문제 은행에서 연상학습을 지원하는 퍼지 검색 시스템

(A Fuzzy Retrieval System to Facilitate Associated Learning in Problem Banks)

최 재 훈 † 김 지 숙 †† 조 기 환 †††

(Jae-Hun Choi) (Ji-Suk Kim) (Gi-Hwan Cho)

요 약 본 논문에서는 사용자 질의가 가지는 특정한 의미로부터 개념적으로 서로 연관된 문제들을 문제 은행에서 검색해 줌으로써 학습자의 연상학습을 지원할 수 있는 퍼지 검색 시스템을 설계하고 구현하였다. 특히, 연상학습의 특성인 검색의 일정한 정확률과 높은 재현율을 유지하기 위해 이 검색 시스템은 퍼지 시소러스를 이용하였다. 여기서, 도메인 종속적인 개념들 사이의 관계를 퍼지 정도로 표현하는 시소러스는 질의에 기술된 탐색어와 문제를 대표하는 색인어 사이의 용어 불일치 문제를 해결해 줌으로써 연상학습에 적합한 검색 성능을 나타낼 수 있게 한다. 따라서, 본 논문에서는 퍼지 시소러스를 이용한 문제 검색 시스템의 연상학습에 대한 적합성을 검색의 정확률과 재현율을 통해 평가하였다.

키워드 : 정보통신기술교육, 퍼지 정보 검색, 시소러스, 연상학습

Abstract This paper presents a design and implementation of fuzzy retrieval system that could support an associated learning in problem banks. It tries to retrieve some of the problems conceptually related to specific semantics described by user's queries. In particular, the problem retrieval system employs a fuzzy thesaurus which represents relationships between domain dependent vocabularies as fuzzy degrees. It would keep track of characteristics of the associated learning, which should guarantee high recall and acceptable precision for retrieval effectiveness. That is, since the thesaurus could make a vocabulary mismatch problem resolved among query terms and document index terms, this retrieval system could take a chance to effectively support user's associated learning. Finally, we have evaluated whether the fuzzy retrieval system is appropriate for the associated learning or not, by means of its precision and recall rate point of view.

Key words : ICT Education, Fuzzy Information Retrieval, Thesaurus, Associated Learning

1. 서 론

최근, 인터넷 기술의 발전으로 학습의 시공간적인 제약이 극복됨으로써 많은 사이트들이 다양한 교육 컨텐츠들을 서비스할 수 있게 되었다[1][2]. 이들의 대부분은 어학이나 전문 자격증과 관련된 시험 문제들을 문제 은행(Problem Bank)으로 구축한 다음, 사용자가 요구하는 문제들을 적절히 검색하여 학습할 수 있도록 지원하고 있다[3]. 이 서비스의 핵심 기술은 문제 은행에서 사용자가 요구하는 특정한 의미와 관련된 문제들을 효과적으로 검색하여 학습할 수 있게 함으로써 학습자의 연상학습을 지원하는 것이다. 이를 위해 기존 사이트들은 문제들을 분야 또는 출제 일자에 따라 분류하여 사용자에게 제시하는 방법을 사용하고 있다[4]. 그러나, 이 방법은 의미적으로 서로 연관된 문제들만을 동적으로 검색하여 학습하려는 사용자들의 연상학습에 대한 빈번한 요구를 충분히 만족시킬 수 없다는 문제점을 가진다.

· 본 연구는 전북대학교 영상정보신기술연구센터의 지원으로 수행되었음.

† 비 회 원 : 한국전자통신연구원 선임연구원
jhchoi@etri.re.kr
†† 비 회 원 : 전북대학교 교육대학원
jskim@cs.chonbuk.ac.kr
††† 정 회 원 : 전북대학교 전자정보공학부 교수
ghcho@dcs.chonbuk.ac.kr

논문접수 : 2001년 8월 3일
심사완료 : 2002년 1월 17일

대표적으로 어학 학습에 관련된 문제들을 서비스하고 있는 참고 사이트 [4]에서는 학습자의 요구와 상관없이 서버가 특정 분야의 문제들을 일방적으로 제공하고 있다. 따라서, 학습자가 특정 분야의 문제를 선별적으로 검색하여 학습할 수 없다. 또한, 참고 사이트 [3]에서는 ASP와 같은 고급 IT 기술을 이용하여 모든 문제들을 각각 분야에 따라 계층적으로 분류한 다음, 사용자가 특정 분야를 선택하면 해당 분야의 문제들을 제시해 준다. 그러나, 이 방법 역시 사용자가 요구하는 특정한 의미로부터 연상되는 다른 문제들을 대화형 방식으로 검색하여 학습할 수 없다는 문제점을 가지고 있다.

이를 해결하기 위해 사용자 질의와 개념적으로 관련을 가지는 문서들을 대화형 방식으로 검색할 수 있게 하는 [5]와 같은 문서 정보 검색 기술이 이용될 수 있다. 여기서, 문서들은 이들을 대표할 수 있는 용어들로 색인되며, 사용자 질의는 중요한 의미를 가지는 탐색어들로 기술된다. 또한, 검색 시스템은 사용자 질의에 기술된 탐색어들로 색인된 문서들을 검색함으로써 사용자가 요구하는 관련 문서들을 제시하게 된다. 이 방법은 대학 도서관의 문헌 검색 시스템이나 인터넷 검색 사이트(야후, 네이버, 엠파스 등)에서 실제 사용하고 있다. 그러나, 이 검색 시스템들은 문헌 또는 문서를 대상으로 한다는 점에서 문제 컨텐츠 검색에 직접 이용하기 어렵다. 따라서, 본 논문의 목적은 문서 정보 검색 방법을 문제 검색 환경에 적합한 형태로 변형하여 사용자의 연상학습을 효과적으로 지원하려는데 있다. 이를 위해 검색 대상이 되는 문제 컨텐츠에 대한 특성을 분석하여 연상학습을 위한 검색 전략으로 활용하였다.

