Cybernetics and Systems: An International Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucbs20

A KNOWLEDGE-BASED METHOD FOR FUZZY QUERY PROCESSING FOR DOCUMENT RETRIEVAL

SHYI-MING CHEN

Published online: 29 Oct 2010.

To cite this article: SHYI-MING CHEN (1997) A KNOWLEDGE-BASED METHOD FOR FUZZY QUERY PROCESSING FOR DOCUMENT RETRIEVAL, Cybernetics and Systems: An International Journal, 28:1, 99-119, DOI: 10.1080/019697297126272

To link to this article: http://dx.doi.org/10.1080/019697297126272

A KNOWLEDGE-BASED METHOD FOR FUZZY QUERY PROCESSING FOR DOCUMENT RETRIEVAL

SHYI-MING CHEN
WEN-HOAR HSIAO
YIH-JEN HORNG

Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan, Republic of China

This paper presents a knowledge-based method for processing fuzzy queries and weighted-fuzzy queries for document retrieval, where fuzzy concept matrices are used for knowledge representation and the elements in a fuzzy concept matrix are represented by trapezoidal fuzzy numbers parameterized by quadruples (a, b, c, d), where $0 \leqslant a \leqslant b \leqslant c \leqslant d \leqslant 1$. Intelligent retrieval capability and flexible user's queries are consequently provided for.

Salton and McGill (1983) pointed out that information retrieval is concerned with the representation, storage, organization, and accessing of information items. Most commercial information retrieval systems still adopt the Boolean logic model for information retrieval. However, the information retrieval systems based on the Boolean logic model are

The authors would like to thank the referees for providing very helpful comments and suggestions. Their insight and comments led to a better presentation of the ideas expressed in this paper.

This work was supported in part by the National Science Council, Republic of China, under grant NSC 83-0408-E-009-041.

Address correspondence to Shyi-Ming Chen, Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan, Republic of China.

Cybernetics and Systems: An International Journal, 28:99–119, 1997

0196-9722/97 $12.00 + .00

rather restricted in applications because these systems cannot process fuzzy queries. Several fuzzy information retrieval methods based on fuzzy set theory (Zadeh, 1965) have been proposed for improving the disadvantage of the Boolean logic model, such as those in Her and Ke (1983), Kraft and Buell (1983), Miyamoto (1990), Murai et al. (1989), Radechi (1979), Tahani (1976), and Zemankova (1989). Although these fuzzy information retrieval methods have the fuzzy query processing capability, the efficiency and effectiveness of these methods are not satisfactory. Lucarella and Morara (1991) presented a fuzzy information retrieval system FIRST based on concept networks. In Chen and Wang (1993, 1995) and Wang and Chen (1993), we have presented some methods for dealing with document retrieval using knowledge-based fuzzy information retrieval techniques. The methods we presented in Chen and Wang (1993, 1995a) and Wang and Chen (1993) allow the system's users to perform simple queries, weighted queries, interval queries, and weighted-interval queries. In this paper, we extend the works of Chen and Wang (1993, 1995a), Lucarella and Morara (1991), and Wang and Chen (1993) to present a knowledge-based method for dealing with fuzzy queries and weighted-fuzzy queries for document retrieval, where fuzzy concept matrices are used for knowledge representation, the elements in a fuzzy concept matrix represent fuzzy relevant values between concepts, and the fuzzy relevant values between concepts are represented by trapezoidal fuzzy numbers. The transitive closure of the fuzzy concept matrix is calculated by the fuzzy number arithmetic operations to evaluate the implicit fuzzy relevant values between concepts. The proposed method is more flexible than the ones presented in Chen and Wang (1993, 1995a), Lucarella and Morara (1991), and Wang and Chen (1993) because it allows the system's users to perform fuzzy queries and weighted-fuzzy queries. Intelligent retrieval capability and flexible user's queries are consequently provided for.

BASIC CONCEPTS OF FUZZY SET THEORY

In 1965, Zadeh proposed the theory of fuzzy sets. In the following, we briefly review some basic definitions of fuzzy sets from Chen (1992a, b, c; 1994), Chen et al. (1991), Kandel (1986), Kaufman and Gupta (1985, 1988), and Zadeh (1965). Let U be the universe of discourse, $U = \{u_1, u_2, \ldots, u_n\}$. A fuzzy set A in U is a set of ordered pairs $\{(u_1, f_A(u_1)), (u_2, f_A(u_2)), \ldots, (u_n, f_A(u_n))\}$, where f_A is the membership function of A, f_A: $U \rightarrow [0, 1]$, and $f_A(u_i)$ indicates the grade of membership of u_i in A. A fuzzy set A is convex if and only if for all u_1, u_2 in U,

$$f_A(\lambda u_1 + (1 - \lambda)u_2) \geqslant \text{Min}(f_A(u_1), f_A(u_2)) \tag{1}$$

where $\lambda \in [0, 1]$. A fuzzy set A of the universe of discourse U is called a normal fuzzy set if $\exists u_i \in U$, $f_A(u_i) = 1$. A fuzzy number is a fuzzy subset in the universe of discourse of U that is both convex and normal.

A fuzzy number M of the universe of discourse U may also be characterized by a trapezoidal distribution parametrized by a quadruple (a, b, c, d) shown in Figure 1.

Let A and B be two trapezoidal fuzzy numbers, where $A = (a_1, b_1, c_1, d_1)$ and $B = (a_2, b_2, c_2, d_2)$. The trapezoidal fuzzy numbers A and B are called equal (i.e., $A = B$) if and only if $a_1 = a_2$, $b_1 = b_2$, $c_1 = c_2$, and $d_1 = d_2$.

