Information Retrieval Estimation via Fuzzy Probability

Zhiheng Huang
Department of Computer Science
The Australian National University
Australia
zhiheng@cs.anu.edu.au

Tamás D. Gedeon
Department of Computer Science
The Australian National University
Australia
tom@cs.anu.edu.au

Abstract *Fuzzy logic has been recognized as a useful approach in support of information retrieval. It helps identify partially matched documents for a given query so that ranking the relevant documents becomes straightforward. Unfortunately, research on fuzzy information retrieval only focuses on the relevant estimation of the retrieved documents but lacks the estimation of imprecise probability of obtaining such relevant documents. This paper makes use of fuzzy probability (FP) to estimate the possibility of probability of retrieving documents with a certain level of relevance. It first reviews the FP calculation method proposed in Huang and Shi [6], then presents an enhanced FP calculation method. In addition, a novel FP calculation method is proposed. Realistic examples are provided to demonstrate the use of the fuzzy probability estimation.*

Keywords Information retrieval, fuzzy logic, fuzzy probability.

1 Introduction

An information retrieval system allows users to efficiently retrieve documents that are relevant to their interests. To measure the efficiency and effectiveness of the retrieval, the *recall* and the *precision* have been used to justify the performance. The former refers to the ratio of the number of correctly retrieved documents to that of the whole available relevant documents, whilst the latter refers to the ratio of the number of correctly retrieved documents to that of all retrieved documents. The goal of the retrieval system is to achieve better recall and higher precision.

The collection of documents from which the selected ones have to be retrieved might be extremely large and the use of terminology might be inconsistent. One of the major problems in information retrieval is that usually it cannot be guaranteed that the user queries include all of the actual relevant words that occur in the desired documents. Also it often happens that words with several meanings occur in a document, but in a rather different context from that expected by the querying person.

Many existing information retrieval systems are expressed with combination between keywords and phrases search according to the direct keyword matching method to get the information which users need. In particular, they can be classified into four categories: the boolean model, vector model, probabilistic model, and fuzzy model.

The boolean model is a simple retrieval model based on set theory and boolean algebra. Since the concept of a set is quite intuitive, the boolean model provides a frame-work which is easy to grasp. The queries in a boolean model are specified as boolean expressions. But the major problem with the boolean model is that the system produces a set of related documents that exactly match the query while rejecting all other partial or non-matching documents [1].

The vector model recognizes that the use of binary weights in the boolean model is too limiting and proposes a framework in which partial matching is possible. This is accomplished by assigning non-binary weights to index terms in queries and in documents. These term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query. By sorting the retrieved documents in a decreasing order of this similarity degree, the vector model takes into consideration documents which they match the query terms only partially. The main resultant effect is that the ranked document answer set is a lot more precise than the document answer set retrieved by the boolean model.

The probabilistic model attempts to capture the information retrieval problem within a probabilistic framework. The fundamental idea is as follows. Given a query q and a document d in the collection, the probabilistic model tries to estimate the probability that the user will find the document d interesting. The model assumes that this probability of relevance depends on the query and the document representations only. Furthermore, the model assumes that there is a subset of all documents which the user prefers as the answer set for the query q.

In order to make the information retrieval more intelligent, fuzzy logic [14] has been applied to automated information retrieval [12] to deal with the imprecise information. For example, information retrieval methods by using the concept matrix, where the elements in a concept matrix represent relevant values between concepts (key

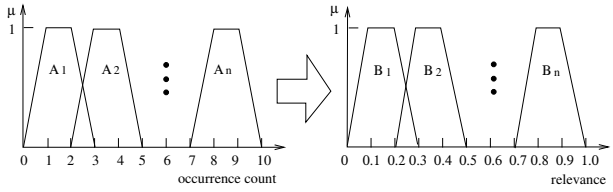

Figure 1: Fuzzy rules representing the relation between occurrence count and relevance of documents

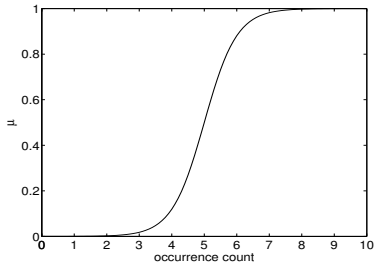

Figure 2: Sigmoid function transforming occurrence counts into membership degrees

words), has been proposed in [4, 11, 2]. Fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques has been presented in [10, 5]. Fuzzy tolerance relation between key words has been introduced in [3, 7, 8, 9]. The technique of transforming occurrence counts into possibilistic fuzzy importance degrees has also been proposed.

Unfortunately, current research on fuzzy information retrieval only focuses on the relevant estimation of the retrieved documents but lacks the estimation of imprecise probability of obtaining such relevant documents. The combination of using fuzzy terms to describe the relevance of documents and using fuzzy probability to describe the possibility of obtaining such relevant documents provides a new way to handle ambiguous and imprecise information. This paper makes use of fuzzy probability to estimate the possibility of probability of retrieving documents with a certain level of relevance. It first reviews the FP calculation method proposed in Huang and Shi [6], then presents an enhanced FP calculation method. In addition, a novel FP calculation method is proposed.

