FCM을 이용한 지적 데이터베이스 검색시스템

정 인[1], 황 승 욱[2]

An Intelligent Database Retrieval System using FCM

Ihn Jeong, Seung-Wook Hwang

ABSTRACT

By means of the advance in the digital technology, computer technology and optical communication technology, the high tech information age has arrived and the technology in hardware has made remarkable progress enough to deal with a large number of information, however, not enough in software. Therefore, it is essential to develope the technique to draw the best data out of a large number 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 draw the alternative data, but it needs the heuristic knowledge according to the target of retrieval. The other system uses the fuzzy membership functions 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

1) 한국해양대학교 대학원 제어계측공학과 석사과정 제어계측공학전공

2) 한국해양대학교 이공대학 제어계측공학과 조교수

membership functions. 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.

In this paper, we propose a retrieval system using knowledge of database expressed linguistically, where the relation between data are constructed by FCM algorithm. In addition, this paper proposes the improved method of adding new cluster and the cooperative response method 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.

제 1 장 서 론

디지탈기술과 컴퓨터기술 및 광통신기술의 비약적인 발전에 힘입어 고도의 정보화 사회가 형성되고 있으며, 폭발적으로 증가하는 정보량을 효율적으로 취급하기 위한 대용량 정보의 데이터베이스(Database)화 및 전달 기술 그리고, 데이터베이스의 검색 및 관리 기술 등의 많은 정보 관련 기술의 개발이 시급한 실정이다.

기존의 데이터베이스 검색시스템은 사용자의 검색조건을 정확히 만족하는 데이터만 사용자에게 제공하고, 데이터베이스에 사용자의 검색조건을 정확히 만족하는 데이터가 없을 경우에는 적절한 데이터를 제공하지 못한다. 이러한 문제점을 해결하기 위해서, 사용자의 검색조건을 완화하여 대체 데이터를 검색하는 시스템이 제안되어 있으나[1], 검색조건의 완화를 위하여 검색대상에 따른 휴리스틱한 지식이 필요하다는 문제점을 갖고 있다. 또한, 검색조건을 멤버쉽함수로 표현하여, 멤버쉽함수의 그레이드가 높은 데이터를 검색결과로 제시하는 퍼지시스템이 있으나[2], 멤버쉽함수의 선정에 따라서는 검색되지 않는 조건도 발생한다는 문제점이 있다. 이러한 문제점을 해결하기 위해서, 퍼지클러스터링 수법을 이용하여, 데이터베이스의 데이터를 복수의 클러스터로 표현하고, 미리 정의된 언어적 레이블(Linguistic label)을 이용하여, 데이터의 분포 상태를 언어적으로 표현하여, 사용자에게 제시하는 방법이 제안되어 있다[3]. 그러나, 정량적인 속성과 정성적인 속성을 동시에 가지는 데이터에는 적용할 수 없는 문제점이 있다. 본 논문에서는 데이터의 정성적인 속성을 정량화하여 데이터 간의 관계지식을 언어적으로 표현 가능하

고, 검색자의 정량적인 검색요구 및 정성적인 검색요구에도 협조적인 응답이 가능한 데이터베이스 검색시스템을 구축하고자 한다[4][5].

본 논문에서는 검색시스템의 구축을 위해서 기존의 데이터들의 관계를 추출하여 지식화하는 FCM(Fuzzy C Means)법 및 언어적인 레이블을 클러스터에 할당하는 방법을 이용한다. 기존의 퍼지클러스터링의 클러스터 추가 알고리즘의 개선 및 사용자와의 대체응답 알고리즘을 구현하였다. 그리고, 본 알고리즘을 우체국에서 실시하고 있는 우편주문책자의 선물고르기에 적용하여, 그 효용성을 확인하였다.

