Research Journal of Applied Sciences, Engineering and Technology 11(8): 860-864, 2015
DOI: 10.19026/rjaset.11.2096
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: May 20, 2015 Accepted: July 14, 2015 Published: November 15, 2015

Research Article

Similarity Measurements of Vector Space Model on Arabic Text

Ahmad M. Odat
Faculty of Science and Information Technology, Irbid National University, Irbid, Jordan

Abstract: This study presented an effective retrieval model through appling a successful comparison between sets of measurment within Vector Space Model (VSM) and to proof that, we use two mechanism in inverted file, the first is word-oriented mechanism for indexing a text collection and the second is block-oriented mechanism, after removing the stop words. This study use 242 collection of arabic abstract and 60 building collection of arabic queries. During building an inverted file as index file, time and space factors are computed. And after running the system, recall and precision calculated to compare the retrieval efficiency of using inverted file. The VSM have many measurment: Cosine measure, Dice measure, Jaccard measure and Inner product similarity. The study achived an effective retrieval system through appled VSM with jaccard measure comparison with the other measurements, jaccard mesure obtain a good result particularly when using the arabic collection documents. The study also obtained a good result from block-oriented mechanism rather than word-oriented mechanism. As a conclusion, the best information retrival model for arabic documents is VSM with jaccard measure using block-oriented technique.

Keywords: Block-oriented mechanism, inverted file, jaccard measure, precision, recall, Vector Space Model (VSM)

INTRODUCTION

Information Retrieval (IR) can be defined broadly as the study of how to determine and retrieve from a corpus of stored information the portions which are responsive to particular information needs (Tengku and Tengku, 2003). IR is also concerned with text representation, text storage, text organization and the retrieval of stored information items that are similar in some sense to information requests received from users. The term IR covers a wide range of disciplines and have some similarities with many other areas of information processing, e.g., management information systems, database management systems, decision support systems, question-answering systems, natural language processing, as well as document retrieval systems.

The Internet is growing with an increasing rate and it is obvious that it will be difficult to search for information in this gigantic digital library. The estimated size of the Internet, from February 1999, indicates that there are about 800 million pages on the World-Wide Web, on about 3 million servers (Katz and Autor, 1999).

The main objectives from using information retrieval is to enable end-users to perform searching effectively and efficiently. There are two main research areas in information retrieval, one of these areas is to use a knowledge-based approach in information retrieval systems which uses expert systems techniques to encode the expertise possessed by a trained intermediary. The other research area focuses on the development of algorithmic procedures which allow the computer to undertake the functions of a trained intermediary (Salton and McGill, 1983).

We using in this study an arabic corpus, the corpus contains: 242 arabic documents as a collection and we building 60 collection of arabic queries, all these document are built manually. We applied all measures belongs to vector space model: Cosine measure, Dice measure, Jaccard measure and Inner product similarity. We also, implement two kinds of orientation for searching: word and block oriented. We obtained a good result by comparing process done between all these measures.

The most important objectives of this study, to help the arabic end-users to use information retrieval techniques for arabic documents (texts), help researchers and end-users through finding the best IR model for dealing with Arabic texts among multiple information retrieval models, determine the appropriate measure among multiple measures and finally to determine the best retrieval orientation, word- oriented or block-oriented.

VECTOR SPACE MODEL

The vector space model uses non-binary weights that are assigned for the documents and queries index terms (Salton, 1988). This will suggest a partial matching retrieval instead of the relevant/non-relevant matching. The non-binary weights assigned for both the queries and documents are ultimately used to measure the degree of similarity between each of the documents

in store in the system and the user query. Hence, the vector model will also take into consideration documents which match the query terms partially.

The vector model uses the t-dimensional vectors to represent both document and query. For a document dj (where j is the document number) and a query q, their t-dimensional representations are dj and q as follows:
The query q representations is:

$$\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

And the document dj representation is:
where, $w_{i,q} \geq 0$ and t is a total number of index terms in the system.

$$\vec{dj} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

The vector model proposes to evaluate the degree of similarity of the document dj with regard to the query q as the correlation between the vectors dj and q. This correlation can be quantified, for instance, by the cosine of the angle between these two vector (Salton, 1988), That is:

$$sim(d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$$

Inverted file: To describes the concept of the inverted file type of index. Assume a set of documents. Each document is assigned a list of keywords or attributes, with optional relevance weights associated with each keyword (attribute). An inverted file is then the sorted list (or index) of keywords (attributes), with each keyword having links to the documents containing that keyword. The use of an inverted file improves search efficiency by several orders of magnitude, a necessity for very large text files.