The rest of the paper is organized as follows: Section 2 describes a generic information retrieval model using fuzzy logic, with extra attention paid to the work proposed by Kóczy et al. [9]. Section 3 reviews the FP calculation method proposed by Huang and Shi [6] and then extends it to an enhanced one. A novel FP calculation method is also been proposed. Section 4 gives examples to show the fuzzy probability estimation. Finally, Section 5 concludes the paper and points out important further work.

2 Generic fuzzy based information retrieval

Fuzzy rules are widely used to model complicated and non-linear systems due to their capability of obtaining arbitrary accuracy for the problem on hand and their transparency in terms of interpretability. Therefore, they can be used to represent the relation between the occurrence count of key word in a document and the relevance of this document to the key word. Generally, the connection between occurrence count and relevance is monotonic, i.e., the more counts the key word appears in the document, the more relevant the document to the query specified by such a key word. This can be represented by fuzzy rules listed as:

If the occurrence count is A_i, then the relevance of the document is B_i,

where A_i and B_i (as shown in Fig. 1), $i = 1, 2, \ldots, n$, are fuzzy terms defined in the universe of discourse of occurrence count and relevance respectively. Such a fuzzy rule base can assign a certain level of relevance to a particular document.

For simplicity, only a few fuzzy rules may be used in a rule base to focus on the high occurrence count and high relevance parts in Fig 1. In the extreme case, Kóczy et al. [9] implicitly adopt one fuzzy rule. In particular, it simply uses a sigmoid function to transform the occurrence counts into possibilistic fuzzy importance degrees. The typical characteristics of such a sigmoid function can be seen in Fig 2. Depending on the length of the document and the occurrence counts of particular key words, the characteristics of the sigmoid curve can change. Membership degrees generated by the occurrence count transformation can be interpreted as possibility measures of a certain document being relevant for a query. The query can thus be carried out by using such degrees.

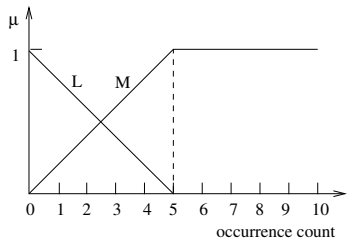

Figure 3: Fuzzy membership function of occurrence counts

Table 1: Occurrence counts of key words in the collection of documents of the example

W/D	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
"damage"	0	0	0	0	0	1	21	2	0	0	2	0	0	0	0	0	1	6	0	1
"bedroom"	0	0	0	0	0	0	2	3	0	0	4	9	1	0	0	1	0	0	2	0
"carpet"	2	0	9	4	0	0	4	8	0	0	29	0	0	0	0	0	0	0	0	51

Table 2: Possibilistic relevance degrees of key words in the selected collection of documents of the example

W/D	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
"damage"	0	0	0	0	0	0.2	1	0.4	0	0	0.4	0	0	0	0	0	0.2	1	0	0.2
"bedroom"	0	0	0	0	0	0	0.4	0.6	0	0	0.8	1	0.2	0	0	0.2	0	0	0.4	0
"carpet"	0.4	0	1	0.8	0	0	0.8	1	0	0	1	0	0	0	0	0	0	0	0	1

The same query examples as in [9] will be presented in the following. The difference is that for simplicity, a piecewise fuzzy membership function M representing "more" occurrence count (as shown in Fig 3) is used here to replace the sigmoid function used in [9]. An additional fuzzy membership function L representing the "less" occurrence count is also defined for FP calculation purpose (see section 3).

The examples are based on the legal data base *http://www.AustLII.edu.au*. A small representative examples of 20 documents $D_1, D_2, \ldots, D_{20}$ have been selected for the experiment. According to the occurrence counts of words, three key words *damage*, *bedroom* and *carpet* have been used for these 20 documents and the occurrence counts of in the collection of 20 documents are shown in Table 1. Based on the "more" fuzzy membership function M (or the occurrence count - relevant degree transformation) defined in Fig 3, the occurrence counts in Table 1 are transformed into possibilistic relevance degrees as shown in Table 2.

As with [9], ad hoc categories of retrieved documents are defined to illustrate the fuzzy relevant degree: *Very important documents* ($\mu = 1$), *Rather important documents* ($1 > \mu \geq 0.9$), *Reasonably important documents* ($0.9 > \mu \geq 0.7$), *Somewhat important documents* ($0.7 > \mu \geq 0.4$), *Tangentially important documents* ($0.4 > \mu > 0$). As a matter of course, the threshold values can be adapted to any concrete applications.

Query 1. "Damage"
Very important documents: D_7, D_{18}.
Rather important documents: $\varnothing$
Reasonably important documents: $\varnothing$
Somewhat important documents: D_8, D_{11}.
Tangentially important documents: D_6, D_{17}, D_{20}.