일반적으로 문서들은 텍스트 길이가 길고 많은 색인어들을 가지는 반면, 문제는 텍스트 길이가 매우 짧고 적은 색인어들을 가진다. 따라서, 사용자 질의와 상당히 관련이 있는 문서들은 사용자 질의에 표현된 탐색어들과 정확히 일치하는 색인어를 대부분 가지고 있지만, 문제들은 오히려 탐색어와 유사한 의미의 몇 개의 색인어들만을 가지게 된다. 또한, 문제를 기술하는 텍스트는 매우 구체적이며 축약된 의미의 색인어들로 표현되기 때문에 검색 과정에서 용어 불일치 문제(vocabulary mismatch problem)가 빈번히 발생한다. 즉, 사용자 질의에는 표현되지 않았지만 의미적으로 사용자의 요구와 매우 일치하는 색인어들을 가지는 문제들을 검색될 수 없게 된다[6][7]. 또한, 학습의 관점에서 문서는 매우 긴 텍스트를 가지기 때문에 매우 많은 학습 시간을 요구한다. 따라서, 사용자 요구에 정확히 일치하는 문서만을 검색해야하기 때문에 검색의 재현율보다는 정확률 향상이 매우 중요한 연구 과제이다. 반면, 짧은 텍스트로 표현되는 문제를 통한 학습은 비교적 적은 시간을 요구하기 때문에 사용자들이 특정한 의미와 서로 관련이 있는 여러 문제들을 함께 검색하여 학습하려는 경향이 있다. 따라서, 문제 검색을 통한 연상학습에서는 검색의 정확률 향상보다는 재현율 향상이 매우 중요하다.

본 논문에서는 문제 검색을 통한 연상학습을 효과적으로 지원하기 위해 퍼지 검색 시스템을 개발하였다. 이 검색 시스템은 퍼지 시소러스를 이용하여 용어 불일치 문제를 해결함으로써 검색에 대한 일정한 정확률을 보장하면서 재현율을 향상시킬 수 있게 한다. 즉, 도메인 용어들 사이의 의미 관계를 퍼지 정도로 표현한 시소러스를 이용함으로써 사용자 질의의 탐색어와 정확히 일치하는 색인어를 가지는 문제들뿐만 아니라 개념적으로 서로 연관된 유사한 의미의 색인어들을 가지는 문제들까지 검색할 수 있게 한다.

이를 설명하기 위해 본 논문을 다음과 같이 구성하였다. 먼저, 2장에서는 교육 지원 시스템의 문제 검색에서 이용될 수 있는 문서 정보 검색 방법들에 대해 살펴본다. 3장에서는 연상학습을 지원하는 퍼지 문제 검색 모델을 설계하고, 이 검색에 이용되는 문제 은행 데이타베이스와 퍼지 시소러스에 대해 설명한다. 4장과 5장에서는 퍼지 문제 검색 시스템을 구현하고, 이 시스템의 연상학습에 대한 성능을 재현율과 정확률을 통해 평가한다. 마지막으로, 6장에서는 본 연구에 대한 결론과 향후 연구 과제들을 제시한다.

2. 관련연구

일반적으로 교육 지원 시스템은 사용자들이 특정한 분야의 학습을 능동적으로 수행할 수 있도록 일정한 학습 절차에 따라 교육 컨텐츠들을 제공해주는 정보 시스템이다[8]. 본 논문은 교육 지원 시스템에서 제공되는 다양한 문제 컨텐츠를 통해 학습자가 효과적인 연상학습을 수행할 수 있도록 지원하는 퍼지 검색 시스템의 개발을 목적으로 하고 있다. 여기서, 문제를 통한 연상학습이란 학습자들이 일정한 의미의 질의 또는 현재 학습하고 있는 특정한 문제로부터 의미적으로 연상되는 다른 문제들을 검색하여 학습하는 것으로 정의한다.

기존의 교육 지원 시스템에서는 도메인 전문가가 문제 컨텐츠들을 해당 분야에 따라 분류하여 문제 은행 데이타베이스를 구성하면, 사용자는 특정 분야의 분류 항목을 선택하여 해당 문제들을 검색하고 이들을 통해 학습하는 방식을 사용하고 있다. 그러나, 이 방식은 특정한 의미를 가지는 문제들만을 검색할 수 없기 때문에

학습자의 연상학습에 대한 요구를 효과적으로 지원할 수 없다는 단점을 가진다. 이 단점을 해결하기 위해 본 논문에서는 문서 정보 검색 방법을 채용한 새로운 문제 검색 모델을 제안한다. 즉, 이 검색 모델은 사용자 질의나 문제 질의로부터 개념적으로 연상되는 다른 문제들을 검색할 수 있게 하여 효과적으로 연상학습을 지원할 수 있게 한다. 또한, 문제를 통한 연상학습의 특성인 검색의 일정한 정확률과 높은 재현율을 유지시키기 위해 퍼지 시소러스를 이용한다.

문서 정보 검색은 사용자 질의에 기술된 탐색어들로 색인된 문서들을 데이타베이스로부터 검색하고, 이들을 질의와의 관련 정도에 따라 제시한다[9]. 따라서, 문서 정보 검색 모델은 사용자 질의와 문서 사이의 관련 정도를 평가하는 방법에 따라 불리언 모델, 벡터 모델 그리고 확률 모델로 구분된다. 불리언 모델은 다시 퍼지 불리언 모델(fuzzy boolean model)과 확장된 불리언 모델(extended boolean model)[10], 벡터 모델은 일반화된 벡터 모델(generalized vector model), 잠재 의미 색인 모델(latent semantic indexing model) 그리고 뉴럴 네트워크 모델(neural network model), 확률 모델은 추론 네트워크 모델(inference network model)과 확신 네트워크 모델(belief network model)로 각각 세분된다[11].

불리언 모델은 문서를 색인어들의 집합으로 표현하며, 질의의 탐색어들에 대한 불리언 연산을 통해 관련 문서를 검색한다[12]. 벡터 모델은 문서와 질의를 n-차원 공간 벡터로 표현하며, 두 벡터의 관련 정도는 대수 연산을 통해 평가된다[13]. 확률 모델은 색인어들의 종속 정도를 이용하여 문서를 표현하고, 확률 연산자 또는 베이지언 추론 네트워크 등을 이용하여 질의와 문서 사이의 관련 정도를 평가한다[14]. 그러나, 벡터 모델과 확률 모델은 문서에 나타나는 색인어가 비교적 많을 경우 유용하기 때문에 색인어가 적은 문제 검색에는 적합하지 않다는 특징을 가지고 있다[5]. 따라서, 본 논문에서는 연상학습을 효과적으로 지원하기 위해 퍼지 불리언 모델을 채용하였다.