Let A and B be two trapezoidal fuzzy numbers, where $A = (a_1, b_1, c_1, d_1)$ and $B = (a_2, b_2, c_2, d_2)$. Based on Chen (1992a) and

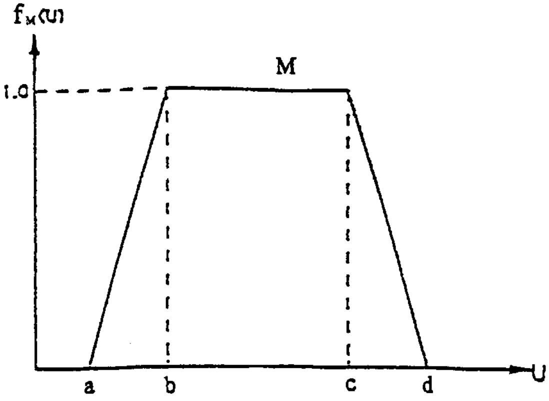

Figure 1. A trapezoidal fuzzy number.

Kaufman and Gupta (1985, 1988), the addition, subtraction, multiplication, division, AND, OR, and ratio operations of the trapezoidal fuzzy numbers A and B can be defined shown as follows.
Fuzzy number addition $\oplus$:

$$A \oplus B = (a_1, b_1, c_1, d_1) \oplus (a_2, b_2, c_2, d_2)$$

$$= (a_1 + a_2, b_1 + b_2, c_1 + c_2, d_1 + d_2) \tag{2}$$

Fuzzy number subtraction $\ominus$:

$$A \ominus B = (a_1, b_1, c_1, d_1) \ominus (a_2, b_2, c_2, d_2)$$

$$= (a_1 - d_2, b_1 - c_2, c_1 - b_2, d_1 - a_2) \tag{3}$$

Fuzzy number multiplication $\otimes$:

$$A \otimes B = (a_1, b_1, c_1, d_1) \otimes (a_2, b_2, c_2, d_2)$$

$$\fallingdotseq (a_1 \times a_2, b_1 \times b_2, c_1 \times c_2, d_1 \times d_2) \tag{4}$$

Fuzzy number division $\oslash$:

$$A \oslash = (a_1, b_1, c_1, d_1) \oslash (a_2, b_2, c_2, d_2)$$

$$\fallingdotseq (a_1/ d_2, b_1/ c_2, c_1/ b_2, d_1/ a_2) \tag{5}$$

Fuzzy number AND $\textcircled{$\wedge$}$:

$$A \textcircled{$\wedge$} B = (a_1, b_1, c_1, d_1) \textcircled{$\wedge$} (a_2, b_2, c_2, d_2)$$

$$= (\mathrm{Min}(a_1, a_2), \mathrm{Min}(b_1, b_2), \mathrm{Min}(c_1, c_2), \mathrm{Min}(d_1, d_2)) \tag{6}$$

Fuzzy number OR $\textcircled{$\vee$}$:

$$A \textcircled{$\vee$} B = (a_1, b_1, c_1, d_1) \textcircled{$\vee$} (a_2, b_2, c_2, d_2)$$

$$= (\mathrm{Max}(a_1, a_2), \mathrm{Max}(b_1, b_2), \mathrm{Max}(c_1, c_2), \mathrm{Max}(d_1, d_2)) \tag{7}$$

Fuzzy number ratio $\textcircled{\%}$:

$$\begin{aligned} A \textcircled{\%} B &= (a_1, b_1, c_1, d_1) \textcircled{\%} (a_2, b_2, c_2, d_2) \\ &= (a_1/\ a_2, b_1/\ b_2, c_1/\ c_2, d_1/\ d_2) \end{aligned} \tag{8}$$

Let k be a real number between zero and one and A be a trapezoidal fuzzy number, $A = (a_1, b_1, c_1, d_1)$. Then, we can see that

$$\begin{aligned} k \otimes A &= (k, k, k, k) \otimes (a_1, b_1, c_1, d_1) \\ &= (k \times\ a_1, k \times\ b_1, k \times\ c_1, k \times\ d_1) \end{aligned} \tag{9}$$

In the following, we introduce a defuzzification technique for trapezoidal fuzzy numbers (Kaufmann & Gupta, 1988; Chen, 1994a). Let us consider the trapezoidal fuzzy number shown in Figure 2, where e is a defuzzification value of the trapezoidal fuzzy number. From Figure 2, we can see that

$$\begin{aligned} &(e - b)(1) + \tfrac{1}{2}(b - a)(1) = (c - e)(1) + \tfrac{1}{2}(d - c)(1) \\ &\quad \Rightarrow (e - b) + \tfrac{1}{2}(b - a) = (c - e) + \tfrac{1}{2}(d - c) \\ &\quad \Rightarrow (e - b) - (c - e) = \tfrac{1}{2}(d - c) - \tfrac{1}{2}(b - a) \\ &\quad \Rightarrow 2e = \frac{a + d - b - c}{2} + \frac{2b + 2c}{2} \\ &\quad \Rightarrow 2e = \frac{a + b + c + d}{2} \\ &\quad \Rightarrow e = \frac{a + b + c + d}{4} \end{aligned} \tag{10}$$

In the following, we present a similarity measure (Chen, 1995b) for measuring the degree of similarity between two trapezoidal fuzzy num-

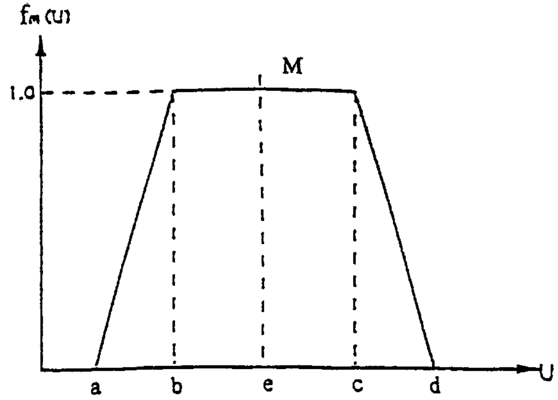

Figure 2. Defuzzification of a trapezoidal fuzzy number.