In the following simple joint and conjunction queries are discussed. A simple joint and a simple conjunction queries have two key words, connected by "or" and "and" operators respectively. The queries of "bedroom or carpet" and "bedroom and carpet" are provided as examples. First, the single queries for each key word are listed as follows.

Query 2. "Bedroom"
Very important documents: D_{12}
Rather important documents: $\varnothing$
Reasonably important documents: D_{11}
Somewhat important documents: D_7, D_8, D_{19}.
Tangentially important documents: D_{13}, D_{16}.

Query 3. "Carpet"
Very important documents: D_3, D_8, D_{11}, D_{20}.
Rather important documents: $\varnothing$

Table 3: Occurrence counts and relevance degrees of the queries of "bedroom or carpet" and "bedroom and carpet"

W/D	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
"or" occurrence	2	0	9	4	0	0	6	11	0	0	33	9	1	0	0	1	0	0	2	51
"or" degree	0.4	0	1	0.8	0	0	1	1	0	0	1	1	0.2	0	0	0.2	0	0	0.4	1
"and" occurrence	0	0	0	0	0	0	2	3	0	0	4	0	0	0	0	0	0	0	0	0
"and" degree	0	0	0	0	0	0	0.4	0.6	0	0	0.8	0	0	0	0	0	0	0	0	0

Reasonably important documents: D_4, D_7.
Somewhat important documents: D_1
Tangentially important documents: $\varnothing$

If "bedroom or carpet" is queried, it means any of those two words equally contributes to fining the relevant documents. The occurrence count of such a joint query can thus be calculated as the *sum* of the occurrence counts of two key words. The relevance degrees can thus be obtained using such summed key word count and the "more" fuzzy term defined in Fig 3. It is worth noting that the relevance degree of a document in a joint query is not the sum of the relevance degrees of two individual queries. If however, "carpet and bedroom" is queried, the conjunction occurrence count and relevance degree are calculated by choosing the *min* of two occurrence counts and relevance degrees. Table 3 shows the occurrence counts and the relevance degrees of the queries of "bedroom or carpet" and "bedroom and carpet" respectively. These queries can thus be summarized in the following:

Query 4. "Bedroom or carpet"
Very important documents: D_3, D_7, D_8, D_{11}, D_{12}, D_{20}.
Rather important documents: $\varnothing$
Reasonably important documents: D_4
Somewhat important documents: D_1, D_{19}.
Tangentially important documents: D_{13}, D_{16}.

Query 5. "Bedroom and carpet"
Very important documents: $\varnothing$
Rather important documents: $\varnothing$
Reasonably important documents: D_{11}
Somewhat important documents: D_7, D_8.
Tangentially important documents: $\varnothing$

3 Fuzzy probability estimation

Probability is the most common theory in uncertainty. It has been used to model uncertainty for decades. However, circumstances arise in which precise values for the probability estimation of events are unavailable, these situations often occur in three situations: 1) probabilities are supplied by human beings; 2) the number of sample events is too small to give a sound probability assessment; and 3) the events to be assessed using probability are fuzzy events, i.e., they are represented by fuzzy sets.

This paper deals with the third case as the queries involve fuzzy events. One example is to estimate the probability of retrieving "very relevant" documents. As "very relevant" can be represented by a fuzzy set, the probability of such documents is hence imprecise. This cannot be modelled by the traditional probability theory and it therefore gives rise to the use of fuzzy sets to represent the imprecise probabilities, which is denoted as *fuzzy probability*. Fig. 4 shows an example of fuzzy probability of obtaining "very important" documents. As can be seen, it is certain (with a possibility value of 1) to obtain "very important" documents with probability 30%, and it is quite possible (with a possibility value of 0.8) to obtain "very important" documents with a range of probability from 20% to 35%. The higher possibility value leads to the narrower estimation of probability. In the extreme case, the peak point in the fuzzy probability set corresponds to the traditional probability estimation. That is, the traditional probability estimation is just a specific case of fuzzy probability estimation.

Fuzzy probability can provide a range of probabilities of fuzzy events if a possibility value is given. This is achieved by using the α-cut level ($\alpha \in [0, 1]$) from the horizontal point of view. If the fuzzy probability is interpreted from the vertical point of view, the term *possibility of probability* to obtaining fuzzy events can be used. For example, one can say, the possibility of probability of 20% in obtaining "very important" is 0.8, so is the probability of 35%, but not the probability of 40%, as 40% gives a possibility of 0.4 to obtaining "very important" documents. In summary, fuzzy probability is used to estimate ranges (rather than singleton values) of obtaining fuzzy events, from the horizontal perspective, whilst possibility of probability is used to provides the

