지식기반 데이터베이스 검색 시스템의 구축

박계각* 서기열** 임정빈*

*목포해양대학교 해사정보전산학전공

**목포해양대학교 해양산업연구소

Building of Database Retrieval System based on Knowledge

Gyei-Kark Park* Ki-Yeol Seo** Jung-Bin Lim*

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

**Dept. of Marine Industrial Research Institute, Mokpo National Maritime Univ.

E-mail : vito@hanmir.com

요 약

본 논문에서는 사용자와 시스템의 DB, 이미지데이터 및 언어적인 지식표현을 통해 구축된 지식기반 데이터베이스(KDB)와의 인터페이스 역할을 위한 협조적인 검색시스템을 구축하였다. 기존의 데이터베이스 검색시스템은 사용자의 요구에 정확히 일치하는 데이터가 시스템 내에 존재할 경우에만 해당 데이터를 제공하지만 그렇지 않은 경우에는 아무런 데이터도 제공할 수 없다. 이러한 문제점을 해결하기 위해 사용자의 요구에 일치하는 데이터가 데이터베이스 내에 존재할 때 인터페이스를 통해 해당 데이터를 제공하고, 만일 사용자의 요구에 일치하는 데이터가 존재하지 않을 경우에는 퍼지 클러스터링과 언어 레이블의 할당을 통한 지식기반 데이터베이스를 구축하여 사용자의 요구에 가장 근접한 데이터 및 이미지정보를 제시하도록 하였다.

ABSTRACT

In this paper, the cooperative retrieval system to interface between users and DB, image data and knowledge-based database(KDB), being formed in a linguistic knowledge expression, of system is presented. Conventional database retrieval systems provide the data only in case that the data exactly corresponding with users' requirements exist in these systems, but don't in other cases. In order to resolve this problem, if the data users require are not in existence, this system shows the data and image information which are approximate with knowledge-based database materialized by fuzzy clustering and allocation of linguistic label.

Ⅰ. 서 론

최근에는 정보에 대한 인간 욕구의 증가와 디지털, 컴퓨터 기술 및 광통신 기술의 비약적인 발전에 힘입어 고도의 정보화 사회가 형성되고 있다. 폭발적으로 증가하는 정보량을 효율적으로 활용하기 위한 데이터베이스 구축과 전달 기술, 그리고 데이터베이스 검색 및 관리 기술 등 많은 정보 관련 기술의 개발이 시급하다. 그 중에서 대량의 데이터로부터 원하는 데이터를 검색하는 기술의 개발이 요구되고 있다. 기존의 데이터베이스 검색시스템은 사용자의 검색 조건에 정확히 일치하는 데이터가 데이터베이스 내에 존재할 경우에만 사용자에게 제공하였고, 사용자의 검색조건을 정확히 만족하는 데이터가 없을 경우에는 적절한 데이터를 제공 할 수 없었다. 이러한 문제점을 해결하기 위해 사용자의 요구에 일치하는 데이터가 데이터베이스 내에 존재 할 때는 해당 데이터를 제공하고, 만일 사용자의 요구에 일치하는 데이터가 존재하지 않을 경우에는 지식기반 데이터베이스(Knowledge-base Database)를 구축하여 사용자의 요구에 가장 근접한 데이터와 이미지 데이터베이스(Image Database)내에서 검색된 해당 데이터의 이미지 정보를 보여주는 협조적인 지식기반 데이터베이스 검색시스템을 구축하고자 한다.

II. 지식기반 데이터베이스 구축

데이터들의 관계를 추출하여 지식화하기 위해 정성적인 속성을 가진 데이터를 클러스터링에 사용하기 위해 정성적인 속성을 경험적 지식을 토대로 각각의 레이블과 데이터와의 관계를 분석하여 정량화 한다.

2.1 FCM법

Bezdek이 제안한 FCM법은 어떤 개체 X_k가 오직 한 클러스터에만 속한다고 보는 HCM(Hard C-Means)법에 퍼지이론의 특성을 포함시켜, 복수개의 클러스터에 서로 다른 정도로 속한다고 정의하는 클러스터링 방법이다. 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 식(2.1)과 같이 유크리트 거리로 표현한다.

$$d_{i,k} = \| X_k - V_i \| \qquad (2.1)$$

이때, 중심벡터 V_i는 식(2.2)와 같이 표현한다.

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

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

여기서, $U_{i,k}$는 X_k가 클러스터 i에 속하는 정도를 나타내고, V_i는 X_k의 멤버쉽 정도의 m차원 가중평균이다. FCM법의 알고리즘은 기본적으로는 통상의 C-Means법의 U와 V를 갱신하기 위한 루틴을 추가한 것이다.

