ISE-TR-88-69

TWO APPROACHES FOR INFORMATION RETRIEVAL THROUGH FUZZY ASSOCIATIONS

by

Sadaaki MIYAMOTO

June 3, 1988

INSTITUTE
OF
INFORMATION SCIENCES AND ELECTRONICS

UNIVERSITY OF TSUKUBA

Two approaches for information retrieval through fuzzy associations

S. Miyamoto

Institute of Information Sciences and Electronics
University of Tsukuba, Tsukuba, Ibaraki 305, Japan

ABSTRACT

The aim of the present paper is to show two approaches for formulation of information retieval through fuzzy associations. A fuzzy association is introduced as a fuzzy relation defined on a set of indexes to a database. The fuzzy association is a generalization of a concept of fuzzy thesauri and methods of generating fuzzy thesauri are applicable to generation of fuzzy associations. One approach for the fuzzy retrieval is extension of fuzzy indexes to a database using fuzzy associations. An algorithm for the fuzzy information retrieval based on this approach is developed. The other approach represents the retrieval process as a block diagram. Maximum and minimum operations are used intead of the ordinary sum and product operations on the diagram. Studies on advanced indexing such as clustering of articles are represented as feedbacks on the diagram and properties on fuzzy information retrieval such as level fuzzy sets and set operations on responses of the retrieval system are discussed using the diagram representation.

1. Introduction

Although information retrieval through associations has been studied by many authors (e.g.. [1], [2], and references therein), fundamental problems such as a mathematical model for associations have not yet studied throughly. Associations in information retrieval mean that we retrieve not only documents that have keywords specified by a user but also other documents that are associated with the former documents or have indexes associated with the specified keywords. A typical example of the association is a thesaurus for information retrieval. Another well-known example is a retrieval through bibliographic citations. Moreover, citation analysis has been carried out by different researchers (e.g., [3]).

The author studied a mathematical model of thesauri and developed a method of generating automatically fuzzy pseudothesauri [4]. Fuzzy information retrieval through a fuzzy thesaurus has also been formulated and an algorithm for the fuzzy retrieval has been given [5]. In the present paper two approaches for fuzzy information retrieval through fuzzy associations are considered. One approach is extension of indexes based on a combination of an association and a fuzzy index. This approach is a generalization of our method of retrieval through a fuzzy thesaurus [5] to a broader framework. The other approach is a block diagram representation of information retrieval that includes feedbacks in association retrieval. Heaps [2] discussed already block diagrams and feedbacks in information retrieval. The difference between his method and the method herein is that we use here maximum and minimum operations instead of the ordinary sum and product operations in matrix calculations. An advantage of the method herein is that the concept of feedback in information retrieval naturally leads to clustering of documents. Clustering of ducuments has been frequently studied in the field of bibliographic information analysis. (See, e.g., [3].) Therefore studies in document clustering can be discussed in the present framework of fuzzy retrieval, which leads to a concept of fuzzy cluster retrieval. Another advantage of the method herein is that algorithms for a large number of documents are developed by the present approach.

Although these two approaches are closely related and indeed lead to the same algorithms and implementations, we compare these two mathematical models for information retrieval through fuzzy associations and discuss relative advantages as frameworks for future researches in this area.

2. Preliminaries

Let $D=\{d_1,d_2,\ldots,d_m\}$ be a set of documents and $X=\{x_1,x_2,\ldots,x_n\}$ be a set of indexes. X may be a set of keywords or may be other kinds of indexes, e.g., the citation indexes. A function T defined on D that maps each document to its corresponding indexes is assumed to be given. A typical example of T gives keywords attached to each document d. In a previous paper we assumed that Td ($d\in D$) is a crisp subset of X. Here we assume that Td may be a fuzzy subset. Therefore the function T is considered to be a fuzzy relation defined on X x D,

$T: X \times D \rightarrow [0,1]$. We consider another relation $U: D \times X \rightarrow [0,1]$ that is the inverse relation of T: $U(d,x) = T(x,d)$, for any $x \in X$ and $d \in D$.

We also use matrix notation for a relation, e.g., $\underline{T}=(t_{ij})$, $1 \leq i \leq n$, $1 \leq j \leq m$, where $t_{ij} = T(x_i, d_j)$. We obtain $\underline{U}=(u_{ij})$, $1 \leq i \leq m$, $1 \leq j \leq n$, $u_{ij} = U(d_i, x_j)$, therefore $\underline{U}$ is the transpose of $\underline{T}$: $\underline{U} = \underline{T}^T$.

We assume a fuzzy association as a fuzzy relation F defined on $X \times X$ ($F: X \times X \rightarrow [0,1]$). A typical example of a fuzzy association is a fuzzy thesaurus [6],[7]. In a previous paper we proposed a method of generating a fuzzy thesaurus or a fuzzy pseudothesaurus [4]. Here we assume that the relation F is obtained by the method of automatic generation of a fuzzy thesaurus, or the relation can be obtained by other procedures including manual determination of a fuzzy thesaurus or by an aggregation of different crisp thesauri into a fuzzy thesaurus. Another type of a fuzzy association is developed on citation indexes using the same method of generating associations [8],[9].

The matrix representation of F is denoted as $\underline{F} = (f_{ij})$: $f_{ij} = F(x_i, x_j)$, $1 \leq i,j \leq n$. The matrix may be symmetric or asymmetric. When the matrix is symmetric, the relation shows proximity of two indexes x_i and x_j. Although $F(x_i, x_i)$ is not defined in many procedures of generating fuzzy thesauri, we assume $F(x_i, x_i)=1$ for simplicity. Note that all the elements in these matrices $\underline{T}$, $\underline{U}$, $\underline{F}$ are nonnegative.

