Comparison between Different Global Weighting Schemes

Mohammad Othman Nassar, Ghassan Kanaan, and Hussain A.H Awad

***Abstract*—the goal in information retrieval is to enable users to automatically and accurately retrieve data relevant to their queries. One possible approach to this problem is to use the vector space model, which models documents and queries as vectors in the term space. The components of the vectors are determined by the term weighting scheme. This paper compared between a selected set from the available term weighting schemes to determine which weighting method is the best one to be used with Arabic data collections. Our results shows that the best method is the probabilistic inverse (IDFP) method; and we recommend using it as a global weighting method for Arabic data collections.**

***Index Terms*—Information systems; Information retrieval; Vector space model; term weighting schemes evaluation.**

I. Introduction

Recently, people have started dealing with an increasing number of electronic documents in information networks. Finding specific documents that users need from among all available documents is an important issue. Information Storage and Retrieval Systems make large volumes of text accessible to people with information needs [2, 6]. The user provides an outline of his requirement perhaps a list of keywords relating to the topic in the form as a question, or even an example document. The system searches its database for documents that are related to the user's query and presents those which are most relevant.

Most document retrieval systems use keywords to retrieve documents. These systems first extract keywords from documents and then assign weights to the keywords by using different approaches. Such systems have a major problem which is how to decide the weight of each keyword [10, 3]. Gerard Salton was a pioneer in developing techniques for term weighting schemes. He and Christopher Buckley summarize the results of the previous 20 years in their paper [4], which was reprinted in [11].

Most people have used some type of information retrieval system in the form of Internet search engines. Search engines are based on information retrieval models such as the Boolean system, the probabilistic model, or the vector space model [7]. We focus on the vector space model, which models documents and queries as vectors and computes similarity scores using different methods such as cosine, dice, and inner product. We are going to use the inner product in our paper to compute the similarity between documents and queries. The performance of the vector space model depends on the term weighting scheme, that is, the functions that determine the components of the vectors [4]. The weighting scheme is composed of three different types of term weighting: local, global, and normalization, our goal is to compare different global weights to decide which is more useful when used with Arabic data collection. We are going to use an Arabic data collection which was presented for the first time by [14]; this data set is composed from 242 documents and 59 queries, the correct answer for each query (relevant documents) is also known in advanced.

Manuscript received December 15, 2009.

Mohammad Othman Nassar is an assistant professor in the computer information systems department in the university of banking and financial sciences, Amman, Jordan (e-mail: moanassar@yahoo.com, mnassar@aabfs.org).

G. Kanaan is a full professor and he is the head of management information systems department in the university of banking and financial sciences, Amman, Jordan (e-mail: ghkanaan@aabfs.org).

Hussain A.H Awad is an assistant professor in the management information systems department in the university of banking and financial sciences, Amman, Jordan (e-mail: hawad@aabfs.org).

II. Arabic Language Overview

Arabic is the official language of twenty two Middle East and African countries, and is spoken by millions of people all over the world. Arabic language belongs to Semitic group of languages, unlike English language which belongs to the Indo–European language group. The Arabic language orientation is from right-to-left. Arabic alphabets are used in several languages such as Persian, Malay, and Urdu [1]. The characters are consisting of letters, numbers, punctuation marks, space and special symbols (e.g. mathematical notations). It is different from English language in its vowels and diacritic marks; which are normally special marks placed above or under the Arabic letters. However, most recent written Arabic texts are non-vowelized. Arabic language is considered a member of a highly sophisticated category of natural languages, having a very rich morphology where one root can generate several different words of different meanings [13]. The previous mentioned factors about Arabic language make it unique language that needs more investigations; this is exactly what motivates us to study the different weighting schemas, then applying them to Arabic document collection to decide which one of them is the most suitable to be used with the Arabic data collections.

III. Term Weighting

Proper term weighting can greatly improve the performance of the vector space method [13]. A weighting scheme is composed of three different types of term

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

weighting: local, global, and normalization. The term weight is given by LijGiNj; where Lij is the local weight for term i in document j, Gi is the global weight for term i, and Nj is the normalization factor for document j. Local weights are functions of how many times each term appears in a document, global weights are functions of how many times each term appears in the entire collection, and the normalization factor compensates for discrepancies in the lengths of the documents.