제 2 장 언어적인 지적 데이터베이스 검색시스템 구축

기존의 데이터베이스 검색시스템에서는 사용자의 요구에 정확히 일치하는 데이터가 데이터베이스 내에 존재할 경우에는 해당 데이터를 제공하나, 그렇지 않은 경우에는 적절한 응답이 불가능하였다. 이러한 문제점을 해결하기 위해 본 논문에서는 Fig. 2.1.과 같이 FCM법과 언어적인 지식 표현을 통해 언어적인 지식 데이터베이스(KDB; Knowledge DataBase)를 구축하여, 사용자의 요구에 일치하는 데이터가 있을 경우에는 해당 데이터를, 그렇지 않은 경우, 대체 응답을 제공한다. 그리고, 사용자와 시스템의 DB 및 KDB 간의 중계역할을 수행하는 협조적인 응답시스템을 구축한다.

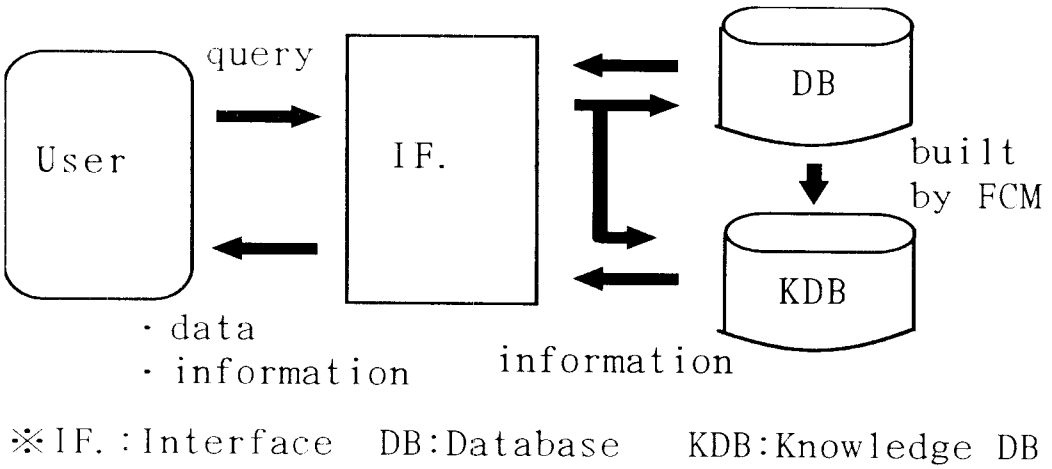

Fig. 2.1. Framework for building Intelligent Retrieval System

2.1 퍼지클러스터링

2.1.1 FCM법

Bezdek이 제안한 FCM법은 어떤 개체 X_k가 오직 한 클러스터에만 속한다고 보는 HCM(Hard C-Means)법에 복수 개의 클러스터에 서로 다른 그레이드로 속한다는 퍼지이론의 특성을 포함시킨 클러스터링 방법이다[6][7]. n개의 t차원의 데이터벡터 $X_k = x_{k,p}$ $(p = 1,2,\cdots,t)$ $(k = 1,2,\cdots,n)$를 c개의 클러스터로 분류할 때, 각 클러스터의 중심벡터 $V_i(i=1,2,\cdots,c)$와 데이터 X_k와의 비유사도 $d_{i,k}$를

$$d_{i,k} = \| X_k - V_i \| \tag{2.1}$$

와 같이 유크리드 거리로 표현한다. 이때, 중심벡터 V_i는

$$V_i = \sum_{k=1}^{n}(U_{i,k})^m X_{kj} \Big/ \sum_{k=1}^{n}(U_{i,k})^m \tag{2.2}$$

와 같이 표현한다. 여기서, 식(2.3)의 $U_{i,k}$는 X_k가 클러스터 i에 속하는 그레이드를 나타내고, V_i는 X_k의 멤버쉽그레이드의 m차원 가중평균이다.

$$U_{i,k}{}^{(l+1)} = 1 \Big/ \sum_{j=1}^{c}(d_{i,k}/d_{j,k})^{1/(m-1)} \tag{2.3}$$

FCM법의 알고리즘은 기본적으로는 통상의 C-means법의 U와 V를 갱신하기 위한 루틴을 추가한 것으로 다음과 같은 순서로 실행한다.

step 1: 클러스터 갯수 c $(2 \le c < n)$, 가중치 m $(1 < m < \infty)$, 수렴판정치 ε (threshold), c개의 분할행렬인 $U^{(l)}$의 초기값 $U^{(l)}{}_{l=0}$을 적당히 정한다.

