퍼지 추론에 의한 자연언어 정보 검색

박 현 규[†]·오 종 훈[††]·김 명 호[†††]·최 기 선[††††]·이 광 형[††††]

요 약

인터넷 전자 상거래 시스템에서 주로 일어나는 정보 검색은 사용자의 상품정보 요구라고 할 수 있다. 이와 같이 사용자가 원하는 상품 정보를 웹 환경에서 검색하기 위해서는 편리한 검색 환경의 제공뿐만 아니라, 검색 성능의 효율성 또한 우수해야 한다. 인터넷 인구와 온라인 쇼핑몰의 급격한 증가로 인해 다양한 조건 검색에 의한 상품검색 요구가 증대되고 있다. 또한, 이러한 상품의 검색 결과는 사용자의 의도와 의미상으로 밀접한 관계를 가져야 한다. 자연언어 정보검색은 이러한 요구의 중요한 대안으로 대두되고 있으나, 자연언어 자체가 가지는 애매한 의미의 해석 등으로 인하여 상용 시스템에 적용하는데 많은 어려움이 있다. 본 논문에서는 이러한 문제점을 해결하기 위하여 퍼지추론을 이용한다. 입력된 자연언어 질의에서 형태소 분석을 통하여 데이터베이스 질의에 사용될 수 있는 의미어(content word)를 추출한 후, 의미어들을 재구성하여 템플릿을 작성한다. 작성된 템플릿은 퍼지 추론을 통하여 의미의 애매성을 해소하고 데이터베이스 질의로 변환하여 사용자의 질의 의도와 부합되는 검색 결과를 제시한다.

Natural Language Information Retrieval by Fuzzy Inference

Hyun K Park[†]·Jong Hoon Oh[††]·Myoung Ho Kim[†††]·
Key Sun Choi[††††]·Kwang Hyung Lee[††††]

ABSTRACT

The common information retrieval in the e-commerce system is the customer's requests for the merchandise information which is provided by the shopping mall. The rapid increase of online shopping mall and internet users requires more efficient conditional searching method about various goods. Moreover, it is necessary that the set of results should be very relevant to the exact intent of users. To offer relevant information to users, natural language support for information retrieval can be considered. However, the ambiguity of natural language makes difficult to apply to the commercial systems. In this paper, we propose a method for natural language query processing through fuzzy inference for the information retrieval to resolve the ambiguity of natural language. From analysed natural language queries by a morphological analyzer, a template is constructed. Then the template is transformed into a database query using fuzzy inference to offer relevant information to users.

키워드 : 자연언어 처리(Natural Language Processing), 퍼지 추론(Fuzzy Inference), 전자 상거래(E-Commerce), 웹 데이터베이스(Web Database)

1. 서 론

인터넷 쇼핑몰과 인터넷 사용자의 증가로 인해 사용자가 원하는 제품을 효과적으로 검색할 수 있는 검색 시스템의 요구가 증대되고 있다. 하지만, 기존의 상품 검색 시스템은 미리 정해진 속성에 대한 검색만을 지원하고 있어 사용자가 원하는 상품을 효율적으로 검색하는데는 그 한계가 있다. 예를 들어, 사용자가 "값이 싼 29인치 TV"라는 검색 요구가 있을 경우, 기존의 상품 검색 시스템은 "값", "제품 분류", "크기"라는 지정된 속성에 대하여 대화식(Interactive) 또는 일괄(Batch) 입력 방식으로 제품 검색을 지원한다. 하지만 사용자가 원하는 "값이 싸다"라는 특성은 사용자가 해당 조건에 맞는 상품을 검색한 후 사용자가 직접 이를 찾아내는 과정이 필요하다. 이로 인해, 검색 방법이 익숙하지 않은 사용자는 효율적인 검색을 하기 어렵고, 가격과 같은 수치 자료는 검색 범위와 매우 근사한 값을 가지고 있어도 검색 결과에서는 배제되는 등의 문제점이 있다.

이러한 문제점을 해결하기 위하여 자연언어 질의를 이용한 정보검색 시스템에 대한 연구가 있어 왔다[1, 4, 6, 10, 17, 19]. 이들 연구에서는 자연언어 질의를 문법 규칙에 기반하여 분석하고, 분석된 자연언어 질의를 데이터베이스 질의어로 변환하였다. 분석된 자연언어질의를 데이터베이스 질의로 변환하기 위해서 자연언어질의를 미리 정의된 중간단계

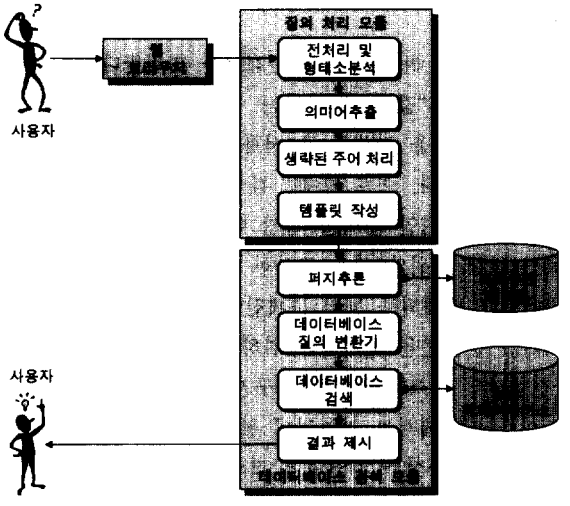
† 정 회 원 : 한국과학기술원 대학원 전산학과
†† 준 회 원 : 한국과학기술원 대학원 전산학과
††† 정 회 원 : 한국과학기술원 전산학과 교수
†††† 종신회원 : 한국과학기술원 전산학과 교수
논문접수 : 2001년 3월 15일, 심사완료 : 2001년 5월 31일

