USING NEW DATA STRUCTURE TO IMPLEMENT DOCUMENTS VECTORS IN VECTOR SPACE MODEL IN INFORMATION RETRIEVAL SYSTEM

Dr. KHALAF KHATATNEH M.WEDYAN DR.MOHAMED ALHAM DR.BASEM ALRIFAI

Prince Abdu Allah Bin Ghazi for IT, Al-Balqa Applied University Salt, Jordan

ABSTRACT

In this paper, we present how table memorized semiring structure contributes in the vector space model in information retrieval system. We implement this new structure by generating table for each document and the first row filled with key word and the second row filled with weight for each word. This new structure implements using 242 Arabic documents which were presented in the Saudi Arabian National Computer Conference. The new method (technique) shows a new result which is more efficient than traditional structure and can save space. The results also show that when we used traditional structure system it occupies 204248 × units to implement vectors, but in new data structure system it occupies 5388×unit which means that we saved more than 198860 space units.

Index Terms **:** *Stop Words, Memorized Semirings, Vector Space Model, Stopwords.*

1. INTRODUCTION:

The world witnesses a huge informational revolution that brings out lots of information and researches. Researching for this information without using search engines is a hard task or it is impossible to gain accurate information without them. Besides, the importance of information retrieval sciences has evolved. This science depended only on libraries in the past, but with the widespread of internet, there was an increasing need for programs and software that facilitate accessing the information accurately and quickly.

Upon this science lies the responsibility to organize the search outcome.

Researching in English and other languages are rich in the field of Information retrieval but those written in Arabic are very poor. Here lies the importance of our study, to take part in the development of the Arabic

Language to be a main participant in this age of information. The Arabic language is our identity and our path to the genuine understanding of this age. It Is also the way to achieve an Arabic scientific development that takes part in building the modern civilization.

STUDY SAMPLE.

2. THE IMPORTANCE OF TABLES

Just like learning to walk before you can run, learning multiplication and memorizing the times tables are building blocks for other math topics taught in education and learning such as division, long multiplication, fractions and algebra. Students who have not memorized the times tables will find these levels of math much more difficult than they need to be. There is no time to pull out a calculator or to take 20 seconds to work out a math strategy before coming up with the answer. Learning system dealing with tables will very often fall behind in math and begin to loose confidence. All because they did not memorize the times tables!

Knowing your multiplication facts is helpful not only in academics; we frequently use multiplication in our daily lives. We might need it when doubling a recipe, determining a discount at a store or figuring out our expected arrival time when traveling. Math calculations are subconscious elements in work, play and daily chores. Knowing the times tables can help simple tasks to be performed rapidly and save time and stress.

Because of these issues we working on tables designed with elements in the same column tend to have similar properties. The gaps at the top of the table are there to make the elements in vertical columns all have the same amount of outer-shell electrons helping Arabic language in arab region.

The study sample consisted of 242 Arabic documents which were presented in the Saudi Arabian National Computer Conference. These documents were used in many studies.

3. THE ARABIC LANGUAGE

Arabic is a language that holds the miracle of the Holy Quran, and accomplished all the requirements of Arabic and Islamic civilization in its peak flourishing. Arab books in Medicine and Science had been the main reference books for the West and in most of its important universities until the end of the middle ages. and the beginning of the renaissance [25].

Internationally, it gained full acceptance and recognition and became a credited language in UN agencies along side with the other five languages used.

4. VECTOR SPACE MODEL.

This model is characterized by the use of very little binary weight, but more has partial matching. When the weight is given to the index terms in queries, these weights are used to calculate the degree of similarity of each file in the system and the user query through a descending arrangement after calculating the degree of similarity; they are arranged in a descending way [7]

This model is characterized by:- [7] Enforcing system efficiency by giving weight to the idioms.

- As it is dealing with partial matching, it allows for documents near to the query to be retrieved.