step 2: 클러스터의 중심벡터 $V_i{}^{(l)}$ $(i=1,2,\cdots,c)$를 식(2.2)에 의해 $U^{(l)}$을 이용하여 구한다.

step 3: $X_k \neq V_i{}^{(l)}$일때, 식(2.3)에 의해 $U_{i,k}^{(l+1)}$로 갱신한다. $X_k = V_i{}^{(l)}$일때는 식(2.4)를 이용하여 갱신한다.

$$U_{i,k}{}^{(l+1)} = \begin{Bmatrix} 1 & i \in I_k \\ 0 & i \notin I_k \end{Bmatrix} \tag{2.4}$$

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

$\forall \; k = 1 \sim n$

step 4: 식(2.5)과 같이 $U^{(l)}$와 $U^{(l+1)}$와의 차가 주어진 수렴판정치 ε 보다 작거나 같으면 종료하고, 그렇지 않으면, step 2로 돌아간다.

$$\| U^{(l+1)} - U^{(l)} \| \le \varepsilon \tag{2.5}$$

2.1.2 클러스터의 증가 및 재초기화 알고리즘

최적의 클러스터 수의 결정은 식(2.6)에 의해 구한 $S(c)$ 를 최소로 하는 클러스터 수 c로 하면 되나, 해석적인 c의 결정법은 아직 알려지지 않고 있다. 본 논문에서는 $S(c) \le S(c+1)$을 만족하면, c를 최적의 클러스터 수로 하는 수법[7]에 조건 $|S(c+1) - S(c)| \le M$ 을 추가하여 두 조건 중 하나만 만족하면 계산을 종료하고, 그렇지 않으면, 클러스터 수를 1개씩 증가시키는 방식을 제안한다.

$$S(c) = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{i,k})^m (\| X_k - V_i \|^2 - \| V_i - \bar{x} \|^2) \tag{2.6}$$

식(2.6)에서 n은 데이터의 수, X_k는 k번째 데이터, $\bar{x}$는 데이터의 평균, V_i는 i번째 클러스터의 중심벡터, $\| \cdot \|$는 노옴(Norm), $\mu_{i,k}$는 k번째 데이터의 i번째 클러스터에 속하는 그레이드, m은 가중치이다. 클러스터 증가시, $U_{i,k}$의 재초기화 알고리즘은 다음과 같다.

step 1: c개의 클러스터 별 중심벡터 V_i에서 해당 클러스터에 속하는 데이터 X_k까지의 거리를 구한다.

step 2: 구한 클러스터 중에서 최대의 거리를 갖는 데이터 X_l을 구한다.

step 3: $U_{c+1,k} = 1 \;\; (k = l)$로 두고, $U_{c+1,k} = 0 \;\; (k \neq l)$, $U_{i,l} = 0 \; (i = 1, \cdots, c)$을 할당하여, $U_{c+1,k}$의 초기치를 설정한다.

2.2 데이터의 언어적인 지식표현

FCM법을 통해 구해진 퍼지클러스터에 언어적인 레이블을 할당하여 언어적인 지식 데이터베이스를 구축한다.

2.2.1 언어적인 레이블

퍼지집합론은 경계가 애매한 특성을 이용하여 Fig. 2.2.과 같이 "Little", "Young", "Middle", "Old" 등의 언어적 레이블을 사용한다. 이러한 언어적 레이블을 이용하여 지식을 표현한다. 퍼지클러스터링을 통해서 구해진 클러스터의 데이터는 각 속성별로 수치적인 값으로 표현된다. 그러나, 제공되는 데이터가 대량인 경우 검색에 있어서 경제적으로 손실이 발생한다. 이러한 점을 개선하기 위해서, 비슷한 특성을 갖는 데이터는 미리 정의된 언어적 레이블을 통해 표현함으로써 보다 효율적인 검색이 가능하다. 언어적 레이블의 개수는 제한이 없으나, 시스템이 갖는 특성에 따라서 적절한 수의 레이블을 상호 간의 관계를 고려해서 선정한다. 다시말하면, Fig. 2.3.에서 데이터 x_k는 속성 P_j에 대해서 언어적 레이블 $L^2{}_{P_j}$과는 그레이드 0.2의 관련성이 있고, 언어적 레이블 $L^3{}_{P_j}$과는 0.7의 관련이 있다. 이와같이, 각각의 데이터는 복수개의 레이블과 관련이 있을 수 있으므로, 각각의 레이블($L^1{}_{P_j}$, $L^2{}_{P_j}$ ⋯, $L^s{}_{P_j}$)은 서로 간의 관계를 고려해서 나열한다.