<u>Remark</u> Although we use matrix notation, we do not keep matrices in the retrieval system, since matrix manipulations in a large scale database require a huge number of computations and vast resources. For example, practical bibliographic databases include

several hundred thousands or several million of records. Therefore the reason why we use the matrix here is purely for simplicity of explanation. If a method of retrieval assumes matrix calculations in a retrieval system. it is very difficult to deal with a large scale database and the method is applicable only to a very small database for personal use, with a poor performance of the retrieval system.[]

One of the main principles in fuzzy sets is maximum and minimum operations in matrix calculations. From now. we assume that the symbols + and $\cdot$ mean maximum and minimum operations, respectively: $a+b = \max(a,b)$, $a \cdot b = \min(a,b)$, $a \geq 0$, $b \geq 0$. Exceptions are stated explicitly.

The following two propositions are well-known in fuzzy sets, therefore the proofs are omitted.

Prop. 1 For any square nonnegative matrix $\underline{A}=(a_{ij})$, $1 \leq i,j \leq p$,

$$\underline{A} + \underline{A}^2 + \underline{A}^3 + \ldots$$

converges. (The opeations + and $\cdot$ are performed by max and min. respectively, as is described above.) Let

$$\hat{\underline{A}} = \underline{A} + \underline{A}^2 + \underline{A}^3 + \ldots$$

Then, $\hat{\underline{A}} = \underline{A} + \underline{A}^2 + \underline{A}^3 + \ldots + \underline{A}^{p-1}$.

(The matrix $\hat{\underline{A}}$ is called a closure of $\underline{A}$.) []

Prop. 2 Assume that a square nonnegative matrix $\underline{A} = (a_{ij})$,

$1 \leq i, j \leq p$ satisfies:

$$a_{ij} = a_{ji} \quad \text{and} \quad a_{ii} = 1. \qquad (1)$$

Then the closure $\hat{\underline{A}} = (\hat{a}_{ij})$ satisfies

$$\hat{a}_{ii} = 1, \qquad \hat{a}_{ij} = \hat{a}_{ji}, \qquad (2)$$
$$\hat{a}_{ij} \geq \max_{k} \min(\hat{a}_{ik}, \hat{a}_{kj}) .$$

(We call the matrix $\underline{A}$ with the property (1) as a proximity matrix and $\hat{\underline{A}}$ satisfying (2) as the transitive closure of $\underline{A}$.) []

Def. 1

Let $x_1 = (x_{11}, \ldots, x_{1n})$, $x_2 = (x_{21}, \ldots, x_{2n})$ be two vectors with real components. Note that inequality $x_1 \geq x_2$ means that $x_{1i} \geq x_{2i}$, $i = 1, 2, \ldots, n$. Let $\underline{A}$ and b be an nxn nonnegative matrix and a nonnegative n-vector, respectively. Consider an equation

$$x = \underline{A} x + b \qquad (3)$$

A solution $\bar{x}$ ($\bar{x} = \underline{A}\bar{x} + b$) is called the minimal solution of the equation (3) if for any other solution x' ($x' = \underline{A}x' + b$), inequality $\bar{x} \leq x'$ holds. []

Remark From Def. 1 it is obvious that the minimal solution is unique. []

Prop. 3 The minimal solution of (3) is given by

$$\bar{x} = \hat{\underline{A}} b + b .$$

Moreover, if $\underline{A}$ satisfies $a_{ii} = 1$, then $\bar{x} = \hat{\underline{A}} b$.

(Proof) Note that

$$\underline{A} + \underline{A}\,\hat{\underline{A}} = \underline{A} + \underline{A}^2 + \underline{A}^3 + \ldots = \hat{\underline{A}}\ .$$

Therefore,

$$\underline{A}\bar{x} + b = \underline{A}(\hat{\underline{A}} + \underline{I})b + b = \hat{\underline{A}}b + b = \bar{x},$$

whence $\bar{x}$ is a solution of (3).

If x' is another solution of (3),

$$x' = \underline{A}x' + b = \underline{A}(\underline{A}x' + b) + b = \underline{A}^2 x' + \underline{A}b + b = \ldots$$

$$= \underline{A}^n x' + (\underline{A}^{n-1} + \underline{A}^{n-2} + \ldots + \underline{I})b = \underline{A}^n x' + (\hat{\underline{A}} + \underline{I})b$$

Hence $x' = \underline{A}\ x' + \bar{x} \geq \bar{x}$. Therefore $\bar{x}$ is the minimal solution of (3). The last relation immediately follows from $\hat{a}_{ii} = 1$. []

Now, note Property 1 in the following that tells an equivalence between the transitive closure and a method of clustering. (See [10] for the proof.)

Property 1 The method of sinle-linkage clustering [11] and the transitive closure of a proximity matrix (denoted as $\hat{\underline{A}}$ here) is equivalent in the following sense (a).

(a) Let $Z=\{z_1, z_2, \ldots, z_p\}$ be a set of objects for cluster analysis. Assume that a measure of similarity between z_i and z_j is given by a_{ij} ($a_{ij} = a_{ji}$) and let $a_{ii} = 1$. (We assume without loss of generality that $0 \leq a_{ij} \leq 1$.) Let $K_k(\alpha)$, $k=1,2,\ldots,N(\alpha)$ with parameter α ($0 \leq \alpha \leq 1$) be clusters formed at the level of similarity α. (i.e., $\bigcup_k K_k(\alpha) = Z$, $K_k(\alpha) \cap K_\ell(\alpha) = \emptyset$, $k \neq \ell$) by the single linkage algorithm. Then two elements z_i, $z_j \in Z$ belong to the same cluster (i.e., there exists $K_\ell(\alpha)$ such that $z_i, z_j \in K_\ell(\alpha)$) that are formed at the level α, if and only if $\hat{a}_{ij} \geq \alpha$, where $\hat{\underline{A}} = (\hat{a}_{ij})$ is the transitive closure of $\underline{A}$. []

It has been proved that algorithms for the minimal spanning tree of a network can be used for the single linkage cluster analysis [11]. Therefore we can summarize in an informal way that three concepts, the single linkage clustering, the minimal spanning tree, and the transitive closure of a proximity matrix with max-min operations are equivalent. This observation underlies methods developed here in the sequel.