의 표현으로 변환한 뒤, 대상 데이터베이스의 질의로 변환한다. 하지만 이들 연구들은 주어진 질의와 데이터베이스에 저장되어 있는 정보와의 상호 불일치성에 대한 문제에 대한 해결보다는 주어진 질의에 대한 단순한 의미어(content word) 추출과, 자연어 질의를 데이터베이스 질의로 변환에 초점이 맞추어져 있다. 따라서 이들 연구에서는 자연언어에서 나타나는 의미를 데이터베이스에 저장되어 있는 의미로의 매핑이 되지 않아 사용자가 원하는 상품에 대한 정보를 찾지 못하는 경우가 발생한다는 점에서 그 한계가 있다.

본 논문에서는 이러한 자연언어가 가지는 의미의 모호성 문제를 해결하기 위하여 퍼지 추론에 의한 자연언어 정보 검색 방법을 제시한다. 본 논문의 구성은 다음과 같다. 제2장에서는 전체적인 시스템의 구성을 설명하고 제3장, 제4장, 제5장에서는 본 논문에서 제안한 방법을 자세히 기술한다. 제6장에서는 실험과 결과를 설명하고 제7장에서는 결론을 맺는다.

2. 시스템 구성

전체적인 시스템 구성은 (그림 1)과 같다. 시스템은 자연언어를 처리하여 템플릿을 만드는 질의 처리 부분과 질의 처리한 결과를 데이터베이스 질의로 변환하여 검색하는 데이터베이스 검색 부분으로 나누어진다. 사용자에 의하여 입력된 자연언어 질의는 질의 처리 부분으로 전달되고 형태소분석기를 통하여 분석된다. 그리고 형태소 분석기에 의해 분석된 질의에 나타나는 의미어를 추출하여 템플릿(template)을 구성한다. 이때, 주어가 생략된 경우 생략된 주어를 처리하는 부분을 통하여 주어를 복원하는 작업이 수행된다. 구성된 템플릿은 데이터베이스 검색 부분으로 전달되고 퍼지 추론을 통하여 템플릿 내에 존재하는 단어의 모호성을 해소한다. 그리고 이들 결과는 데이터베이스 질의어 변환기를 통하여 데이터베이스 질의어인 SQL문으로 변환된다. 이를 이용해 데이터베이스내에 존재하는 상품정보를 검색하고 해당 결과를 사용자에게 제시한다.



(그림 1) 시스템 구성

(그림 1)에서 사용된 웹 브라우저는 일반적인 Internet Explorer를 이용하였고, 데이터베이스 검색을 위한 서버로 Sun Ultra-1을 사용하였다. 사용된 웹 서버는 Netscape Server 3.0를 상품정보를 저장하는 데이터베이스에 대한 DBMS는 Informix Universal Server 9.12[3]를 사용하였다. 그리고 질의처리를 위한 서버는 펜티엄 Linux 서버를 사용하였다. 검색 시스템은 웹기반 시스템으로서, 기존의 CGI기반 방식을 사용하지 않고 Informix Web Datablade를 사용하여 CGI기반의 시스템보다 빠른 수행이 가능하도록 하였다.

현재는 Prototype 시스템이지만 Web Datablade는 표준 SQL을 사용하여 제품 데이터 및 규칙 베이스 등을 일관된 형태로 접근할 수 있으며, 모든 코드를 서버로 집중시켜 실제 전자 상거래 시스템으로 확장 시에도 모든 사용자들의 동시성(Concurrency)과 일관성(Consistency) 제어가 용이하도록 하였다.

3. 퍼지 규칙에 의한 질의 표현(Query Representation in Quantization Fuzzy Rules)

L.A. Zadeh[8, 9]에 의하여 제시된 퍼지 이론을 데이터베이스 분야의 정보 검색에 적용한 연구는 [2] 등이 있으며, 본 논문에서는 정보 검색을 위한 자연언어 질의 처리를 퍼지 집합으로 표현하여 질의어를 추론한다. 자연언어에 의한 질의 특징은 기존의 데이터베이스 질의와는 구별되는 부정확성을 포함한다. 이를 관계형 데이터베이스에서 다룰 수 있도록 데이터베이스의 속성은 퍼지 집합으로 표현하고, 퍼지 튜플 $t_1, \cdots, t_n$은 속성 값 c와 퍼지 영역에서 $0 < \mu \leq 1$의 값을 가지는 집합 $< c_1 : \mu_1, \cdots, c_n : \mu_n >$이 된다. 이를 위하여 퍼지 집합을 <정의 1>과 같이 정의한다.

정의 1.