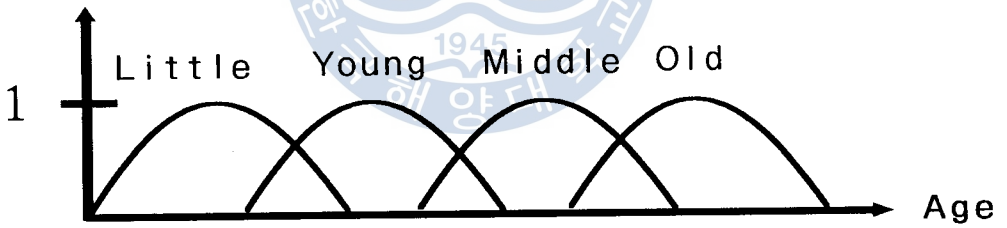

Fig. 2.2. Example of linguistic label in property ' Age'

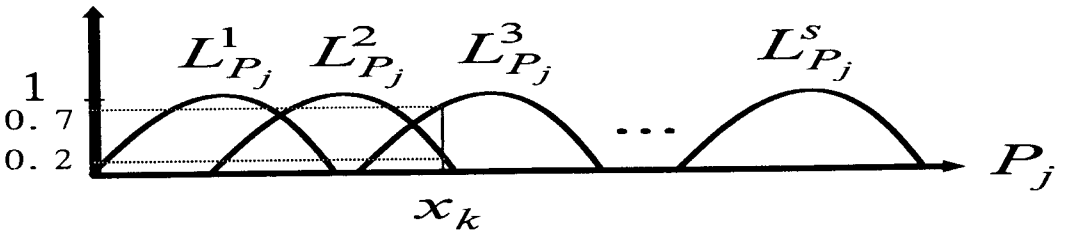

Fig. 2.3. Linguistic labels

2.2.2 언어적 레이블의 할당

결정된 각 퍼지 클러스터를 언어적으로 표현하기 위해, Fig. 2.3.과 같이 데이터 $x_{k,p}$의 j개의 속성별로 적당한 s개의 언어적 레이블 L^s_j을 할당한다[9]. $x_{k,p}$의 i번째 클러스터의 멤버쉽그레이드 $U_{i,k}$를 각 속성에 Fig. 2.4.와 같이 사상시켜, i번째 클러스터에 대한 속성별 멤버쉽함수 $\mu_{L^s_j}$를 구한다. 구해진 속성별 언어적 레이블의 멤버쉽함수 $\mu_{L^s_j}$와 $U_{i,k}$와의 적합도를 식(2.7)에서 구하여, C_s를 최소로 하는 언어적 레이블이 i번째 클러스터에 할당된다. 식(2.7)에서 클러스터에 속하는 그레이드가 미소한 데이터로 인해서 C_s가 증가함을 피하기 위하여 적절한 수렴판정치 α 이상의 멤버쉽그레이드를 갖는 데이터만을 대상으로 하여 적합도를 구한다.