The document vectors and query vectors are weighted using separate schemes. The local weight is computed according to the terms in the given document or the query. The global weight, however, is based on the document collection regardless of whether we are weighting documents or queries. The normalization for the documents is done after the local and global weighting. Normalizing the query vectors is not necessary because it does not affect the relative order of the ranked document list.

Local weighting formulas perform well if they work on the principle that the terms with higher within-document frequency are more pertinent to that document [5].

There are many different local weight schemas available in the literature some of them provided in table (1); we will choose only one method which is the logarithm (log) method because the other methods known to have problems, for example; the BNRY method which presented in [5] does not differentiate between terms that appear frequently and terms that appear only once, while the FREQ method which presented in [5], gives too much weight to terms that appear frequently. So the log method which presented in [8] offers a middle ground and that's why we chose it. Logarithms are used to adjust within document frequency because a term that appears ten times in a document is not necessarily ten times as important as a term that appears once in that document. We have to note that fij in the logarithm method is the frequency of term i in document j. Finally the differences between the local weighs are beyond the scope of our research, and we will concentrate our efforts on global weighting methods.

TABLE (1): LOCAL WEIGHTS FORMULAS

Formula	Name	Abbr.
1 if $f_{ij} > 0$ 0 if $f_{ij} = 0$	Binary	BNRY
f_{ij}	Within-document frequency	FREQ
$1+\log f_{ij}$ if $f_{ij} > 0$ 0 if $f_{ij} = 0$	Log	LOGA

TABLE (2): GLOBAL WEIGHTS FORMULAS

Formula	Name	Abbr..
$\log\left(\frac{N}{n_i}\right)$	Inverse document Frequency	IDFB
$\log\left(\frac{N-n_i}{n_i}\right)$	Probabilistic inverse	IDFP
$\frac{F_i}{n_i}$	Global frequency IDF	IGFF
1	No global weight	NONE
$\log\left(\frac{F_i}{n_i}+1\right)$	Log-global frequency IDF	IGFL
$\frac{F_i}{n_i}+1$	Incremented global frequency IDF	IGFI
$\sqrt{\frac{F_i}{n_i}-0.9}$	Square root global frequency IDF	IGFS

Global weighting tries to give a "discrimination value" to each term. Many schemes are based on the idea that the less frequently a term appears in the whole collection, the more discriminating it is [5]. A number of global weighting schemes presented in table (2) and they will be discussed in the following context.

A commonly used global weight is the inverted document frequency measure, or IDF, derived by Sparck Jones [12]. We have used two variations, IDFB [4] and IDFP [5], there formulas are in table (2), where N is the number of documents in the collection and ni is the number of documents in which term i appears. IDFB is the logarithm of the inverse of the probability that term i appears in a random document. IDFP is the logarithm of the inverse of the odds that term i appears in a random document. IDFB and IDFP are similar in that they both award high weight for terms appearing in few documents in the collection and low weight for terms appearing in many documents in the collection; however, they differ because IDFP actually awards negative weight for terms appearing in more than half of the documents in the collection, and the lowest weight IDFB gives is one.

In addition we used a global frequency IDF weight (IGFF) [8], here if a term appears once in every document or once in one document, it is given a weight of one, the smallest possible weight. A term that is frequent relative to the number of documents in which it appears gets a large weight. This weight often works best when combined with a different global weight on the query vector.

The IGFL, IGFI, and IGFS, provided by [13] will be discussed in the following context; IGFL provided is simply a combination of the IDFA and IGFI weights. Like IGFL, IGFS is a combination of formulas. In this case, the authors observed that square root was an excellent local weight, so they adapted it to be a global weight. They found that subtracting larger numbers from Fi/ni improved

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

performance. They do not subtract one because that could cause a global weight of zero for some terms. Finally we will discuss the incremented global frequency (IGFI), Since IGFF already performed best, they add one to it, and the result was IGFI.

It is important to use a normalization factor which usually used to correct discrepancies in document lengths. The idea is to normalize the document vectors so that documents are retrieved independent of their lengths. In this paper we will use the most familiar form of normalization in the vector space model which is the cosine normalization (COSN) [4], where Gi is the global weight, Li is the local weight, and m is the number of terms in document j.

IV. Experiments and Results

As we discuss in the previous sections we will evaluate a selected set from the available global weighting schemes, to do this we used a test collection that have 242 documents and 59 queries, the correct answer for each query (relevant documents) is known in advanced this test collection is created by [14]. Table 3 presents the methods which we plan to compare between them, to test these weighting formulas, we prepare the term list for the given 242 documents in the test collection as follows; Numbers, punctuation, and stop words are removed, then the remaining words form our set of terms. The same operations are conducted on the 59 available queries; then to compare a document and query, we compute their similarity score by computing their dot product.