불리언 모델은 매우 단순한 구조를 가지며, 사용자 질의에서 탐색어들 사이의 논리적 관계를 AND나 OR로 비교적 자연스럽게 표현할 수 있기 때문에 다른 모델에 비해 사용자의 요구를 정확히 반영한다는 장점을 가진다[12]. 예를 들어, 두 탐색어 t1과 t2에 대해 사용자 질의가 q="t1 AND t2"와 같이 표현될 수 있다. 만약, 이 질의 q에 대해 문서 d가 $t1 \in d \wedge t2 \in d$이면 검색되고, $t1 \notin d \vee t2 \notin d$이면 검색되지 않는다. 따라서, 불리언 모델은 문서와 질의 사이의 관련 정도를 평가할 수 없다는 단점을 가지고 있다. 즉, 문서의 색인어들이 모두 같은 정도로 문서의 의미를 대표하며, 질의와 부분적으로 일치하는 문서를 검색할 수 없기 때문에 높은 재현율을 요구하는 연상학습에 적합하지 않다는 특성을 가진다.

이 단점을 해결하기 위해 d와 q 사이의 퍼지 멤버쉽 함수 $\mu_{\|q\|}(d)$에 따라 관련 정도를 평가할 수 있는 퍼지 불리언 모델이 제안되었다[9][15]. 즉, 사용자 질의의 탐색어 t와 문서 d 사이의 관련 정도를 도메인에 따라 적절히 설계된 퍼지 멤버쉽 함수 $\mu_{\|t\|}(d)$로 계산할 수 있기 때문에 검색된 문서들을 순위화할 수 있다는 장점을 가진다[13]. 또한, 사용자는 자신의 의도에 따라 질의의 탐색어 각각에 대해 서로 다른 관련 정도를 부여할 수 있다. 예를 들어, 두 용어 t1과 t2에 대해 사용자의 의도를 대표할 수 있을 관련 정도가 각각 α와 β라면, 사용자 질의를 q="t1:α OR t2:β"로 표현할 수 있다. 따라서, d와 q 사이의 관련 정도는 $\mu_{\|t1\|}(d) = \max(\min(\mu_{\|t1\|}(d), \alpha), \min(\mu_{\|t2\|}(d), \beta))$와 같이 평가될 수 있다.

또한, 퍼지 불리언 모델은 도메인 개념들 사이의 관계를 정의한 시소러스와 쉽게 통합될 수 있다는 장점을 가지고 있다[12]. 일반적으로 상관 관계 행렬 시소러스가 여기에 사용되며, 이 행렬 요소 $c_{i,j}$는 도메인 문서에서 두 용어 t_i와 t_j의 동시 출현 빈도에 대한 통계 정보로 계산된다[10]. 따라서, 이 시소러스는 도메인에 매우 의존적이며 문헌과 같이 긴 텍스트 검색에 유용한 특성을 가진다. 그러나, 용어의 동시 출현 빈도가 두 용어 사이의 의미적 관계를 나타내지 못하기 때문에 이 시소러스를 텍스트의 길이가 짧은 문제 검색에 이용할 경우 오히려 낮은 검색 성능을 가질 수 있다[6][13][15]. 따라서, 본 논문의 퍼지 문제 검색 시스템은 상관 행렬 시소러스뿐만 아니라 도메인 전문가가 직접 구축한 시소러스 역시 이용될 수 있도록 설계한다.

3. 퍼지 문제 검색 모델

본 논문에서 제안하는 퍼지 문제 검색 모델은 사용자 질의 또는 특정 문제로부터 자동으로 생성된 질의를 통해 학습자의 연상학습을 효과적으로 지원하는 것을 목적으로 하고 있다. 즉, 학습자가 사용자 질의로부터 개념적으로 연상될 수 있는 유사한 문제들을 검색하여 학습할 수 있다. 이 장에서는 사용자의 연상학습을 지원하는 퍼지 문제 검색 모델에 대해 설명한다.

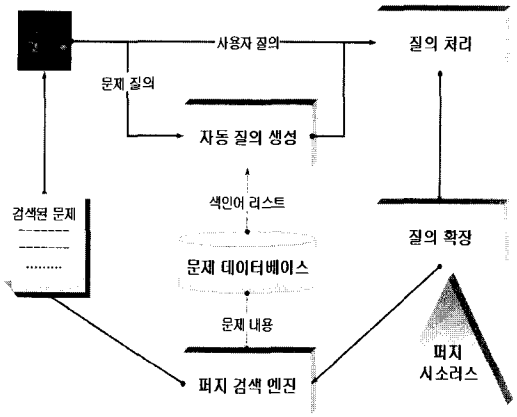

그림 1 퍼지 문제 검색 모델의 구성

그림 1은 본 논문에서 제안하는 퍼지 문제 검색 모델의 구성을 나타내고 있다. 이 모델에서 사용자는 퍼지 불리언 형태의 질의 또는 문제 질의를 통해 자신이 요구하는 관련 문제들을 검색할 수 있도록 지원한다. 여기서, 문제 질의는 자동 질의 생성 과정을 통해 퍼지 불리언 형태의 사용자 질의로 변형되어 처리된다. 퍼지 불리언 질의는 사용자 의도와 탐색어 사이의 관련 정도를 0과 1사이의 퍼지 값으로 표현하고, 이들의 논리적 관계를 불리언 연산자(AND/OR)를 통해 표현한다. 또한, 이 질의에 포함된 탐색어들은 퍼지 시소러스를 통해 확장되며, 검색 엔진을 통해 사용자에게 관련 문제들을 제시한다. 이 문제 검색 모델의 특징은 사용자 질의나 문제 질의를 통해 학습자가 자신이 원하는 문제들을 대화형 방식으로 검색할 수 있으며, 퍼지 시소러스를 이용하여 용어 불일치 문제를 해결하였다는 것이다[7]. 특히, 검색의 재현율을 저하시키는 주된 원인인 용어 불일치 문제는 높은 재현율을 요구하는 연상학습을 위해 반드시 해결되어야 한다. 다음은 이 검색을 위해 문제 은행 데이타베이스와 퍼지 시소러스의 구성에 대해 설명한다.

3.1 문제 은행 데이타베이스

문제 은행은 다양한 도메인에 소속되는 방대한 문제들에 대한 정보를 체계적으로 관리할 수 있도록 지원하는 데이타베이스이다. 본 논문에서는 퍼지 시소러스를 이용한 문제 검색을 위해 문제 은행 데이타베이스를 그림 2와 같이 5개의 관계 테이블로 구성하였다. 이 절에서는 이 테이블들 사이에 관계에 대해 자세히 설명한다.