전체 집합 U 에서 퍼지 집합 F 는 다음과 같은 소속 함수(Membership Function)에 의하여 정의된다.

$\mu_F : U \rightarrow [0, 1]$

이 때, $\mu_F(u)$, $u \in U$는 퍼지 집합 F에서의 소속 정도를 나타낸다.

■

즉 퍼지 관계 데이터베이스에서 퍼지 Relation $R = \{R_1, \cdots, R_n\}$은 퍼지 데이터베이스 스키마(Schema) $r = \{r_1, \cdots r_n\}$의 개체인 퍼지 튜플 $t_1, \cdots, t_n$ 집합이다.

검색은 다수의 퍼지 영역에 대하여 이루어지므로 각각 다른 소속함수의 값을 가지는 영역에서의 퍼지 관계 대수는 <정의 2>를 따른다.

정의 2. Fuzzy Relational Model

합집합 규칙(Conjunction Rule) : $\mu_{A\cup B}(u) = max(\mu_A(u), \mu_B(u))$

교집합 규칙(Disjunction rule) : $\mu_{A\cap B}(u) = min(\mu_A(u), \mu_B(u))$

부정 규칙(Negation rule) : $1 - \mu_A(u)$

■

따라서 자연언어에 의한 질의 표현은 '매우', '적당한'과 같이 다수의 영역(Domain)에 서술어를 포함하는 다음과 같은 형태가 된다.

$$\mu_t : \mathrm{dom}(A_1) \times \mathrm{dom}(A_2) \times \cdots \times \mathrm{dom}(A_n) \rightarrow [0,1] \quad (1)$$

그러므로 <정의 1>과 <정의 2>에 의하여 (1)과 같은 관계 대수 질의 결과는 <예제 1>과 같이 소속 함수에 의한 결과의 조합으로 변환하여 구할 수 있다.

예제 1.

<표 1>과 같은 관계(Relation)이 있다고 하면, 여기서 속성 집합 dom(item), dom(class)는 일반 집합(Crisp Set)이지만, dom(price), dom(weight), dom(size)는 전체 집합 U_{price}, U_{weight}, U_{size}의 각각의 퍼지 집합이 된다.

〈표 1〉 실험 데이터베이스의 ITEM Relation 예

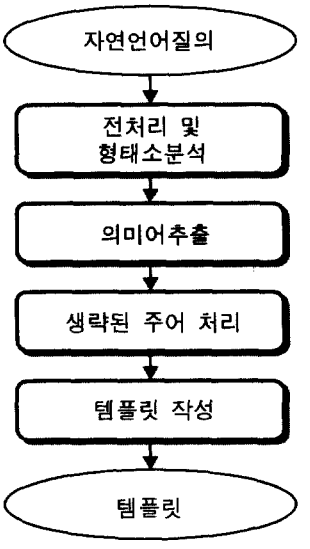
item	class	price	weight	size
CT15A8D	TV	258000	1500	15
D-FJ65	CDP	148000	200	32
MZ-R91	MDP	320000	153	20

■

4. 자연언어 해석 및 질의어 생성(Generation of Queries from Natural Language)

4.1 자연언어 질의 분석기의 구조

일반적으로 자연언어에 의한 데이터베이스 질의어의 경우에는 형태소 분석과 구문 분석, 그리고 의미 분석에 의해 자연언어 질의문을 분석하여 SQL문으로 변환한다[10, 14]. 하지만 전자거래 시스템의 상품검색에 대한 질의는 대부분 다른 질의문에 비하여 단순하며, 상품검색이라는 도메인이 한정되어 있기 때문에 사용되는 용어나 언어형태가 일정한 형식을 가지게 된다. 또한 문장의 형태도 복문보다는 단문이나 중문으로 이루어진 경우가 많다. 따라서 본 논문에서는 주어진 질의가 단문이나 중문이라고 가정하고 형태소 분석기만을 이용하여 상품검색 질의 템플릿을 구성한다. (그림 2)는 이러한 자연언어 질의 분석기의 구조를 나타낸다. 사용자에 의해 주어진 자연언어 질의는 전처리기를 통하여 단문으로 분할되고, 형태소 분석기를 거쳐 질의어에 나타난 의미어를 추출한다. 그리고 주어가 생략된 경우, 생략된 주어를 복원한다. 마지막으로 의미어를 재구성하여 템플릿을 작성한다.



(그림 2) 자연언어 질의 분석기

4.2 전처리기 및 형태소 분석기

전처리기는 주어진 질의를 단문의 형태로 변환하는 작업을 수행한다. 일반적으로 중문은 여러 단문이 대등적 연결 어미인 '고', '며', '거나' 등으로 연결되어 있거나, 대등 접속사로 연결되어 있는 문장을 지칭한다. 주어진 문장이 이러한 중문인 경우에는 정보를 추출하는 것이 용이하지 않기 때문에, 이를 단문으로 변환하여 주는 작업이 필요하다. 예를 들어, "값이 싸고 무게가 가벼운 CDP"라는 질의의 경우 "값이 싸다", "무게가 가볍다", "CDP"로 분할된다.