bers. Let A and B be two trapezoidal fuzzy numbers, where $A = (a_1, b_1, c_1, d_1)$ and $B = (a_2, b_2, c_2, d_2)$. The degree of similarity between the trapezoidal fuzzy numbers A and B can be measured by the similarity function S,

$$S(A, B) = 1 - \frac{|a_1 - a_2| + |b_1 - b_2| + |c_1 - c_2| + |d_1 - d_2|}{4} \tag{11}$$

where $S(A, B) \in [0, 1]$. The larger the value of $S(A, B)$, the greater the similarity between the trapezoidal fuzzy numbers A and B. It is obvious that if $A = (1, 1, 1, 1)$ and $B = (0, 0, 0, 0)$, then

$$S(A, B) = 1 - \frac{|1 - 0| + |1 - 0| + |1 - 0| + |1 - 0|}{4} = 0 \tag{12}$$

Furthermore, if $A = B = (a_1, b_1, c_1, d_1)$ (i.e., A and B are identical trapezoidal fuzzy numbers), then

$$S(A, B) = 1 - \frac{|a_1 - a_1| + |b_1 - b_1| + |c_1 - c_1| + |d_1 - d_1|}{4} = 1 \tag{13}$$

Let x and y be two real values between zero and one. It is obvious that x and y can be represented by trapezoidal fuzzy numbers, that is, $x = (x, x, x, x)$ and $y = (y, y, y, y)$. Based on formula (11), the degree of similarity between x and y can be evaluated shown as follows:

$$S(x, y) = 1 - \frac{|x - y| + |x - y| + |x - y| + |x - y|}{4} = 1 - |x - y| \quad (14)$$

This result is coincident with the one we show in Chen et al. (1989).

CONCEPT NETWORKS AND CONCEPTS MATRICES

Lucarella and Morara (1991) presented concept networks for fuzzy information retrieval. A concept network includes nodes and directed links, where each node represents a concept or a document and each directed link connects two concepts or is directed from one concept C_i to one document d_j and is labeled with a real value between zero and one. If $C_i \overset{\mu}{\rightarrow} C_j$, then it indicates that the degree of relevance from concept C_i to concept C_j is μ, where $\mu \in [0, 1]$. If $C_i \overset{\mu}{\rightarrow} d_j$, then it indicates that the degree of relevance of document d_j with respect to concept C_i is μ, where $\mu \in [0, 1]$. Figure 3 shows a concept network adapted from Lucarella and Morara (1991), where $C_1, C_2, \ldots,$ and C_7 are concepts; d_1, d_2, d_3, and d_4 are documents.

In Chen and Wang (1993, 1995) and Wang and Chen (1993), we have extended the work of Lucarella and Morara (1991) to allow the

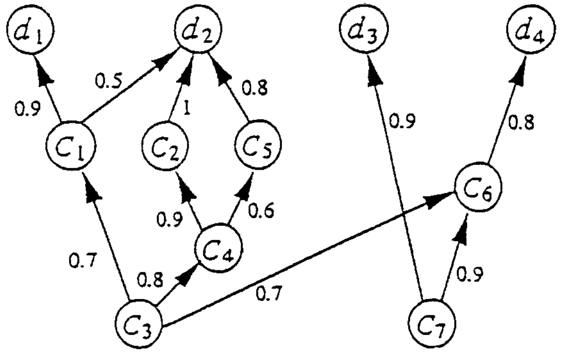

Figure 3. A concept network.

directed links in a concept network to be associated with real intervals in $[0, 1]$. However, if we can allow the directed links in a concept network to be associated with linguistic terms or trapezoidal fuzzy numbers parameterized by (a, b, c, d), where $0 \leqslant a \leqslant b \leqslant c \leqslant d \leqslant 1$, then there is room for more flexibility. Thus, in this paper, we further extend the works of Chen and Wang (1993, 1995a), Lucarella and Morara (1991), and Wang and Chen (1993) to allow the directed links in a concept network to be associated with linguistic terms or trapezoidal fuzzy numbers. The set of linguistic terms we used in this paper and their corresponding trapezoidal fuzzy numbers are shown in Table 1.

In Chen and Wang (1993, 1995a) and Wang and Chen (1993), we have presented the definitions of concept matrices for modeling concept networks. The definitions of concept matrices and the transitive closure of the concept matrices are reviewed as follows.

Definition 1: Let C be a set of concepts, $C = \{C_1, C_2, \ldots, C_n\}$. A concept matrix M is a fuzzy matrix (Kandel, 1986); $M(C_i, C_j)$ represents the relevant value from concept C_i to concept C_j, where $M(C_i, C_j) \in [0, 1]$.

A concept matrix M has the following properties:

1. Reflexivity,

$$M(C_i, C_i) = 1, \qquad \forall\ C_i \in C$$

Table 1. Linguistic terms and their corresponding trapezoidal fuzzy numbers

Linguistic terms	Trapezoidal fuzzy numbers
Nonrelevant	(0, 0, 0, 0)
Very low	(0, 0, 0.02, 0.07)
Low	(0.04, 0.1, 0.18, 0.23)
Medium low	(0.17, 0.22, 0.36, 0.42)
Medium	(0.32, 0.42, 0.58, 0.65)
Medium high	(0.58, 0.63, 0.80, 0.86)
High	(0.72, 0.78, 0.92, 0.97)
Very high	(0.975, 0.98, 1, 1)
Fully relevant	(1, 1, 1, 1)

2. M may not be symmetric,

$$M(C_i, C_j) \neq M(C_j, C_i)$$

3. Transitivity,

$$M(C_i, C_k) \geqslant \underset{C_j \in C}{\text{Max}} \, \text{Min}\big(M(C_i, C_j), M(C_j, C_k)\big)$$