문제 테이블은 데이타베이스에서 각각의 문제를 유일하게 구별할 수 있게 하는 식별자 "ProbID", 문제의 내용 "Prob" 그리고 문제에 대한 정답 "Answer" 애트

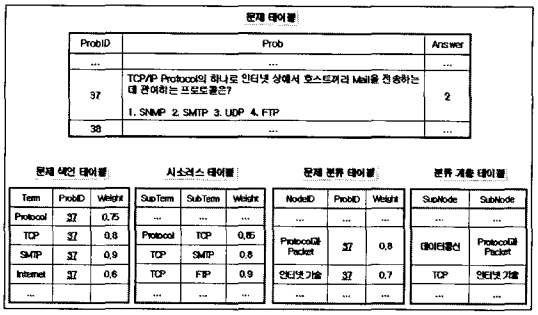

그림 2 문제 은행 데이타베이스

리뷰트로 구성된다. 문제 식별자 "ProbID"는 문제 색인 테이블과 문제 분류 테이블의 "ProbID"과 조인(Join) 연산을 위해 이용된다. 문제 색인 테이블은 하나의 문제 "ProbID"와 이 문제를 대표할 수 있는 색인어 "Term"에 대한 관계를 "Weight"에 0과 1사이의 퍼지 관련 정도로 명시한다. 즉, 37번 문제는 'Protocol : 0.75', 'TCP : 0.8', 'SMTP : 0.9' 그리고 'Internet : 0.6'과 같이 여러 개의 색인 정보를 가지며, 이 색인 정보는 사용자 질의와의 매칭을 통해 관련된 문제를 검색하게 된다. 색인어 t_i와 문제 p_j 사이의 관련 정도 $w_{i,j}$는 t_i의 출현 빈도에 대한 통계 값으로 아래와 같은 수식을 통해 자동으로 계산될 수 있다[13].

$$w_{i,j} = tf_{i,j} \times ipf_i : tf_{i,j} = \frac{freq_{i,j}}{\max_k(freq_{k,j})},\ ipf_i = \log\left(\frac{N}{n_i}\right)$$

여기서, $tf_{i,j}$는 색인어 t_i가 문제 p_j에서 어느 정도 출현했는지를 나타내는 상대적인 출현 비율이고, ipf_i는 전체 문제 집합에서 t_i가 특정 문제들을 어느 정도 변별할 수 있는지를 나타내는 색인 변별력이다. 따라서, $w_{i,j}$는 p_j에서 t_i의 출현 비율 $tf_{i,j}$이 높을수록 그리고 t_i의 변별력 ipf_i이 좋을수록 높은 값을 가진다. 또한, $tf_{i,j}$에서 $freq_{i,j}$는 t_i가 문제 p_j에서 몇 번 출현했는지를 나타내는 절대적인 출현 빈도이다. 따라서, $tf_{i,j}$는 p_j에서 t_i의 절대적 출현 빈도 $freq_{i,j}$에 비례하고, 가장 고빈도의 색인어 t_k의 출현 빈도 $freq_{k,j}$에 반비례한다. 즉, p_j에서 가장 고빈도인 색인어 t_k가 p_j의 내용을 대표하는 정도에 대한 t_i의 대표 정도를 상대적인 비율로 나타낸다.

반면, t_i의 변별력 ipf_i에서 N은 총 문제의 개수이고, n_i는 색인어 t_i를 가지는 문제의 개수이다. 만약, t_i가 모든 문제에서 한번 이상 출현한다면, t_i는 어떠한 문제도 변별해 낼 수 없기 때문에 이 색인어는 전혀 변별력이 없다고 할 수 있다. 반대로, 이 색인어가 단지 하나의 문제에서만 출현한다면, 이 색인어는 가장 큰 변별력을

가진다. 따라서, w_{ij}는 문제 p_j에서 색인어 t_i의 상대적인 출현 비율 tf_{ij}와 그 변별력 ipf_i에 비례한다. 이 자동 색인 방법은 자연어 문장에서 색인어들을 추출할 수 있는 형태소 분석기를 반드시 요구하게 된다. 또한, 이 형태소 분석기를 위해서는 대용량의 사전 데이타가 구축되어야 하는 방대한 작업을 요구한다. 이런 관점에서 본 논문의 주된 관점이 문제 자동 색인보다는 색인된 데이타를 이용한 퍼지 문제 검색 모델에 있기 때문에 현재 본 논문의 평가를 위해 사용된 문제들은 수작업으로 직접 색인되었다.

시소러스 테이블은 색인어로 사용되는 용어들 사이의 의미적 계층 관계를 상의어 "SupTerm"와 하의어 "SubTerm"에 대한 퍼지 관련 정도 "Weight"로 표현한다. 즉, 'Protocol'에 대해 'TCP'는 의미적으로 0.85 정도로 하의어이며, 'SMTP'와 'FTP'는 'TCP'에 대해 모두 0.9 정도로 하의어이다. 이 시소러스는 질의 확장을 위해 이용되며, 다음절에서 보다 자세히 설명된다. 문제 분류 테이블에서 37번 문제는 "Protocol과 Packet" 그리고 "인터넷 기술"이라는 주제 노드에 각각 0.8과 0.7 정도로 분류되어 있다. 또한, 이 노드들은 각각 "데이타통신"과 "TCP"의 하위 주제 분류임을 분류 계층 테이블로부터 알 수 있으며, 이 분류 계층은 도메인 전문가에 의해 동적으로 관리될 수 있다.

3.2 퍼지 시소러스

시소러스는 도메인 종속적인 용어들 사이의 의미 관계를 명시하는 도메인 지식이며, 질의 처리에 이용되어 검색 성능을 향상시킨다[16]. 즉, 사용자 질의와 의미적으로 관련된 시소러스 용어들을 탐색어로 추가함으로써 검색의 재현율을 향상시킬 수 있으며, 일반적인 의미의 탐색어를 구체적인 의미의 시소러스 용어들로 대치함으로써 검색의 정확률을 향상시킬 수 있다[17][18]. 따라서, 검색 성능 향상을 위해 시소러스는 반드시 요구되는 도메인 지식이라고 할 수 있다[19]. 시소러스는 구조적인 측면에서 도메인 용어들을 나타내는 노드와 이들 사이의 의미 관계를 나타내는 링크로 구성된다. 이때, 시소러스 용어들 사이의 관계는 상/하의어(BT/NT: Broader/Narrower Term) 그리고 관련어(RT: Related Term) 관계로 표현한다[20]. 본 논문의 퍼지 시소러스는 BT/NT 관계만을 이용하며, 시소러스 용어들 사이의 관계를 퍼지 관련 정도로 표현하였다. 그 이유는 특정 용어와 RT 관계를 가지는 용어들이 너무 많을 경우, 문제 검색 과정이 너무 복잡해지며 도메인에 따라 검색 성능이 오히려 저하될 수 있기 때문이다.

