Comparative Study on Inverted File versus Signature File performance in Information Retrieval System used by Arabic Language

Mohamed Abdeldaiem Abdelhadi

***Abstract*---**In this research paper we have presented a comparison among two Information Retrieval models namely, Inverted file and Signature file for investigating their performance in Arabic Information Retrieval Systems. We have studied both models as to judge the models performance and their effectiveness.

***Keywords*---**Information Retrieval Systems, Arabic Language, Performance Evaluation.

I. INTRODUCTION

Information Retrieval (IR) evaluation has increased in importance and is an active area of research and development. For example, research funding agencies now more than ever require IR research, including digital library projects, to illustrate their applicability and utility to real world problems. This requires evaluation. Furthermore, as advanced information retrieval systems move from research to the real world of commercial competition, designers and developers, vendors, and sales representatives of new information products, such as electronic (or digital) books, search engines, and personal Internet filters want to know whether their products offer potential users and purchasers competitive advantages.

Information Retrieval (IR) refers to the processing of user requests, commonly referred to as queries, to obtain relevant information from collection of documents [1]. Due to historical reasons, documents in a collection are frequently represented through a set of indexing terms or keywords, these keywords could be extracted directly from the text of the document or might be specified by human specialist. An inverted file index consists of a record, or inverted list, for each term that appears in the document. A term's record contains an entry for every occurrence of the term in the document collection identifies the documents and, possibly, gives the location of the occurrences or a weight associated with the occurrences [7].

II. RELATED WORKS

Most researches concentrate on the stemming in Arabic Information Retrieval, because there have been possibilities to deal within its roots or stems as the desired level of analysis for information retrieval .Considerable research on stemming and morphological analysis of the Arabic Language, but no standard IR-oriented algorithm has yet emerged. Four different approaches to Arabic stemming can be identified manually constructed dictionaries, algorithmic light stemmers which remove prefixes and suffixes, morphological analyses which attempt to find roots, and statistical stemmers, which group word variants using clustering techniques. Manually constructed dictionaries of words with stemming information are in surprisingly wide use. Al-Kharashi and Evens worked with small text collections, for which they manually built dictionaries of roots and stems for each word to be indexed [5],[6],[7].

Mohamed Abdeldaiem Abdelhadi, Faculty of Science at Hoon / University of Sirt, Libya. Email id: mabdeldaiem2009@gmail.com

III. INVERTED FILE MODEL

As a matter of fact, there are many models implemented to deal within Inverted files scheme and Signature file too. We have chose those two pseudo codes which can be used to build any sort of an inverted file and signature file as well:

1. 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 [1].
2. Applying stemming algorithm on the list that is created in step1, since stems are useful to improve retrieval performance because they reduce variants of the same root word to a common concept [1].
   The terms created in step 2 are the terms used for indexing.
3. Storing the stem list created in step 2 in an array along with the first character position of each word to represent the location.
4. 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.
5. 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.
6. For each term, add a new entry which contains the number of documents in which that term appears
7. Fori = 1 to number of documents

```
    Findmaxfreqi
    For j = 1 to number of terms within document i
    Wi, j = (freqi, j / maxfreqi) *Log2 (N/nj).
    End { for}
    End {for}
    Where:
    Wi, j: weight of the termi in documentj
    freqi,j: the frequency of termj in documenti
```

Maxfreqi: the maximum frequency over all the term in documneti
N: number of documents
ni: number of documents the termj appear.

IV.SIGNATURE FILE MODEL

The following describes the steps followed to build up the signature file which has to be started by:

```
1. Removing stop-words from the document collection
2. Applying stemming algorithm to the list created in the
   previous step.
3. Computing the weight for all terms in the collection
4. For i = 1 to Number of Documents.
5. Remove duplicate words from document i and
6. Compute term frequency.
7. Sort the document's terms according to the term frequency.
8. Split document i terms to blocks according to term
   frequency and block size:-
For i = 1 to number of blocks
For j = 1 to number of terms in the block
9. Compute the signature of termjas such;
Signature (i) = Signature (i) OR termj_Signature
10. Block_weight (i) = Block_weight (i) + term_weight (j).
11. End {for} End {for}.
```

V. OUR PROPOSED MODEL

Is based on the simplest model construction to understand how it works. For Arabic words stemming, we have used the light stemming method [6] which was very good for our model. The model itself is based on standard query operation methods.