Definition 2: Let M be a concept matrix,

$$M = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix}$$

where n is the number of concepts in a concept network, $f_{ij} \in [0, 1]$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant n$, and let

$$M^2 = M \quad M$$

$$= \begin{bmatrix} \bigvee\limits_{i=1,\ldots,n} (f_{1i} \wedge f_{i1}) & \bigvee\limits_{i=1,\ldots,n} (f_{1i} \wedge f_{i2}) & \cdots & \bigvee\limits_{i=1,\ldots,n} (f_{1i} \wedge f_{in}) \\ \bigvee\limits_{i=1,\ldots,n} (f_{2i} \wedge f_{i1}) & \bigvee\limits_{i=1,\ldots,n} (f_{2i} \wedge f_{i2}) & \cdots & \bigvee\limits_{i=1,\ldots,n} (f_{2i} \wedge f_{in}) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \bigvee\limits_{i=1,\ldots,n} (f_{ni} \wedge f_{i1}) & \bigvee\limits_{i=1,\ldots,n} (f_{ni} \wedge f_{i2}) & \cdots & \bigvee\limits_{i=1,\ldots,n} (f_{ni} \wedge f_{in}) \end{bmatrix} \tag{15}$$

where $\vee$ and $\wedge$ are the maximum and minimum operators, respectively. Then, there exists an integer $p \leqslant n - 1$, such that $M^P = M^{P+1} = M^{P+2} = \cdots$ (please see Kandel, 1986, p. 117). Let $Q = M^P$. Q is called the transitive closure of the concept matrix M.

In this paper, we allow the directed links in a concept network to be associated with linguistic terms shown in Table 1 or trapezoidal fuzzy numbers parameterized by (a, b, c, d), where $0 \leqslant a \leqslant b \leqslant c \leqslant d \leqslant 1$, and we use fuzzy concept matrices to modeling the concept networks. The definition of fuzzy concept matrices is presented as follows.

Definition 3: Let C be a set of concepts, $C = \{C_1, C_2, \ldots, C_n\}$, and F be a fuzzy concept matrix. $F(C_i, C_j) = (a_{ij}, b_{ij}, c_{ij}, d_{ij})$ indicates that the fuzzy relevant value from concept C_i to concept C_j is represented by trapezoidal fuzzy number $(a_{ij}, b_{ij}, c_{ij}, d_{ij})$, where $0 \leqslant a_{ij} \leqslant b_{ij} \leqslant c_{ij} \leqslant d_{ij} \leqslant 1$.

Definition 4: Let F be a fuzzy concept matrix,

$$F = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}$$

where n is the number of concepts, A_{ij} are trapezoidal fuzzy numbers, $A_{ij} = (a_{ij}, b_{ij}, c_{ij}, d_{ij})$, $0 \leqslant a_{ij} \leqslant b_{ij} \leqslant c_{ij} \leqslant d_{ij} \leqslant 1$, $1 \leqslant i \leqslant n$, and $1 \leqslant j \leqslant n$, and let

$$F^2 = F \odot F$$

$$= \begin{bmatrix} \underset{i=1,\ldots,n}{\ovee} \left(A_{1i} \owedge A_{i1} \right) & \underset{i=1,\ldots,n}{\ovee} \left(A_{1i} \owedge A_{i2} \right) & \cdots & \underset{i=1,\ldots,n}{\ovee} \left(A_{1i} \owedge A_{in} \right) \\ \underset{i=1,\ldots,n}{\ovee} \left(A_{2i} \owedge A_{i1} \right) & \underset{i=1,\ldots,n}{\ovee} \left(A_{2i} \owedge A_{i2} \right) & \cdots & \underset{i=1,\ldots,n}{\ovee} \left(A_{2i} \owedge A_{in} \right) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \underset{i=1,\ldots,n}{\ovee} \left(A_{ni} \owedge A_{i1} \right) & \underset{i=1,\ldots,n}{\ovee} \left(A_{ni} \owedge A_{i2} \right) & \cdots & \underset{i=1,\ldots,n}{\ovee} \left(A_{ni} \owedge A_{in} \right) \end{bmatrix} \tag{16}$$

where $\textcircled{\wedge}$ and $\textcircled{\vee}$ represent the AND and OR operators of the trapezoidal fuzzy numbers, respectively. Then there exists an integer p, $p \leqslant n - 1$, such that $F^P = F^{P+1} = F^{P+2} = \cdots$. Let $Q = F^P$. Q is called the transitive closure of the fuzzy concept matrix F.

FUZZY QUERY PROCESSING TECHNIQUES FOR DOCUMENT RETRIEVAL

Let D be a set of documents, $D = \{d_1, d_2, \ldots, d_m\}$, and C be a set of concepts, $C = \{C_1, C_2, \ldots, C_n\}$. A document in a document retrieval system is generally described by a set of concepts with each concept representing a topic. The relations between documents and concepts can be represented by a document descriptor matrix D shown as follows:

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ d_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_n \end{array} \\ \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ B_{m1} & B_{m2} & \cdots & B_{mn} \end{bmatrix} \end{array}$$

where m is the number of documents, n is the number of concepts, B_{ij} is a trapezoidal fuzzy number parameterized by a quadruple representing the degree of relevance of document d_i with respect to concept C_j, $1 \leqslant i \leqslant m$, and $1 \leqslant j \leqslant n$. In a document descriptor matrix D, the degree of relevance of each document with respect to a specific concept is determined by experts. However, an expert may possibly neglect the degree of relevance of a certain document with respect to some specific concepts. Because concepts may be not independent of each other, the transitive closure Q of the fuzzy concept matrix F can be used to evaluate the implicit relevant values of each document with respect to specific concepts to improve this. Let D be a document descriptor matrix and Q be the transitive closure of the fuzzy concept matrix F, where

