Intelligent Database Retrieval System using FCM

°Ihn JEONG*, Gyei-Kark PARK**, Seung-Wook HWANG*

* Dept. of Control and Instrumentation Eng., Korea Maritime Univ.
1 Dongsam-dong Youngdo-ku Pusan 606-791, KOREA
E-mail : micro@soback.kornet.nm.kr

** Dept. of Nautical Science, Mokpo National Maritime Univ.
571 Chukkyo-dong Mokpo 530-729, KOREA
E-mail : gkpark@quagga.kaist.ac.kr

ABSTRACT

In this paper, we propose a retrieval system using knowledges of database expressed linguistically, where the relation between data are constructed by FCM. Several algorithms have been proposed to solve the major problem in the conventional retrieval system that the system doesn't reply in case of no data equal to user's query, and to express knowledge of database linguistically. This paper proposes the improved method of adding new cluster and the method of retrieving database from user's query. The validity of this retrieval system is shown by applying its algorithm to an example : the mail order service in Post office.

I . Introduction

The hegemony of world economy has been changed from the 20 century-typed industry economy to the 21 century-typed information economy. By means of the advance of the digital technology, computer technology and optic-commnication technology, the high-tech information age has arrived and thus the technology in hardware has made remarkable progress enough to deal with a large amount of information, however, not enough in software. Therefore, it is essential the technique to obtain the best data out of a large amount of data. In the conventional database retrieval system, only the data that satisfy the user's query has been served, otherwise no data has been served. In order to solve such a problem, many systems have been proposed. One system loosens the condition of retrieval to retrieve the alternatiive data, but it needs the heuristic knowledge according to the target of retrieval[1]. The other system uses the fuzzy membership function to express the condition of retrieval and draws the data with the high grade of fuzzy membership function, but it also has a problem that it happens not to retrieve according to the selection of fuzzy membership functions[2]. To solve these problems, another system has been proposed, which uses FCM(Fuzzy C-Means) method to express data of database as the multiple clusters and the distribution of data is described linguistically by using linguistic labels defined[3]. This paper proposes the improved method of adding new cluster and interface part between user and system. The

validity of this retrieval system is shown by applying its algorithm to an example : the mail order service in Post office.

Ⅱ. Construction of Linguistic Intelligent Database Retrieval System

For constructing the system, FCM is used to draw the relation between data and to make it knowledge.

2.1 Fuzzy Clustering

FCM method proposed by Bezdek regards that the data X_k is classified into multiple clusters with different grade repectively and is a fuzzy version of HCM(Hard C-Means) method which says that data is contained in only one cluster. Then FCM method is described briefly.

When n data vectors with t-dimension $X_k = x_{k,p}$ $(p=1,2,\cdots,t)$ $(k=1,2,\cdots,n)$ is classified into c clusters, the dissimilarity $d_{i,k}$ between each central vector of cluster V_i $(i=1,2,\cdots,c)$ and data X_k is expressed by using Euclid distance like equation (1).

$$d_{i,k} = \| X_k - V_i \| \qquad (1)$$

Then, central vector V_i is expressed such as equation (2) and $U_{i,k}$ is like eqation (3),

$$V_i = \sum_{k=1}^{n} (U_{ik})^m X_{ki} \Big/ \sum_{k=1}^{n} (U_{ik})^m, \qquad (2)$$

$$U_{i,k}^{(l+1)} = 1 \Big/ \sum_{j=1}^{c} (d_{i,k}/d_{j,k})^{1/(m-1)} \qquad (3)$$

Here, $U_{i,k}$ means the grade how much X_k is contained in cluster i, V_i is the m dimensional weight mean of X_k's membership function value. FCM method put the routine renewing U and V and is as following.

Procedure of FCM:

step 1: Number of clusters c($(2 \le c < n)$, weight m $(1 < m < \infty)$, ε (threshold), cs divided matrix $U^{(0)}$ of initial U is assigned and put $l=0$.

step 2: central vectors of cluster $V_i^{(l)}$ $(i=1,2,\cdots,c)$ are solved by equation (2) using $U^{(l)}$.

step 3: If $X_k \neq V_i^{(l)}$, $U_{i,k}^{(l+1)}$ is renewed by equation (3). If $X_k = V_i^{(l)}$, put

$$U_{i,k}^{(l+1)} = \begin{Bmatrix} 1 & i \in I_k \\ 0 & i \notin I_k \end{Bmatrix}, \qquad (4)$$

$$I_k = \{ i \mid 1 \le i \le c,\ d_{i,k} = |X_k - V_i| = 0 \}$$

$$\forall k = 1 \sim n. \qquad (5)$$

step 4: For threshold given ε , if equation (6) is satisfied, terminate. Otherwise put $l = l+1$ and return step 2.

$$\| U^{(l+1)} - U^{(l)} \| \le \varepsilon. \qquad (6)$$

2.2 Construction of Linguistic Knowledge Expression

The algorithms of increasing cluster and expressing the relation between data linguistically by matching linguistic labels onto fuzzy clusters obtained through FCM is described.