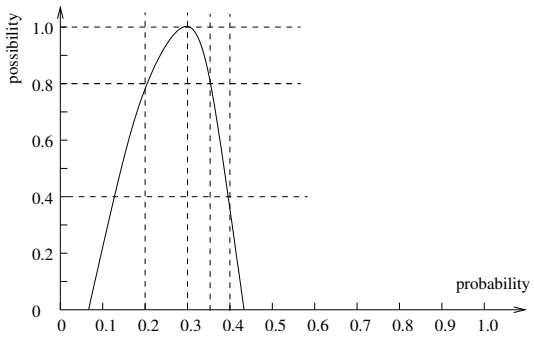

Figure 4: A fuzzy probability estimation of obtaining "very important" documents

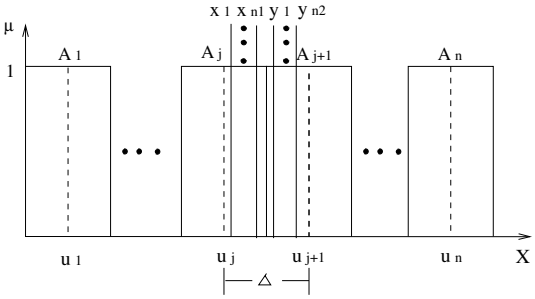

Figure 5: Exclusive intervals of the universe of discourse

possibilities of given probabilities to obtaining fuzzy events, from the vertical perspective. They are two different interpretations of the same fuzzy probability figure (such as Fig. 4) and they may be interchangeable.

There are considerable studies on fuzzy probability. For example, the possibility-probability distribution [6] has been recently proposed to estimate fuzzy risk in an example of floods where the probability of exceeding losses from floods is not one value but a fuzzy number. The problem of decision making [13], selecting a best alternative action, in the face of a fuzzy probability assessment is investigated.

This paper first gives a brief review of the FP calculation method proposed in Huang and Shi [6]. Such a calculation divides events (flood loss in that case) into mutually exclusive categories, then the possibility of probability of such categories is computed as a fuzzy number. The second step is creative but the first step may not well represent the imprecise information existing in the real world. Instead of using mutually exclusive categories, fuzzy division of events is adopted in this paper, resulting in an enhanced FP calculation method. In addition, a novel and intuitive method of calculating fuzzy probability via fuzzy membership values is proposed.

3.1 Original FP calculation method

This subsection reviews the FP calculation method proposed by Huang and Shi [6]. The definition of *fuzzy probability* is first introduced, which is interchangeable with *possibility of probability* in this paper.

Definition 1 *Let $X = \{A_1, A_2, \ldots, A_n\}$ be n fuzzy terms or intervals in the space of interest, $P = [0, 1]$ be the universe of discourse of probability, then $\pi_{A_i}(p)$, where $i \in \{1, \ldots, n\}$, $p \in P$, be fuzzy probability of A_i $(i = 1, \ldots, n)$ occurring being p.*

In [6], A_i, $i = \{1, \ldots, n\}$, are evenly distributed in the space of interest as shown in Fig 5. $U = \{u_1, \ldots, u_n\}$ are the middle values of A_i and u_i ($i = \{1, \ldots, n\}$) are called *controlling points*. Let $u_{j+1} - u_j \equiv \triangle$, $j = \{1, \ldots, n-1\}$. For a compact demonstration, only the space between u_j and u_{j+1} is considered in calculating the FP of fuzzy term A_j. That is, the FP calculation presented here only concerns the right part of fuzzy set A_j, although the left part should be also considered in the same manner to compute the whole fuzzy term A_j's fuzzy probability. For simplicity, the fuzzy term A_j used hereafter only considers the right part of A_j as shown in Fig 5. Assume n_1 data points $X = \{x_1, \ldots, x_{n_1}\}$ fall in $[u_j, u_j + \frac{\triangle}{2}]$ and n_2 data points $Y = \{y_1, \ldots, y_{n_2}\}$ fall in $(u_j + \frac{\triangle}{2}, u_{j+1}]$. Let $n = n_1 + n_2$. Hence, the fuzzy probability of $\frac{n_1}{n}$ for A_j occurring can be defined as 1. That is,

$$\pi_{A_j}(\frac{n_1}{n}) = 1. \tag{1}$$

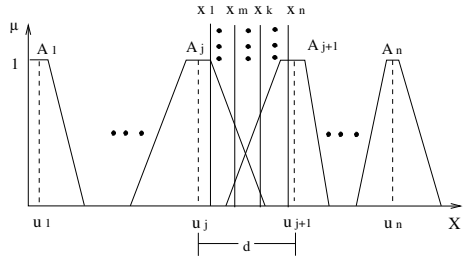

Figure 6: Fuzzy membership functions of the universe of discourse

Now considering the possibility that one of n_1 data may leave interval A_j. It is obvious that the maximal value of x_i, $i = \{1, \ldots, n_1\}$, is mostly possible to leave. Without losing generality, let assume x_{n_1} as shown in Fig 5 be such a data point, then the possibility of probability of A_j occurring being $\frac{n_1-1}{n}$ can be calculated as:

$$\pi_{A_j}(\frac{n_1-1}{n}) = \frac{x_{n_1} - u_j}{\triangle}. \tag{2}$$

Similarly, if two data points are considered to move away from A_j, it is high likely that this happens to the two data points which have the maximal values, say x_{n_1-1} and x_{n_1}. As x_{n_1-1} is less likely to move away than x_{n_1} due to it being closer to the controlling point of interval A_j. The possibility of probability of A_j occurring being $\frac{n_1-2}{n}$ is thus calculated based on x_{n_1-1}. That is,

$$\pi_{A_j}(\frac{n_1-2}{n}) = \frac{x_{n_1-1} - u_j}{\triangle}. \tag{3}$$

Similar processes are carried out until all the n_1 data points move away from interval A_j. It is worth noting that the controlling points u_i, $i = \{1, \ldots, n\}$, are the barriers which prevent the data from moving too far away from their original positions.