3. Extension of indexes by fuzzy associations

Let us suppose for the moment that the relation T is binary valued. i.e., for any $d \in D$, $T(\cdot,d)$ is a crisp subset. In a previous paper we defined a fuzzy extension $Tf: X \times D \rightarrow [0.1]$ of the crisp index T through a fuzzy thesaurus F:

$$Tf(y,d) = \max_{x \in T(\cdot,d)} F(x,y) \qquad y \in X, \quad d \in D \qquad (4)$$

($T(\cdot,d)$ means the subset of X determined by T given d.)

The above definition means that the original crisp index is extended through a fuzzy thesaurus and in the extended index Tf the grade of relevance of a keyword y to a document is determined as the maximum of grades of relevance of y to keywords that are directly indexed to the document d.

Now, consider a question that if $T(x,d)$, the index function, itself is fuzzy, how we can define the extended index $Tf(y,d)$ through the fuzzy thesaurus. For this purpose we use max-min operations to represent (4) as follows:

$$Tf(y,d) = \max_{x \in T(\cdot,d)} F(x,y) = \sum_{\text{all } x \in X} F(x,y)\, T(x,d) \qquad (5)$$

$$(= \max_{x \in X} \min (F(x,y), T(x,d)))$$

The relation (5) is used when the index T is fuzzy, $T: X \times D \rightarrow [0,1]$. By the matrix form, $\underline{Tf} = \underline{F}^T \underline{T}$, where $\underline{Tf} = (tf_{ij})$, $tf_{ij} = Tf(x_i, d_j)$, $x_i \in X$, $d_j \in D$.

The extended index Tf is also represented by the fuzzy integral by Sugeno [12]:

$$Tf(y,d) = \int_X T(\cdot,d) \circ \bar{F}(\cdot,y) = \int_X F(\cdot,y) \circ \bar{T}(\cdot,d) \qquad (6)$$

where fuzzy measures $\bar{F}$ and $\bar{T}$ are defined as:

$$\bar{F}(K,y) = \max_{x \in K} F(x,y)$$

$$\bar{T}(K,d) = \max_{x \in K} T(x,d)$$

(K: any crisp subset of X) .

The proof that the representations (5) and (6) are equivalent is easy and is omitted here (See [13].)

An algorithm for the fuzzy retrieval uses the inverse Uf of Tf. Let

$$Uf(d,y) = Tf(y,d) .$$

The fuzzy retrieval function is expressed in three ways using U(d,x):

$$Uf(d,y) = \sum_{\text{all } x \in X} F(x,y) U(d,x) \qquad (7)$$

$$Uf(d,y) = \int_X U(d,\cdot) \circ \bar{F}(\cdot,y) = \int_X F(\cdot,y) \circ \bar{U}(d,\cdot) \qquad (8)$$

$$\underline{Uf} = \underline{Tf} = (\underline{F}^T \underline{U}^T)^T = \underline{UF} \qquad (9)$$

$$(\bar{U}(d,\cdot) = \bar{T}(\cdot,d))$$

Algorithm FR (fuzzy retrieval)

Assumption the fuzzy retrieval system has two files that keep grades of relevance. One is a fuzzy inverted file (FIF). FIF gives for each index x the corresponding documents d with the grade $U(d,x) \neq 0$. The other is the fuzzy association file (FAF). FAF provides for each index y the associated indexes x with the grades $F(x,y) \neq 0$.

Input: a given index $y \in X$.

Output: a sequence of records $\{(d, Uf(d,y))\}$ that are ordered according to the decreasing order of Uf.

FR1

```
For a given y
see FAF
for all x such that $F(x,y) \neq 0$
    see FIF
    for all d such that $U(d,x) \neq 0$
        $v(d,y) \leftarrow \min(F(x,y), U(d,x))$
        make record $(d, v(d,y))$
```

FR2 Sort the set of records made by FR1 according to the decreasing order by the first key d and the second key v. Scan the sorted sequence of records. For each subsequence of records for a particular occurrence of d, take the first record and delete the rest. (It is easy to see that the first record of such a subsequence represents the value of Uf(d,y), cf.[5].)

FR3 Sort the resulting sequence according to the decreasing order by the key v.

End of FR.

As is stated above. it is straightforward to apply the concept of fuzzy thesauri to other type of indexes such as the citation index. If we apply the method of automatic generation of a fuzzy pseudothesaurus to a set of bibliographic citation, we obtain a fuzzy association on a citation index. Application of the above method of a fuzzy retrieval on an extended fuzzy index to citation index is straightforward by considering that $F(x,y)$ is defined on a set of citations.

Clustering of citations has been studied by different authors. Two well-known studies are called a method of bibliographic coupling by Kessler [14] and a method of co-citation by Small [15]. The latter was applied to a large amount of bibliography based on the database SCIENCE CITATION INDEX (SCI) by the Institute for Scientific Information (ISI) [3]. As a result, clusters by co-citation are used as an advanced index of SCI. (See [3].)

If we assume here that the index set X means a set of citations and $\underline{F}$ be a proximity matrix which is a fuzzy association on citations. Clustering on citations (co-citation) is expressed simply by the transitive closure $\hat{\underline{F}}$. (Indeed, the clustering by ISI based on co-citation has been done by the minimal spanning tree algorithm for the single linkage method.) The bibliographic coupling by Kessler is a clustering on documents through citations, therefore a result of clustering documents is expressed as $\widehat{\underline{T}^T \underline{F}^T \underline{T}}$. An index by clusters on citations are obtained simply by replacing $\underline{F}$ by $\hat{\underline{F}}$ in equation (9). Thus,

$$\underline{Uf} = \underline{U}\,\hat{\underline{F}} \qquad (10)$$

is the extended fuzzy retrieval function through clustering on citations. Thus, we obtain an extended fuzzy cluster retrieval function defined by (10). For another type of retrieval on citation index, see [16].

