Expert Systems with Applications 42 (2015) 1223–1234

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# A new fuzzy logic based ranking function for efficient Information Retrieval system

Yogesh Gupta, Ashish Saini *, A.K. Saxena

*Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Agra 282110, Uttar Pradesh, India*

ARTICLE INFO

*Article history:*
Available online 19 September 2014

*Keywords:*
Information Retrieval
Fuzzy Logic Controller
Precision
Recall
Ranking function

ABSTRACT

The relevant documents from large data sets are retrieved with the help of ranking function in Information Retrieval system. In this paper, a new fuzzy logic based ranking function is proposed and implemented to enhance the performance of Information Retrieval system. The proposed ranking function is based on the computation of different terms of term-weighting schema such as term frequency, inverse document frequency and normalization. Fuzzy logic is used at two levels to compute relevance score of a document with respect to the query in present work. All the experiments are performed on *CACM* and *CISI* benchmark data sets. The experimental results reveal that the performance of our proposed ranking function is much better than the fuzzy based ranking function developed by Rubens along with other widely used ranking function *Okapi-BM25* in terms of precision, recall and F-measure.

## 1. Introduction

In recent times, Information Retrieval (IR) has become an important area of research in computer science. IR systems are used in several application domains such as web search, digital library search, blog search, information filtering, recommender system and social search, etc. The major concern of IR is to find "relevant" information or documents with respect to user need, modeled through a query from large data corpus in appropriate time interval (Salton & McGill, 1983; Yates & Berthier, 1999). IR system uses ranking function to retrieve relevant documents by computing relevance score between a query and a document. Although the conventional statistical ranking functions such as *Cosine, Jaccard* (Salton, 1998), *Euclidean* and *Okapi* (Robertson, Walker, & Beaulieu, 1999) have been extensively used but these measures fail to capture inherent features of documents and queries due to subjectivity involved in natural language text.

Natural language is often vague and uncertain (Subtil, Mouaddib, & Faucout, 1996). It is very difficult to determine something that is uncertain and vague with crisp formulas and crisp logics. Therefore, fuzzy logic (Zadeh, 1965) is found very suitable, to handle this uncertainty, vagueness and impreciseness. It transforms vagueness and uncertainty of documents, queries and their characteristics into fuzzy membership functions (Zadeh, 1997). The documents are retrieved by query with the help of the rules framed in Fuzzy Inference System (FIS) (Abraham, Lihong, & Zhiqiang, 1992; Jang & Sun, 1997; Ross, 1997; Sugeno, 1985a; Zadeh, 1997). Fuzzy logic uses degrees of memberships to express relevance unlike the Binary/Boolean model which is based on binary decision criterion i.e. {relevant, not relevant}.

In the present paper, a new fuzzy logic based ranking function is proposed. The performance of proposed ranking function is compared with *Okapi-BM25* and Rubens' ranking function (Rubens, 2006). Vector Space Model (VSM) is used as an IR model to develop proposed ranking function due to its strengths over other models, which are explained in Section 2. The main contributions of this paper are following:

- The proposed ranking function is based on composite *FIS*, which has two levels: *first level FIS* and *second level FIS*. *First level FIS* consists of two Fuzzy Logic Controllers (*FLCs*). First *FLC* is for structuring the features of documents and second *FLC* is for structuring the features of queries. *Second level FIS* consists of one *FLC*.
- The proposed ranking function retrieves the relevant documents on the basis of different variables; those capture the features of documents and queries as well.
- New fuzzy rules are framed in this paper for each *FLC* at each level of *FIS*.
- New linguistics variables are used to transform existing knowledge and information into fuzzy rules.

The rest of the paper is structured as follows. In Section 2, a brief description of VSM and work related to the already developed

* Corresponding author. Tel.: +91 562 2801224; fax: +91 562 2801226.
*E-mail address:* ashish7119@gmail.com (A. Saini).

http://dx.doi.org/10.1016/j.eswa.2014.09.009
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

ranking functions model are presented to form the necessary theoretical foundation for this work. The details of proposed fuzzy logic based ranking function and comparison of its important features with Rubens' approach are presented in Section 3. In Section 4, the experimental results and analysis are discussed. Finally, conclusion and future directions are drawn in Section 5.

## 2. Related work and theoretical foundation

There are different factors, which affect the performance of an IR system, but ranking function is one which affects the most (Lancaster & Warner, 1993). Ranking functions match the documents or information to a user's query and rank them according to the relevance score in descending order. The documents and queries both need to be transformed into a model that can be effectively processed by computers to facilitate this relevance estimation process. VSM (Cordon, Moya, & Zarco, 2004; Haase, Steinmann, & Vejda, 2002; Harman, 1993; Jones & Furnas, 1987; Mercier & Beigbeder, 2005; Robertson, 1997; Salton & Buckley, 1988; Witten, Moffat, & Bell, 1999; Yap & Wu, 2005; Yates & Berthier, 1999) is considered as one of the most successful IR models.

This section describes various advantages of VSM using as an IR model and its important features followed by the discussion on the literature of ranking functions.