Using ranking measures as (Cosine ranking), which allows the documents to be arranged according to their relation with the query.
The disadvantage of this model is that it requires mathematical operations that may affect the performance of the system.
The vector space model has been named so because each document and term has a vector

Figure (1) shows the tradition structure for implementing vector space model

	T_1	T_2	T_3	…	T_n
D_1	W_{11}	W_{12}	W_{13}	…	W_{1n}
D_2	W_{21}	W_{22}	W_{23}	…	W_{2n}
D_3	W_{31}	W_{32}	W_{33}	…	W_{3n}
:	…	…	…	…	:
:	…	…	…	…	:
D_m	W_{m1}	W_{m2}	W_{m3}	…	W_{mn}

Figure1

All keyword and all documents will appear in the matrix

5. NEW DATA STRUCTURE USING MEMORIZED SEMIRINGS

In the new structure, the tables are special kinds of two rows arrays. The first row is filled with a word, and the second with some coefficients. This structure generalized the (finite) k-sets sets of Eilenberg[6], it is versatile (one can vary the letters, the words and the coefficient), easily implemented and fast computable. Varying the scalars and the operations on them, one can obtain many different structures and among them semirings.

The new structure can be implemented by generating a table for each document, and the first row filled with a key word and the second row filled with weight for each key word.

D1	T_1	T_2	T_3	….	…..	T_n
	W_{11}	**W_{12}**	**W_{13}**	**…**	**…**	**W_{1n}**

Figure2

Only the keyword appear in the document is appear the table not all keyword

6. COMPUTE THE WEIGHT

We can compute the weight by these equations:

$$W_{i,j} = f_{i,j} * \log N/n_i \text{ ---[7]}$$

$W_{i,j}$: the weight of the term i in the document j.
N: number of documents in the system.
n_i: the number of documents that term i appear in it.

$$f_{i,j} = freq_{i,j} / MAX_L\ freq_{L,j} \text{ ---[7]}$$

$Freq_{i,j}$ →the number of times the term i appeared in the text of the document j.

MAX L freq L,j→ the maximum is compute over all terms which are mentioned in the text of the documents dj.

but in a way to surpass the threshold. This is a value given to discriminate between relevant and irrelevant documents. Those above the threshold are relevant ones.

ELIMINATING STOPWORDS

Stop words are those words that are repeated in every document, so they are considered as weak to be distinguished, we can not distinguish the content of a text depending on them [16].

There are other benefits from eliminating stopword as "shortening indexing structure" [7] and are useful in making the process faster and doesn't have information Retrieval and the degree of the efficiency of recalling system. [7] It doesn't also burden the system with unnecessary information (Swaine', 1994)

It is not clear which words can be considered stopwords and which cannot. Traditional methods consider that words that are repeated many times are stop words, but there are some words that are repeated in a certain document and considered as important words "indexing terms". But when the subjects are more specialized, as to say a subject specialized in database. Then the use of repeated words, as "index terms" such as the word "computer" are useless to be "index terms". [11].

In the first phase, stop words have been deleted. Those were collected by Al-Shalabi (2004, et al) and they gained 98% success in distinguishing stopwords in addition to deleting some signs that appeared.

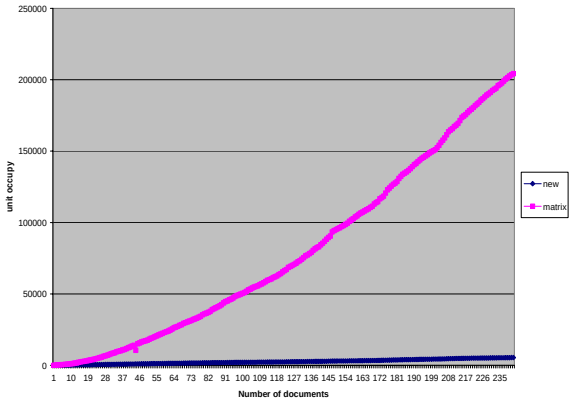

Figure3:
The result show the save space when we used memorized semirings structure

EXPLAIN THE RESULT:

The result may appear overstatement!, this is because the complexion of the first structure is a "matrix" so any document added means the rows increase, and any new term added to the matrix means the number of column increase by one. So, matrix size equal rows × columns, so matrix size will increase in a geometric series. But in the memorized semirings structure any document added means the new table increases and any new terms added to table not make this gap so memorized semirings structure size increase in an arithmetic series.

