INFORMATION RETRIEVAL ESTIMATION VIA FUZZY PROBABILITY

Zhiheng Huang, The Australian National University, Australia, zhiheng@cs.anu.edu.au

Tamás D. Gedeon, The Australian National University, Australia, tom@cs.anu.edu.au

ABSTRACT

Fuzzy logic has been recognized as a useful approach for 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 estimation of relevancy of the retrieved documents, but lacks the estimation of imprecise probability of obtaining such relevant documents. This paper makes an attempt to the latter. It extends the fuzzy probability (FP) calculation method proposed by Huang and Shi to an enhanced FP calculation method. Furthermore, it proposes a novel FP calculation method. An example is provided to illustrate the use of fuzzy probability estimation.

Keywords Fuzzy Probability, Possibility of Probability, Fuzzy Logic, Information Retrieval.

1 INTRODUCTION

Information retrieval systems allow users to efficiently retrieve documents that are relevant to their interests. These systems are 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. 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 provides a framework in which partial matching is possible. The probabilistic model attempts to capture the information retrieval problem within a probabilistic framework.

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. To resolve this problem, fuzzy logic [11] has been applied to information retrieval [9] to deal with the imprecise information. For example, information retrieval using the concept matrix, where the elements in a concept matrix represent relevant values between concepts (key words), has been proposed in [2, 8]. Information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques has been presented in [4, 7]. Fuzzy tolerance relation based information retrieval has been introduced in [3, 6].

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 relevancy 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 this attempt. It first extends the FP calculation method proposed by Huang and Shi [5] to an enhanced FP calculation method, and then proposes a novel FP calculation method.

The rest of the paper is organized as follows: Section 2 outlines the FP calculation method proposed by Huang and Shi [5] and then extends it to an enhanced one. Furthermore, it proposes a novel FP calculation method. Section 3 presents an example to illustrate the calculation of fuzzy probability. Finally, Section 4 concludes the paper and points out important further work.

2 FUZZY PROBABILITY ESTIMATION

Probability has been used to model uncertainty for decades. However, there are circumstances in which precise values for the probability estimation of events are not available. These 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 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 obtaining such documents is hence imprecise, which gives rise to *fuzzy probability*.

There are considerable studies on fuzzy probability. For example, Zadeh [12] has proposed the framework of fuzzy probability. Yager [10] has investigated how to select a best alternative action in the face of a fuzzy probability assessment. Recently, Huang and Shi [5] have presented a method to calculate the possibility-probability distribution to estimate fuzzy risk.

Section 2.1 first outlines the FP calculation method proposed by Huang and Shi [5]. 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. Instead of using mutually exclusive categories, Section 2.2 extends the work of [5] to an enhanced FP calculation method using fuzzy intervals. Furthermore, Section 2.3 proposes a novel method to calculate fuzzy probability via fuzzy membership values.

2.1 Original FP Calculation Method

First, the definition of *fuzzy probability* is introduced, which is interchangeable with *possibility of probability* in this paper.

Definition 1 Let $X = \{A_1, A_2, \dots, A_n\}$ 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 $i = 1, \dots, n$, $p \in P$, be fuzzy probability of A_i occurring being p .

In [5], A_i , $i = 1, \dots, n$, are evenly distributed in the space of interest as shown in Fig 1. $U = \{u_1, \dots, u_n\}$ are the middle values of A_i and u_i , $i = 1, \dots, n$, are called *controlling points*. Let $u_{j+1} - u_j \equiv \Delta$, $j = 1, \dots, n - 1$. For simplicity, only the region between u_j and u_{j+1} is considered in

