FCM을 이용한 지식기반 데이터베이스 검색 시스템의 구축

Building of Database Retrieval System based on Knowledge using FCM

서기열* 박계각* 천대일* 양원재**

*목포해양대학교 해상운송시스템학부 해사정보전산학 전공

**한국해양대학교 해사수송과학부 해사정보공학 전공

Ki-Yeol Seo* Gyei-Kark Park* Dea-Il Cheon* Won-Jea Yang**

*Dept. of Maritime Information & Computer Science, Mokpo National Maritime Univ.

**Dept. of Maritime Transportation Science, Korea Maritime Univ.

E-mail : g000011@mail.mmu.ac.kr

ABSTRACT

Conventional database retrieval system have problems of being able to select data out of database only if the data exactly equal to retrieval conditions offered by users. If there are no data in database which exactly equal to user's retrieval conditionals, the system can not provide adequate data. To solve these problems, cluster increase of FCM and re-initialization of algorithm were suggested in this study.

And by interlocking knowledge-based database, built with FCM, to image database, new retrieval system was built to provide the data which are most appropriate for the requirement of users. We applied this new retrieval system to gift selection database system in pamphlet of mail order, and confirmed its effectiveness.

Ⅰ. 서 론

기존의 데이터베이스 검색시스템은 사용자의 검색조건에 정확히 일치하는 데이터가 데이터베이스 내에 존재할 경우에만 사용자에게 해당 데이터를 제공할 수 있고, 사용자의 검색조건을 정확히 만족하는 데이터가 없을 경우에는 적절한 데이터를 제공 할 수 없는 문제점이 있다. 이러한 문제점을 해결하기 위하여 퍼지클러스터링(FCM:Fuzzy C-means)을 이용하여 DB 내의 데이터를 복수의 클러스터로 표현하고, 미리 정의된 언어레이블을 이용함으로써, 데이터의 분포상태를 언어적으로 표현하여 사용자에게 제시하는 검색방법의 연구사례가 있다[2-4]. 그러나 데이터를 분류하는 FCM 알고리즘 적용 시 클러스터의 증가 및 재 초기화 알고리즘과 이미지데이터를 취급하는 기법이 제안되어 있지 않다는 문제점이 있다. 따라서, 본 논문에서는 FCM의 클러스터증가 및 재 초기화 알고리즘을 제안하고, FCM을 이용하여 DB 내의 데이터로부터 구축된 지식기반 데이터베이스(KDB)와 이미지 데이터베이스(Image DB)와의 연동을 통하여, 사용자의 요구에 가장 근접한 데이터를 제시해주는 검색시스템을 제안하고자 한다[5].

본 연구에서 제안된 수법을 우체국 우편주문책자를 이용한 선물고르기 DB 시스템에 적용하여 그 유효성을 확인하고자 한다.

Ⅱ. 지식기반 데이터베이스 구축

DB내 데이터간의 패턴, 유사성 등의 관심있는 지식을 찾아내는 데이터 마이닝(Data Mining)의 수법으로는 통계분석, 클러스터링, 신경망, 규칙추론 및 의사결정 트리 등이 이용되고 있으나[6], 본 논문에서는 매크로한 지식추출의 유효성이 높은 FCM을 이용하여 지식기반 DB(Knowledge-based Database)를 구축하고자 한다.

2.1 FCM법

FCM법은 어떤 개체 X_k가 오직 한 클러스터에만 속한다고 보는 HCM(Hard C-Means)법에 퍼지이론

의 특성을 포함시켜, 복수개의 클러스터에 서로 다른 정도로 속한다고 정의하는 클러스터링 방법이다[7]. n개의 t차원의 데이터 벡터 $X_k = x_{kp}$ (여기에서 $p = 1, 2, \cdots, t$ $k = 1, 2, \cdots, n$)를 c개의 클러스터로 분류할 때, 각 클러스터의 중심벡터 V_i (여기에서 $i = 1, 2, \cdots, c$)와 데이터 X_k와의 비유사도 d_{ik}를 식(2.1)과 같이 유크리트 거리로 표현한다.

$$d_{ik} = \| X_k - V_i \| \quad (2.1)$$

이때, 중심벡터 V_i는 식(2.2)와 같이 표현한다.

$$V_i = \sum_{k=1}^{n} (U_{ik})^m X_{kj} \; / \; \sum_{k=1}^{n} (U_{ik})^m \quad (2.2)$$

$$U_{ik}^{(l+1)} = 1 \; / \; \sum_{j=1}^{c} (d_{ik}/d_{jk})^{1/(m-1)} \quad (2.3)$$