4. Block diagram representation based on max-min operations

4A. Feedbacks in information retrieval and max-min operations

Feedbacks in information retrieval have been extensively discussed in the SMART retrieval system [17]. Heaps [2] showed block diagrams to represent information retrievals through associations and feedbacks. In this section we show that a block diagram based on fuzzy sets and max-min operations provides a good framework for considering various studies on advanced indexes and for showing possibilities of future researches.

Consider a simple example of a block diagram in Fig. 1, where q is a query and r is a response from the retrieval system. The retrieval system is represented by a rectangle with the symbol U which is also used to represent an inverted file. (or, it is called a retrieval function.) In general we need not distinguish between functions (or matrices) and symbols on rectangles that show components of a diagram. Therefore we can write r = U q.

As U is represented by a matrix $\underline{U}$, we assume in general that q is represented as an n-vector $q=(q_1,\ldots,q_n)$ and r is represented as an m-vector $r=(r_1,\ldots,r_m)$. That is, q is a fuzzy query in which index x_i is included with the membership q_i; r is the fuzzy response in which a document d_j is retrieved with the grade of relevance r_j. The query q might also be represented as a combination of fuzzy logical operations on elements in X, the index set, for example,

$$q = x_1 \text{ OR } (x_2 \text{ AND } x_3) \text{ OR } (x_4 \text{ AND } x_5) \text{ OR } \ldots \qquad x_1, x_2, \ldots \in X.$$

We consider a problem on operations on queries in section 4C.

Now, a simple fuzzy retrieval through a fuzzy thesaurus or a fuzzy association is represented as Fig. 2: $r = U F q$, which is nothing but the relations (7), (8), or (9). In Fig. 2 we used an additional symbol $q' = F q$: q' shows an extended query using F, and q' serves as an input to the retrieval system U ($r = Uq'$). Now, consider a feedback system in which the input is q and the output is q'. The feedback system is shown as Fig. 3. In Fig. 3 the extensin of q through F is performed many times.

$$q' = q + F q + F^2 q + \ldots..$$

An equation for q is obtained:

$$q' = F q' + q .$$

Prop. 3 is applied and we have a minimal solution $q' = (\underline{\hat{F}} + \underline{I})q$ In particular, when F is a proximity relation, $q' = \underline{\hat{F}}\, q$. From now we use a notation $\hat{F}$ that shows the fuzzy equivalence relation obtained from the transitive closure matrix $\underline{\hat{F}}$, since we need not distinguish between a relation and a matrix, therefore we have $r = U \hat{F} q$, which is nothing but (10). Thus, we are led to the concept of fuzzy cluster retrieval again by considering a feedback.

There is another way of feedback in a retrieval system, i.e., feedback from output. Consider Fig. 4a, where indexes of directly retrieved documents $r_1 = U\,q$ are used as a secondary input to the retrieval system and the secondary output $r_2 = U\,T\,U\,q$ is added to the direct response r :

$$r = (\,U\,T\,U\ +\ U\,)q\ .$$

If we use once again the indexes of the retrieved documents r_3 as an input, we have

$$r = (\,U\,T\,U\,T\,U\ +\ U\,T\,U\ +\ U\,)q.$$

Continuing in this way, we are led to a feedback from the output shown in Fig. 4b, where

$$r = U\,(\,T\,r\ +\ q\,)\ .$$

If we denote the closure of the relation U T as $\widehat{U\,T}$, we have a minimal solution

$$r = (\,\widehat{U\,T}\ +\ I\,)\,U\,q\ .$$

A variation of Fig. 4a is shown as Fig. 5a, where the fuzzy association is used in the processing of the secondary input:

$$r = (\,U\,F\,T\,U\ +\ U\,)\,q\ .$$

In the same way, a minimal solution of the feedback system with the fuzzy association (Fig. 5b) is obtained:

$$r = (\,\widehat{U\,F\,T}\ +\ I\,)\,U\,q\ .$$

Note that if we use the matrix form of the relations, T is represented as $\underline{U}^T$.

We have the following proposition that shows a relation between the feedback system and the clustering.

Prop. 4

(i) $U\,T$ is a proximity relation if for any document d_i, $i=1,2,\ldots,m$, there exists an index $x_j \in X$ such that $T(x_j, d_i) = 1$ (that is, x_j is a crisp index to d_i.) In this case $\widehat{U\,T}$ agrees with the result of single linkage clustering based on the measure of similarity $s(d_i, d_j)$ for cluster formation:

$$s(d_i, d_j) = \max_{k=1,\ldots,n} \min\ (\ T(x_k, d_i),\ T(x_k, d_j)\ )$$

(ii) Assume that $F(x_i, x_i) = 1$ for $i=1,\ldots,n$ and $F(x_i, x_j) = F(x_j, x_i)$ for $i,j=1,2,\ldots,n$, $i \neq j$. Then $U\,F\,T$ is a proximity relation if for any document d_i, $i=1,2,\ldots,m$. there exists an index $x_j \in X$ such that $T(x_j, d_i) = 1$. In this case $\widehat{U\,F\,T}$ agrees with the result of single linkage clustering based on the measure of similarity $s'(d_i, d_j)$ for cluster formation:

$$s'(d_i, d_j) = \max_{\substack{k=1\ldots,n \\ \ell=1,\ldots,m}} \min\ (\ T(x_k . d_i),\ F(x_k, x_\ell),\ T(x_\ell, d_j)\ )$$

(Proof) It is easy to prove the above relations from the definition of the max-min operations and from Property 1. Therefore the proof is omitted. []

Remark We have discussed an equation $x = \underline{A}\, x + b$ for which the minimal solution is $x = \hat{\underline{A}}\, b + b$. A natural question is that whether or not there is another solution that are not minimal. The answer is, in general, the affirmative one. If there is a nontrivial answer of the equation $x' = \underline{A}\, x'$, the solution of $x = \underline{A}\, x + b$ is not unique, since $x'' = x + x'$ satisfies