2.2.1 Selection of Appropriate Label

Linguistic labels are defined by membership function $\mu_{L^s_j}(x_{k,j})$. The linguistic label L^s_j that minimizes equation (7) is selected as the linguistic label according to each cluster on the j-th property. Only data

with membership grade greater than threshold value α are applied to equation (7) so as to avoid the increase of C_s due to data with little membership grade,

$$C_s = \sum_{k=1}^{n} e_k ,$$

$$e_k = \begin{cases} U_{i,k} - \mu_{L^s_j}(x_{k,j}) & (U_{i,k} \geq \mu_{L^s_j}(x_{k,j}), \alpha) \\ 0 & (U_{i,k} < \mu_{L^s_j}(x_{k,j}) . \end{cases} \quad (7)$$

$U_{i,k}$: membership grade of k-th data in i-th cluster.

L^s_j: s-th linguistic label on j-th property.

$\mu_{L^s_j}(x_{k,j})$: membership function of k-th data's s-th linguistic label on j-th property.

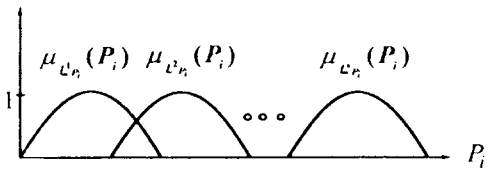

Fig. 1 Membership functions

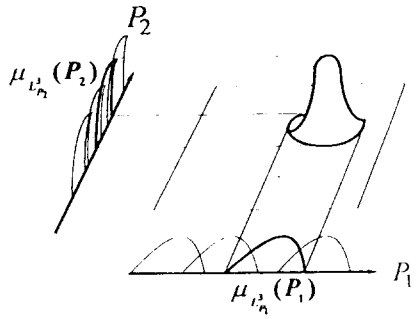

Fig. 2 Projection of fuzzy cluster

2.2.2 Increasing Number of Cluster and Re-initialzing Increased Cluster

The solution for determining optimal number of cluster has not been known. So, the number of cluster c that minimizes $S(c)$ in the equation (8) is determined as the optimal number of cluster. That is, when $S(c) \leq S(c+1)$ is available, c is determined as the optimal number of cluster[8].

In addition , in this paper, when $S(c) \leq S(c+1)$ or $| S(c+1) - S(c) | \leq M$ is available, where M is threshold value and c is determined as the optimal number of cluster and clustering is terminated. Otherwise, the number of cluster is increased.

$$S(c) = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{i,k})^m (\| X_k - V_i \|^2 - \| V_i - \bar{x} \|^2) \quad (8)$$

In equation (8), n is number of data, X_k is k-th data, $\bar{x}$ is mean of data, V_i is central vector in i-th cluster, $\| \cdot \|$ is norm, $\mu_{i,k}$ is membership grade of k-th data in i-th cluster, m is weight.

For increasing number of cluster, re-initialzing of $U_{i,k}$ is as following.

step 1: Calculate the distance between V_i and X_k.

step 2: Slove the data X_l with the smallest distance step 1.

step 3: Put $U_{c+1,k} = 1 \quad (k = l)$,

$U_{c+1,k} = 0 (k \neq l)$,

assign $U_{i,l} = 0 (i = 1, \cdots, c)$,

re-initialize $U_{c+1,k}$.

2.3 Cooperative Response System

User inputs quantitatively on the quantitative property and selects input from examples given by the system on the

qualitative property.

The procedure on quantitative properties is as following.

Step 1: Calculate $\mu_{L'_j}(k)$ $(l=1,\cdots,s)$.

Step 2: Solve the number of label l with the biggest membership function grade $\mu_{L'_j}(k)$.

Step 3: Compare l with the number of fuzzy clusters f^n_j $(n=1,\cdots,c)$ $(j=1,\cdots,p)$.
If $l = f^n_j$, output L'_j.
Otherwise, calculate $avg(k)$ $(k=1,\cdots,s)$ and solve the label L^k_j with the smallest distance between k and $avg(k)$.

$avg(k)$: mean of k-th label's 4 parameters.
f^n_j : label number of n-th fuzzy cluster on j-th property.

The procedure on qualitative properties is as following.

Step 1: Compare the number of label selected by user l with f^n_j.

Step 2: If $l = f^n_j$, output L'_j.
Otherwise, calculate $avg(l)$ and $avg(n)$ $(n=1,\cdots,c)$ and solve the label L^k_j with the smallest distance between $avg(l)$ and $avg(n)$.