Inverted file construction: The structure used for implementing the inverted file is sorted array; we follow the same method described by Harman (1993). in their survey (Inverted Files) where they describe the various structures that can be used in building inverted files (AL-Kharashi, 1991). The following are the steps we precede:

- Removing stop-words from the documents collection, because words which occur in 80% of the documents in the collection are useless for the purpose of retrieval, like articles, punctuations and conjunctions (Yates and Neto, 1999).
- Applying stemming algorithm on the list that is created in step 1, since stems are useful to improve retrieval performance because they reduce variants of the same root word to a common concept, Optional step (Yates and Neto, 1999).
- Sorting the array created in step 3, different sort algorithm could be used in this step, but the best one is the quick sort which is O(n log(n)) time complexity.
- Removing duplication, during this process same words in the same document are regarded as one word, a new column denoting the frequency of a word should be added.
- For each term, add a new entry which contains the number of documents in which that term appears
- for i = 1 to number of documents

  Find maxfreqi
  For j = 1 to number of terms within document i
  $W_{i,j} = (freq_{i,j}/maxfreq_i)$
  $*Log_2(N/n_j)$…… (Yates and Neto, 1999)
  End {for}
  End {for}

where,

$W_{i,j}$	:	Weight of the $term_i$ in $document_j$
$freq_{i,j}$	:	The frequency of $term_j$ in $document_i$
$maxfreq_i$	:	The maximum frequency over all the term in $documnet_i$
N	:	Number of documents
n_i	:	Number of documents the $term_j$ appear.

LITERATURE REVIEW

Vector Model is the most popular model among the research community in information retrieval. Much of this popularity is due to the long term research of Salton (1988). Most of this research revolved around the SMART retrieval system developed at Cornell University (Salton and McGill, 1983). Term weighting for the vector model has also also been investigated thoroughly. Simple term weighting was used early by Salton (1988). further studied the effects of term weighting in the final ranking. Salton (1988) summarize 20 years of experiments in term weighting with the SMART system.

AL-Kharashi (1991) compared between three similarities (cosine, Dice and Jacard) for binary weight vector model and found out that they produced the same ranking for all the queries.

Proposed a speedy method of finding the roots of Arabic. He limited them in one day. This method depends on automatic and statistical methods. The system is called Subway. It is based on entering a couple of words and roots. The program uses these entered words to decide affixes and suffixes. Then the word whose root is to be produced is entered so that the system produces possible roots arranged. 9606 words have been entered to the system. They were extracted from texts called “ZAD”. The second list includes 560, 000 words extracted from (Linguistic Data Consortium

(LCD) Arabic collection). The researcher has been able to enter 270,000 words of them as he used a program called ALPNET to produce the roots. The word which the program ALPNET could not produce its root was ignored by the researcher. There, 290,000 words remained to researcher. These words were used to assess the program.

Jaffer *et al*. (2014) used stemming algorithms to retrieve a greater number of document related to the users query and they aim to evaluate the impact the three different Arabic stemmer on Arabic information retrieval performance for arabic language. The evaluation of the three different stemmers ranked the best performance was achieved by light 10 stemmer in term mean the average precision.

Marie-Claire and Dan (2005) describe a stemmer which is designed to stem conservatively to orthographically correct word forms and recognizing words which do not need to be stemmed, such as proper nouns. He compare the performance of thir stemmer with several other stemmers and propose further work to make this stemmer more effective for information retrieval, topic detection and other linguistic applications.

EXPERIMENT OF THE SYSTEM

Experiment of vector model: We ran the 60 queries against the 242 Arabic documents using the vector model for the four similarity measurement, for each run we enter the query automatically and specify the similarity measurement and then the system retrieve documents then ascending ordered these documents according their similarity value. Table 1 show that the precision value for jaccard measure get best result than the other measures.

Table 2 show that the Average Recall and Precision Values for jaccard measure get best result than the other measures.

Experiment of the inverted file: The researchers used here two mechanism in inverted file the first is word-oriented mechanism for indexing a text collection and the second is block-oriented mechanism in order to see what is better for searching. As we show in the Table 3 the value retrieved in block mechanism get best result than the word mechanism.

Also As we show in the Table 4 and Fig. 1 the value of average retrieved in block mechanism get best result than the word mechanism.