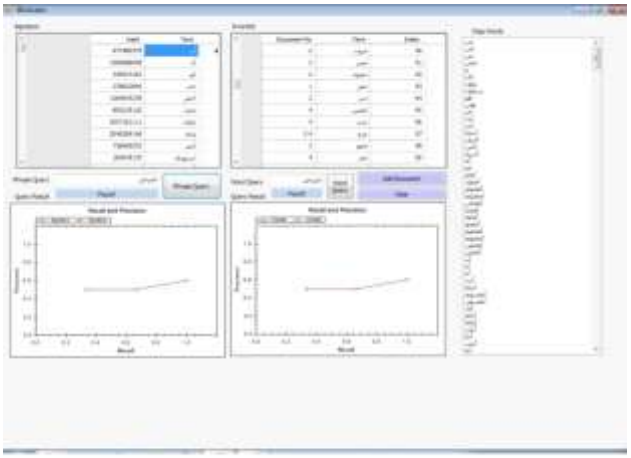


Fig. 1 Inverted vs. Signature file Model construction for Arabic-IRs.

The Inverted file model deals with words query techniques and the signature file should deals with phrases query technique; after doing the classical procedures as (removing stop words, stemming, construct Inverted file, Signature file, storing documents).Thenafter,the performances of both can be compared directly by calculating the Recall and Precision for the query operations on both (Inverted via Signature).

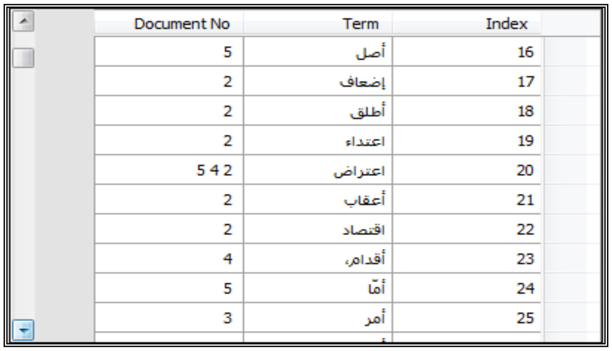
VI.ARABIC DATA TEST COLLECTION

To compare the inverted file versus signature file, we have used in our Model small amount of Arabic documents. The comparisons were totally based on the best recall, and precision ratio. Our system was implemented in C# language. To measure the efficiency of retrieval evaluation; for instance, Recall and precision evaluation measurement were used to determine the performance of retrieved documents. High recall means high number of relevant documents retrieved while high precision means high number of relevant retrieved documents, where the optimal is to have a retrieval model within high recall and high precision as well.

VII. RESULTS ANALYSIS

We have selected just 50 small documents to test the program correctness and the performance of both Models. The results were exact and have no faults. For big collection we need real Corpus to see how the performance of both Models will be...The results of our experiment Model has showed that both models proved its validation, and we have found that, the results of the Inverted Model as shown in Table-1, Training Data, and Data Test set are totally considerable.

TABLE I
ARABIC DATA STOED IN IVERTED FILE



Document No	Term	Index
5	أصل	16
2	إضعاف	17
2	أطلق	18
2	اعتداء	19
5 4 2	اعتراض	20
2	أعقاب	21
2	اقتصاد	22
4	أقدام،	23
5	أمّا	24
3	أمر	25

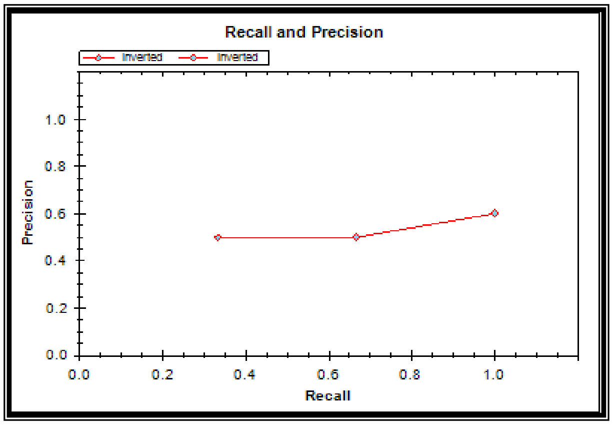
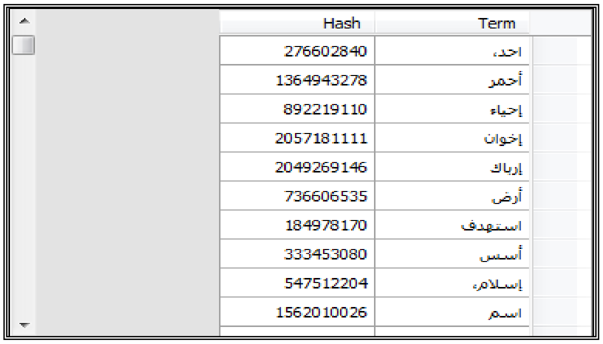


Fig. 2 Recall and Precision Result

In Table-2, we have evaluated small Arabic Data Test after its conversion into Words, which are belongs to Modern Standard Arabic Language Data Base [4]. Our Model was created upon the baseline of a developed Model for Information Retrieval System based on Knowledge Base System [8]. Figure3 showed the performance evaluation by its words matching and queries retrieval used by our proposed Model.