On the other hand, one or more of the data points from $Y = \{y_1, \ldots, y_{n_2}\}$ may move to A_j when there is a disturbance in the experiment. Again, it is safe to assume that the minimal value of $Y = \{y_1, \ldots, y_{n_2}\}$, say y_1, is most likely to move due to the fact that it is has the closest distance to the controlling point of fuzzy term A_j. The possibility of probability of A_j occurring being $\frac{n_1+1}{n}$ can thus be calculated as:

$$\pi_{A_j}(\frac{n_1+1}{n}) = \frac{u_{j+1} - y_1}{\triangle}. \tag{4}$$

The possibility of probability of A_j occurring being $\frac{n_1+2}{n}$ can be similarly calculated as:

$$\pi_{A_j}(\frac{n_1+2}{n}) = \frac{u_{j+1} - y_2}{\triangle}. \tag{5}$$

Similar processes are carried out until all the data points in $Y = \{y_1, \ldots, y_{n_2}\}$ are moved into A_j.

3.2 Enhanced FP calculation method

Instead of using mutually exclusive intervals, fuzzy divisions of data points are employed (as shown in Fig 6) as an extension of the method proposed in [6]. In this figure, the trapezoidal fuzzy membership functions are adopted for computational simplicity. $U = \{u_1, \ldots, u_n\}$ are the middle core of A_i (i.e., the middle point of the α-cut with $\alpha = 1$). Note that $(u_{j+1} - u_j)$ vary for $j = \{1, \ldots, n-1\}$. Again, attention is drawn into the range of $[u_j, u_{j+1}]$. Let the distance of this range be d. Assume n data points $X = \{x_1, \ldots, x_n\}$ fall in $[u_j, u_{j+1}]$. Let the fuzzy membership value of X with respect to fuzzy functions A_j and A_{j+1} be $\mu_{A_j}(x_i)$, $\mu_{A_{j+1}}(x_i)$ respectively, where $i = \{1, \ldots, n\}$. Let $S_1 = \sum_{i=1}^{n} \mu_{A_j}(x_i)$, $S_2 = \sum_{i=1}^{n} \mu_{A_{j+1}}(x_i)$, and $S = S_1 + S_2$, the fuzzy probability of $\frac{S_1}{S}$ for A_j occurring can be defined as 1. That is,

$$\pi_{A_j}(\frac{S_1}{S}) = 1. \tag{6}$$

Now considering the possibility of that one of n data may leave from fuzzy term A_j. It is worth noting that here the move causes the graduate fuzzy membership value changes for A_j and A_{j+1}, rather than the sudden jump from one interval to another (as described in subsection 3.1). The datum which has the minimal positive value of fuzzy term A_j, say, x_k, $k \in [1, n]$, is mostly possible to leave, then the possibility of probability of A_j occurring being $\frac{S_1 - \mu_{A_j}(x_k)}{S}$ can be calculated as:

$$\pi_{A_j}(\frac{S_1 - \mu_{A_j}(x_k)}{S}) = \frac{x_k - u_j}{d}. \tag{7}$$

Similarly, two data points, say x_k and x_{k-1}, $k \in [1, n]$ and $k-1 \in [1, n]$, which have the minimal positive membership values with fuzzy term A_j, are most likely to move away from A_j. As x_{k-1} is less likely than x_k due to it being closer to the controlling point of fuzzy set A_j. The possibility of probability is thus calculated based on x_{k-1}. That is,

$$\pi_{A_j}(\frac{S_1 - \mu_{A_j}(x_k) - \mu_{A_j}(x_{k-1})}{S}) = \frac{x_{k-1} - u_j}{d}. \tag{8}$$

Similar processes are carried out until all the S_1 membership values of data points x_i, $i = \{1, \ldots, k\}$, move away from fuzzy term A_j.