$x'' = \underset{\sim}{A}\, x'' + b$. Moreover, if the diagonal elements of $\underset{\sim}{A} = (a_{ij})$ are unity: $a_{ii} = 1$, then it is easy to see that $x' = (1,1,\ldots,1)$ satisfies $x' = \underset{\sim}{A}\, x'$. Therefore in the case of a proximity matrix, the uniqueness of the solution fails. In case of information retrieval with feedbacks, $x' = (1,1,\ldots,1)$ means all the documents in the database. Therefore we are interested only in the minimal solution, since another solution which is not minimal may contain unnecessary informations. []

4B. Filters and level fuzzy sets

Practical considerations on fuzzy information retrieval should include consideration and implementation of some filters to exclude excessive numbers of retrieved documents with low grades of relevance. Moreover a filter may include other kinds of functions such as modification of grades of relevance that reflects current interest of a particular user, in which case a filter should use a profile of the user. Thus, a filter performs supplementary functions which can not be incorporated into fuzzy indexes or fuzzy associations (Fig. 6).

A simple example of a filter is a level fuzzy set [18] that cuts off elements whose values of membership are less than alpha. A level fuzzy set $\tilde{A}_\alpha$ [18] is defined as $\tilde{A}_\alpha = \{(x, m_A(x)),\ x \in A_\alpha\}$, where A_α is the alpha-cut of A and m_A is the membership of A. Here we introduce an operator $C(\alpha)$ that maps a fuzzy set A to $\tilde{A}_\alpha$: $C(\alpha)A = \tilde{A}_\alpha$. If a fuzzy association has a large number of nonzero entries $F(x_i, x_j)$, the processing in the whole system needs a large amount of computation. Therefore $C(\alpha)$ is introduced as in Fig. 7a or in Fig. 7b (feedback on F).

The relations between the queries and the responses in these figures are

$$r = C(\alpha)\ U\ F\ q \tag{11}$$

in Fig. 7a and

$$r = C(\alpha)\ U\ (\ \hat{F} + I\ )\ q \tag{12}$$

in Fig. 7b.

Remark The level fuzzy set in information retrieval has been introduced by Radecki [7]. $C(\alpha)$ can not be expressed as a fuzzy relation nor a matrix. Therefore it is a "nonlinear" element in the diagram representation. []

Prop. 5 The following equations are valid.

$$C(\alpha)\ (\ U\ F\ ) = (\ C(\alpha)\ U\ )\ (\ C(\alpha)\ F\ ) \tag{13}$$

$$C(\alpha)(\ U(\ \hat{F} + I\ ))^{\cdot} = (\ C(\alpha)\ U\ )\ ((\ \widehat{C(\alpha)\ F}\ ) + I\ ) \tag{14}$$

To prove this proposition, note the following lemma.

Lemma 1 For two fuzzy relations A and B defined on X x X,

$$C(\alpha)(\ A + B\ ) = (\ C(\alpha)A\ ) + (\ C(\alpha)B\ )$$

$$C(\alpha)(A\ B) = (\ C(\alpha)A\ )\ (\ C(\alpha)B\ )$$

(Proof of Lemma 1) Let $\tilde{A}_\alpha = C(\alpha)A$ and $\tilde{B}_\alpha = C(\alpha)B$. Then

$$\tilde{A}_\alpha(x_i, x_j) = \begin{cases} A(x_i, x_j) & (A(x_i, x_j) \geq \alpha) \\ 0 & (\text{ otherwise }) \end{cases}$$

Therefore,

$$[(C(\alpha)A) + (C(\alpha)B)](x_i,x_j) = \max(\tilde{A}_\alpha(x_i,x_j),\ \tilde{B}_\alpha(x_i,x_j))$$

$$= \begin{cases} \max(A(x_i,x_j),B(x_i,x_j)) & (\max(A(x_i,x_j),B(x_i,x_j)) \geq \alpha) \\ 0 & (\text{ otherwise }) \end{cases}$$

$$= [C(\alpha)(A + B)](x_i,x_j)$$

$$[(C(\alpha)A)(C(\alpha)B)](x_i,x_j) = \max_k \ \min(\tilde{A}_\alpha(x_i,x_j),\tilde{B}_\alpha(x_i,x_j))$$

$$= \begin{cases} \max\limits_k \min(A(x_i,x_j),B(x_i,x_j)) & (\max\limits_k \min(A(x_i,x_j),B(x_i,x_j)) \geq \alpha) \\ 0 & (\text{ otherwise }) \end{cases}$$

$$= \ [C(\alpha)(A\ B)](x_i,x_j)\ . \qquad []$$

(Proof of Prop. 5) Equation (13) directly follows from Lemma 1. To Prove (14), note that $\hat{F} = F + F^2 + \ldots + F^{n-1}$. Therefore,

$$C(\alpha)(U(\hat{F} + I)) = (C(\alpha)U)(C(\alpha)(\hat{F} + I))$$

$$= (C(\alpha)U)(C(\alpha)\hat{F} + C(\alpha)I) = (C(\alpha)U)(C(\alpha)\hat{F} + I)$$

$$= (C(\alpha)U)(C(\alpha)(F + F^2 + \ldots + F^{n-1}) + I)$$

$$= (C(\alpha)U)((C(\alpha)F) + (C(\alpha)F)^2 + \ldots + (C(\alpha)F)^{n-1} + I)$$

$$= (C(\alpha)U)(\ \widehat{C(\alpha)F}\ + I\ ) \qquad []$$

A significance of the relations (13) and (14) is that if we wish to cut off documents with grade of relevance below alpha, we can apply $C(\alpha)$ not only on the response but also on the fuzzy association and on the fuzzy index. Since the relation $(C(\alpha)U)(C(\alpha)F)$ (resp. $(C(\alpha)U)(\widehat{C(\alpha)F} + I)$) has a fewer number of nonzero entries than $C(\alpha)(UF)$ (resp. $(C(\alpha)(U(\hat{F} + I))$), we can reduce the amount of computation using $C(\alpha)$ on F and U beforehand. Note also that the following corollary holds.