### 2.1. Vector Space Model

Vector Space Model is used as an IR model in present paper to develop the proposed ranking function because of following advantages:

- It is simple and fast model as documents and queries are represented in the form of vectors in $n$-dimensional space, where $n$ is the number of unique terms used to describe the contents of documents and queries (Cordon, Viedma, Pujalte, Luque, & Zarco, 2003). Therefore, the properties of these vectors such as similarity and closeness can be studied easily.
- It can handle weighted terms.
- It produces a ranked list as output and that the indexing process is automated which means a significantly lighter workload for the administrator of the collection.
- It is easy to modify individual vectors, which is essential for the query expansion technique and logic based ranking functions.

VSM is based on the assumption that the relevance of a document with respect to a query is correlated with the distance between that query and document. A block schematic of queries and documents represented as vectors in VSM, is shown in Fig. 1.

The representation of documents and queries can be extended by including their features. An empirically validated document feature is the number of term occurrences within a document (term frequency or *tf*) (Salton, 1968). The intuitive justification for this feature is that a document that notifies a term more often is more likely to be relevant for that term. Another important feature is the potential for a term to discriminate between documents, named as inverse document frequency (or *idf*) (Jones, 1972). This particular feature (*idf*) has been observed to be inversely proportional to the number of term occurrences in a data corpus. The terms, those are common in a corpus, less likely to be used to discriminate relevant and irrelevant documents.

### 2.2. Ranking function

Many researchers have developed different ranking functions using VSM as IR model in the past. The major contributions in developing such type of ranking functions are categorized and discussed under following subsections.

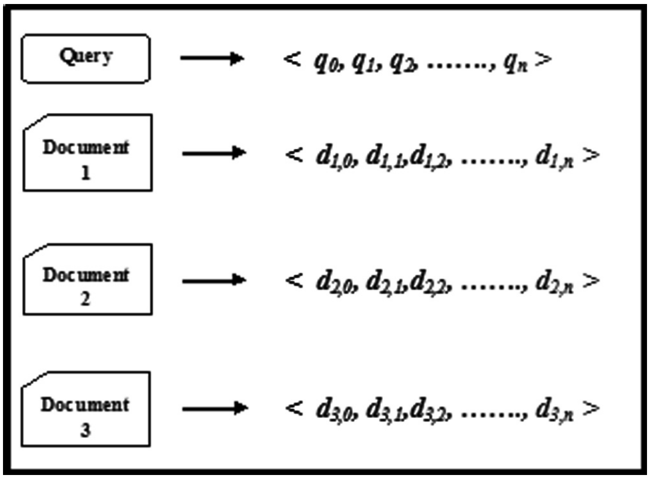
Query
$< q_0, q_1, q_2 \ldots\ldots., q_n >$
Document 1
$< d_{1,0}, d_{1,1}, d_{1,2}, \ldots\ldots., d_{1,n} >$
Document 2
$< d_{2,0}, d_{2,1}, d_{2,2}, \ldots\ldots., d_{2,n} >$
Document 3
$< d_{3,0}, d_{3,1}, d_{3,2}, \ldots\ldots., d_{3,n} >$

**Fig. 1.** Vector Space Model.

#### 2.2.1. Statistical ranking functions

There are different conventional ranking functions in literature such as *Cosine*, *Jaccard* (Salton, 1998) and *Okapi* (Robertson, Walker, & Beaulieu, 1999), etc. *Cosine* ranking function computes cosine of the angle between the query and document vector. The assumption used in the *Cosine* is that the document length has no impact on relevance but later on Singhal, Salton, Mitra, and Buckley (1996) found that more documents judged to be relevant actually were found in longer documents. *Jaccard* is defined as the intersection of document and query vectors divided by the union of document and query vectors. Subsequently *Okapi* is developed as ranking function to overcome shortcomings of *Cosine* and *Jaccard*. This ranking function not only considers the term frequency, but also the length of the document and average length of the whole collection. *Okapi-BM25* (Christopher, Raghavan, & Schutze, 2009) is another latest variant of *Okapi* which enhances the performance of IR system. The mathematical representation of *Okapi-BM25* is given by (1)–(3).

$$Okapi - BM25(Q, D_i) = \sum_{T \in Q} W \frac{(k_1 + 1)tf}{K + tf} \times \frac{(k_3 + 1)qtf}{k_3 + qtf} \tag{1}$$

where,

$$K = (k_1(1 - b) + b.dl/avdl) \tag{2}$$

$$W = \log(N - n + 0.5)/(n + 0.5) \tag{3}$$

$Q$ is a query that contains the words $T$. $D_i$ is a document in data set D. $k_1$, $b$ and $k_3$ are constant parameters. *tf* is the term frequency of the term with a document, *qtf* is the term frequency in the query. $N$ is the number of documents and $n$ is the number of documents containing the term. *dl* and *avdl* are the document length and average document length respectively.

Unfortunately, the ranking functions mentioned above are not able to capture all the features of queries and documents due to the reasons already explained in previous section.

#### 2.2.2. Evolutionary algorithm based and/or hybrid ranking functions

Some researchers used evolutionary algorithms such as Genetic Algorithm (GA) and Genetic Programming (GP) to construct ranking function for enhancement of IR system. Pathak, Gordon, and Fan (2000) propose a new weighted matching function to overcome the limitations of statistical ranking functions, which is