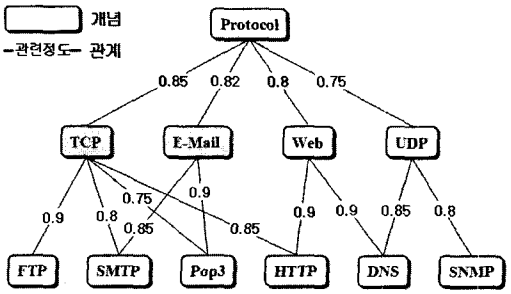

그림 3 퍼지 시소러스

그림 3은 "데이타통신"이라는 도메인에서 사용되는 용어들 사이의 관계를 표현한 퍼지 시소러스의 일부이다. 여기서, 시소러스 용어는 노드로 표현되며, 이들 사이의 관계는 링크로 표현된다. 링크에 레이블(label)된 0과 1사이의 값은 두 용어 사이의 퍼지 관련 정도를 나타낸다. 예를 들어, 'TCP'는 'Protocol'의 하의어로 0.85로 관련을 가지며, 'TCP'는 하의어로 'FTP', 'SMTP', 'Pop3' 그리고 'HTTP'가 있으며, 각각 0.9, 0.8, 0.75 그리고 0.85의 퍼지 관련 정도를 가지고 있다. 따라서, 모든 도메인 용어들의 집합 C에 대해 퍼지 시소러스 T(C)는 다음과 같은 관계 집합으로 표현될 수 있다.

$$T(C)=\{\langle c_1, c_2, w_{1,2}\rangle \mid c_1, c_2 \in C, w_{1,2} \in [0,1]\}$$

여기서, 관계 $\langle c_1, c_2, w_{1,2}\rangle \in T(C)$는 c_1은 하의어로 c_2를 가지며, 두 개념들 사이의 퍼지 관련 정도가 $w_{1,2}$임을 의미한다. 예를 들어, 그림 3에서 <Protocol, TCP, 0.85>, <Protocol, E-Mail, 0.82>, <TCP, SMTP, 0.8> 그리고 <E-Mail, SMTP, 0.85> $\in$ T(C)이다. 다음은 이 관계들로부터 'SMTP'가 의미적으로 'Protocol'의 하의어가 되는 전이적 성질에 대해 설명한다.

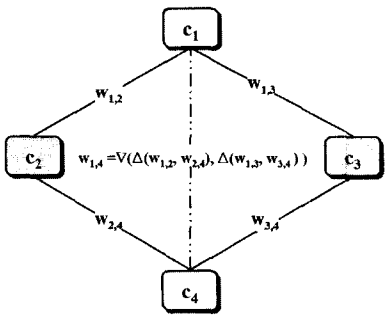

그림 4 퍼지 시소러스의 전이적 성질

그림 4는 퍼지 시소러스의 전이적 성질에 대한 설명이다. 즉, $<c_1, c_2, w_{1,2}>$, $<c_1, c_3, w_{1,3}>$, $<c_2, c_4, w_{2,4}>$, $<c_3, c_4, w_{3,4}> \in T(C)$이면 $<c_1, c_4, w_{1,4}> \in T(C)$가 성립한다. 이때, c_1과 c_4 사이의 퍼지 관련 정도 $w_{1,4}$는 자데(Zadeh)의 퍼지 확장 원리를 통해 $w_{1,4}=\nabla(\triangle(w_{1,2}, w_{2,4}), \triangle(w_{1,3}, w_{3,4}))$로 평가된다. 즉, 시소러스 용어 c_1로 확장될 수 있는 퍼지 집합 $EXP(c_1)$에서 $c_4 \in_{w1,4} EXP(c_1)$이다. 여기서, $\triangle$(t-norm)와 ∇(s-norm)는 두 퍼지 값에 대해 "AND/OR" 연산을 수행하기 위해 사용되는 함수이다. 현재, 응용 도메인에 따라 $\triangle/\nabla$에 대한 다양한 함수들이 개발되고 있으나 [21], 본 논문에서는 퍼지 관계에 대한 설명을 단순화하기 위해 가장 일반적인 min/max 함수를 사용한다.

예를 들어, 그림 3에서 'SMTP'는 'Protocol'의 일종이기 때문에 두 용어는 서로 일정한 관련을 가지며, 그 관련 정도는 $w_{Protocol,SMTP}=\nabla(\triangle(0.85, 0.8), \triangle(0.82, 0.85))$으로 평가될 수 있다[21]. 그 이유는 시소러스에서 관계 <Protocol, TCP, 0.85>, <TCP, SMTP, 0.8>, <Protocol, E-Mail, 0.82> 그리고 <E-Mail, SMTP, 0.85>의 의미가 "SMTP는 TCP와 0.8 정도로 관련을 가지며(AND), TCP는 Protocol과 0.85 정도로 관련을 가진다. 또한(OR), SMTP는 E-Mail과 0.85 정도로 관련을 가지며(AND), E-Mail은 Protocol과 0.82 정도로 관련을 가진다"로 해석되기 때문에 <Protocol, E-Mail, $w_{Protocol,SMTP}$>인 관계가 성립한다. 여기서, AND/OR를 min/max 함수로 해석한다면, $w_{Protocol,SMTP}=\max(\min(0.85, 0.8),\min(0.82, 0.85))=0.82$로 평가될 수 있다. 따라서, 'Protocol'로 확장될 수 있는 퍼지 용어 집합은 EXP('Protocol')={Protocol:1.0, TCP : 0.85, E-Mail : 0.82, Web:0.8, UDP : 0.75, FTP : 0.85, SMTP : 0.82, Pop3 : 0.82, HTTP : 0.85, DNS : 0.8, SNMP : 0.75}이다. 이 두 개념 사이의 관련 정도 평가는 다음 3.3절에서 설명되는 사용자 질의 확장에 이용된다.

3.3 퍼지 시소러스를 이용한 문제 검색

본 논문의 퍼지 문제 검색 모델은 특정한 의미의 질의 또는 현재 학습하고 있는 문제로부터 개념적으로 연상될 수 있는 다른 문제들을 검색해 줌으로써 특정한 주제에 대해 사용자들이 연상학습을 효과적으로 수행할 수 있도록 지원한다. 이때, 퍼지 불리언 형태로 표현되는 사용자 질의는 퍼지 시소러스를 통해 확장되어 문제 검색에 이용된다. 따라서, 이 절에서는 퍼지 불리언 질의를 통한 문제 검색 방법에 대해 먼저 기술하고, 다음으로 퍼지 시소러스를 이용한 질의 확장과 이 확장된 질의에서 연상되는 문제들을 검색하는 방법에 대해 설명한다.