The result in Fig. 2 show that the size requirement for the block mechanism is less than the size requirement for word mechanism.

Figure 2 illustrates the space requirement of the original documents inverted file (block mechanism, word mechanism).

Table 1: Query number 1"علوم الحاسب والمعلومات" search output using VSM

Doc	Similarity (cosine)	Similarity (Dice)	Similarity (Inner)	Similarity (Jaccard)
201	0.30462	0.02363	0.71622	0.02363
171	0.22371	0.00908	0.53448	0.00908
234	0.18403	0.00243	0.47044	0.00243
5	0.11324	0.00202	0.3924	0.00202
202	0.10601	0.00161	0.38481	0.00161
203	0.09009	0.00158	0.31373	0.00158
220	0.08891	0.00143	0.25311	0.00143
169	0.8744	0.00128	0.24949	0.00128
93	0.07276	0.00124	0.23522	0.00124

Table 2: Average recall and precision values for 60 query using VSM

Recall	Precision	Cosine	Dise	jaccard	Inner
0.1	1.0	0.132	0.131	0.130	0.132
0.2	1.0	0.14	0.172	0.170	0.178
0.4	0.75	0.147	0.262	0.261	0.265
0.4	0.67	0.151	0.214	0.213	0.221
0.5	0.71	0.156	0.357	0.355	0.376
0.6	0.67	0.178	0.379	0.335	0.381
0.7	0.53	0.183	0.383	0.385	0.391
0.8	0.44	0.234	0.388	0.389	0.394
1.0	0.4	0.241	0.431	0.434	0.456
1.0	0.36	0.255	0.444	0.440	0.467

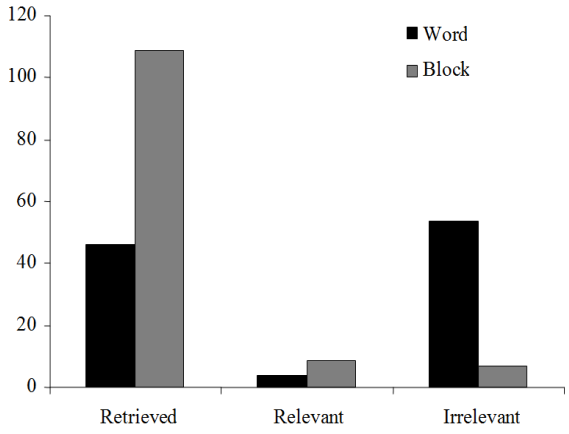

Fig. 1: Average retrieval graph of 60 queries in the inverted file (block, word mechanism)

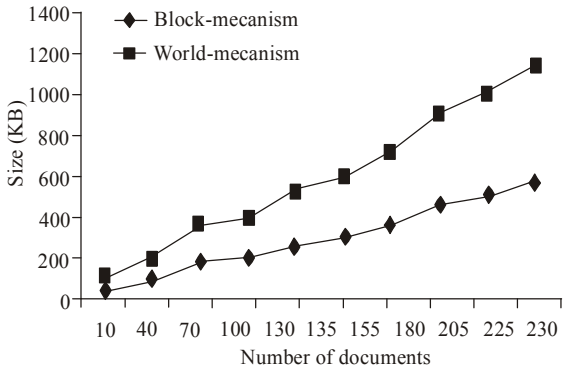

Fig. 2: Space requirement

Figure 3 illustrates the time requirement of the original documents inverted file (block mechanism, word mechanism).

Table 3: Automatic retrieval system results of 60 queries against 242 abstracts using block mechanism and word mechanism retrieval methods (inverted file)