$$C_s = \sum_{k=1}^{n} e_k \qquad (2.7)$$

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

$U_{i,k}$: k번째 데이터 x_k가 i번째 클러스터에 속하는 그레이드

L^s_j : j번째 속성에 대한 s번째 언어적 레이블

$\mu_{L^s_j}(x_{k,j})$: j번째 속성에 대한 k 번째 데이터 x_k의 s번째 언어적 레이블에 대한 멤버쉽함수

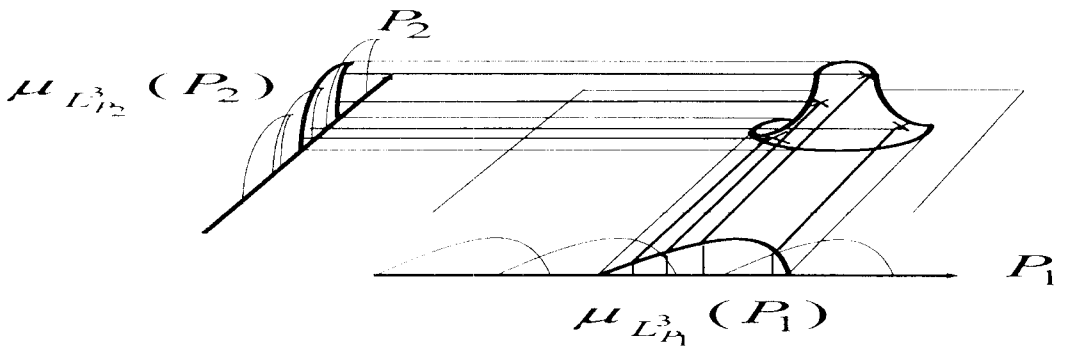

Fig. 2.4. Projection of linguistic label onto cluster

2.3 협조적인 응답시스템 구축

사용자의 요구와 일치하는 데이터가 데이터베이스 내에 있으면, 해당 데이터를 제공하고, 그렇지 않으면, 대체응답으로 사용자의 요구에 가장 근접한 데이터 및 정보를 데이터베이스 및 언어적인 지식 데이터베이스로부터 검색하여 제공할 수 있는 응답시스템이 필요하다. 이를 위한 대체 응답 알고리즘은 다음과 같다.

Step 1 : 사용자는 j번째 정량적 속성에 대해서는 수치 K_j를 입력하고, k번째 정성적인 속성에 대해서는 시스템이 제공한 언어적 레이블 중에서 l번째 레이블 L^l_k를 선택한다.

Step 2 : 정성적 속성에 대해 입력된 언어적 레이블 L^l_k의 중심값 avg_k를 계산한다.

Step 3 : 사용자의 입력 즉, 벡터 $V_{input}(K_j, avg_k)$ ($j=(1,\cdots,n), k=(1,\cdots,m)$)와 퍼지클러스터링을 통해 생성된 퍼지클러스터의 중심벡터 V_i와의 유클리드 거리를 구한다.

Step 4 : 최소의 거리를 갖는 퍼지 클러스터의 언어적 레이블을 출력하고, 해당 클러스터의 데이터를 출력한다.

제 3 장 선물고르기에의 응용 예

3.1 선물고르기 데이터베이스 구축

선물을 고를 때 보편적으로 고려하는 4가지 항목인 연령, 대상, 용도, 가격을 Table 1.와 같이 시스템의 속성으로 지정하고, Table 2.와 같이 105개의 상품을 데이터로 선정한다. 선택된 시스템의 4개의 속성은 Fig. 3.1.에서와 같이 각각의 언어적 레이블의 범위로 정의한다. Fig. 3.1.에서 속성 '연령', '대상', '용도'에 대해서는 '0~70'의 범위에서, 속성 '가격'에 대해서는 '0~100'까지의 범위에서 값을 할당받는다. 각 데이터의 입력된 값은 Table 2.와 같다. 각 속성별 각 언어적 레이블의 범위와 순서는 언어적 레이블 상호간의 관계와 데이터의 속성별 복수 개의 언어적 레이블에 대한 관련성을 고려하여 결정한다. 정량적 속성의 경우는 수치와 언어적 레이블의 관계로 결정하고, 정성적 속성의 경우는 경험적 지식을 이용하여 결정한다.

Table 1. Linguistic labels of 4 properties

속성	언어적 레이블
연령	Little, Young, Middle, Old
대상	Child, Lover, Friend, Colleague, Senior
용도	Entrance, Birthday, Ceremony, Wedding, Thank
가격	Man_1_2, Man_3_4, Man_5, Man_7_8, Man_10

* Man_1_2 : 1~2만원 정도 Man_3_4 : 3~4만원 정도

Man_5 : 5만원 정도 Man_7_8 : 7~8만원 정도

Man_10 : 10만원 정도

Table 2. List of data value

P / ND	Value on each Property				Data
	Age	Object	Use	Price(×1000)	
1	35	53	41	15	사과
2	60	53	42	20	곶감
3	51	60	51	15	호두
4	52	61	60	20	땅콩
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
102	37	37	43	12	
103	36	34	42	39	전기장판 · 요
104	51	70	41	55	은수저세트
105	52	70	42	100	돗자리

* ND : Number of data, P : Property .