일반적인 정보 검색 시스템에서 사용자는 정형화된 형태의 질의를 통해 자신이 요구하는 정보를 검색한다. 특히, 문제 검색에서 대부분의 사용자들은 특정 도메인에 대해 상당한 지식을 가지고 있다고 할 수 있기 때문에 포괄적인 의미의 검색이나 구체적인 의미의 검색을 모두 요구할 수 있다. 따라서, 사용자 질의는 다양한 사용자의 요구를 정확히 표현할 수 있도록 설계되어야 한다. 그러나, 일반적인 정보 검색의 사용자 질의는 모두 동일한 의미적 중요성을 가지는 탐색어들에 대한 불리언 형태로만 표현된다. 그러나, 실제 많은 사용자 질의는 특정한 의미를 나타내는 핵심 탐색어들을 가질 수 있다. 이를 위해 본 논문에서는 아래와 같은 퍼지 불리언 형태의 질의를 사용하여, 각각의 탐색어들에 대한 의미적 중요성을 표현할 수 있도록 하였다.

$$Q=(\text{AND} \mid \text{OR})_{i=1}^{n}[t : \alpha]_i,\ 0 \le \alpha \le 1$$

여기서, AND와 OR는 불리언 연산자이며, t는 탐색어이고 α는 이 탐색어와 사용자 의도 사이의 퍼지 관련 정도이다. 예를 들어, "Internet에서 사용하는 Protocol과 관련된 문제"에 대해 'Internet' 보다 'Protocol'이 중요한 의미를 가질 경우, "Internet : 0.7 AND Protocol : 0.9"와 같이 사용자 질의를 표현할 수 있다. 반면, 포괄적인 의미로 질의를 표현하기 위해 α를 생략할 수 있으며, 이 경우 묵시적으로 $\alpha=1.0$으로 간주한다.

이 퍼지 불리언 질의는 단일 질의(mono-query), 이접 질의(disjunctive query) 그리고 연접 질의(conjunctive query)로 분류하여 표준화할 수 있다. 즉, 단일 질의는 탐색어 t에 대해 $Qm=t:\alpha$와 같이 표현되며, 단일 질의 Qm_j, j=1,..,m에 대해 이접 질의는 $Q=OR_{j=1}^{m} Qm_j$이고, 이접 질의 Q_i, i=1,..,n에 대해 연접 질의는 $Q=AND_{i=1}^{n} Q_i$로 정의될 수 있다. 따라서, 모든 퍼지 불리언 질의는 단일 질의의 이접 질의에 대한 연접 질의로 표준화될 수 있다. 다음은 사용자 질의로부터 연상되는 문제를 검색하기 위한 질의 평가 방법을 설명한다.

먼저, 문제 p에 대한 퍼지 색인어 집합이 IDX(p)라면, 색인어 t에 대해 $\mu_{IDX(p)}(t)=\alpha$, $0 \le \alpha \le 1$이다. 또한, $p \in_{\beta} \|Q\|$은 p가 사용자 질의 Q를 β, $0 \le \beta \le 1$ 정도로 만족함을 나타낸다. 따라서, 단일 질의 $Qm=t:\alpha_1$에 대해 $\mu_{IDX(p)}(t)=\alpha_2$이면, $p \in_{\alpha} \|Qm\|$, $\alpha=\triangle(\alpha_1, \alpha_2)$로 평가

된다. 또한, Qm_j, j=1,..,m에 대해 이접 질의가 $Q=OR_{j=1}^{m} Qm_j$이고 $p\in_{a_1}\|Qm_1\| \vee p\in_{a_2}\|Qm_2\| \vee ... \vee p\in_{am}\|Qm_m\|$이면, $p\in_{a}\|Q\|, \alpha=\nabla(\alpha_1,\alpha_2,...,\alpha_m)$로 평가된다. 한편, 이접 질의 Q_i, i=1,..,n에 대해 연접 질의가 $Q=AND_{i=1}^{n} Q_i$이고 $p\in_{a_1}\|Q_1\| \wedge p\in_{a_2}\|Q_2\| \wedge ... \wedge p\in_{a_n}\|Q_n\|$일 때, $p\in_{a}\|Q\|, \alpha=\triangle(\alpha_1,\alpha_2,...,\alpha_n)$로 평가된다. 여기서, 두 퍼지 함수 △/▽는 도메인에 적합한 다양한 함수가 적용될 수 있으나 앞 절과의 일관성을 위해 *min/max* 함수를 각각 사용한다.

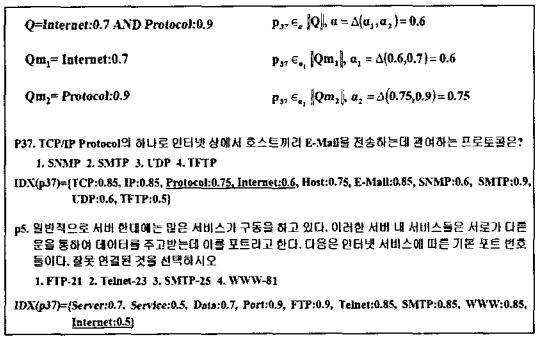

그림 5 사용자 질의 평가

예를 들어, 그림 5에서 사용자 질의가 Q="Internet : 0.7 AND Protocol : 0.9"일 때, 다음과 같은 과정을 통해 $p_{37}\in_{0.6}\|Q\|$로 평가된다. 즉, Qm_1="Internet : 0.7"이고 Qm_2="Protocol : 0.9"라면, $p_{37}\in_{0.6}\|Qm_1\|$이고 $p_{37}\in_{0.75}\|Qm_2\|$이다. 따라서, $p_{37}\in_{\alpha}\|Q\|$에서 대해 $\alpha=\min(0.6, 0.75)=0.6$으로 평가된다. 그러나, p_5와 같은 문제는 사용자 질의 Q와 의미적으로 상당히 관련이 있지만 $\mu_{IDX(p_5)}$(Protocol)=0이기 때문에 검색될 수 없는 단점을 가지고 있다. 이 단점은 검색의 재현율을 감소시키는 주된 원인으로 작용하기 때문에 사용자의 효과적인 연상학습을 저해하게 된다. 본 논문에서는 이 단점을 해결하기 위해 퍼지 시소러스를 통한 질의 확장 방법을 이용하였다.