일반적으로 형태소 분석기라 함은 주어진 문장에 대하여 문장내의 형태소를 분석하여 가능한 형태소 후보들을 제시하고 이들 중에서 가장 올바르다고 생각되는 후보열을 제시하는 품사 태깅(Part-Of-Speech tagging) 도구를 총칭한다. 단문으로 재구성된 질의문에 대하여 통계 기반 품사 태깅 시스템은 학습된 통계치를 이용하여 품사태깅을 실시한다. 품사 태깅 방법에는 크게 규칙 기반 모델과 통계정보 기반 모델의 두 가지 방법이 있는데[12, 13, 16, 18], 본 논문에서 사용하는 품사 태깅은 통계기반 모델을 이용하여 수행한다.

통계기반 품사태깅 모델은 주로 은닉마르코프 모델(HMM : Hidden Markov Model)[7]에 기반하여 모델링된다. 주어진 문장 W에 대하여 식 (1)을 만족하는 품사열 T를 찾는 것

이 통계 기반 품사 태깅 방법의 기본적인 모델이다.

$$\phi(W) \equiv \text{argmax}_T \Pr(T|W) = \text{argmax}_T \frac{\Pr(T, W)}{\Pr(W)} = \text{argmax}_T \Pr(T, W) \quad (1)$$

식 (1)을 연쇄 규칙(chain rule)과 마르코프 가정을 이용하여 전개하면, 식 (2)를 얻을 수 있다.

$$\phi(W) = \text{argmax}_{t_{1,|W|}} \prod_{i=1,|W|} \Pr(t_i|t_{i-h,i-1}) \Pr(w_i|t_i) \quad (2)$$

여기서, |W|는 문장 |W|의 길이를 의미하고, $t_{1,|W|}$는 품사열 $t_1 t_2 \cdots t_{|W|}$를 나타내고, $t_{i-h,i-1}$은 품사열 $t_{i-h} t_{i-h+1} \cdots t_{i-1}$을 의미한다. 이와 같은 모델을 품사 태깅을 위한 은닉 마르코프 모델(Hidden Markov Model, HMM)이라고 한다. 여기서, $\Pr(W_i|t_i)$을 어휘확률(lexical probability) 혹은 어휘정보(lexical information)라고 하고, $\Pr(t_i|t_{i-h,i-1})$을 문맥확률(contextual probability) 혹은 문맥정보(contextual information)라고 한다. h는 문맥정보의 크기를 조절하는 상수이며, h = 1일 경우 1차 은닉 마르코프 모델이라고 하고, h = 2 일 경우를 2차 은닉 마르코프 모델이라고 한다[11]. 어휘 확률과 문맥확률은 대량의 품사 태깅된 문서로부터 학습하여 획득한다.

4.3 템플릿 생성

템플릿 생성기는 형태소 분석된 결과에서 술어구조를 분석하여 원하는 정보를 템플릿(Template) 형태로 구성하는 부분이다. 전처리에 의해 단문으로 재구성된 질의어에서 주어에 해당하는 부분이 데이터베이스의 속성인지를 검사하고 해당 단문의 술어는 해당 속성의 값을 결정하는 역할을 하게 된다. 주어와 해당 속성은 형태소 분석기를 이용하여 분석된 결과에서 주어는 명사와 주격조사로 파악되고 술어는 동사 및 형용사로 파악된다. 데이터베이스의 속성은 "가격", "무게", "크기", "두께", "배터리수명" 등이며 이들 주어는 주격조사와 같이 나타나는 명사를 파악함으로써 추출된다. 예를 들어, "가격이 싸다"의 경우, 형태소 분석기에 의해 "가격/명사 + 이/주격조사 싸/형용사 + 다/어미"로 분석되는데 주격조사와 같이 나타나는 명사인 "가격"은 속성으로 추출되고 해당 주어에 대한 형용사인 "싸다"는 속성인 "가격"에 대한 술어로 파악된다. 따라서 "가격 : 싸다"라는 템플릿을 구성할 수 있다. 하지만 술어가 어떠한 속성을 명확히 제한하는 경우에는 술어의 주어가 생략되는 경우가 많다. 예를 들어 "싼 CDP"의 경우 "싸다"라는 술어에 대한 "가격"이라는 주어(속성)가 생략되어 있다. 따라서 이러한 생략현상을 효과적으로 처리하기 위해서는 주어를 명확하게 지시하는 술어에 대한 처리가 가능하여야 한다. 즉, 어떠한 술어가 주어가 생략된 형태로 나왔을 경우에도 해당 템플릿을 구성할 수 있도록 처리할 수 있는 규칙을 구성하여야 한다. <표 2>는 주어가 생략되었을 때 처리할 수 있는 규칙을 기술하였다. 예를 들어 술어가 "얇다" 또는 "두껍다"일 경우에는 주어는 두께가 된다. 이러한 규칙을 이용하여 주어진 질의 "저렴하고, 두께가 얇으며, 무게가 가벼운 CDP"는 "가격 : 싸다", "두께 : 얇다", "무게 : 가볍다", "제품 : CDP"라는 템플릿을 구성한다.