Cor. 1 For $0 < \alpha_1 \leq \alpha_2 \leq 1$,

$$C(\alpha_1)(\ U\ F\ ) = C(\alpha_1)(\ C(\alpha_2)U\ )\ (\ C(\alpha_2)F\ )$$

$$C(\alpha_1)(\ U\ (\ \hat{F} + I\ )\ ) = C(\alpha_1)(\ C(\alpha_2)U\ )\ (\ \widehat{C(\alpha_2)F} + I\ )$$

(Proof) It is sufficient to note that for any fuzzy relation A,

$$C(\alpha_1)(\ C(\alpha_2)A\ ) = C(\alpha_1)A \qquad []$$

A similar result holds on output feedbacks. We omit the detail.

4C Simple queries and composite queries

In section 3 we assumed that a query to a fuzzy information retrieval system is a simple index x_i, whereas in section 4 a query q is a composite fuzzy query in which many keywords may be included with their memberships. We discuss here how these two are related without ambiguity nor contradiction.

First, note that it is straightforward to perform a fuzzy retrieval by giving a single index y with a grade of relevance w ($0 \leq w \leq 1$). It is sufficient to retrieve documents by a variation of algorithm FR in which the statement $v(d,y) \leftarrow \min(F(x,y),U(d,x))$ in FR1 is replaced by $v(d,y) \leftarrow \min(w,F(x,y).U(d,x))$.

Next, consider a composite query. It appears that a composite query is represented like

$$q = x_1 \text{ OR } (\ x_2 \text{ AND } x_3\ ) \text{ OR } x_4 \ldots$$

using AND/OR operations, where x_i, $i=1,2,\ldots$ may be a single fuzzy index. First of all, it should be noted that logical operations of the above form are not performed on the queries but performed on the responses. For example. a query $q = x_i$ AND x_j should be interpreted as $U q = (U x_i)$ AND $(U x_j)$. We use the minimum for fuzzy AND, therefore the response is $(U x_i) \cdot (U x_j)$ (cf. Table 1).

On the other hand, Consider OR operation using the maximum. A query $q = x_i$ OR x_j is interpreted as $U q = (U x_i)$ OR $(U x_j)$. Since OR is expressed as '+', we have

$$U q = U (x_i + x_j) = U x_i + U x_j \qquad (15) .$$

In other words. $(U x_i)$ OR $(U x_j) = U (x_i$ OR $x_j)$. This relation means that if a composite query is represented as a combination of simple queries by the OR operation, the response to a composite query is equal to the result of the same form of combination of responses to the simple queries. In short, OR operation by the maximum has a property of linearity represented by (15).

Thus, the expression $q = x_i$ AND x_j is not a mathematical one. whereas $q = x_i$ OR x_j can be interpreted as a logical operation on queries owing to (15). Thus, for a composite queries, we decompose it into elementary queries, and operations are performed on responses by minimum and maximum. For OR operation, we can modify the algorithm FR for composite queries. Namely, FR1 is modified as follows:

```
FR1' (Assumption: Input is $q=(q_1,q_2,\ldots,q_n)$, where $q_i$ is the
      value of membership corresponding to $y_i$.)
    For all $y_i$ such that $q_i \neq 0$
        see FAF
        for all x such that $F(x,y_i) \neq 0$
            see FIF
            for all d such that $U(d,x) \neq 0$
                $v(d,q) \leftarrow \min(q_i,F(x,y_i),U(d,x))$
                make record $(d,v(d,q))$
```

The other steps of the modified algorithm are the same as FR.

5. Conclusions

Two methods of representation of information retrieval through fuzzy associations have been described. One is a calculation of an extended index and the other is a block diagram representation based on max-min algebra. A block diagram representation of information retrieval based on the ordinary addition, subtraction, multiplication, and division has been proposed by Heaps [2]. Comparing the former approach by Heaps with the method herein, we point out two reasons why the present method is better than the former method.

(1) The former method of block diagram is based on the ordinary type of matrix calculation. Since the numbers of indexes and documents are huge in practical databases, the matrix calculation, even if the matrices are sparse, needs great amount of computation. On the other hand, the present approach uses

sorting and manipulation of linear files for the max-min calculation. such as algorithm FR and network algorithms for the minimum spanning tree (i.e., single likage clustering). Therefore the method herein is more appropriate for large amount of data in bibliographic databases.

(2) Clustering of documents has been considered to be an important subject of advanced indexing. The present method represents clustering of documents by the feedback of indexes or outputs. whereas the former method can not represent clusterings in its block diagram. Thus, the framework of fuzzy relation covers wider area of studies in bibliographic information retrieval in a unified representation by diagrams.

When the two methods of the extended indexes and the block diagram representation, we find advantages of each method over the other:

(a) The block diagram shows in a compact way how queries are processed into a response. Thus, the filter and level fuzzy sets are shown on the diagrams, and behavior of the functions can be studied by calculations on the diagram. Moreover. composite queries are studied and linearity of OR operation is shown.

(b) The extended indexes do not seem to represent information flows and behavior of the system in a macroscopic way. Instead, detailed calculation on fuzzy indexes needs consideration such as those in section 3. Thus, the representation of extended indexes is appropriate for consideration of properties of fuzzy indexes in a microscopic scale and for developing algorithms such as FR.

In a previous paper [4] we showed a method of generating fuzzy thesauri. The concept of fuzzy thesauri is generalized to fuzzy associations through the method of automatic generation [4],[8],[9]. Here methods of fuzzy extended indexes and a block diagram representation are shown. Throughout these studies, we observe that foregoing and current studies on advanced indexing and on bibliographic information analysis can be discussed within the framework of fuzzy sets, as if the foregoing and current studies like the automatic generation of thesauri [17] and the clustering of documents [3] unconsciously assumed fuzzy sets. Thus, the method of fuzzy sets is one of the best frameworks for a theory of information retrieval and will be useful in future studies in this area.