퍼지 시소러스를 이용한 질의 확장은 각각의 탐색어에 대해 수행되기 때문에 사용자 질의 $Q=(AND\ |OR)_{i=1}^{n} Qm_i$에서 하나의 단일 질의 $Qm_i=[t:\alpha]_i$에 대해 확장된 질의 Q_i'는 t와 관련된 모든 시소러스 용어들을 OR 연산자로 연결한 이접 질의 형태로 표현될 수 있다. 즉, 자데의 퍼지 확장 원리를 통해 시소러스 T(C)에 대해 t와 관련된 시소러스 용어들에 대한 퍼지 집합을 F=EXP(t)라고 하면, c∈C에 대해 $\mu_F(C)=\beta$이다(3.2절 참조). 여기서, c=t이면 $\beta=1.0$이고, <t, c, γ>∈T(C)이면 $\beta=\gamma$이다. 따라서, 초기 질의에 대한 요소 $Qm_i=[t:\alpha]_i$을 $Q_i'=OR_{j=1}^{s}[c:\alpha']_j$, $\alpha'=\triangle(\alpha,\beta)$으로 확장할 수 있다. 또한, 이 확장 과정에서 임의의 임계값 ω에 대해 $\beta\geq\omega$인 c만을 이용함으로써 검색의 정확률과 재현율을 사용자가 직접 조절할 수 있다. 이를 위해 본 논문에서는 사용자 인터페이스로부터 ω를 입력받아 질의 확장 조건으로 이용하였다. 이 확장된 질의를 통한 문제 검색은 앞 절에서 설명한 방법과 동일하다.

예를 들어, 사용자 초기 질의 Q="Internet : 0.7 AND Protocol : 0.9" 그림 3의 시소러스를 통해 Q'=Q1' AND Q2' 확장될 수 있다. 여기서, Qm1 = "Internet : 0.7"과 Qm2="Protocol:0.9" 각각에 대해 시소러스를 통해 확장된 질의는 Q1'과 Q2'이다. 만약, Q_1'="Internet : 0.7 OR ... OR LAN : 0.63"라고 가정하고, Q2'는 다음과 같은 과정에 의해 확장될 수 있다. 즉, 퍼지 시소러스에서 'Protocol'과 관련이 있는 퍼지 용어 집합은 F=EXP('Protocol')={Protocol : 1.0, TCP : 0.85, E-Mail : 0.82, Web : 0.8, UDP : 0.75, FTP : 0.85, SMTP : 0.82, Pop3 : 0.82, HTTP : 0.85, DNS : 0.8, SNMP : 0.75}으로 구성된다. 따라서, Qm2="Protocol : 0.9"에 의해 확장된 질의는 Q2'="Protocol : 0.9 OR TCP : 0.85 OR E-Mail : 0.82 OR Web : 0.8 OR UDP : 0.75 OR FTP : 0.85 OR SMTP : α OR Pop3 : 0.82 OR HTTP : 0.85 OR DNS : 0.8 OR SNMP : 0.75"로 표현된다. 여기서, α는 Qm2="Protocol : 0.9"이고 μ_F(SMTP)=▽(△(0.85,0.8),△(0.82,0.85))=0.82임으로 $\alpha=\triangle(0.9, 0.82)=0.82$이다. 즉, 'SMTP'는 0.82의 퍼지 관련 정도로 확장된 질의 Q2'에 참여한다. Q2'의 다른 탐색어에 대해 같은 방법으로 질의 참여 정도를 평가할 수 있다.

문제 p_5는 퍼지 색인 집합 IDX(p_5)={Internet : 0.5, Server : 0.7, Service : 0.5, Data : 0.7, Port : 0.9, FTP : 0.9, Telnet : 0.85, SMTP : 0.85, WWW : 0.85}에서 $\mu_{IDX(p5)}$(Protocol)=0이기 때문에 사용자 초기 질의 Q="Internet : 0.7 AND Protocol : 0.9"에 의해 검색될 수 없었다. 그러나, Qm2="Protocol : 0.9"는 Q2'="Protocol : 0.9 OR TCP : 0.85 OR E-Mail : 0.82 OR Web : 0.8 OR UDP : 0.75 OR FTP : 0.85 OR SMTP : 0.82 OR Pop3 : 0.82 OR HTTP : 0.85 OR DNS : 0.8 OR SNMP : 0.75"로 확장될 수 있다. 즉, Qm1="Internet : 0.7"이고 $\mu_{IDX(p5)}$(Internet)=0.5이기 때문에

$p_5 \in_{\alpha_1} \| Q1' \|, \alpha_1 = \triangle(0.5,\ 0.7) = 0.5$로 평가된다. 또한, $\mu_{IDX(p5)}('FTP') = 0.9$ 그리고 $\mu_{IDX(p5)}('SMTP') = 0.85$이며, Q_2'에서 탐색어 'FTP' 그리고 'SMTP'가 각각 0.85와 0.82의 관련 정도를 가지기 때문에 $p_5 \in_{\alpha_2} \| Q2' \|, \alpha_2 = \nabla(\triangle(0.9,\ 0.85),\ \triangle(0.85,\ 0.82)) = 0.85$ 평가된다. 따라서, p_5는 Q'에 의해 $p_5 \in_{\alpha} \| Q' \|$, $\alpha = \triangle(\alpha_1, \alpha_2) = 0.5$로 평가되어 검색될 수 있다. 이와 같은 방법은 "문제 질의 검색" 과정에 그대로 적용된다.

4. 구 현

이 장에서는 퍼지 시소러스를 이용한 문제 검색 시스템을 구성하는 문제 관리 컴포넌트와 문제 검색 컴포넌트의 구현 내용에 대해 설명한다. 특히, 검색 컴포넌트의 문제 질의 검색과 사용자 질의 검색에 대한 구현 예를 자세히 설명한다. 이 두 컴포넌트는 Windows 운영체제 환경에서 Visual Basic 6.0 언어로 구현되었다. 또한, 문제 정보 데이타베이스는 관계형 데이타베이스(RDBMS)에서 관리되며, 이들은 ODBC (Open Database Connectivity)를 통해 접근된다.

4.1 문제 관리 컴포넌트

문제 관리 컴포넌트는 도메인 전문가들이 문제들을 특정 주제에 따라 계층적으로 관리할 수 있도록 지원한다. 즉, 특정한 문제를 데이타베이스에 등록하여 해당 계층 노드에 분류할 수 있게 하며, 분류된 문제들을 참조하여 변경하거나 삭제할 수 있다. 이 절에서는 문제를 등록하고 분류하는 과정에 대해 설명한다.