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

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

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

step 3 : $X_k \neq V_i^{(l)}$ 일 경우, 식(2.3)에 의해 $U_{ik}^{(l+1)}$로 갱신한다. $X_k = V_i^{(l)}$ 일 경우에는 식(2.4)를 이용하여 갱신한다.

$$U_{ik}^{(l+1)} = \left\{ \begin{array}{ll} 1 & i \in I_k \\ 0 & i \notin I_k \end{array} \right\} \quad (2.4)$$

단, $I_k = \{ i \mid 1 \leq i \leq c, d_{ik} = |X_k - V_i| = 0 \}$, $\forall k = 1 \sim n$

step 4 : 식(2.5)와 같이 $U^{(l)}$과 $U^{(l+1)}$의 차가 주어진 수렴 판정치 ε보다 작거나 같으면 종료하고, 그렇지 않으면 step2로 되돌아간다.

$$\| U^{(l+1)} - U^{(l)} \| \leq \varepsilon \quad (2.5)$$

2.2 클러스터 증가 및 재초기화 알고리즘

최적의 클러스터 수의 결정은 식(2.6)에 의해 구한 $S(c)$를 최소로 하는 클러스터 수 c로 하면 되지만 해석적인 c의 결정법은 아직 알려지지 않고 있다. 기존에는 $S(c) \leq S(c+1)$을 만족하면, c를 최적의 클러스터 수를 결정하는 방법이 사용되었으나, $S(c)$ 값의 미묘한 변화로 인한 클러스터 수의 증가로 클러스터링에 불합리한 점이 발생한다.

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

본 논문에서는 $S(c)$값의 차이가 임계값 M이하일 경우 즉, 조건 $| S(c+1) - S(c) | \leq M$을 추가하여 두 조건 중 하나만 만족하면 해당 c를 최적의 클러스터 수로 결정하고, 그렇지 않으면 클러스터 수를 1개씩 증가시키는 방식을 제안하고자 한다.

2.3 언어레이블에 의한 클러스터의 표현

언어레이블은 데이터가 갖는 특성에 따라서 적절한 수의 레이블을 상호 간의 관계를 고려하여 선정한다. 그림 2.1에서 데이터 x_k는 속성 P_j에 대해서 언어 레이블 $L^2{}_{P_j}$과는 0.2의 관련성이 있고, 언어 레이블 $L^3{}_{P_j}$과는 0.7의 관련성이 있다. 이와 같이 각각의 데이터는 복수개의 레이블과 관련이 있을 수 있으므로, 각각의 언어레이블 ($L^1{}_{P_j}$, $L^2{}_{P_j}$, $\cdots$, $L^s{}_{P_j}$)의 배치는 서로간의 관계를 고려하여 결정한다.

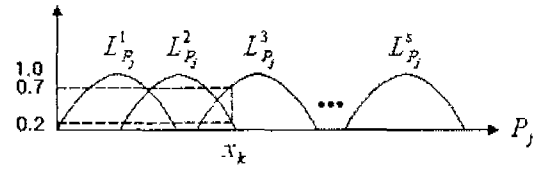

그림 2.1 언어레이블을 위한 멤버쉽 함수

결정된 각 퍼지 클러스터를 언어적으로 표현하기 위해 그림2.1과 같이 데이터 x_{kp}의 j개의 속성별로 적당한 s개의 언어적 레이블 $L^s p_j$을 할당한다. x_{kp}의 i번째 클러스터에 속하는 정도 U_{ik}를 각 속성에 사상시켜, i번째 클러스터에 대한 속성별 멤버쉽 함수를 구한다. 구해진 속성별 언어 레이블의 멤버쉽 함수와 U_{ik}와의 적합도를 식(2.7)에서 구하여, C_s를 최소로 하는 언어 레이블이 i번째 클러스터에 할당된다. 식(2.7)에서 클러스터에 속하는 정도가 미소한 데이터로 인하여 C_s가 증가함을 피하기 위해 적절한 임계치 α이상의 소속도를 갖는 데이터만을 대상으로 하여 적합도를 구한다.

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

Ⅲ. 대체응답 알고리즘

클러스터링을 통해서 구한 퍼지 클러스터에 언어적인 레이블을 할당하여 구축한 지식 데이터베이스로부터 사용자의 입력을 검색하여, 대응하는 데이터가 있으면 해당 데이터를 출력하고, 그렇지 않으면 대체응답을 제공하는 협조적인 응답시스템의 구축을 위한 알고리즘은 다음과 같다.