References

1. F. W. Lancaster. Vocabulary Control for Information Retrieval. Washigton DC. Information Resources. 1972.

2. H. S. Heaps, Information Retrieval: Computation and Theoretical Aspects. New York. Academic Press, 1978.

3. E. Garfield, Citation Indexing: Its Theory and Application in Science, Technology, and Humanities, New York, Wiley, 1979.

4. S. Miyamoto. T. Miyake, and K. Nakayama, Generation of a pseudothesaurus for information retrieval based on cooccurrences and fuzzy set operations. IEEE Trans., Syst., Man, and Cybern., Vol.SMC-13. No.1, 62-70. 1983.

5. S. Miyamoto and K. Nakayama. Fuzzy information retrieval based on a fuzzy pseudothesaurus, IEEE Trans., Syst., Man, and Cybern., Vol. SMC-16. No.2, 178-282. 1986.

6. L. Reisinger. On fuzzy thesauri. in COMPSTAT 1974, G. Bruckman et al., Eds.. Vienna, Physica-Verlag. 1974, 119-127.

7. T. Radecki, Mathematical model of information retrieval system based on the concept of fuzzy thesaurus. Information Processing and Management, Vol.12. 313-318. 1976.

8. S. Miyamoto, T. Miyake, and K. Nakayama. Structure generation on bibliographic databases with citations based on a fuzzy set model, Proceeding, IFAC Symp.. on Fuzzy Information, Marseille. 225-230, 1983.

9. S. Miyamoto, Algorithms for a fuzzy association retrieval. Working Paper. International Institute for Applied Systems Analysis. WP-87-20. 1987.

10. J. C. Dunn, A graph theoretic analysis of pattern classification via Tamura's fuzzy relation. IEEE Trans.. Syst.. Man. and Cybern.. Vol.SMC-4. 310-313. 1974.

11. M. R. Anderberg. Cluster Analysis for Applications. New York. Academic Press. 1973.

12. M. Sugeno, Theory of fuzzy integral and its applications, Ph. D. Thesis, Tokyo Inst. of Technology, 1974.

13. D. Dubois and H. Prade, Fuzzy Sets and Systems: Theory and Applications, New York, Academic Press, 1980.

14. M. M. Kessler. Bibliographic coupling between scientific papers, American Documentation, Vol.14. No.1, 10-25. 1963.

15. H. Small. Co-citation in the scientific literature: a new measure of the relationship between two documents. Jounal of the American Society for Information Science, Vol.24. No.4, 265-269, 1973.

16. K. Nomoto, S. Wakayama, T. Kirimoto, and M. Kondo. A fuzzy retrieval system based on citations, Preprint, Second IFSA Congress, Tokyo, 723-726, 1987.

17. The SMART Retrieval System, Experiments in Automatic Document Processing, G. Salton, Ed., Englewood Cliffs, NJ: Prentice-Hall, 1971.

18. A. Kandel, Fuzzy Mathematical Techniques with Applications, Reading, Massachusetts, Addison-Wesley, 1986.

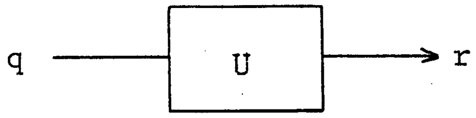

Fig. 1 A block diagram reprentation of information retrieval process.

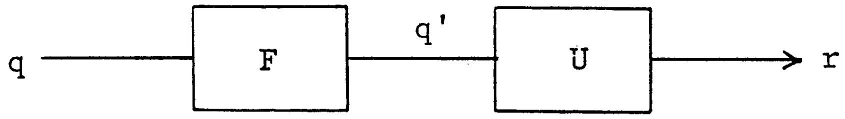

Fig. 2 A block diagram of information retrieval through fuzzy association.

