COMPARISON BETWEEN VECTOR SPACE MODEL AND FUZZY SET MODEL RETRIEVAL OF ARABIC COLLECTION

Eman I. ElAjrami*

Mohammed AbuLamddi*

ABSTRACT

Information Retrieval Systems should provide the users with more relevant documents for their queries. There are many models for calculating the similarity between a query and the documents. We used Vector Space model using cosine similarity and Fuzzy Model then evaluated the results for each on an Arabic collection. We used the inverted file for indexing, stemming the Arabic words using suffix, infix and prefix removal. We observed that Fuzzy can give higher recall than Vector Space but in general Vector Space gives better precision and it depend on the threshold value. We used MS visual basic to implement our project and connected it with MS Access which we used to build the data base for our work.

Keywords: *Vector Space model, Boolean Model, Fuzzy model*

*Faculty of Information Technology,Palestine University, Gaza, Palestine.

INTRODUCTION

This paper presents a short survey of fuzzy and neural approaches to Information Retrieval. The goal of such approaches is to define flexible Information Retrieval Systems able to deal with the inherent vagueness and uncertainty of the retrieval process [2].

David Parry used the fuzzy ontology for medical document retrieval [3]. A mapping between query terms and members of an ontology is usually a key part of any ontology enhanced searching tool. However the relative importance of a particular mapping to an overloaded term may be different for different users, and this information is vital for accurate satisfaction of a query [3].

One way of overcoming this problem is the postulation of a "fuzzy ontology". [3]. QARAB implemented by Hammo et.al. [1], provides short answers to questions expressed in the Arabic language.

The system Processes the input question then retrieves the candidate documents containing answers from the IR system. They used Vector Space model in retrieving the documents. Finally the system Processes each one of the candidate documents in the same way as the question is processed and returning sentences that may contain the answer.

METHODOLOGIES

We applied in our system both of Vector Space and Fuzzy set models (which we discussed in details in the previous section) to compare between there results evaluation.

Our Approach in retrieval has the following steps:

1. Separate the documents words.
2. Removing the stop words from the documents.
3. Removing the words that have lowest and highest frequency.
4. Stemming for the query terms and file terms.(prefix, suffix and infix removal).
5. Building the inverted file which includes the index terms.
6. Using Vector Space model or fuzzy model for calculating similarity.
7. Ranking the documents after suitable threshold value (we used 0.5) to display the documents that have similarity more than 0.5.
8. Evaluate the values for precision and recall for each query.

EVALUATION

We tested our system of an Arabic corpus includes 165 documents and for 29 queries on those documents. We will show here the evaluation of 10 query samples.

We observed that vector model give as an average high precision but less recall. See figure 1.1 and figure 1.2.

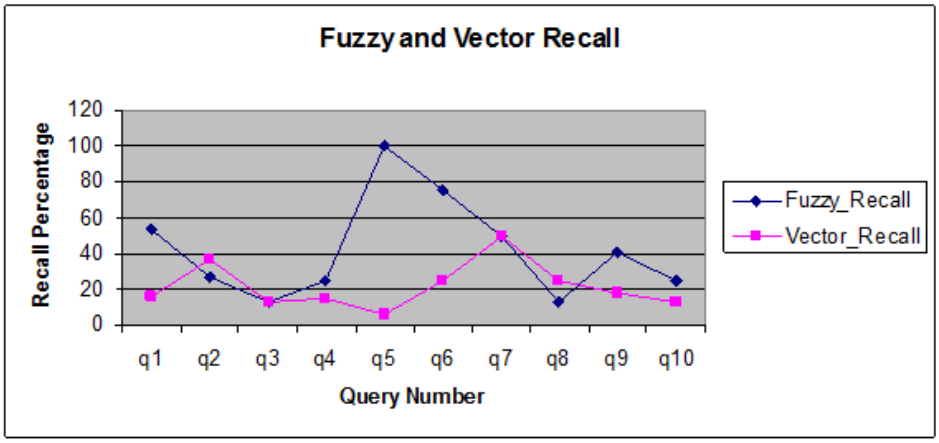

Figure 1. 1 Fuzzy and Vector Recall evaluation.

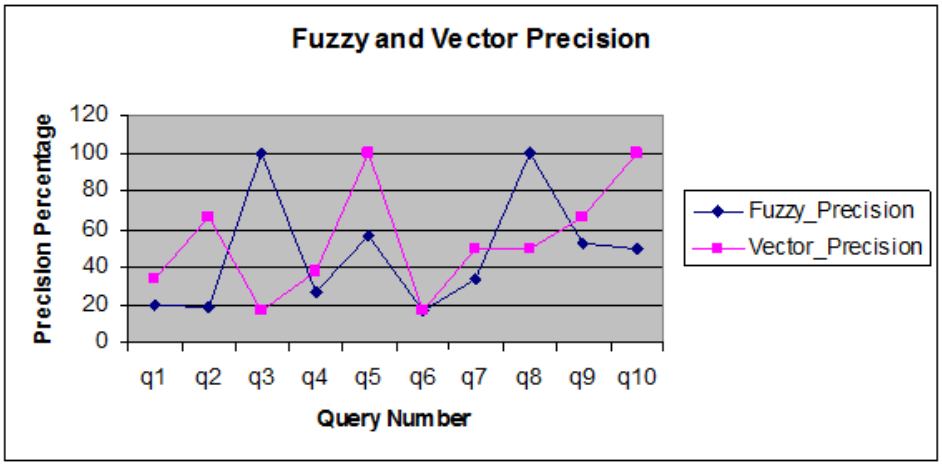

Figure 1. 2 Fuzzy and Vector Precision evaluation.

DISCUSSION

According to our implementation and testing we concluded that using Fuzzy Model in Arabic collection give higher recall than using Vector Space model but sometimes it gives lower precision than vector.

The percentage of precision and recall depends on threshold value, when threshold value is high it gives better precision but less recall and if it is small it gives high recall and low precision.

Using stemming for the query and documents improve the recall values and gives more relevant documents.

REFERENCES

[1] Abu-Salem H., Hammo B., and Lytinen S.,"QARAB: A Queston Answering System to Support the Arabic Language", Depaul Unversity, Chicago IL 60604.

[2] Bordogna, G. and G. Pasi. Application of Fuzzy Set Theory to extend Boolean Information Retrieval. Soft Computing in Information Retrieval: Techniques and Applications. F. Crestani and G. Pasi. Heidelberg, Physica-Verlag. 2000.

[3] David Parry, A fuzzy ontology for medical document retrieval, Australian Computer Society, Inc. vol. 32, 2004.