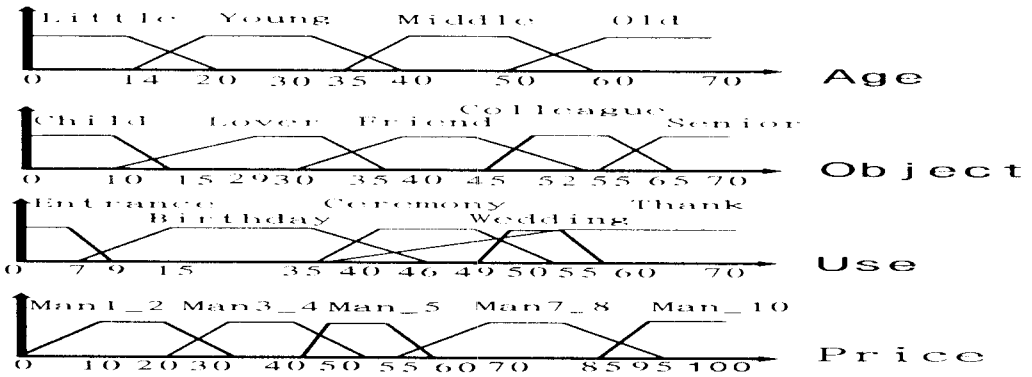

Fig. 3.1. Expression of each property's labels

각각의 입력 값을 할당받은 데이터들을 FCM법을 사용하여 퍼지클러스터를 구함으로써, 데이터들을 데이터베이스화하여 데이터들 간의 관련성을 정의하고, 사용자와의 상호응답시스템에 의해 사용자에게 효율적인 결과를 제공한다.

각각의 퍼지클러스터는 식(2.7)을 이용하여 Fig. 2.4.와 같이 언어적 레이블을 할당받아 Table 3.과 같이 언어적인 지식 데이터베이스를 구한다. Table 3.에서 두 번째 클러스터는 속성 '연령'에 대해서 '노년', 속성 '대상'에 대해서 '친구', 속성 '용도'에 대해서 '생일', 속성 '가격'에 대해서 10만원 그레이드의 선물 40개를 보유하고 있음을 나타낸다.

Table 3. Macro expression of Database

NC	속 성				ND
	연령	대상	용도	가격	
1	Young	Friend	Birthday	Man_10	39
2	Old	Friend	Birthday	Man_10	40
3	Young	Friend	Birthday	Man_7_8	14
4	Middle	Lover	Thank	Man_3_4	13

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

3.2 협조적인 응답의 예

퍼지클러스터링과 언어적 레이블의 할당을 통해서 구축된 데이터베이스 및 언어적인 지식 데이터베이스를 검색해서 사용자의 입력과 일치하는 데이터가 있으면, Fig. 3.2.와 같이 해당 데이터를 출력하고, 그렇지 않으면, Fig. 3.3.과 같이 사용자의 입력에 가장 근접한 데이터를 대체응답으로 제공한다. 즉, Fig. 3.2.에서 사용자의 입력은 '25세 정도의 친구에게 생일의 용도로 75,000원 그레이드의 선물'을 하고자 할 때, 시스템은 데이터베이스를 검색해서, 입력에 대응되는 클러스터를 출력하고 사용자가 원할 경우 해당 클러스터의 13개의 데이터를 출력하게 됨을 알 수 있다. 이와는 달리, Fig. 3.3.에서는 사용자가 '40세 정도의 동료에게 감사용으로 50,000원 그레이드의 선물'을 입력할 때, 시스템은 입력을 정확히 만족하는 데이터는 없지만, 그에 가까운 대체응답을 구하여 출력한다.