Figure 1: Exclusive intervals of the universe of discourse

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 . Assume n_1 data points $X = \{x_1, \dots, x_{n_1}\}$ fall in $[u_j, u_j + \frac{\Delta}{2}]$ and n_2 data points $Y = \{y_1, \dots, y_{n_2}\}$ fall in $(u_j + \frac{\Delta}{2}, u_{j+1}]$. Let $n = n_1 + n_2$. The fuzzy probability of $\frac{n_1}{n}$ for A_j occurring can be defined as 1. That is,

$$\pi_{A_j}\left(\frac{n_1}{n}\right) = 1. \quad (1)$$

Now consider the possibility that one of n_1 data may leave interval A_j . It is obvious that the maximal value of x_i , $i = 1, \dots, n_1$, say x_{n_1} , is mostly possible to leave, then the possibility of probability of A_j occurring being $\frac{n_1-1}{n}$ can be calculated as:

$$\pi_{A_j}\left(\frac{n_1-1}{n}\right) = \frac{x_{n_1} - u_j}{\Delta}. \quad (2)$$

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,

$$\pi_{A_j}\left(\frac{n_1-2}{n}\right) = \frac{x_{n_1-1} - u_j}{\Delta}. \quad (3)$$

Similar processes are carried out until all n_1 data points move away from interval A_j . It is worth noting that the controlling points $u_i, i = 1, \dots, n$, 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 $Y = \{y_1, \dots, y_{n_2}\}$ may move to A_j when there is a disturbance in the experiment. Again, it is safe to assume that the minimal value of $Y = \{y_1, \dots, y_{n_2}\}$, that is y_1 , is most likely to move due to the fact that it 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:

$$\pi_{A_j}\left(\frac{n_1+1}{n}\right) = \frac{u_{j+1} - y_1}{\Delta}. \quad (4)$$

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

$$\pi_{A_j}\left(\frac{n_1+2}{n}\right) = \frac{u_{j+1} - y_2}{\Delta}. \quad (5)$$

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

2.2 Enhanced FP Calculation Method

Instead of using mutually exclusive intervals, this section extends the work in Section 2.1 to an enhanced FP calculation method which uses fuzzy intervals. Fig. 2 shows such an example in which trapezoidal

Figure 2: Fuzzy intervals of the universe of discourse

fuzzy membership functions are employed. $U = \{u_1, \dots, u_n\}$ are the middle cores of A_i (i.e., the middle points of the α -cuts with $\alpha = 1$). Note that $(u_{j+1} - u_j)$ may vary for $j = 1, \dots, n - 1$. Again, attention is drawn into the range of $[u_j, u_{j+1}]$. Assume the distance of this range is d and n data points $X = \{x_1, \dots, x_n\}$ 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 $\mu_{A_j}(x_i), \mu_{A_{j+1}}(x_i), i = 1, \dots, n$, respectively. 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,

$$\pi_{A_j}\left(\frac{S_1}{S}\right) = 1. \quad (6)$$

Now consider the possibility of that one of n data may leave from fuzzy term A_j . It is worth noting that 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 Section 2.1). 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. Thus, the possibility of probability of A_j occurring being $\frac{S_1 - \mu_{A_j}(x_k)}{S}$ can be calculated as:

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

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,

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

Similar processes are carried out until all the S_1 membership values of data points $x_i, i = 1, \dots, k$, 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 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:

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

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:

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

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

2.3 A Novel FP Calculation Method

The original and enhanced FP calculation methods presented in Section 2.1 and 2.2 make use of the ratios of distances in calculating the possibility of probability for each fuzzy term. This section proposes a new method which calculates the fuzzy probability using the ratios of the fuzzy membership values.

Equation (6) holds here in the initial status. Now consider 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:

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

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,

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

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

Conversely, it is possible for some data points moving towards fuzzy term A_j so that the total membership associated with it 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:

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

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:

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

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 interpretable fuzzy values, the transformation function $T : \pi' \in [0, +\infty) \rightarrow \pi \in [0, 1]$ is applied as follows:

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

3 AN ILLUSTRATIVE EXAMPLE

It is generally assumed that the connection between occurrence count and relevancy is monotonic in information retrieval systems. That is, the more counts a key word appears in a document, the more relevant the document to the query specified by such a key word. Kóczy et al. [6] use a sigmoid function to transform the occurrence counts into possibilistic fuzzy importance degrees. Membership degrees generated by the occurrence count transformation can be interpreted as possibility measures of a certain document being relevant for a query.