7. RESULTS:

REFERENCES:

[1] Abu Salem, H., A Microcomputer Based Arabic Bibliographic Information Retrieval system With Relational Thesauri, Ph.D. Thesis, University of Illinois,Chicago,USA,1992.

[2] Adriani, M., and Croft, W., Retrieval Effectiveness of Various Indexing Techniques on Indonesian News Articles,1997.

[3] Aljaly M., Frieder O., On Arabic Search: Improving the Retrieval Effectiveness via light Stemming Approach, in 11th International Conference on Information and Knowledge Mangement (CIKM), Virginia, USA, 2002, pp 340-347.

[4] Aljlaly, M., Frieder, O., and Grossman, D., On Arabic-English Cross-Language Information Retrieval: A Machine Translation Approach, IEEE Third Int'l Conf. On Information Technology:Coding and Computing (ITCC), Las Vegas,Nevada, 2002.

[5] Al-Shalabi, R., Kanaan, G., Al-Jaam, J., Hasnah, A., and Helat, E., Stop-word Removal Algorithm for Arabic Language, proceeding of the 1st International Conference on Information & Communication Technologies: from theory to Applications-ICTTA, Damascus, Syria,2004.

[6] Al-Zoman, A., Al-Faiz, R., Assiri, A., Domain Names: When is it be in Arabic, The International Arab Conference on Information Technology, PP 364-372, 2001.

[7] Baeza-yates, R.,and Rierio-neto, B., Modern Information Retrieval , Addison-
[8] Wesley,New-York,1999.

[9] Carlberger, J., Dalianis, H., Hassel, M., and Knutsson, O., Improving Precision in Information Retrieval for Swedish Using Stemming, In proceeding of NODALIDA '01-13 th Nordic Conference on computational linguistics, Uppsala, Sweden, 2001.

[10] Darwish, k., Building a Shallow Arabic Morphological Analyzer in one Day, Acl Workshop on Computational Approaches to Semitic Language, 2002, PP. 47-57.

[11] Frakes, W., and Baeza-yates, R., Information Retrieval Data Stractures & Algorithms, P T R Prentice Hall, New Jersey, 1992.

[12] Hmeidi, I., Kanaan, G., and Evens, M., Desing and Implementation of Automatic Indexing for Information Retrieval with Arabic Documents, JASIS 48(10):869-881,1997.

[13] Ingwersen, P., Information Retrieval interaction, free download from www.db.dk/pi/iri.

[14] Kanaan, G., Comparing Automatic Statistical and Syntactic Phrase Indexing for Arabic Information Retrieval, Ph.D. Thesis, University of Illinois, Chicago, USA,1997.

[15] Lassi, M., Automatic Thesaurus Construction, university collage of boras, 2002

[16] Salton, G., and McGill, M., Introduction to Modern Information Retrieval, McGraw-Hill, New-York, 1983.

[17] Salton, G., and Buckley, C ., Term-Weighting Approaches in Automatic Text Retrieval, Informatin Processing and Management, 24, pp 513- 523, 1988

[18] salton, G., Automatic Text Processing: the Translation Analysis. And Retrieval of Information by Computer, Addison –wesley Publishing , USA, 1988.

[19] Xu, j., Fraser, A., and Weischedel, R., Empirical Studies in Strategies for Arabic Retrieval, Proceeding of 25th Annual International Conference on Research and Development in Information Retrieval,SIGR,Augest 11-15, PP 269-274, 2002.