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

step 2 : 정성적 속성에 대해 입력된 언어 레이블 L^l_k의 중심값 avg_k를 계산한다.

step 3 : 사용자의 입력 즉, 벡터 $V_{input}(K_j, avg_k)$ $j=1,2,\cdots,n$ $k=1,2,\cdots,m$ 와 퍼지 클러스터링을 통해 생성된 퍼지 클러스터의 중심벡터 V_i 의 유크리트 거리(Euclidean Distance)를 구한다.

step 4 : 최소의 거리를 갖는 퍼지 클러스터의 언어 레이블을 출력하고, 해당 클러스터의 데이터를 출력한다.

Ⅳ. 데이터베이스 및 검색시스템 구축

데이터베이스의 효율적인 검색을 위해 일반 데이터베이스와 이미지 데이터베이스 그리고 지식기반 데이터베이스로 구분하여 DB 검색시스템을 구축하였다. 지식기반 검색시스템의 구성을 그림 4.1에 나타내었다. 그림 4.1에서 보여주는 것처럼 사용자의 요구에 일치하는 데이터가 데이터베이스 내에 존재할 때, 해당 데이터의 내용과 이미지를 제공하고, 만일 사용자의 요구에 일치하는 데이터가 존재하지 않을 때는 지식기반 데이터베이스(KDB)를 구축하여 사용자의 요구에 가장 근접한 데이터와 해당 데이터의 이미지 정보를 제시하도록 하였다. 일반 DB 검색과 FCM에 의해 구축된 지식기반 DB 검색이 가능하도록 하였으며, 화상 DB와의 연동이 가능하도록 설정하였다.

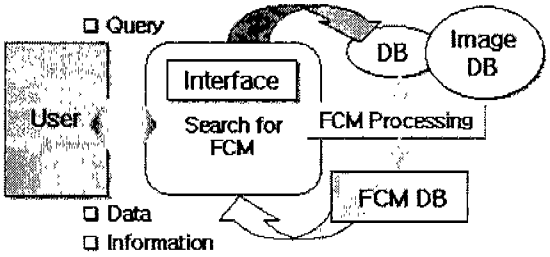

그림 4.1 검색시스템의 구성

4.1 데이터베이스 구축

데이터베이스 구축을 위해 우체국에서 발매하는 우편주문판매상품 안내서를 바탕으로 데이터베이스를 구축하였다[8]. 우편주문안내책자의 데이터를 M/S 액세스를 이용하여 작성하였고[9], 상품번호, 상품명, 지역, 상품내용, 가격, 공급처 등의 텍스트 데이터를 설정하여 총 312개의 데이터를 입력하였다.

4.2 이미지 데이터베이스 구축

데이터베이스에 이미지를 저장하는 방법으로는 이미지를 따로 파일에 저장하고 그 파일의 파일명만을 데이터베이스에 문자열로 저장하는 방법과 이미지 데이터 자체를 데이터베이스에 저장하는 방법이 있으나 이미지 파일을 따로 관리해야 하는 번거로움이 있기 때문에, 보다 정확하고 효율적인 이미지 검색을 위해 이미지파일을 해당 레코드에 삽입하는 방식을 취했다. 데이터베이스 내에 삽입을 위해 긴 이진 파일(Long Binary File)로 변환하여 해당 레코드에 삽입하여 데이터베이스를 구축하였다.

4.3 속성에 대한 언어레이블과 멤버쉽 함수

본 검색시스템에서는 정성적인 속성을 경험적 지식을 토대로 정성적 속성의 각각의 레이블과 데이터와의 관계를 분석하여 정량화 하였다. 각 속성에 대한 언어레이블은 표 4.1과 같이 설정하였고, 멤버쉽 함수는 그림 4.2와 같이 구분하였다.

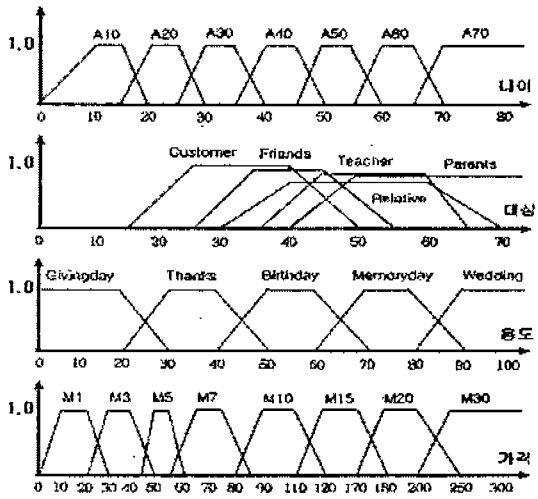

그림 4.2 언어레이블의 멤버쉽 함수

<표 4.1> 속성의 언어레이블 표현

Property	Linguistic Labels
Object	Customer, Friends, Teacher, Relatives, Parents
Age	A10, A20, A30, A40, A50, A60, A70
Use	Givingday,Thanks,Birthday,Memoryday, Wedding
Price	M1, M3, M5, M7, M10, M15, M20, M30