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ d_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_n \end{array} \\ \left[\begin{array}{cccc} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ B_{m1} & B_{m2} & \cdots & B_{mn} \end{array} \right] \end{array}$$

$$Q = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ S_{n1} & S_{n2} & \cdots & S_{nn} \end{bmatrix}$$

where B_{ij} and S_{ij} are trapezoidal fuzzy numbers parametrized by quadruples, $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant n$. Let

$$D^* = D \odot Q$$

$$= \begin{bmatrix} \underset{i=1,\ldots,n}{\textcircled{$\vee$}} \left(B_{1i} \textcircled{$\wedge$} S_{i1} \right) & \underset{i=1,\ldots,n}{\textcircled{$\vee$}} \left(B_{1i} \textcircled{$\wedge$} S_{i2} \right) & \cdots & \underset{i=1,\ldots,n}{\textcircled{$\vee$}} \left(B_{1i} \textcircled{$\wedge$} S_{in} \right) \\ \underset{i=1,\ldots,n}{\textcircled{$\vee$}} \left(B_{2i} \textcircled{$\wedge$} S_{i1} \right) & \underset{i=1,\ldots,n}{\textcircled{$\vee$}} \left(B_{2i} \textcircled{$\wedge$} S_{i2} \right) & \cdots & \underset{i=1,\ldots,n}{\textcircled{$\vee$}} \left(B_{2i} \textcircled{$\wedge$} S_{in} \right) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \underset{i=1,\ldots,n}{\textcircled{$\vee$}} \left(B_{ni} \textcircled{$\wedge$} S_{i1} \right) & \underset{i=1,\ldots,n}{\textcircled{$\vee$}} \left(B_{ni} \textcircled{$\wedge$} S_{i2} \right) & \cdots & \underset{i=1,\ldots,n}{\textcircled{$\vee$}} \left(B_{ni} \textcircled{$\wedge$} S_{in} \right) \end{bmatrix}, \tag{17}$$

where $\textcircled{$\wedge$}$ and $\textcircled{$\vee$}$ are the AND and OR operators of the trapezoidal fuzzy numbers, respectively. The document descriptor matrix D^* indicates the degrees of relevance of each document with respect to specific concepts and is used as a basis for similarity measures between user's queries and documents as described later.

In a fuzzy information retrieval system, the user's query can be described by a query descriptor Q represented by a query descriptor matrix $\bar{q}$, that is,

$$Q = \{(C_1, V_1), (C_2, V_2), \ldots, (C_n, V_n)\}$$

$$\bar{q} = [V_1, V_2, \ldots, V_n]$$

where V_i is a trapezoidal fuzzy number parameterized by a quadruple, $1 \leqslant i \leqslant n$, representing the degree of strength that the desired documents contain concept C_i. If the user considers that certain concepts may be neglected, then the user does not have to assign the degrees of strength with respect to such concepts in the query descriptor vector $\bar{q}$. The symbol $-$ is used for labeling a neglected concept. Thus, if $V_i = -$, it indicates that concept C_i is a neglected concept. In this case, the concept C_i would not be considered in the document retrieval process.

Let d_i denote the ith row of the document descriptor matrix D^*, $\bar{d}_i = [P_{i1}, P_{i2}, \ldots, P_{in}]$, and let $\bar{q}$ be the query descriptor matrix, $\bar{q} = [V_1, V_2, \ldots, V_n]$, where P_{ij} and V_j are trapezoidal fuzzy numbers,

$$P_{ij} = (a_{ij}, b_{ij}, c_{ij}, d_{ij})$$

$$V_j = (w_j, x_j, y_j, z_j)$$

$1 \leqslant j \leqslant n$, $1 \leqslant i \leqslant m$, n is the number of concepts, and m is the number of documents. Let $q(j)$ denote the jth component of the query descriptor matrix $\bar{q}$. If $q(j) = -$, it indicates that the concept C_j is a neglected concept with respect to the fuzzy query. Based on formula (11), the degree of similarity between $\bar{d}_i$ and $\bar{q}$ can be evaluated as follows:

$$RS(\bar{d}_i) = \frac{\sum\limits_{q(j) \neq \text{"}-\text{" and } j=1,\ldots,n} S(P_{ij}, V_j)}{k} \tag{18}$$

where

$$S(P_{ij}, V_j) = 1 - \frac{|a_{ij} - w_j| + |b_{ij} - x_j| + |c_{ij} - y_j| + |d_{ij} - z_j|}{4} \tag{19}$$

$RS(d_i) \in [0,1]$, $1 \leqslant i \leqslant m$, and k is the number of concepts not neglected in the query. The retrieval status value $RS(d_i)$ indicates the degree of similarity between the query and the document d_i, where $1 \leqslant i \leqslant m$. The larger the value of $RS(d_i)$, the higher the similarity between the query and the document d_i.

Consider the following OR-connected query:

$$\overline{q_1} \text{ OR } \overline{q_2}$$

where $\overline{q_1}$ and $\overline{q_2}$ are query descriptor matrices. In this case, the degree of similarity between the query and the documents can be evaluated as follows:

$$RS^*(d_i) = \text{Max}(RS_1(d_i), RS_2(d_i)) \tag{20}$$

where $RS_1(d_i)$ represents the degree of similarity between the query descriptor matrix $\overline{q_1}$ and the ith row of the document descriptor matrix D^*, $RS_2(d_i)$ represents the degree of similarity between the query descriptor matrix $\overline{q_2}$ and the ith row of the document descriptor matrix D^*, the retrieval status value $RS^*(d_i)$ represents the degree of similarity of the query with respect to the document d_i, and $1 \leqslant i \leqslant m$. The information retrieval system would display every document having a retrieval status value greater than a threshold value λ, where $\lambda \in [0,1]$, in a sequential order from the document with the highest degree of retrieval status value to that with the lowest one.