<표 2> 주어가 생략된 경우 주어 처리 규칙

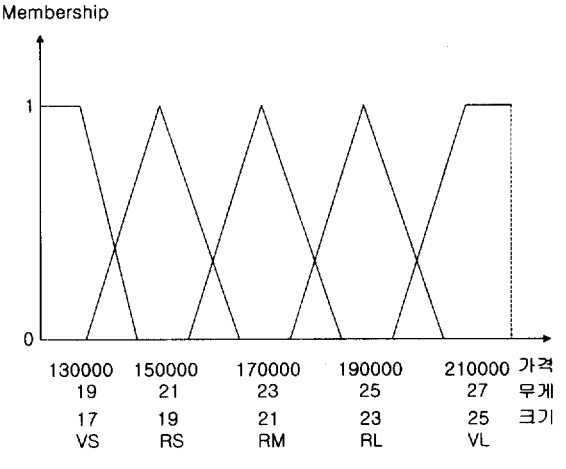
술 어	처리 결과
"고가이다", "저가이다", "저렴하다", "비싸다", "싸다"	주어는 가격
"소형이다", "대형이다", "크다", "작다"	주어는 크기
"얇다", "두껍다"	주어는 두께
"무겁다", "가볍다", "경량이다", "중량이다"	주어는 무게
"길다", "짧다"	주 어 배터리수명

5. 추론 엔진(Inference Engine) 및 규칙 베이스(Rule-Base)

자연언어 처리기에서 생성된 템플릿은 SQL로 변환하여, 추론 엔진에서는 규칙 베이스의 정보를 이용하여 질의 결과를 제시한다.

퍼지 집합인 질의문의 술어를 5단계로 구분하는 것은 실제 제품 정보의 가격, 무게, 크기와 같은 대부분 데이터 속성이 5단계 구분시 각 단계의 경계부분이 적절히 겹치게 되며 5단계 이상 구분이 별다른 의미를 갖지 못하기 때문이다. 따라서 사용자의 의미 구분은 인간 공학적 구별 단계로 제시되는 7±2 단계의 구분이 적절하므로[15], 5단계로서 충분한 의미 반영이 가능하다.

다양한 제품으로 구성된 쇼핑몰에서는 자연언어 처리기에서 추출된 템플릿에 포함되어 있는 서술어로부터 해당되는 규칙 베이스의 정보를 획득하여 퍼지 추론을 하게 된다.



(그림 2) 퍼지 소속 함수(Membership Function)

퍼지 추론을 위한 규칙 베이스는 제품 종류와 서술어의 주어에 해당하는 영역별로 t-norm을 따르는 해당 값의 범위를 설정하며, 이를 기반으로 퍼지 추론을 관계 연산을 통하여 이루어진다. (그림 2)는 퍼지 추론을 위하여 샘플 데이터베이스에서 제품에 대한 가격, 무게, 크기 등의 정규화된 속성(Attribute) 값의 예이다.

5.1 규칙 베이스 (Linguistic Rule Base)

규칙 베이스의 정보는 제품의 종류, 각각의 속성에 따라 질의문의 서술어가 가지는 의미를 퍼지 소속 함수로 표현한 것이다. 규칙 베이스의 정보 검색은 각 제품별, 질의 영역별로 정규화된 형태의 레코드를 검색하므로 서술어가 속하는 범위를 보통 집합(Crisp Value)들의 관계 연산에 의한 정보 검색 결과가 해당 규칙이 된다.

사용되는 규칙은 서술어의 Certainty를 가장 보편적으로 신뢰하는 95%의 구간에서 데이터 값이 가지는 범위에 대한 평균 x로부터 $(x-1.96\cdot\frac{\sigma}{\sqrt{n}}, x+1.96\cdot\frac{\sigma}{\sqrt{n}})$의 범위를 갖는 값을 적용하였다. 이를 적용하여 "price=cheap"과 같은 질의 등에 대한 값의 범위를 다음과 같은 관계대수 연산을 실시하고, 범위에 해당하는 튜플 집합이로 퍼지 추론의 규칙으로 제시된다.

$$\sigma_{c(A_1,\cdots,A_j)}(R)=\{\langle c_1:\mu_1,\cdots,c_n:\mu_n\rangle | c_{\mu_1,\cdots,\mu_j}\} \quad (3)$$

즉, 전체 상품 데이터베이스를 D, 각 속성을 $A_1, A_2, \cdots, A_n$ 그리고 A_i에 대한 퍼지 집합을 $F_1, F_2, \cdots, F_n$이라고 하면, 규칙 베이스에서 추출하는 규칙들은 $\{A_{i1}=F_{i1j1}, A_{i2}=F_{i2j2}, \cdots\}$의 데이터 Q는 Certainty μ를 포함하는 {tuple = T_k}로 정형화 할 수 있다.

5.2 관계 연산을 위한 퍼지 추론(Fuzzy Inference for Relational Operation)

퍼지 추론 과정에서는 규칙 베이스의 결과와 질의어를 추론 함수에 의하여 결과를 제시할 수 있도록 표준 SQL문으로 표현하게 된다. 즉 추론 함수가 f_θ일 때, t-norm을 따르는 퍼지 추론은 다음 관계 대수로 표현된다.

$$\sigma_{c(\mu_1,\cdots,\mu_j)}(R)=\{\langle c_1:\mu_1,\cdots,c_n:\mu_n\rangle | c_{\mu_1,\cdots,\mu_j}\} \quad (4)$$

추론 엔진은 사용자 질의문에 대한 템플릿과 규칙 베이스의 결과를 제품 정보 데이터베이스에서 해당되는 튜플들을 검색하여 해당되는 레코드를 결과로서 제시한다. 추론을 통한 결과는 기존 데이터베이스 조건 검색과 비교하여 다음 예제와 같은 차이를 제시한다.