For each weighting scheme in table 3 we can see two parts, the first part is for the query, and it is composed from local weighting scheme, and global weighting scheme, no normalization is included. The second part is for the document, and it is composed from local weighting scheme, global weighting scheme, and normalization scheme.

TABLE (3): RESULTS FOR THE ARABIC TEST COLLECTION

Document weight	Query weight	Top ten	Scheme name
LOGA IDFB COSN	LOGA IDFB	5.3	Scheme (1)
LOGA IDFP COSN	LOGA IDFP	5.8	Scheme (2)
LOGA IGFF COSN	LOGA IGFF	4.2	Scheme (3)
LOGA IGFL COSN	LOGA IGFL	5.2	Scheme (4)
LOGA IGFI COSN	LOGA IGFI	3.8	Scheme (5)
LOGA IGFS COSN	LOGA IGFS	4.1	Scheme (6)

we implemented the vector space model using visual basic for access applications (VBA) as a programming language; then we carried out our experiments using the weighting schemes in table (3), Then for a given weighting scheme, we computed the similarity between the documents and each query (using dot product see reference [13] for more details) in the test collection and returned a list of documents ranked in order of their similarity scores. To evaluate our results we used a method called Top Ten [13]; Top Ten is the average number of relevant documents in the first ten documents retrieved for a given query. Table (3) presents the Top Ten results.

V. Conclusion

Our results show that the best global weighting scheme was IDFP, so we recommend using it with the Arabic data collections since it can return more relevant documents. As a future work we can compare the local and normalization methods to decide which method is better to be used with the Arabic data collections. Also as a future work we are planning to use other Arabic data sets such as [15, 9] to evaluate local, global, and normalization methods.

References

[1] S.S. Al-Fedaghi, and H.B. Al-Sadoun, 1990. "Morphological Compression of Arabic Text. Information Processing & Management," 26(2): pp. 303-316.

[2] C. J. Van Rijsbergen, "*Information Retrieval*," Butterworths, London, second edition, 1979.

[3] D. Lewis, R. Shapire, J.P. Callan, and R. Papka, "Training algorithms for linear text classifiers," ACM SIGIR'96, Zurich, Switzerland, pp. 298–306, 1996.

[4] G. Salton, and C. Buckley,"Term weighting approaches in automatic text retrieval," Information Processing and Management, pp. 513-523, 1988.

[5] W. B. Croft, and D. J. Harper. Using probabilistic models of document retrieval without relevance information. J. Documentation, 35(4): pp. 285-295, 1979.

[6] G. Salton, and M. J. McGill, "*Introduction to modern information retrieval*," Englewood Cli.s, NJ: Prentice-Hall, 1983a.

[7] T. G. Kolda. "Limited-Memory Matrix Methods with Applications," PhD thesis, Applied Mathematics Program, University of Maryland, College Park, Maryland, 1997.

[8] S. Dumais. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments, and Computers, 23: pp. 229-236, 1991.

[9] A. Goweder, and A. De Roeck, "assessment of significant Arabic corpus," Presented at the Arabic NLP workshop at ACL/EACL 2001, Toulouse, france, 2001.

[10] M. Gordon, "Probabilistic and genetic algorithms in document retrieval," Communications of the ACM, Vol. 31, No. 10, pp.1208-1218, 1988.

[11] K. S. Jones, and P. Willett (Eds.), Readings in information retrieval. San Francisco: Morgan Kaufman, pp. 323 -328, 1997.

[12] K. S. Jones. "A statistical interpretation of term specicity and its application in retrieval," J. Documentation, 28(1): pp. 1-21, 1972.

[13] E. Chisholm, and T. G. Kolda, "new term weighting formulas for the vector space model in information retrieval," Computer Science and Mathematics Division, Oak Ridge National Laboratory, March 1999.

[14] I. Hmedi, and G. Kanaan and M. Evens, "desiegn and implementation of automatic indexing for information retrieval with Arabic documents," Journal of American society for information science, pp. 867-881, 1997.

[15] A. Abdelali, J. Cowie, and H. Soliman, "Building a modern standard Arabic corpus," workshop on computational modeling of lexical acquisition. The split meeting. Croatia, 25th to 28th of july 2005.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)