Conversely, it is possible for some data points moving towards fuzzy term A_j so that the total membership associated with it (i.e., S_1) increases. Suppose x_m has the minimal positive value of fuzzy term A_{j+1} and it is thereby mostly possible to leave A_{j+1}, converting the fuzzy membership value of A_{j+1} to that of A_j. The possibility of probability of A_j occurring being $\frac{S_1 + \mu_{A_{j+1}}(x_m)}{S}$ can be calculated as:

$$\pi_{A_j}(\frac{S_1 + \mu_{A_{j+1}}(x_m)}{S}) = \frac{u_{j+1} - x_m}{d}. \tag{9}$$

Similarly, the possibility of probability of A_j occurring being $\frac{S_1 + \mu_{A_{j+1}}(x_m) + \mu_{A_{j+1}}(x_{m+1})}{S}$ can be calculated as:

$$\pi_{A_j}(\frac{S_1 + \mu_{A_{j+1}}(x_m) + \mu_{A_{j+1}}(x_{m+1})}{S}) = \frac{u_{j+1} - x_{m+1}}{d}. \tag{10}$$

Similar processes are carried out until all the S_2 membership values associated with fuzzy term A_{j+1} move away.

3.3 A novel FP calculation method

The original and enhanced FP calculation methods presented in subsection 3.1 and 3.2 make use of the ratios of distances in calculating the possibility of probability for each fuzzy term. The newly proposed method calculates the fuzzy probability using the ratios of the fuzzy membership values.

Equation (6) holds here as with the enhanced FP calculation method. Now considering the possibility that one of n data may leave from fuzzy term A_j. The datum x_k, $k \in [1, n]$, which has the minimal positive value of fuzzy term A_j, is mostly possible to leave, the possibility of probability of A_j occurring being $\frac{S_1 - \mu_{A_j}(x_k)}{S}$ can be calculated as:

$$\pi'_{A_j}(\frac{S_1 - \mu_{A_j}(x_k)}{S}) = \frac{\mu_{A_{j+1}}(x_k)}{\mu_{A_j}(x_k)}. \tag{11}$$

Similarly, two data points, say x_k and x_{k-1}, $k \in [1, n]$ and $k-1 \in [1, n]$, which have the minimal positive membership values with fuzzy term A_j, are most likely to move away from A_j. As x_{k-1} is less likely to move away than x_k due to it being closer to the controlling point of fuzzy set A_j. The possibility is thus calculated based on x_{k-1}. That is,

$$\pi'_{A_j}(\frac{S_1 - \mu_{A_j}(x_k) - \mu_{A_j}(x_{k-1})}{S}) = \frac{\mu_{A_{j+1}}(x_{k-1})}{\mu_{A_j}(x_{k-1})}. \tag{12}$$

Similar processes are carried out until all the S_1 membership values of data points x_i, $i = \{1, \ldots, k\}$, move away.

Conversely, it is possible for some data points moving towards fuzzy term A_j so that the total membership associated with it (i.e., S_1) increases. Suppose x_m has the minimal positive value of fuzzy term A_{j+1} and it is thereby mostly possible to leave A_{j+1}, converting the fuzzy membership value of A_{j+1} to that of A_j. The possibility of probability of A_j occurring being $\frac{S_1 + \mu_{A_{j+1}}(x_m)}{S}$ can be calculated as:

$$\pi'_{A_j}(\frac{S_1 + \mu_{A_{j+1}}(x_m)}{S}) = \frac{\mu_{A_j}(x_m)}{\mu_{A_{j+1}}(x_m)}. \tag{13}$$

Similarly, the possibility of probability of A_j occurring being $\frac{S_1 + \mu_{A_{j+1}}(x_m) + \mu_{A_{j+1}}(x_{m+1})}{S}$ can be calculated as:

$$\pi'_{A_j}(\frac{S_1 + \mu_{A_{j+1}}(x_m) + \mu_{A_{j+1}}(x_{m+1})}{S}) = \frac{\mu_{A_j}(x_{m+1})}{\mu_{A_{j+1}}(x_{m+1})}. \tag{14}$$

Similar processes are carried out until all the S_2 membership values associated with fuzzy term A_{j+1} move away.

The proposed method results in the calculated fuzzy probability values in a range of $[0, +\infty)$. In order to obtain valid fuzzy values, the transformation function $\mathbb{T} : \pi' \in [0, +\infty) \rightarrow \pi \in [0, 1]$ as follows is applied to every calculated π'.

$$\pi = \frac{2}{1 + e^{-\pi'}} - 1. \tag{15}$$

Table 4: Fuzzy probability of fuzzy term L in three FP calculation schemes

$\pi_A(p)$	$\frac{0}{20}$	$\frac{12}{20}$	$\frac{13}{20}$	$\frac{13.8}{20}$	$\frac{14.6}{20}$	$\frac{15.4}{20}$	$\frac{16}{20}$	$\frac{16.6}{20}$	$\frac{16.8}{20}$	$\frac{17}{20}$	$\frac{17.2}{20}$	$\frac{17.6}{20}$	$\frac{18}{20}$	$\frac{19}{20}$	$\frac{20}{20}$
Enhanced 1	0	0	0.2	0.2	0.2	0.4	0.4	1	0.8	0.8	0.8	0.6	0.6	0	0
Enhanced 2	0	0	0.2	0.2	0.2	0.4	0.4	1	0.95	0.95	0.95	0.9	0.9	0.71	0
newly proposed	0	0	0.12	0.12	0.12	0.32	0.32	1	0.96	0.96	0.96	0.64	0.64	0	0