예제 2.

사용자 질의문 "가격이 20만원 이상 30만원 이하이고, 화면 크기가 30인치 이하인 TV"에 대한 SQL표현은 "select TV from ITEM where price >= 200000 and price <= 300000 and size <= 30 ; "이 되며, 결과는 <표 3>과 같이 집합으로 제시된다.

〈표 3〉 일반 질의에 대한 검색 결과

item	price	weight	width	length	height	size
TV	210000	23	541	500	460	25
TV	250000	26	644	500	460	27

반면 자연어 질의인 "값이 싸고 가벼운 중형 TV"는 SQL로 표현하면 다음과 같다.

"select TV from ITEM where price = cheap and weight = light and tvsize = medium ; "

그리고 이에 대한 결과는 <표 4>와 같다.

〈표 4〉 퍼지 질의에 대한 검색 결과

item	price	weight	width	length	height	size
TV	199000	23	541	500	460	25
TV	210000	23	541	500	460	25
TV	250000	26	641	500	460	27

■

자연어 질의문에 대한 퍼지추론 결과는 예제에서와 같이 제품가격이 '199000원'인 경우 "값이 싸다"라는 사용자의 의도를 만족하지만 기존의 관계 연산 질의 방식에서와 같이 범위를 보통 값(Crisp Value)으로 입력하는 경우 <표 3>과 같이 결과에 나타나지 않게 되며, 자연어 질의를 하는 경우에는 이를 포함한 결과를 <표 4>와 같이 제시함으로써 보다 사용자 의도를 충족할 수 있다.

6. 실 험

6.1 실험 데이터

실험 데이터는 인터넷상에 존재하는 상품정보를 수집하여 사용하였다. 수집된 상품정보는 휴대용 카세트와 휴대용 CDP와 같은 음향기기 상품과 TV와 같은 영상기기 상품에 관한 것이다. 데이터베이스에 사용된 속성은 "상품분류", "모델명", "가격", "무게", "두께", "색상", "제조사" 등이다. 실험에 사용된 질의는 자연언어처리의 단계에 따른 질의처리의 효율성을 검증하기 위하여 문장으로 구성된 질의와 명사구로 구성된 질의를 각각 5개씩 구성하여 사용하였다. 형태소 분석기는 [13]을 이용하였다.

6.2 질의 처리 실험결과

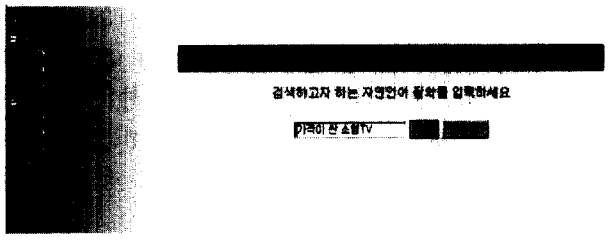
질의 처리를 평가하기 위하여 <표 5>와 같은 10개의 질의를 사용한다. 질의는 자연언어처리의 난이도에 따른 질의처리의 성능을 평가하기 위하여 문장으로 구성된 질의와 단순 명사구로 구성된 질의로 나누어 평가한다. 평가는 결과의 품질에 따라 "상", "중", "하"로 평가한다. <표 5>에 나타난 평가 결과는 본 논문에서 사용한 질의처리기가 자연언어 질의를 효율적으로 처리함을 보여주고 있다. 특히 자연언어 질의가 문장의 형태로 나타난 경우에도 비교적 자연언어 처리가 쉬운 명사구 형태의 경우와 마찬가지로 좋은 성능을 보여줌을 알 수 있다. 또한 생성된 템플릿은 일관된 형태의 용어를 사용함으로써 데이터베이스 질의로 변환할 때의 일관성을 유지하여 데이터베이스 검색을 용이하게 할 수 있다.