2.2 클러스터 증가 및 재 초기화 알고리즘

최적의 클러스터 수의 결정은 식(2.4)에 의해 구한 $S(c)$를 최소로 하는 클러스터 수 c로 하면 되지만 해석적인 c의 결정법은 아직 알려지지 않고 있다. 기존에는 $S(c) \le S(c+1)$을 만족하면, c를 최적의 클러스터 수를 결정하는 방법이 사용되었으나, $S(c)$값의 미묘한 변화로 인한 클러스터 수의 증가로 클러스터링에 불합리한 점이 발생한다.

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

$$S(c) = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{i,k})^m (\| X_k - V_i \|^2 - \| V_i - \bar{x} \|^2) \qquad (2.4)$$

식(2.4)에서 n은 데이터의 수, X_k는 k번째 데이터, $\bar{x}$는 데이터의 평균, V_i는 i번째 클러스터의 중심벡터, $\| \cdot \|$는 노움(Norm), $\mu_{i,k}$는 k번째 데이터의 i번째 클러스터에 속하는 정도, m은 가중치이다. 클러스터 증가시, U의 c분할행렬 $U_{i,k}$의 초기 설정이 필요하다. 본 논문에서 사용된 $U_{i,k}$의 재초기화 알고리즘은 다음과 같다.

step 1 : c개의 각 클러스터의 중심벡터 V_i에서 해당 클러스터에 속하는 데이터 X_k까지의 거리를 구한다.

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

step 3 : $k = r$인 경우, $U_{c+1,k} = 1$로 두고, $k \neq r$인 경우, $U_{c+1,k} = 0$, $U_{i,r} = 0$ $i = 1, 2, \cdots, c$을 할당하여, $U_{c+1,k}$의 초기치를 설정한다.

2.3 언어레이블에 의한 클러스터의 표현

결정된 각 퍼지 클러스터를 언어적으로 표현하기 위해 그림2.1과 같이 데이터 $x_{k,p}$의 j개의 속성별로 적당한 s개의 언어적 레이블 $L^s p_j$을 할당한다. $x_{k,p}$의 i번째 클러스터에 속하는 정도 $U_{i,k}$를 각 속성에 사상시켜, i번째 클러스터에 대한 속성별 멤버쉽 함수를 구한다. 구해진 속성별 언어 레이블의 멤버쉽 함수와 $U_{i,k}$와의 적합도를 식(2.5)에서 구해, C_s를 최소로 하는 언어 레이블이 i번째 클러스터에 할당된다. 식(2.5)에서 클러스터에 속하는 정도가 미소한 데이터로 인하여 C_s가 증가함을 피하기 위해 적절한 역치 α이상의 소속도를 갖는 데이터만을 대상으로 하여 적합도를 구한다.

$$C_s = \sum_{k=1}^{n} e_k \qquad (2.5)$$

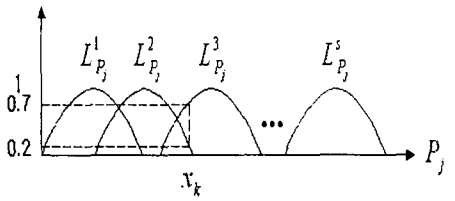

그림 2.1 언어레이블을 위한 멤버쉽 함수

III. 대체응답 알고리즘

클러스터링을 통해서 구한 퍼지 클러스터에 언어적인 레이블을 할당하여 구축한 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 언어적 레이블 중에서 l번째 레이블 L'_k를 선택한다.

step 2 : 정성적 속성에 대해 입력된 언어적 레이블 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_i의 유크리트 거리를 구한다.

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

IV. 검색시스템 구축

데이터베이스의 효율적인 검색을 위해 일반 데이터베이스와 이미지 데이터베이스 그리고 지식기반 데이터베이스로 구분하여 데이터베이스를 구축한다. 그림 4.1과 같이 사용자의 요구와 일치하는 데이터가 데이터베이스 내에 존재할 때, 인터페이스를 통해 해당 데이터 내용과 이미지를 제공하고, 관일 사용자의 요구에 일치하는 데이터가 존재하지 않을 때는 지식기반 데이터베이스(KDB)를 구축하여 사용자의 요구에 가장 근접한 데이터 및 이미지정보를 제시하도록 한다.

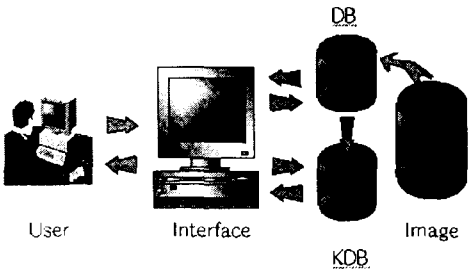

그림 4.1 지식기반 검색시스템의 개요

4.1 데이터베이스 구축

'98우편주문책자의 지역 특산물 데이터는 M/S 액세스를 이용하여 작성하였다. 상품번호, 상품명, 지역, 상품내용, 가격, 공급처 등의 텍스트 데이터를 설정하여 총 312개 정도의 특산물 데이터베이스를 구축하였다. 또한, 이미지 데이터의 설정을 위해 각 상품의 사진을 스캐닝하여 이미지 데이터베이스를 구축하였다.