Table 5: Fuzzy probability of fuzzy term M in three FP calculation schemes

$\pi_A(p)$	$\frac{0}{20}$	$\frac{1}{20}$	$\frac{2}{20}$	$\frac{2.4}{20}$	$\frac{2.8}{20}$	$\frac{3}{20}$	$\frac{3.2}{20}$	$\frac{3.4}{20}$	$\frac{4}{20}$	$\frac{4.6}{20}$	$\frac{5.4}{20}$	$\frac{6.2}{20}$	$\frac{7}{20}$	$\frac{8}{20}$	$\frac{20}{20}$
Enhanced 1	0	0	0.6	0.6	0.8	0.8	0.8	1	0.4	0.4	0.2	0.2	0.2	0	0
Enhanced 2	0	0.71	0.9	0.9	0.95	0.95	0.95	1	0.4	0.4	0.2	0.2	0.2	0	0
newly proposed	0	0	0.64	0.64	0.96	0.96	0.96	1	0.32	0.32	0.12	0.12	0.12	0	0

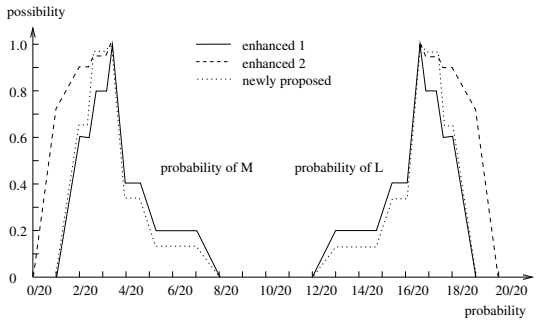

Figure 7: Fuzzy probability of fuzzy term L and M respectively

4 Illustrative examples

In this section, the examples presented in section 2 are used to show the fuzzy probability estimation. In order to calculate the fuzzy probability, the complementary fuzzy set "less" L of the "more" fuzzy set M is first constructed (see Fig 3). The fuzzy probability of each fuzzy set is then calculated according to the enhanced and newly proposed FP calculation methods. As fuzzy partition is used in the example (otherwise, it is a boolean information retrieval model), the original FP calculation method proposed in [6] is not applicable here.

FP estimation of query 1. Given the query word "damage", the fuzzy probabilities of fuzzy term L and M are obtained (as shown in table 4 and table 5 respectively) using three schemes, namely, the enhanced FP calculation with fuzzy term M's controlling point being 5 (*enhanced 1* in the table), being 21, which is the maximal occurrence count for all documents (*enhanced 2*), and the newly proposed fuzzy probability calculation (*newly proposed*). Fig. 7 shows the fuzzy probability of fuzzy term L and M with three different schemes.

The fuzzy probabilities generated here can be used to estimate the FP of retrieving documents with a certain level of relevance. In the following, the *enhanced 1* scheme is used for such estimations (other schemes apply in the same manner). For example, apply the α-cut of 1 associated with *very important documents* to the fuzzy probability of M, a probability of $\frac{3.4}{20}$ is obtained which is the probability of obtaining such *very important documents*. Similarly, if the α-cut of 0.4 which is associated with the *somewhat important documents* is applied to fuzzy probability of M, the range of $[1.67/20, 4.6/20]$ is estimated as the probability range of retrieving such *somewhat important documents*. The combination of using fuzzy terms to describe the relevance of documents and using fuzzy probability to describe the possibility of obtaining such relevant documents provides a new way to handle ambiguous and imprecise information.

Note that in each scheme, the fuzzy probability terms of M and L are complemented with each other on every α-cut level, $\alpha \in [0, 1]$. This is in accordance with the probability theory in which the sum of probability of all possible events is equal to 1.

FP estimation of query 4. In this example, only the fuzzy probability of fuzzy term M which accounts for the relevance of the documents is considered, whilst the fuzzy term L which accounts for the non-relevance of the documents is omitted. For simplicity, only *enhanced 1* scheme is chosen to estimate query 4 (of course, other schemes can be applied). First the fuzzy probabilities of fuzzy term "more" M for key words "bedroom" and "carpet" are calculated respectively (as shown in Fig. 8).