System : Input for Property **'Age'**. User : *25*
System : Select Input for Property **'Object'**. User : *3*
(1)Child (2)Lover (3)Friend (4)Colleague (5)Senior
System : Select Input for Property **'Use'**. User : *2*
(1)Entrance (2)Birthday (3)Ceremony (4)Wedding (5)Thank
System : Input for Property **'Price'**. User : *75000*
System: There is a cluster which is the same as your query. There are 14 data in the cluster. User : *1*
Type '1' & return if you want to see data.

System : Goods in the cluster are as following.

No. 12 더덕	No. 18 난	No. 19 육포
No. 22 메주	No. 28 수저세트	No. 42 해산물세트
No. 47 랜턴	No. 57 전기요	No. 63 홍삼
No. 75 에어매트	No. 78 대금	No. 89 둥근찻상
No. 95 다기세트	No. 98 호박엿	

Fig. 3.2. User's input and system's output(Case Ⅰ)

System : Input for Property **'Age'**. User : *40*
System : Select Input for Property **'Object'**. User : *4*
(1)Child (2)Lover (3)Friend (4)Colleague (5)Senior
System : Select Input for Property **'Use'**. User : *5*
(1)Entrance (2)Birthday (3)Ceremony (4)Wedding (5)Thank
System : Input for Property **'Price'**. User : *50000*
System : There is no cluster which is the same as your query. But there are 14 data near to your query. User : *1*
Type '1' & return if you want to see data.
System : Goods in the cluster are as following.

No. 14 알로에	No. 15 민속주	No. 16 한과
No. 17 꿀	No. 23 곶감	No. 39 젓갈
No. 42 건어물	No. 46 명주	No. 47 랜턴
No. 57 전기매트	No. 58 돌그릇	No. 63 인삼
No. 79 죽예품	No. 101인삼즙	

Fig. 3.3. User's input and system's output (Case Ⅱ)

제 4 장 결 론

본 논문에서는 기존의 퍼지클러스터링의 클러스터 추가 방법의 개선 및 사용자의 입력에 따른 협조적인 응답시스템을 구축하였다. 그리하여, FCM법과 퍼지클러스터에 대한 언어적 레이블의 할당을 통해 언어적인 지식 데이터베이스를 구축하여 사용자의 입력에 정확히 만족하는 데이터가 없을 경우에도, 그 입력에 가장 근접한 데이터 및 정보를 제시하는 지적 데이터베이스 검색시스템을 구축하였다. 그리고, 본 알고리즘을 우체국에서 실시하고 있는 우편주문책자의 선물고르기에 적용하여, 그 효용성을 확인하였다.

앞으로 보다 실용적인 시스템의 구축을 위해서 본 시스템의 특성상 발생하는 데이터의 편중성 등의 시스템 관련 문제점의 해결과 클러스터링을 위한 정성적인 속성의 정량화 수법 및 시스템과 사용자 간의 보다 원활한 인터페이스 방법 그리고, 사물에 대한 인간의 개념적 사고와 같은 개념적 클러스터링 등의 연구가 필요하다.

참 고 문 헌

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

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

[3] 정 인, 박계각, 황승욱, "FCM을 이용한 데이터베이스의 언어적인 지식표현 을 통한 검색시스템," *'95 대한전자공학회 하계학술대회 논문집*, pp.682-685, 1995.

[4] I. Jeong, G.K. Park and S.W. Hwang, "Intelligent Retrieval System using FCM," *Proc. of Korea Fuzzy Logic and Intelligent Systems Society Fall Conference '95*, Vol. 5, No. 2, pp. 40-44, 1995.

[5] Jun Ozawa and Koichi 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, pp.560-564,

[6] J. C. Bezdek, *Pattern Recognition wirh Fuzzy Objective Function Algorithm*, Plenum Press, New York, 1981.

[7] 坂和正敏, *ファジィ理論の基礎と應用*, 森北出版, 1990.

[8] Jun Ozawa and Koichi Yamada, "Cooperative Answering with macro expression of a database," *The 10th Fuzzy System Symposium*, pp.101 104, 1994.

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

[10] *우편주문책자*, 우체국, 1994.