Weighted-fuzzy queries can also be processed by our method. In weighted-fuzzy queries, a query expression can be represented by a query descriptor matrix $\overline{q}$ shown as follows:

$$\overline{q} = [(V_1, W_1), (V_2, W_2), \ldots, (V_n, W_n)]$$

where V_j and W_j are trapezoidal fuzzy numbers parameterized by quadruples, W_j represents the weights of concept C_j, and $1 \leqslant j \leqslant n$. Let the ith row of the document descriptor matrix D^* be $[P_{i1}, P_{i2}, \ldots, P_{in}]$, where $P_{i1}, P_{i2}, \ldots,$ and P_{in} are trapezoidal fuzzy numbers parameterized by quadruples. Then the degree of similarity between the weighted fuzzy query and the document d_i can be calculated as follows:

$$RS_w(d_i) = \left(\sum_{q(j) \neq \text{“}-\text{”} \text{ and } j=1,\ldots,n} S(P_{ij}, V_j) \otimes W_j \right) \textcircled{\%} \left(\sum_{Q(j) \neq \text{“}-\text{”} \text{ and } j=1,\ldots,n} W_j \right) \tag{21}$$

where $\otimes$ and $\textcircled{\%}$ are the multiplication operator and the ratio operator of the trapezoidal fuzzy numbers, respectively, and Σ denotes the summation of the trapezoidal fuzzy numbers. The retrieval status value $RS_w(d_i)$ is a trapezoidal fuzzy number indicating the degree of similarity between the weighted-fuzzy query and the document d_i, where $1 \leqslant i \leqslant m$. Assume that

$$\begin{aligned} RS_w(d_1) &= (a_1, b_1, c_1, d_1) \\ RS_w(d_2) &= (a_2, b_2, c_2, d_2) \\ &\vdots \\ RS_w(d_m) &= (a_m, b_m, c_m, d_n) \end{aligned} \tag{22}$$

Based on formula (10), the retrieval status value $RS_w(d_i)$ can be defuzzified into a crisp real value, where $1 \leqslant i \leqslant m$. In this case, the defuzzified value of $RS_w(d_i)$ is equal to $(a_i + b_i + c_i + d_i)/4$, where $1 \leqslant i \leqslant m$. Let the defuzzified value of $RS_w(d_i)$ be equal to k_i, where $k_{2'} \in [0, 1]$ and $1 \leqslant i \leqslant m$, and let λ be a threshold value, $\lambda \in [0, 1]$. The information retrieval system would display every document having a defuzzified retrieval status value k_i greater than the threshold value λ in a sequential order from the document with the highest degree of defuzzified retrieval status value to that with the lowest one.

In the following, we use an example to illustrate the weighted-fuzzy query processing process for document retrieval.

Example: Assume that the retrieval threshold value λ is 0.65, and assume that there are four concepts C_1, C_2, C_3, C_4 and five documents

d_1, d_2, d_3, d_4, and d_5. Furthermore, assume that the document descriptor matrix D and the fuzzy concept matrix F have the following forms:

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & C_3 & C_4 \end{array} \\ \left[\begin{array}{cccc} (0.2, 0.3, 0.4, 0.5) & (0.5, 0.6, 0.7, 0.8) & (1, 1, 1, 1) & (0, 0, 0, 0) \\ (1, 1, 1, 1) & (0.6, 0.7, 0.8, 0.9) & (0.3, 0.4, 0.5, 0.6) & (0.5, 0.6, 0.7, 0.8) \\ (0.5, 0.6, 0.7, 0.8) & (1, 1, 1, 1) & (0.1, 0.2, 0.3, 0.4) & (0.7, 0.7, 0.7, 0.7) \\ (0, 0, 0, 0) & (0.3, 0.4, 0.5, 0.6) & (0.5, 0.6, 0.7, 0.8) & (1, 1, 1, 1) \\ (0.3, 0.4, 0.5, 0.6) & (0.4, 0.5, 0.6, 0.7) & (0, 0, 0, 0) & (0.5, 0.6, 0.7, 0.8) \end{array}\right] \end{array}$$

$$F = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & C_3 & C_4 \end{array} \\ \left[\begin{array}{cccc} (1, 1, 1, 1) & (0.975, 0.98, 1, 1) & (0, 0, 0, 0) & (0, 0, 0, 0) \\ (0, 0, 0, 0) & (1, 1, 1, 1) & (0.58, 0.63, 0.80, 0.86) & (0.975, 0.98, 1, 1) \\ (0, 0, 0, 0) & (0, 0, 0, 0) & (1, 1, 1, 1) & (0, 0, 0, 0) \\ (0, 0, 0, 0) & (0, 0, 0, 0) & (0, 0, 0, 0) & (1, 1, 1, 1) \end{array}\right] \end{array}$$

In this case, the transitive closure Q of the fuzzy concept matrix F can be calculated as follows:

$$Q = \begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & C_3 & C_4 \end{array} \\ \left[\begin{array}{cccc} (1, 1, 1, 1) & (0.975, 0.98, 1, 1) & (0.58, 0.63, 0.80, 0.86) & (0.975, 0.98, 1, 1) \\ (0, 0, 0, 0) & (1, 1, 1, 1) & (0.58, 0.63, 0.80, 0.86) & (0.975, 0.98, 1, 1) \\ (0, 0, 0, 0) & (0, 0, 0, 0) & (1, 1, 1, 1) & (0, 0, 0, 0) \\ (0, 0, 0, 0) & (0, 0, 0, 0) & (0, 0, 0, 0) & (1, 1, 1, 1) \end{array}\right] \end{array}$$