When user wants all of the data in the selected cluster, system outputs them that satisfies eq.(9),

$$u_{i,k} > \varepsilon, \quad (0<i<c,\ 0<k<n). \tag{9}$$

Ⅲ. Application to Choosing Gift

3.1 Construction of Database Retrieval System

The linguistic intelligent database retrieval system is applied to the mail order service in Post office to verify its usefulness. In the database, there are 105 data with 4 kinds of properties(to be considered when choosing gift ; Age, Object, Use, and Price).

The qualitative properties 'Object' and 'Use' are arranged and expressed quantitatively with respect of the relations between labels on each property.

Table 1 shows linguistic labels defined in each property. The result of relation between data acquired by matching linguistic labels onto fuzzy clusters obtained through FCM is shown in Table 2.

For example, the 2nd cluster means 'Old' for property 'Age', 'Friend' for property 'Object', 'Birthday' for propery 'Use', 'about ₩100,000' for property 'Price', and 40 data is available.

3.2 Example of Cooperating Response

When user's input is identical with data in database, system outputs result like Fig. 3. Otherwise, system outputs the alternative data like Fig. 4.

```
System : Input for Property 'Age'.
User   : 45
System : Select Input for Property 'Object'.
(1)Child (2)Lover (3)Friend (4)Colleague (5)Senior
User   : 2
System : Select Input for Property 'Use'.
(1)Entrance (2)Birthday (3)Ceremony (4)Wedding (5)Thank
User   : 5
System : Input for Property 'Price'.
User   : 15000

There is a cluster which is the same as your query.
Ther is 13 data in the cluster.
Type '1' & return if you want to see data.
```

Fig. 3 User's input and system's output 1

```
System : Input for Property 'Age'.
User   : 45
System : Select Input for Property 'Object'.
(1)Child (2)Lover (3)Friend (4)Colleague (5)Senior
User   : 3
System : Select Input for Property 'Use'.
(1)Entrance (2)Birthday (3)Ceremony (4)Wedding (5)Thank
User   : 2
System : Input for Property 'Price'.
User   : 35000

There is no cluster which is the same as your query.
But there are 14 data near to your query.
Type '1' & return if you want to see data.
```

Fig. 4 User's input and system's output 2

Table 1 Linguistic labels

Property	Linguistix labels
Age	Little, Young, Middle, Old
Object	Child, Lover, Friend, Colleague, Senior
Use	Entrance, Birthday, Ceremony, Wedding, Thank
Price	Man_1_2, Man_3_4, Man_5, Man_7_8, Man_10

Table 2 Macro expression of database

NC.	Age	Object	Use	Price	ND
1	Young	Friend	Birthday	Man_10	39
2	Old	Friend	Birthday	Man_10	40
3	Young	Friend	Thank	Man_7_8	13
4	Middle	Lover	Thank	Man_3_4	14

*NC : Number of cluster, ND : Number of data

Ⅳ. Conclusion

This paper has improved the conventional method for adding new cluster and proposed the method of retrieving database from user's query. Thus, we have constructed the retrieval system that extracts knowledge between data by using FCM algorithm and provides user with the alternative data even if no data exactly satisfying user's query. In addition, the validity of this retrieval system has been shown by applying its algorithm to an example : the mail order service in Post office. The better method of quantification for qualitative property has to be studied for the more practical system.

[Reference]

[1] T. Gaasterland, P. Godfrey and J. Minker, "An Overview of Cooperative Answering", Journal of Intelligent Information System, 1, 1992, pp.123-157.

[2] S. Miyamoto : "Fuzzy Sets in Information Retrieval and Cluster Analysis", Theory and Decision Library, Series D, (Kluwer Academic Publishers, 1990).

[3] J. Ozawa and K. Yamada, "Generating a fuzzy model from a database and using it to find alternative data", Proc. of First Australian and New Zealand Conference on Intelligent Information Systems, ANZIIS-93, 1993, pp.560-564.

[4] J. C. Bezdek, "Pattern Recognition wirh Fuzzy Objective Function Algorithm," (Plenum Press, New York, 1981).

[5] 坂和正敏, "ファジイ理論の基礎と應用", (森北出版, 日本, 1990).

[6] J. Ozawa and K. Yamada, "Cooperative Answering with macro expression of a database", the 10th Fuzzy System Symposium, 1994, pp.101-104.

[7] M. Sugeno, and T. Yasukawa, "A Fuzzy-Logic-based Approach to Qualitative Modeling", IEEE Trans. on Fuzzy systems, Vol.1, No.1, 1993, pp.7-31.

[8] "Mail order catalog", Post office, 1994.

[9] I. Jeong, G.K. Park, S.W. Hwang, "Retrieval system through linguistic knowledge expression of database using FCM", Proceedings of KITE Summer Conference '95, 1995, pp.682-685.