		Block		Word	
Qry No.	Relv. Jdg.	Ret.	Relv.	Ret.	Relv.
Q1	22	78	15	111	26
Q2	11	56	10	134	12
Q3	12	22	11	108	14
Q4	4	3	2	20	5
Q5	13	123	14	156	16
Q6	4	5	0	21	1
Q7	5	2	3	19	3
Q8	2	45	2	118	3
Q9	3	3	2	1	3
Q10	1	32	1	166	2
Q11	5	11	7	26	8
Q12	10	109	9	146	10
Q13	3	95	3	129	3
Q14	9	36	5	165	9
Q15	11	49	6	151	11
Q16	11	50	1	152	11
Q17	24	44	10	157	24
Q18	4	81	2	171	4
Q19	11	65	7	123	9
Q20	3	61	1	96	3
Q21	16	44	3	177	16
Q22	5	106	5	165	5
Q23	4	34	2	121	4
Q24	7	92	6	106	7
Q25	4	36	0	132	4
Q26	26	24	8	46	23
Q27	44	59	20	170	43
Q28	4	43	3	119	4
Q29	12	55	10	128	12
Q30	6	34	3	91	6
Q31	5	35	2	121	5
Q32	5	52	2	72	5
Q33	0	34	0	130	0
Q34	0	33	0	117	0
Q35	2	33	0	117	2
Q36	10	35	2	111	10
Q37	1	57	0	93	1
Q38	2	36	1	142	2
Q39	2	33	0	121	2
Q40	7	38	2	122	7
Q41	4	33	0	117	4
Q42	4	89	3	118	4
Q43	8	77	7	115	8
Q44	19	79	16	160	19
Q45	6	41	5	131	6
Q46	2	55	2	71	2
Q47	0	79	0	153	0
Q48	26	54	16	112	26
Q49	2	66	1	85	2
Q50	3	33	0	118	3
Q51	7	33	0	126	7
Q52	1	67	1	96	1
Q53	2	39	2	69	2
Q54	5	3	2	100	5
Q55	9	33	1	138	9
Q56	3	3	1	122	3
Q57	6	33	1	101	6
Q58	5	60	3	123	5
Q59	10	65	2	161	10
Q60	11	109	3	145	11

The result in Fig. 3 show that the Time requirement for the implement the block mechanism is less than the time requirement for the word mechanism.

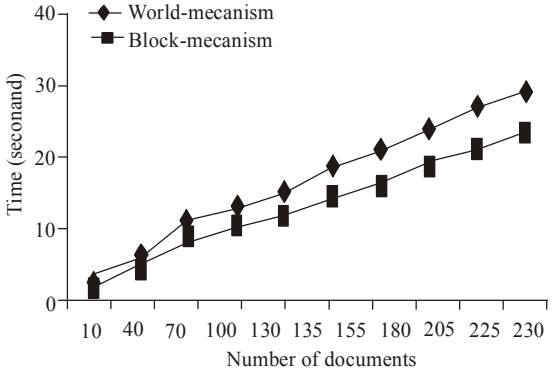

Fig. 3: Time requirement

Table 4: Average retrieval of 60 queries using inverted file (block, word mechanism)

Mechanism	Retrieved	Relevant	Irrelevant
Word	46.13	3.5	53.67
Block	108.7	8.65	6.78

CONCLUSION

This study apply different measures over Vector Space Model and also, use two mechanism in inverted file, first is word-oriented mechanism for indexing a text collection and the second is block-oriented mechanism. The study compares between different measures of VSM and two different mechanism orientation, to calculate and evaluate the recall and precision formula. We noticed that the VSM with Jaccard measure get best strategy result over others VSM measures. And in the second hand we noticed that the search with block-oriented mechanism get best result rather than word-oriented mechanism. This result obtained after making a comprehensive comparative between all VSM measures. These good results we have obtained, dedicated only for arabic documents.

REFERENCES

Al-Kharashi, I.A., 1991. Micro-airs: Microcomputer based Arabic information retrieval system, comparing words, stems, roots as index terms. Ph.D. Thesis, Computer Science Department, Illinois Institute of Technology, Chicago.

Harman, D.K., 1993. The first text retrieval conference (TREC-1) Rockville, MD, U.S.A., 4-6 November, 1992. Information processing and management, 29(4): 411-414.

Jaffer, A., M. Masnizah, K. Ghassan and B. Qusay, 2014. Impact of stemmer on arabic text retrieval. In: Jaafar, A. *et al*. (Eds.), AIRS, 2014. LNCS 8870, Springer International Publishing, Switzerland, pp: 314-326.

Katz, L.F. and D.H. Autor, 1999. Changes in the Wage Structure and Earnings Inequality. In: Ashen-felter, O. and D. Card (Eds.), Handbook of Labor Economics. North Holland Press, Amsterdam, pp: 1463-1555.

Marie-Claire, J. and S. Dan, 2005. Conservative stemming for search and indexing. Proceeding of the SIGIR, 2005.

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

Salton, G., 1988. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, Reading, Mass.

Tengku, M. and S. Tengku, 2003. Character strings to natural language processing in information retrieval. Proceeding of the 6th International Conference on Asian Digital Libraries (ICADL, 2003). Kuala Lumpur, Malaysia.

Yates, R.B. and B.R. Neto, 1999. Modern Information Retrieval. Addison-Wesley, New York.