〈표 5〉 자연언어 질의 처리 평가를 위한 질의와 평가 결과

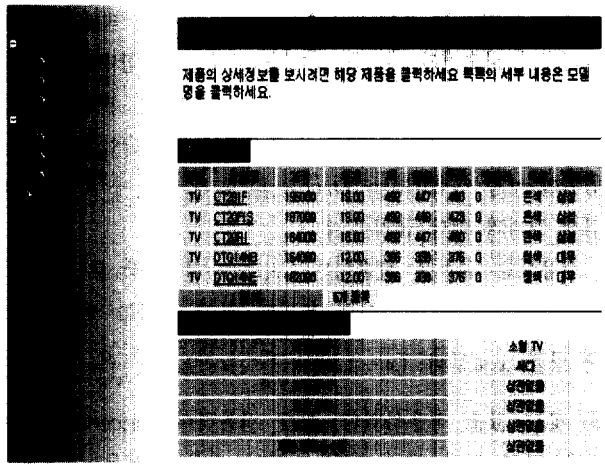
질의 내용	생성된 템플릿	평가결과
소형TV 중에서 가격이 싸고 무게가 무거운 것을 찾아주세요.	제품 : 소형TV 무게 : 가볍다 가격 : 싸다	상
무게는 무겁지만 저렴한 워크맨이 필요해요.	제품 : 워크맨 무게 : 무겁다 가격 : 싸다	상
가볍고 싼 CDP가 필요해요.	제품 : CDP 무게 : 가볍다 가격 : 싸다	상
무거워도 괜찮으니까 가격이 매우 싼 MP3플레이어 사고 싶어요	제품 : MP3P 무게 : 무겁다 가격 : 매우 싸다	상
얇고 가벼우며 크기도 작은 휴대용 카세트 원해요.	제품 : 워크맨 크기 : 작다 무게 : 가볍다 두께 : 얇다 크기 : 작다	상
싸고 큰 TV	제품 : TV 크기 : 크다 가격 : 싸다	상
가격이 적당한 대형 TV	제품 : TV 크기 : 크다 가격 : 보통	상
경량의 CDP	제품 : CDP 무게 : 가볍다	상
매우 저렴하고 크기도 적당한 워크맨	제품 : 워크맨 크기 : 보통 가격 : 매우 싸다	상
가벼운 MDP	제품 : MDP 무게 : 가볍다	상

6.3 데이터베이스 검색 실험결과

구현된 시스템에서는 자연언어 질의의 입력을 위하여 (그림 4)와 같은 형태로 사용자 인터페이스를 제공한다. 사용자에 의해 입력된 자연언어 질의는 질의 처리를 통하여 이를 "가격 : 싸다", "제품 : 소형TV"로 템플릿이 생성되고 퍼지추론에 의하여 사용자의 의도에 가장 부합되는 질의문을 생성한다. 추론에 의한 최종 결과는 (그림 5)와 같이 나타나게 된다.



(그림 4) 자연어에 의한 질의 화면



(그림 5) 퍼지 추론에 의한 제품 검색 결과 화면

본 논문에서 제시한 기법은 실험에서 나타난 결과와 같이 데이터베이스의 일반적인 값들의 범위를 모르는 사용자들에게 가격과 같은 자료를 구간 검색을 반복 수행하지 않고 의미를 만족하는 결과를 획득할 수 있는 장점이 있다. 특히, 자연어 질의를 계량화된 데이터를 다루는 제품 검색과 같은 분야에서 규칙 베이스의 결과는 다른 분야보다 적용이 용이하고 효율적인 시스템 구현이 가능하다.

7. 결론 및 향후 과제

논문에서는 최근 빠른 속도로 증가하는 전자 상거래 시스템에서의 데이터베이스 검색에 적용할 수 있는 자연어 퍼지 질의 검색 시스템을 제시하였다. 입력된 자연언어 질의에서 형태소 분석을 통하여 데이터베이스 질의에 사용될 수 있는 의미어를 추출한 후, 의미어들을 재구성하여 템플릿을 작성한다. 작성된 템플릿은 퍼지추론을 통하여 의미의 애매성을 해소하고 데이터베이스 질의인 SQL문으로 변환하여 사용자의 질의 의도와 부합되는 검색 결과를 제시한다.

이를 적용한 효과로서 데이터베이스 속성 중심으로 사용자들이 질의를 하도록 되어 있는 기존의 쇼핑몰 시스템을 사용자 중심의 의미를 내포한 질의 검색이 이루어 질 수 있도록 하였으며, 전자 상거래 시스템에서 직접 활용 가능한 Prototype을 구현하였다.

향후 연구과제로서 보다 다양한 질의 형태를 적절히 처리할 수 있는 자연언어 처리기의 개선과 퍼지 추론이 데이터베이스 내용 검색과 이미지와 같은 멀티미디어 검색 추론 기능을 갖출 수 있도록 확장하는 것과 기능을 Agent화하여 인터넷의 특성을 활용한 결과 검색을 고려할 수 있다.

참 고 문 헌

[1] J. Chae and S. Lee, "Identifying Basic Patterns of Korean Natural Language Query," In NLPRS'95, pp.606-611, 1995.

[2] R. Fagin, "Combining Fuzzy Information from Multiple Systems," J. of Computer and System Sciences, Vol.58, pp. 83-99, 1999.

[3] Informix Technical Brief, "Informix Web Datablade Module," Informix Corp.

[4] A. Klein, J. Matiasek, and H. Trost, "The treatment of noun phrase queries in a natural language database access system," In COLING-ACL'98 workshop on the computational treatment of nominals, pp.39-45, 1998.

[5] KM. Lee, H. Lee-Kwang, "Fuzzy Information Processing for Expert Systems," Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol.3, No.1, pp.93-109, 1995.

[6] R. Nelken and N. Francez. "Querying Temporal Databases Using Controlled Natural Language," In COLING'2000, pp 1076-1080, 2000.

[7] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition," Proceedings of the IEEE, Vol.77, No.2, pp.257-286.

[8] L. Zadeh, "Fuzzy Sets," Inf. Control, Vol.8, pp.338-353, 1965.