The document descriptor matrix D^* can be obtained based on the document descriptor matrix D and the transitive closure Q of the fuzzy concept matrix F as follows:

$$D^* = D \odot Q$$

$$= \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & C_3 & C_4 \end{array} \\ \left[\begin{array}{cccc} (0.2, 0.3, 0.4, 0.5) & (0.5, 0.6, 0.7, 0.8) & (1, 1, 1, 1) & (0.5, 0.6, 0.7, 0.8) \\ (1, 1, 1, 1) & (0.975, 0.8, 1, 1) & (0.58, 0.63, 0.80, 0.86) & (0.975, 0.98, 1, 1) \\ (0.5, 0.6, 0.7, 0.8) & (1, 1, 1, 1) & (0.58, 0.63, 0.80, 0.86) & (0.975, 0.98, 1, 1) \\ (0, 0, 0, 0) & (0.3, 0.4, 0.5, 0.6) & (0.5, 0.6, 0.7, 0.8) & (1, 1, 1, 1) \\ (0.3, 0.4, 0.5, 0.6) & (0.4, 0.5, 0.6, 0.7) & (0.4, 0.5, 0.6, 0.7) & (0.5, 0.6, 0.7, 0.8) \end{array}\right] \end{array}$$

If the user's weighted-fuzzy query is represented by the query descriptor matrix $\bar{q}$ shown as follows:

$$\bar{q} = [((0.6, 0.7, 0.8, 0.9), (0.6, 0.7, 0.8, 0.9)), -, -,$$
$$((0.9, 0.95, 0.95, 1), (0.5, 0.6, 0.7, 0.8))]$$

Then based on formula (21), we can get the following results:

$$\begin{aligned} RS_w(d_1) = {} & \Bigg[\left(1 - \frac{|0.2 - 0.6| + |0.3 - 0.7| + |0.4 - 0.8| + |0.5 - 0.9|}{4}\right) \\ & \otimes (0.6, 0.7, 0.8, 0.9) \\ & \oplus \left(1 - \frac{|0.5 - 0.9| + |0.6 - 0.95| + |0.7 - 0.95| + |0.8 - 1|}{4}\right) \\ & \otimes (0.5, 0.6, 0.7, 0.8)\Bigg] \\ & \oslash [(0.6, 0.7, 0.8, 0.9) \oplus (0.5, 0.6, 0.7, 0.8)] \\ = {} & (0.64545, 0.64615, 0.64667, 0.64706) \end{aligned}$$

$$\begin{aligned} RS_w(d_2) = {} & \Bigg[\left(1 - \frac{|1 - 0.6| + |1 - 0.7| + |1 - 0.8| + |1 - 0.9|}{4}\right) \\ & \otimes (0.6, 0.7, 0.8, 0.9) \\ & \oplus \left(1 - \frac{|0.975 - 0.9| + |0.98 - 0.95| + |1 - 0.95| + |1 - 1|}{4}\right) \\ & \otimes (0.5, 0.6, 0.7, 0.8)\Bigg] \\ & \oslash [(0.6, 0.7, 0.8, 0.9) \oplus (0.5, 0.6, 0.7, 0.8)] \\ = {} & (0.84603, 0.8475, 0.84859, 0.84941) \end{aligned}$$

$$
\begin{aligned}
RS_w(d_3) = {} & \Bigg[\left(1 - \frac{|0.5-0.6|+|0.6-0.7|+|0.7-0.8|+|0.8-0.9|}{4}\right) \\
& \otimes (0.6, 0.7, 0.8, 0.9) \\
& \oplus \left(1 - \frac{|0.975-0.9|+|0.98-0.95|+|1-0.95|+|1-1|}{4}\right) \\
& \otimes (0.5, 0.6, 0.7, 0.8)\Bigg] \\
& \oslash [(0.6, 0.7, 0.8, 0.9) \oplus (0.5, 0.6, 0.7, 0.8)] \\
= {} & (0.92785, 0.92827, 0.92859, 0.92882)
\end{aligned}
$$

$$
\begin{aligned}
RS_w(d_4) = {} & \Bigg[\left(1 - \frac{|0-0.6|+|0-0.7|+|0-0.8|+|0-0.9|}{4}\right) \\
& \qquad \otimes (0.6, 0.7, 0.8, 0.9) \\
& \oplus \Bigg[\left(1 - \frac{|1-0.9|+|1-0.95|+|1-0.95|+|1-1|}{4}\right) \\
& \qquad \otimes (0.5, 0.6, 0.7, 0.8)\Bigg] \\
& \oslash [(0.6, 0.7, 0.8, 0.9) \oplus (0.5, 0.6, 0.7, 0.8)] \\
= {} & (0.56818, 0.57308, 0.57667, 0.57941)
\end{aligned}
$$

$$
\begin{aligned}
RS_w(d_5) = {} & \Bigg[\left(1 - \frac{|0.3-0.6|+|0.4-0.7|+|0.5-0.8|+|0.6-0.9|}{4}\right) \\
& \otimes (0.6, 0.7, 0.8, 0.9) \\
& \oplus \left(1 - \frac{|0.5-0.9|+|0.6-0.95|+|0.7-0.95|+|0.8-1|}{4}\right)
\end{aligned}
$$

$$\otimes (0.5, 0.6, 0.7, 0.8) \Big] \, ⁒ \, [(0.6, 0.7, 0.8, 0.9) \oplus (0.5, 0.6, 0.7, 0.8)]$$

$$= (0.7, 0.7, 0.7, 0.7)$$

Based on formula (10), we can get the following results: The defuzzified value of $RS_w(d_1)$ is equal to

$$\frac{0.64545 + 0.64615 + 0.64667 + 0.64706}{4} \doteq 0.64333$$

The defuzzified value of $RS_w(d_2)$ is equal to

$$\frac{0.84603 + 0.8475 + 0.84859 + 0.84941}{4} \doteq 0.84788$$

The defuzzified value of $RS_w(d_3)$ is equal to

$$\frac{0.92785 + 0.92827 + 0.92859 + 0.92882}{4} \doteq 0.92838$$

The defuzzified value of $RS_w(d_4)$ is equal to

$$\frac{0.56818 + 0.57308 + 0.57667 + 0.57941}{4} \doteq 0.57434.$$

The defuzzified value of $RS_w(d_5)$ is equal to

$$\frac{0.7 + 0.7 + 0.7 + 0.7}{4} = 0.7.$$

Because the retrieval threshold value λ is 0.65, the documents d_1 and d_4 will not be retrieved because the retrieval status values of the documents d_1 and d_4 are less than the threshold value. From these results, we also can see that the document d_3 is the most suitable to the user's weighted-fuzzy query because it has the largest retrieval status value.