Since both of the two key words contribute to finding the relevant documents, the probability of finding joint relevant documents may be calculated as the *sum* of the two probability ranges. In particular, let the probability

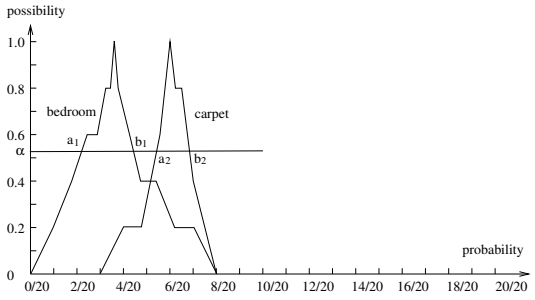

Figure 8: Fuzzy probability of fuzzy terms M for "bedroom" and "carpet"

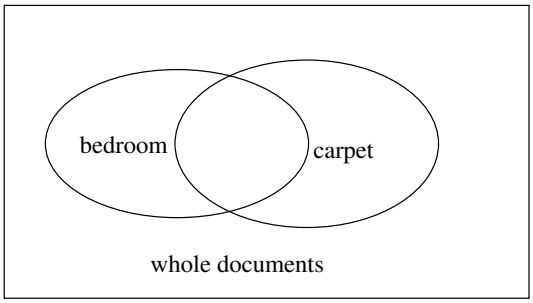

Figure 9: Set representation of documents

range of "bedroom" and "carpet" on α-cut level ($\alpha \in [0, 1]$) be $[a_1, b_1]$ and $[a_2, b_2]$ respectively. The probability range of the jointly query in the same α-cut is calculated as $[a_1+a_2, b_1+b_2]$. However, this raise a potential problem in which some documents may be repeatedly counted. This can be explained in Fig 9. The ellipses including "bedroom" and "carpet" stand for the documents relevant to "bedroom" and "carpet" queries respectively. Such ellipses may have blur boundary to represent the fuzziness of relevance. The area of the ellipse is proportional to the number of documents it contains. The *sum* operator is a *bold* estimation as it always gives a wider estimation range. Similarly, the *max* operator is too *conservative*. These two operators can estimate the lower and upper bounds of the probability of the joint query, but neither of them provides an accurate estimation. One alternative is to calculate $[a_1 + a_2 - a_1 * a_2, b_1 + b_2 - b_1 * b_2]$ as the estimated probability range.

FP estimation of query 5. This example attempts to estimate the probability range of a conjunction query, in which both "bedroom" and "carpet" are included. The *min* operator may be used to estimate the probability range of conjunction query, but it may be too *bold*, i.e., offers a wider probability range. A better way may adopt the *product* operator.

5 Conclusion

Based on the work of [6], an enhanced FP calculation method has been proposed in this paper. Furthermore, a novel FP calculation method has been proposed via the use of membership values. Realistic examples have been provided to demonstrate the use of the FP estimation.

For single key word queries, FP can be successfully used to estimate the fuzzy probability of finding documents with a certain level of relevance, providing a useful way of modelling imprecise knowledge involving both possibility and probability uncertainty. However, much work is required to improve the accuracy of FI estimation on joint, conjunction or even more complex queries. In order to achieve this goal, it is essential to integrate the knowledge of document relevance (relation) into fuzzy probability, which is one of the important future tasks.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley, 1999.

[2] S. M. Chen and J. Y. Wang. Document retrieval using knowledge-based fuzzy information retrieval techniques. *IEEE Transactions on Systems, Man, and Cybernetics*, Volume 25, pages 793 – 803, 1995.

[3] T. D. Gedeon and L. T. Kóczy. Hierarchical co-occurence relations. In *IEEE International Conference on Systems, Man, and Cybernetics*, Volume 3, pages 2750 – 2755, 1998.

[4] G. T. Her and J. S. Ke. A fuzzy information retrieval system model. In *Proc. 1983 National Computer Symposium*, Volume 1, pages 147 – 155, 1983.

[5] Y. J. Horng, S. M. Chen, Y. C. Chang and C. H. Lee. A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. *IEEE Transactions on Fuzzy Systems*, Volume 13, pages 216 – 228, 2005.

[6] C. F. Huang and P. J. Shi. Fuzzy risk and calculation. In *18th International Conference of the North American on Fuzzy Information Processing Society*, pages 90 – 94, 1999.

[7] L. T. Kóczy and T. D. Gedeon. Information retrieval by fuzzy relations and hierarchical co-occurences. Technical Report Part 1, TR97 – 01, 1997.

[8] L. T. Kóczy and T. D. Gedeon. Information retrieval by fuzzy relations and hierarchical co-occurences. Technical Report Part 2, TR97 – 03, 1997.

[9] L. T. Kóczy, T. D. Gedeon and J. A. Kóczy. Fuzzy tolerance relations and relational maps applied to information retrieval. *Fuzzy sets and systems*, Volume 126, pages 49 – 61, 2002.

[10] D. H. Kraft, J. Chen and A. Mikulcic. Combining fuzzy clustering and fuzzy inferencing in information retrieval. In *IEEE International Conference on Fuzzy Systems*, Volume 1, pages 375 – 380, 2000.

[11] D. Lucarella and R. Morara. First: Fuzzy information retrieval system. *J. Information Sci.*, Volume 17, pages 81 – 91, 1991.

[12] S. Miyamoto. *Fuzzy sets in information retrieval and cluster analysis*. Kluwer, Dordrecht, 1990.

[13] R. R. Yager. Decision making with fuzzy probability assessments. *IEEE Transactions on Fuzzy Systems*, Volume 7, 1999.

[14] L. A. Zadeh. Fuzzy sets. *Information and Control*, Volume 8, 1965.