TABLE II
ARABIC DATA STORED IN SIGNATURE FILE



Hash	Term
276602840	احدء
1364943278	أحمر
892219110	إحياء
2057181111	إخوان
2049269146	إرباك
736606535	أرض
184978170	استهدف
333453080	أسس
547512204	إسلامء
1562010026	اسم

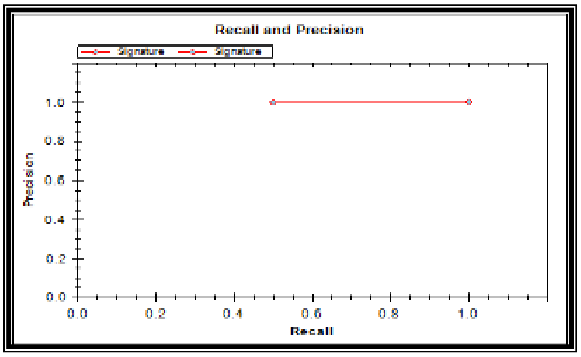


Fig. 3 Recall and Precision Result

VIII. CONCLUSION

Query Operations in Arabic Information Retrieval is not easy as to English, because of its morphology and its unique Formulation. We have found that it is possible to build any Information Retrieval System whenever the system requirements are 100% clear and the proposed model is not a phenomena designed.

ACKNOWLEDGMENT

I would like to thank all of my co-others who helped me in the practical sessions to fulfill this research work.

REFERENCES

[1] Salton, G., "Automatic Text Processing: the Translation Analysis and Retrieval of Information by Computer." Addison-Wesley Publishing, USA, 1988.

[2] Paul McNa, JHU/APL at TREC 2002:‖ Experiments in Filtering and Arabic Retrieval‖, 2002.

[3] Baeza-Yates, R., and Ribeiro-Neto, B. (1999). "Modern Information Retrieval". Addison Wesley.

[4] Abu-Salem, H., Al-Omari, M., and Evens, M.Stemming MethodologiesoverindividualquerywordsforArabicinformationretrieval, *JASIS*, (6), pp.524-529, 1999.

[5] Al-Fedaghi,S.S.andAl-Anzi,F.S.Anewalgorithmto GenerateArabicroot-patternforms.In*Proceedingsofthe11*th*nationalcomputerconference*.KingFahdUniversityofPetroleum&Minerals,Dhahran,SaudiArabia,pp.391-400,1989.

[6] Khoja, S. and Garside, R. Stemming Arabic text. Computing Department, Lancaster University, Lancaster, 1999.

[7] Al-Kharashi, I. and Evens, M.W. Comparing words, stems, and roots as index terms in Arabic information retrieval system. *JASIS*, 45 (8), pp-548-560,1994.

[8] Mohamed. A Abdelhadi, TiruveedulaGopi Krishna ,GhassanKanann," A New Developed Model for Arabic Information Retrieval System based on Knowledge Base System", International Journal of Emerging Research in Management &Technology ,ISSN: 2278-9359 (Volume-2, Issue-11),November 2013.

[9] Mohamed Abdelhadi, TiruveedulaGopi Krishna, GhassanKanann,"Developed Taxonomy for Information Retrieval Systems Based on Arabic Language", International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, Volume-3, Issue-6, pp135-137, January 2014.

About Author

Dr. Mohamed Abdeldaiem Abdelhadi received his B.Sc degree in Computer Science from Sebha University at Sebha-Libya in 1988, M.Sc degree in Computer Engineering in 1991 from Humboldt University-Faculty of TeschnischeInformatik at Berlin-Germany and PhD degree in Computer Information Systems from the University of Banking and Financial Sciences-Faculty of Information Technology Sciences at Amman-Jordan in 2010; current research interests; Information Retrieval Data Mining.More than10 Research papers have been published in various reputed and high impact factor cited International Journals.

http://dx.doi.org/10.15242/IJCCIE.D0315010