4.2 속성에 대한 언어레이블과 멤버쉽 함수

데이터의 표현 방식은 정량적 속성과 정성적 속성으로 구분하여 설정하였다. 여기에서 '나이'나 '가격'같은 정량적 속성은 클러스터링에 사용하기 용이하나 정성적 속성은 클러스터링에 사용하기가 쉽지 않으므로 적절히 정량화 하여야 한다. 본 검색시스템에서는 정성적인 속성을 경험적 지식을 토대로 정성적 속성의 각각의 레이블과 데이터와의 관계를 분석하여 정량화 하였다. 각 속성에 대한 언어레이블과 멤버쉽 함수는 그림4.2와 같이 구분하였다.

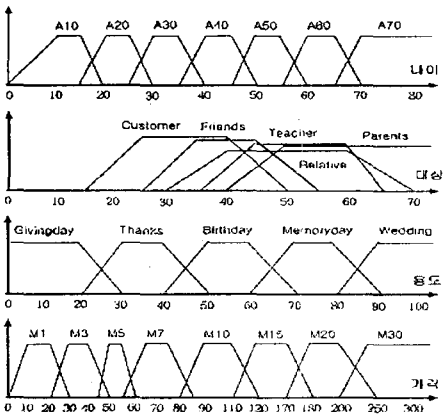

그림 4.2 속성에 대한 언어레이블과 멤버쉽 함수

4.3 검색시스템 구축 예

그림4.3은 지식기반 검색을 위한 입력창을 나타낸다. 즉, 나이가 60세정도의 부모님께 명절선물로 20만원 정도의 선물을 하려고 한다면 그림4.4와 같은 검색결과를 보여준다.

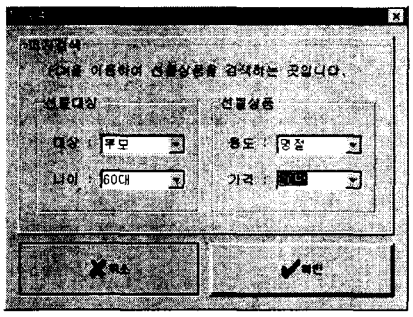

그림 4.3 지식기반 검색 입력창 예

그림4.4는 사용자의 요구에 검색된 상품을 출력하여 보여주는 예이다.

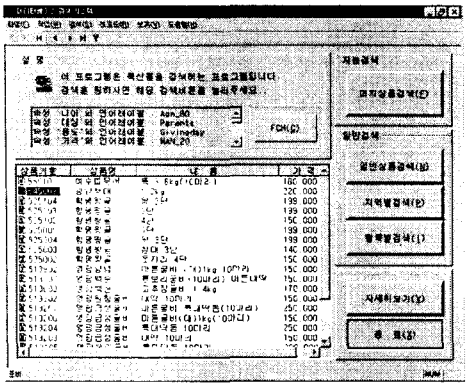

그림 4.4 검색된 상품출력 예

그림4.5는 검색된 상품 중에서 하나의 상품을 자세히 보여주는 경우이며, 또한 선택된 상품의 정보와 이미지를 보여준다.

그림 4.5 이미지 출력 예

V. 결 론

본 논문에서는 기존의 데이터베이스 검색시스템의 문제점인 사용자의 입력에 정확히 일치되는 데이터가 없을 때, 적절한 응답이 불가능한 점을 개선하였고, 또한 사용자와 시스템의 데이터베이스, 이미지데이터 및 언어적인 지식표현을 통해 구축된 지식기반 데이터베이스(KDB)를 검색할 수 있는 협조적인 검색시스템을 구현하였다.

앞으로 보다 실용적인 검색시스템의 구축을 위해서는 클러스터링을 위한 정성적인 속성의 정량화 기법과 데이터의 편중, 그리고 사물에 대한 인간의 개념적 사고와 같은 개념적 클러스터링의 연구가 필요하다.

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

[5] 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.

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

[7] Jun Ozzawa & Koich Yamada, "Cooperative Answering with macro expression of a database", *the 10th Fuzzy System Symposium*, pp.101-104, 1994.

[8] M. Sugeno & T. Yasukawa, "A Fuzzy-Logic-based Approach to Qualitative Modeling", *IEEE Trans. on Fuzzy systems*, Vol. 1, pp.7-31, 1993.

[9] '98우편주문판매상품안내, *우체국*, 1998.

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