Annals of Library and Information Studies 52, 2; 2005; 65-67

RANKING FUNCTION FOR QUERY BASED DIGITAL INFORMATION RETRIEVAL

J P S Kumaravel
Information Scientist
Dr.T.P.M.Library
Madurai Kamaraj University
Madurai – 625 021
E-mail: jpskumar@yahoo.com

K Sangeetha
School of Mathematics
Madurai Kamaraj University
Madurai – 625 021
E-mail: ksangeethawins@yahoo.co.in

The paper discusses the problems of information retrieval in digital environment and presents a mathematical model for ranking the retrieved records based on the weight of the key words.

INTRODUCTION

Information is the fundamental unit of transaction in the process of communication. In a restricted connotation, information is a sensible statement, opinion, fact, concept or idea or an association of statements. Organisations store large amount of information in manuals, procedures, documentation, email archives, news clippings, technical reports, research reports and numerous other kinds of publications. In libraries, information about publications such as author, title, keyword, publisher, etc., is stored in a database and a program retrieves the relevant information. Such a system that stores and retrieves information is called an information retrieval system.

Computerized information retrieval systems are query based where Boolean operators are used for getting only the relevant information. In a query based retrieval system, there is no concrete methodology to rank the relevance of the results. The present paper proposes to develop an algorithm based on mathematical models to rank the result set in a query based information retrieval system. Here, a ranking function is used to order documents in terms of their predicted relevance to a particular query. However, it is very difficult to design such a ranking function that can be successful for each and every query of a user or a document collection.

DOCUMENTS AND KEYWORDS

It is known that documents are represented by the metadata elements like author, title, keywords, etc. The keywords are those words that represent the theme of the document and are taken from the title as well as the content pages of the document. The documents are represented by $D = \{d_i \quad I = 1 \ldots n\}$ and $K = \{k_i$ where each k_i is the set containing the keywords of document d_i. Since the core theme of the document is represented by keywords, it is evident that among the keywords, the one, which has more frequency of appearance in the document, can be identified as the main key term. Similarly the other keywords representing a document can be ranked according to their frequency. The keywords along with their weights and the document identification numbers are stored in a separate table. This table is based on a network model as given in the Fig. 1.

D_i = Documents, K_i = Keyword set, Kw_{ij} = Keywords

It is presumed that the information retrieval system is designed with a database containing the following fields

Record Id	-	identifying uniquely a document
Keyword	-	the term representing a theme in the document
Rank of Keyword	-	The frequency of the term in the complete text of the document.

In normal systems when there is a query using Boolean operators the system lists all the records

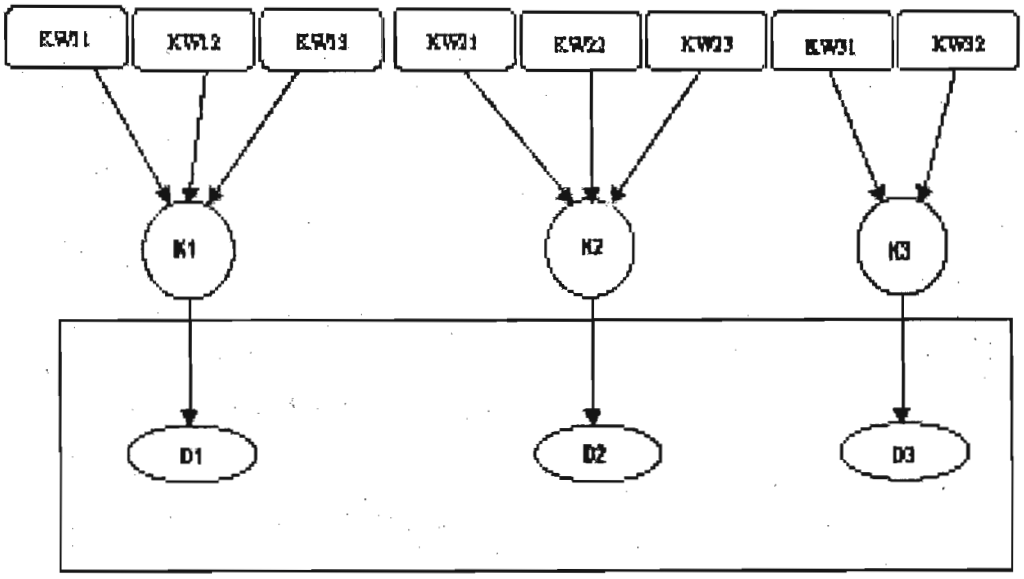

Fig. 1— A network model of documents and keywords

containing the terms given in the query. For example if there is a need for a scholar to retrieve all the documents dealing with the healing of tuberculosis in AIDS patients, here the query terms used are Tuberculosis and AIDS. The Boolean operator used in this query is AND. The search is done in the following manner:

- ❖ system searches those records containing the term tuberculosis
- ❖ system searches those records containing the term AIDS
- ❖ the intersection of the record sets of the two query terms is the result of this query which is a set of records containing the record id and the details of the document.

Here there is no evidence about the relevance of the result. Hence it is proposed to develop a function to rank the relevance of the records retrieved on the basis of the rank or the weight of the terms contained in the database.

RANKING FUNCTION - DEMONSTRATED

Now consider that the following is the resulting record set for the above query:

Step 1

Title id	Term = AIDS	Term = Tuberculosis
T1	30	10
T2	40	40
T3	50	30
T4	60	50
T5	70	20

Now in order to rank the result, the following steps are carried out.

- ❖ Find the average of the weights in the two columns
- ❖ Now apply the averaging function by selecting the Title ids that has the value greater than the averages in the two columns.

Step 2

Now the resultant table will be

Title id	Term = AIDS	Term = Tuberculosis
T2	40	40
T3	50	30
T4	60	50
T5	70	20

In the above table the title id with T2 has lesser value than the average value for the Term AIDS. Similarly T5 has lesser value in the Tuberculosis weight. Hence the two rows T2 and T5 are eliminated and hence we have only two records namely T3 and T4.

Repeat the step 1 and step 2 for T3 and T4. The result table is

Title id	Term = AIDS	Term = Tuberculosis
T4	60	50

As a result, T4 has the maximum value for both the terms AIDS and Tuberculosis. Hence this is the most relevant record for the query. When the same procedure is applied to the remaining records the second ranked title is foun d to be T2 and so on.

Demonstration

Let $T = \{x_n : x_i$ is a hit record for the query term1 and query term 2$\}$

Let a_1 be the average weight for the hit records with query term 1

Let a_2 be the average weight for the hit records with query term 2

The averaging function is defined as

$f(avg) = \{y_n : y_i$ is a hit record for the query term1 and query term 2 such that weight $(v_i > a_1$ and $a_2\}$

The procedure is repeated until the record with maximum weight is reached.

If the above function fails to arrive at the top most ranked document, then the averaging function is modified by dividing the weights averages of the two terms by 2. Now apply a mid function by adding and subtracting this value to the averaging function value. Now apply the step 2.

CONCLUSION

The proposed model ensures maximum hits while reducing noise among the retrieved documents or information.

BIBLIOGRAPHY

1. FAN (Weiguo) *et al.* Discovery of context : Specific ranking functions for effective information retrieval using genetic programming. *IEEE Transactions on Knowledge and Data Engineering.* 16, 4; 2004.

2. GORDON (M) and PATHAK (P). Finding information on the WWW: The retrieval effectiveness of search engines. *Information Processing and Management.* 35; 2, 1999; 141-180.