The same query example as in [6] is outlined. 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 [6].

Figure 3: Fuzzy membership function of occurrence counts

The example is based on the legal data base <http://www.AustLII.edu.au>. A small representative examples of 20 documents $D_1, D_2, \dots, D_{20}$ are selected for the experiment. The key word *damage* is used and its occurrence counts and relevancy measures (based on the *more* fuzzy set M) in the 20 documents are shown in Table 1. As with [6], ad hoc categories of retrieved documents are defined to

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Occurrence	0	0	0	0	0	1	21	2	0	0	2	0	0	0	0	0	1	6	0	1
Relevancy	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

Table 1: Occurrence counts and relevancy measures of key word *more*

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* ($0.4 > \mu > 0$). In so doing, the query can be made as follows:

Query 1. “Damage”

Very important documents: D_7, D_{18} .

Rather important documents: $\emptyset$

Reasonably important documents: $\emptyset$

Somewhat important documents: D_8, D_{11} .

Tangentially important documents: D_6, D_{17}, D_{20} .

To illustrate the fuzzy probability estimation for the query, the complementary fuzzy set *less* of the *more* fuzzy set 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. Note that the original FP calculation method [5] is not applicable here as the events are fuzzy.

FP estimation of query 1. Given the query word “damage”, the fuzzy probabilities of fuzzy term L and M can be calculated. Table 2 shows the fuzzy probability of fuzzy term M . Note that 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*) are employed. Fig. 4 shows the fuzzy probability of fuzzy term L and M for three different schemes.

The fuzzy probabilities generated here can be used to estimate the FP of retrieving documents with a certain level of relevancy. For example, if the *enhanced 1* scheme is used for such estimations, apply

$\pi_A(p)$	0	1	2	2.4	2.8	3	3.2	3.4	4	4.6	5.4	6.2	7	8	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

Table 2: Fuzzy probability of fuzzy term M in three FP calculation schemes

Figure 4: Fuzzy probability of fuzzy term L and M respectively

the α -cut of 1 (membership value 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 relevancy 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.

It is worth noting 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.

4 CONCLUSION

This paper makes an attempt to estimate fuzzy probability for information retrieval. It first extends the fuzzy probability (FP) calculation method proposed by Huang and Shi [5] to an enhanced FP calculation method. Furthermore, it proposes a novel FP calculation method. An example is provided to illustrate the use of the fuzzy probability estimation.

This paper concerns a single key word query only. The work to deal with the FI estimation on joint, conjunction or even more complex queries is worth further investigating.

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*, 25:793 – 803, 1995.
- [3] T. D. Gedeon and L. T. Kóczy. Hierarchical co-occurrence relations. In *IEEE International Conference on Systems, Man, and Cybernetics*, volume 3, pages 2750 – 2755, 1998.
- [4] 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*, 13:216 – 228, 2005.
- [5] 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.
- [6] L. T. Kóczy, T. D. Gedeon, and J. A. Kóczy. Fuzzy tolerance relations and relational maps applied to information retrieval. *Fuzzy sets and systems*, 126:49 – 61, 2002.
- [7] 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.
- [8] D. Lucarella and R. Morara. First: Fuzzy information retrieval system. *J. Information Sci.*, 17:81 – 91, 1991.
- [9] S. Miyamoto. *Fuzzy sets in information retrieval and cluster analysis*. Kluwer, Dordrecht, 1990.
- [10] R. R. Yager. Decision making with fuzzy probability assessments. *IEEE Transactions on Fuzzy Systems*, 7, 1999.
- [11] L. A. Zadeh. Fuzzy sets. *Information and Control*, 8, 1965.
- [12] L. A. Zadeh. Fuzzy probabilities. *Information Processing and Management*, 3, 1984.