[9] L. Zadeh, "The Concept of a Linguistic Variable and Its Application to Approximate Reasoning," Inf. Science, Vol.8, No.3, pp.199-249, 1975.

[10] 강승식, "전자거래 시스템에서 가격지정 연산자의 인식", 제 11회 한글 및 한국어정보처리학회, pp.85-88, 1999.

[11] 김재훈, "오류-보정 기법을 이용한 어휘 모호성 해소", 한국과학기술원 전산학과 박사학위 논문, 1996.

[12] 김진동, 임희석, 임해창, "Twoply HMM : 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델", 정보과학회논문지(B), 제24권 제12호, pp.1502-1512, 1997.

[13] 신중호, 한영석, 박영찬, 최기선, "어절구조를 반영한 은닉 마르코프 모델을 이용한 한국어 품사 태깅", 제6회 한글 및 한국어 정보처리 학술대회, pp.389-394, 1997.

[14] 윤성희, "한국어 자연언어 질의 문장 파싱에서의 중의성 해소", 정보과학회논문지(B), 제24권 12호, pp.1482-1492, 1997.

[15] 이광형, 오길록, "퍼지 이론과 응용", 홍릉과학출판사, 서울, 1992.

[16] 이정규, 이상주, 임희석, 임해창, "규칙기반 한국어 품사 태깅을 위한 어휘 규칙 획득의 수작업 최소화 방안", 제24회 한국정보과학회 춘계학술발표대회논문집, Vol.24, No.1, pp.479-482.

[17] 이호동, 박종철, "결합범주문법을 이용한 자연언어 인터페이스", 한국정보과학회 추계학술발표논문집, Vol.27(II), pp.173-175, 2000.

[18] 임희석, 김진동, 임해창, "어절 태그 변형 규칙을 이용한 한국어 품사 태거", 정보과학회논문지(B), 제 24권 제6호, pp.673-684, 1997.

[19] 채진석, 김성기, 이석호, "한국어 데이터베이스 검색을 위한 질의 시스템의 설계 및 구현", 정보과학회논문지, 제20권 제6호, pp.810-820, 1993.

박 현 규

e-mail : hkpark@dbserver.kaist.ac.kr

1987년 육군사관학교 전산학(학사)

1992년 Naval Postgraduate School 전산학(석사)

1999년~현재 한국과학기술원 전산학과 박사과정

1987년 육군장교 임관

1997년~현재 육군 소령

관심분야 : 시공간 데이터베이스, 이동 컴퓨팅 등

오 종 훈

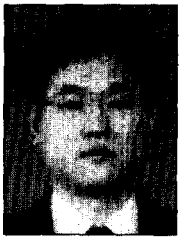
e-mail : rovellia@world.kaist.ac.kr

1998년 성균관대학교 정보공학과 졸업(학사)

2000년 한국과학기술원 전산학과 졸업(공학석사)

2000년~현재 한국과학기술원 전산학과 박사과정

관심분야 : 자연언어처리, 전문용어, 정보검색 등



김 명 호

e-mail : mhkim@dbserver.kaist.ac.kr

1982년 서울대학교 컴퓨터공학과(학사)

1984년 서울대학교 컴퓨터공학과(석사)

1989년 Michigan 주립대 전산학과(박사)

1989년~1993년 한국과학기술원 조교수

1993년~1999년 한국과학기술원 부교수

1999년~현재 한국과학기술원 교수

1993년 개방형 컴퓨터 통신 연구회(OSIA) 분산트랜잭션처리 분과위(TG-TP)의장

1993년~1994년 한국통신기술협회(TTA) 분산트랜잭션처리 실무위원회 의장

관심분야 : 분산데이터베이스, 분산트랜잭션, 멀티미디어 데이터베이스 등

최 기 선

e-mail : kschoi@cs.kaist.ac.kr

1978년 서울대학교 수학과 졸업(학사)

1980년 한국과학기술원 전산학과 졸업 (공학석사)

1986년 한국과학기술원 전산학과 졸업 (공학박사)

1985년~1986년 한국외국어대학교 전산학과 조교수

1987년~1988년 일본 NEC C&C 정보연구소 초빙연구원

1988년~현재 한국과학기술원 전산학과 교수

1998년~현재 한국과학기술원 전문용어언어공학연구센터 소장

관심분야 : 자연언어처리, 기계번역, 정보검색, 전문용어 등

이 광 형

e-mail : khlee@if.kaist.ac.kr

1978년 서울공대 산업공학학사

1980년 한국과학기술원 산업공학 석사

1982년 프랑스 INSA 전산학과 석사(DEA)

1985년 프랑스 INSA 전산학과 공학박사

1988년 프랑스 국가박사(전산학 INSA-LYONI대)

1985년~1995년 한국과학기술원 전산학과 조교수 및 부교수

1995년~현재 한국과학기술원 전산학과 교수

1995년~1996년 미국 Stanford Research Institute International Fellow

2000년~현재 Director of Information Security Center, KAIST

2000년~현재 Direcotr of Hacking and Virus Research Center, KAIST

관심분야 : 퍼지시스템, 인공지능, 페트리 네트, 전문가시스템 등