TREC,2002, pp. 631-639.
[14] I. Hmeidi, G. Kanaan, and M. Evens, "Design and implementation of automatic indexing for information retrieval with Arabic documents," JASIS, vol. 48, no. 10, pp. 867-881, 1997.
[15] M. Franz, and J. S. McCarley, "Arabic Information Retrieval at IBM", IBM, 2002.
[16] H. Abu-Salem, M. Al-Omari, and M. W. Evens, "Stemming methodologies over individual query words for an Arabic information retrieval system," Journal of the American Society for Information Science, vol. 50, no. 6, pp. 524-529, 1999.
[17] M. Wedyan, B. Alhadidi, and A. Alrabea, "The effect of using a thesaurus in arabic information retrieval system," International Journal of Com-puter Science Issues, IJCSI, vol. 9, no. 1, pp. 431-435, 2012.
[18] G. Kanaan, R. Al-Shalabi, M. Abu-Alrub et al., "Relevance Feedback: Experimenting with a Simple Arabic Information Retrieval System with Evaluation," International Journal of Applied Science and Computations, vol. 12, no. 2, 2005.
[19] Q. He, "Neural Network and its Application in IR," Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign Spring, 1999.
[20] M. Kim, and V. V. Raghavan, "Adaptive concept-based retrieval using a neural network", Proc. Of ACM-SIGIR Workshop on Mathematical/Formal Methods in IR, 2000.
[21] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to information retrieval: Cambridge university press Cambridge, 2008.
[22] N. Polettini, "The vector space model in information retrieval-term weighting problem," Entropy, pp. 1-9, 2004.
[23] M. Lassi, "Automatic thesaurus construction," A paper written within the GSLT course Linguistic Resource, pp 2-10,2002.
[24] L. Skovajsova, "Text Document Retrieval by Feed-forward Neural Networks," Information Sciences and Technologies Bulletin of the ACM Slovakia, vol. 2, no. 2, pp. 70-78, 2010.