CONCLUSIONS

In this paper, we have presented a knowledge-based method for processing fuzzy queries and weighted-fuzzy queries for document retrieval, where the fuzzy concept matrices are used for knowledge representation and the elements in fuzzy concept matrices are represented by trapezoidal fuzzy numbers parameterized by quadruples (a, b, c, d), where $0 \leqslant a \leqslant b \leqslant c \leqslant d \leqslant 1$. We also use an example to illustrate the weighted-fuzzy query processing process for document retrieval. From the illustrated example, we can see that the proposed method can be executed very efficiently. The proposed method is a significant improvement over the method based on Boolean algebra because it has the fuzzy query processing capability. Furthermore, the proposed method is more flexible than the ones presented in Chen and Wang (1993, 1995a), Lucarella and Morara (1991), and Wang and Chen (1993) because it allows the system's users to perform fuzzy queries and weighted-fuzzy queries. Intelligent retrieval capability and flexible user's queries are consequently provided for.

REFERENCES

Chen, S. M. 1992a. A fuzzy reasoning technique based on the α-cuts operations of fuzzy numbers. *Proceedings of the Second International Conference on Automation Technology*, Taipei, Taiwan, Republic of China, pp. 147−154.

Chen, S. M. 1992b. An improved algorithm for inexact reasoning based on extended fuzzy production rules. *Cybern. Syst.* 23(5):463−481.

Chen, S. M. 1992c. A new approach to inexact reasoning for rule-based systems. *Cybern. Syst.* 23(6):561−582.

Chen, S. M. 1993. An inexact reasoning technique using linguistic rule matrix transformations. *Proceedings of the IEEE 23rd International Symposium on Multiple-Valued Logic*, Sacramento, CA, pp. 190−195.

Chen, S. M. 1994a. Using fuzzy reasoning techniques for fault diagnosis of the J-85 jet engines. *Proceedings of the Third National Conference of Science and Technology of National Defense*, Taipei, Taiwan, Republic of China, pp. 29−34.

Chen, S. M. 1994b. A new method for handling multicriteria fuzzy decision-making problems. *Cybern. Syst.* 25(3):409−420.

Chen, S. M. and J. Y. Wang. 1993. A new approach for fuzzy information retrieval. *Proceedings of 1993 National Computer Symposium*, Chiayi, Taiwan, Republic of China, vol. 2, pp. 767−774.

Chen, S. M. and J. Y. Wang. 1995a. Document retrieval using knowledge-based fuzzy information retrieval techniques. *IEEE Trans. Syst. Man Cybern.* 25(5):793−803.

Chen, S. M. and S. Y. Lin. 1995b. A new method for fuzzy risk analysis. *Proceedings of 1995 Artificial Intelligence Workshop*, Taipei, Taiwan, Republic of China, pp. 245−250.

Chen, S. M., J. S. Ke, and J. F. Chang. 1989. Techniques for handling multicriteria fuzzy decision-making problems. *Proceedings of the 4th International Symposium on Computer and Information Sciences*, Cesme, Turkey, vol. 2, pp. 919−925.

Chen, S. M., J. S. Ke, and J. F. Chang. 1991. An inexact reasoning algorithm for dealing with inexact knowledge. *Int. J. Software Eng. Knowledge Eng.* 1(3):227−241.

Her, G. T. and J. S. Ke. 1983. A fuzzy information retrieval system model. *Proceedings of 1983 National Computer Symposium*, Taiwan, Republic of China, pp. 147−155.

Kamel, M., B. Hadfield, and M. Ismail. 1990. Fuzzy query processing using clustering techniques. *Inform. Process. Manage.* 26(2):279−293.

Kandel, A. 1986. *Fuzzy mathematical techniques with applications*. Reading, MA: Addison-Wesley.

Kaufman, A. and M. M. Gupta. 1985. *Introduction to fuzzy arithmetic: Theory and applications*. New York: Van Nostrand Reinhold.

Kaufman, A. and M. M. Gupta. 1988. *Fuzzy mathematical models in engineering and management science*. Amsterdam: North-Holland.

Kraft, D. H. and D. A. Buell. 1983. Fuzzy sets and generalized Boolean retrieval systems. *Int. J. Man Machine Stud*. 19(1):45−56.

Lucarella, D. and R. Morara. 1991. FIRST: Fuzzy information retrieval system. *J. Inform. Sci*. 17:81−91.

Miyamoto, S. 1990. Information retrieval based on fuzzy associations. *Fuzzy Sets Syst*. 38:191−205.

Murai, T., M. Miyakoshi, and M. Shimbo. 1989. A fuzzy document retrieval method based on two-valued indexing. *Fuzzy Sets Syst*. 30:103−120.

Radechi, T. 1979. Fuzzy set theoretical approach to document retrieval. *Inform. Process. Manage.* 15:247−259.

Salton, G. and M. J. McGill. 1983. *Introduction to modern information retrieval.* New York: McGraw-Hill.

Tahani, V. 1976. A fuzzy model of document retrieval system. *Inform. Process. Manage*. 12:177−187.

Wang, J. Y. and S. M. Chen, 1993. A knowledge-based method for fuzzy information retrieval. *Proceedings of the First Asian Fuzzy Systems Symposium*, Singapore, November.

Zadeh, L. A. 1965. Fuzzy sets. *Inform. Control* 8:338−353.

Zemankova, M. 1989. FIIS: A fuzzy intelligent information system. *Data Eng.* 12(2):11−20.