4.4 검색시스템의 활용 예

그림 4.3은 지식기반 검색을 위한 대화상자를 나타낸다. 즉, "*나이가 60세 정도의 부모님께 명절 선물로 20만원 정도의 선물을 하려고 한다*"면 그림 4.4와 같은 검색결과를 나타내준다.

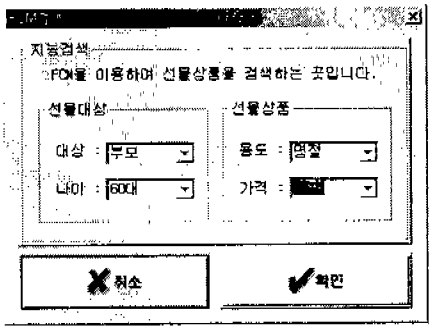

그림 4.3 지식기반 검색 질의 상자

그림 4.4는 사용자의 질의에 의해 검색된 상품을 출력하여 보여주는 예이다.

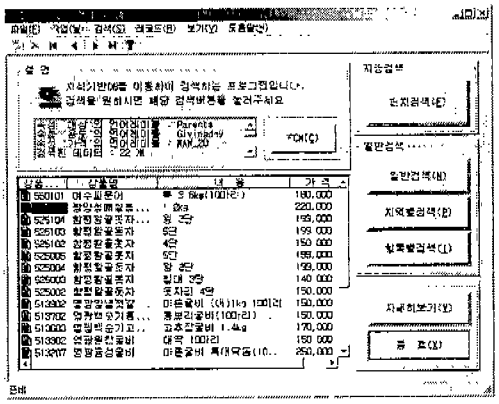

그림 4.4 검색된 데이터 출력 예

그림 4.5는 검색된 상품 중에서 하나의 상품을 자세히 보여주는 경우이며, 또한 선택된 상품의 정보와 이미지를 보여준다.

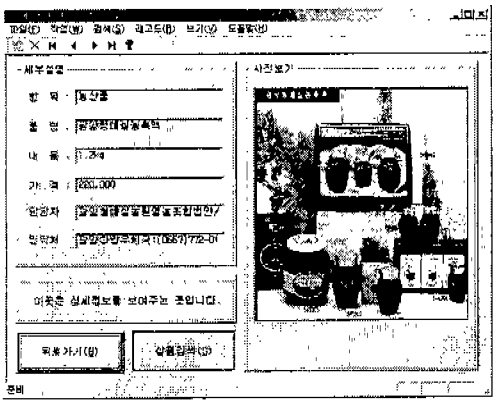

그림 4.5 데이터 정보와 이미지 출력

V. 결 론

본 논문에서는 FCM을 이용하여 지식 DB를 구축하고 이를 언어적으로 표현하여 사용자의 막연한 질문에도 대응할 수 있는 DB 검색시스템의 구축 수법을 제안하였다. 구체적인 연구내용으로는, FCM의 클러스터 증가 및 재 초기화 알고리즘을 제안하고, 이미지 데이터와 연동하는 수법을 제안하였다. 제안된 수법을 우체국 우편주문안내책자를 이용한 선물고르기 DB 시스템에 적용하여 그 실효성을 확인하였다.

앞으로 보다 실용적인 검색시스템의 구축을 위해서는 클러스터링을 위한 정성적인 속성의 효율적인 정량화 기법과 개념적인 속성을 갖는 데이터를 클러스터링 할 수 있는 개념 클러스터링에 관한 연구가 필요하다.

VI. 참고문헌

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

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

[3] 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, 1993.

[4] I. Jung, G. K. Park, 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] 박계각, 서기열, 임정빈, "지식기반 데이터베이스 검색 시스템의 구축", *한국해양정보통신학회 '99 추계종합학술대회지*, pp.450-453, 1999.

[6] 주혜종, 박상원, 이상필, "데이터베이스 총론", *정일*, pp.391-454, 1999.

[7] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithm", *Plenum Press*, New York, 1981.

[8] '98우편주문판매상품안내, *우체국*, 1998.

[9] 김성원, "한글 액세스97", *혜지원*, 1997.

[10] 김용성, "Visual C++ 6.0 완벽가이드", *영진출판사*, 1998.