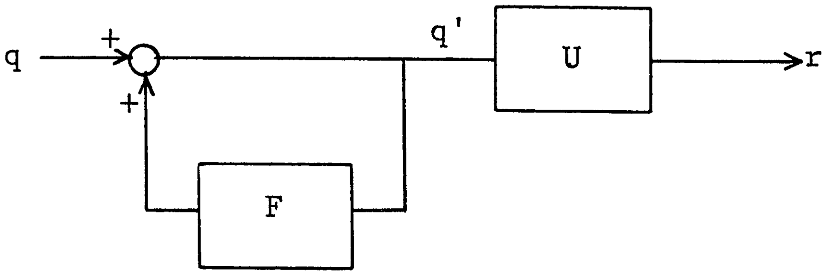

Fig. 3 A feedback on fuzzy association

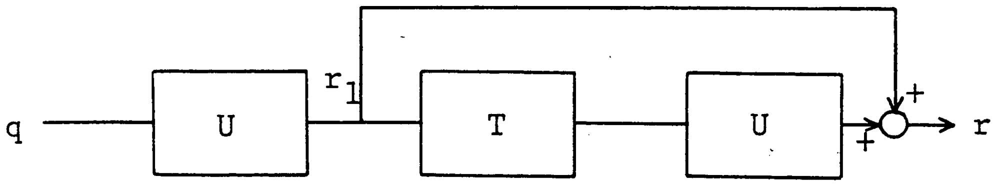

Fig. 4a

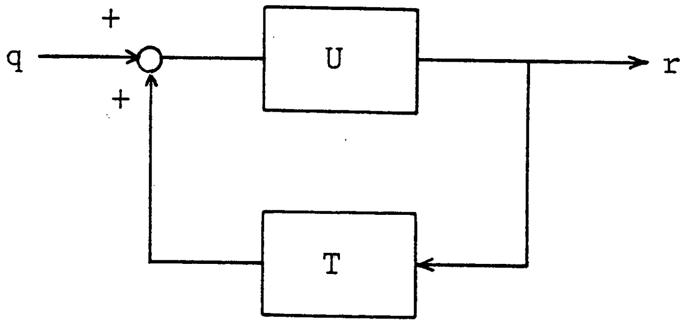

Fig. 4b

Fig. 4 Feedback from output of retrieval system

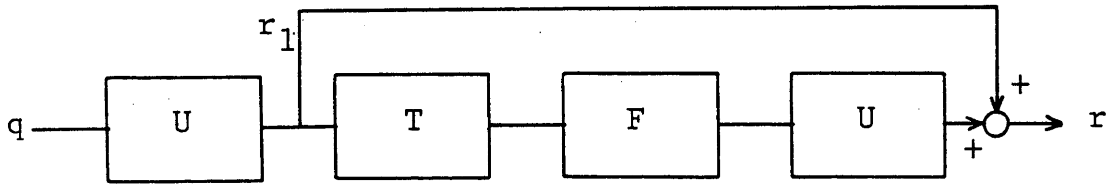

Fig. 5a

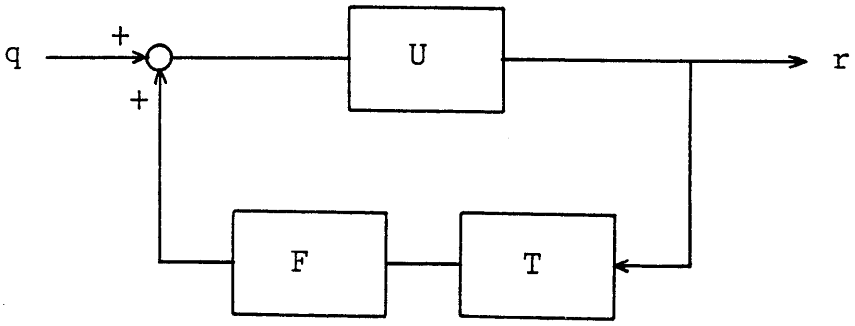

Fig. 5b

Fig. 5 Feedback from output through fuzzy association

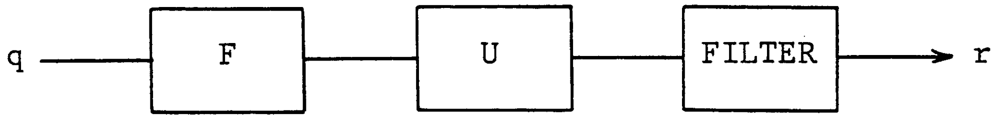

Fig. 6 A filter on block diagram of information retrieval system.

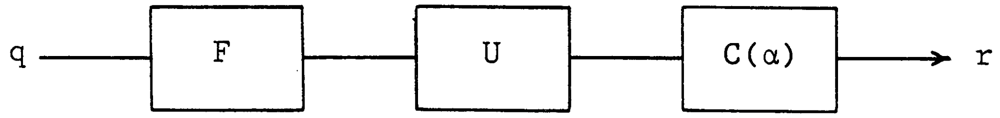

Fig. 7a

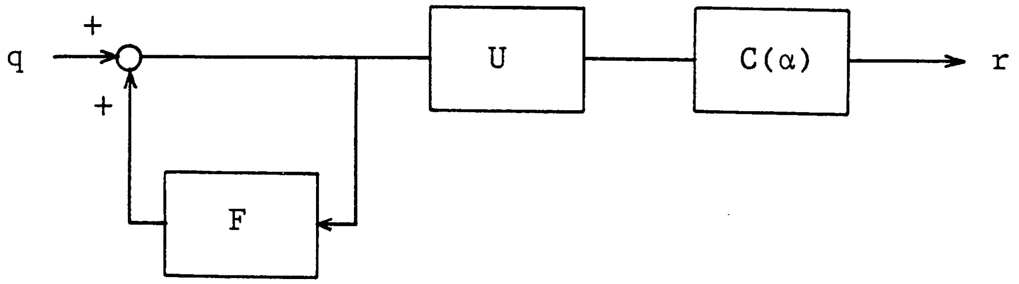

Fig. 7b

Fig. 7 Level fuzzy sets on block diagrams of information retrieval system.

Table 1 AND/OR operations and +/· operations of the system.

x_i OR x_j	$\longrightarrow$	$U x_i + U x_j = U(x_i + x_j)$
x_i AND x_j	$\longrightarrow$	$(U x_i) \cdot (U x_j)$

INSTITUTE OF INFORMATION SCIENCES AND ELECTRONICS
UNIVERSITY OF TSUKUBA
SAKURA-MURA, NIIHARI-GUN, IBARAKI 305 JAPAN

REPORT DOCUMENTATION PAGE	REPORT NUMBER ISE-TR-88-69

TITLE

Two approaches for information retrieval through fuzzy associations

AUTHOR(S)

Sadaaki Miyamoto

REPORT DATE June 3, 1988	NUMBER OF PAGES 25 + figures
MAIN CATEGORY Information Storage and Retrieval	CR CATEGORIES H.3.1, H.3.3

KEY WORDS

fuzzy information retrieval, clustering, block diagrams, fuzzy associations, thesauri.

ABSTRACT

The aim of the present paper is to show two approaches for formulation of information retrieval through fuzzy associations. A fuzzy association is introduced as a fuzzy relation defined on a set of indexes to a database. The fuzzy association is a generalization of a concept of fuzzy thesauri and methods of generating fuzzy thesauri are applicable to generation of fuzzy associations. One approach for the fuzzy retrieval is extension of fuzzy indexes to adatabase using fuzzy associations. An algorithm for the fuzzy information retrieval based on this approach is developed. The other approach represents the retrieval process as a block diagram. Maximum and minimum operations are used instead of the ordinary sum and product operations on the diagram. Studies on advanced indexing such as clustering of articles are represented as feedbacks on the diagram and properties on fuzzy information retrieval such as level fuzzy sets and set operations on responses of the retrieval system are discussed using the diagram representation.

SUPPLEMENTARY NOTES