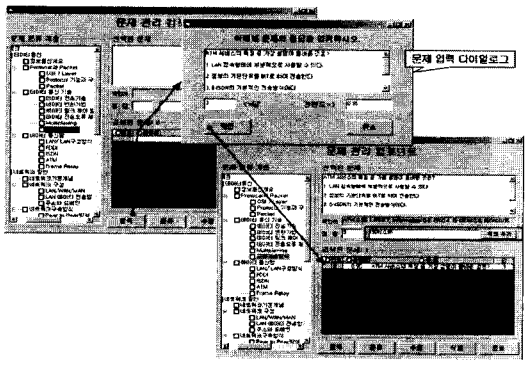

그림 6 문제 등록

그림 6은 계층 노드 '교환기술방식'에 하나의 문제를 등록하는 과정을 나타내고 있다. 문제 입력 다이얼로그에 입력되는 관련 정도 0.95는 등록될 문제가 선택된 노드에 대한 주제와 어느 정도 관련이 있는지를 나타낸다. 또한, "색인 추가" 버튼을 통해 입력된 문제에 대해 색인 정보들이 'ISDN:0.85'와 같은 형태로 입력된다. 이때, '0.85'는 'ISDN'이 10번 문제를 대표할 수 있을 정도이다. 이 색인 정보는 퍼지 시소러스를 이용한 문제 질의 검색과 사용자 질의 검색에 이용된다. 색인 과정은 앞에서 언급했듯이 현재 상용화된 문서 색인 기술을 통해 자동화될 수 있다. 또한, 특정 계층 노드에 등록된 문제는 동적으로 수정되거나 삭제될 수 있다.

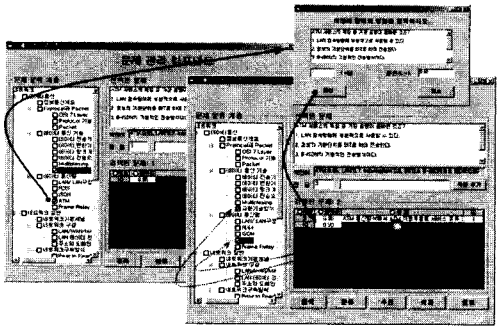

그림 7 문제 분류

그림 7은 '교환기술방식'에 이미 등록된 10번 문제를 다른 노드 'ATM'에 분류하는 과정을 설명하고 있다. 즉, 10번 문제를 선택한 다음, 계층 노드 'ATM'를 체크하고 적절한 관련 정도 0.9를 입력하여 해당 노드에 이 문제를 분류한다. 또한, 10번 문제를 사용자가 선택함으로써 이 문제가 어떤 노드에 분류되었는지를 직접 판별할 수 있다. 문제 계층에서 동적으로 관리되는 노드는 특정 도메인에서 하나의 주제를 나타내기 때문에 상위 노드는 의미적으로 하위 노드를 포괄하게 된다. 따라서, 하위 노드에 속한 모든 문제들은 그 상위 노드에서 참조될 수 있다. 즉, 'ATM'에 분류된 10번 문제는 그 상위 노드 '데이타 통신망'에서 역시 참조될 수 있다.

4.2 문제 검색 컴포넌트

퍼지 시소러스를 이용한 문제 검색 컴포넌트는 크게 사용자 질의 검색 그리고 문제 질의 검색으로 구분된다. 사용자 질의 검색은 특정한 의미를 나타내는 사용자 질의로부터 개념적으로 연상되는 문제들을 직접 검색할 수 있도록 지원하며, 문제 질의 검색은 현재 사용자가 학습하고 있는 문제로부터 연상되는 다른 문제들을 검색할 수 있도록 지원한다.

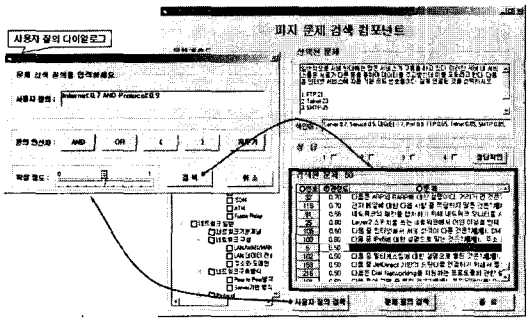

그림 8 사용자 질의 검색에 대한 결과 화면

그림 8은 "인터넷에서 사용되는 프로토콜과 관련된 문제"라는 의미의 사용자 질의 "Q=Internet : 0.7 AND Protocol : 0.9"로부터 연상되는 문제들을 검색한 검색 결과이다. 즉, "사용자 질의 다이얼로그"를 통해 사용자가 자신이 의도한 의미의 질의를 표현하여 관련 문제들을 검색할 수 있다. 이 질의에 의해 검색된 문제들은 'Internet'과 'Protocol'로부터 개념적으로 연상될 수 있는 문제들이다. 특히, IDX(p_5)={Server : 0.7, Service : 0.5, 데이타 : 0.7, Port : 0.9, FTP : 0.85, Telnet : 0.85, SMTP : 0.85, WWW : 0.85, Internet : 0.5}에서 $\mu_{IDX(p5)}$(Protocol)=0이기 때문에 문제 p_5는 이 질의에 의해 검색될 수 없다. 그러나, 퍼지 문제 검색은 시소러스의 관계 정보 <Protocol, FTP, 0.85> 그리고 <Protocol, SMTP, 0.82>∈T(C)를 통해 'FTP'나 'SMTP'가 의미적으로 'Protocol'과 관련을 가지기 있음을 추론할 수 있기 때문에 이들 색인어를 가지는 문제들까지 검색할 수 있다. 이때, 퍼지 확장 정도 0.5는 시소러스의 확장 범위를 나타내며, 이 값을 통해 사용자가 검색의 재현율과 정확률을 적절히 조절할 수 있다.

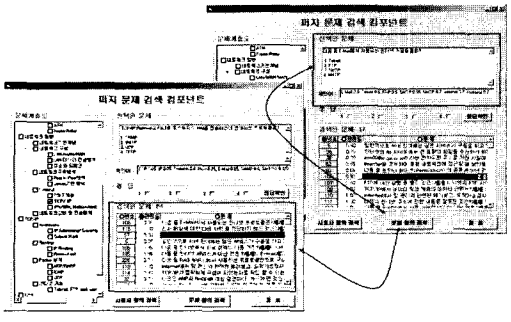

그림 9 문제 질의 검색에 대한 결과 화면

그림 9는 사용자가 현재 학습하고 있는 문제로부터 연상되는 유사한 의미의 다른 문제들을 검색할 수 있도록 지원하는 문제 질의 검색 결과를 나타내고 있다. 이 검색은 선택된 문제의 색인 정보로부터 초기 질의를 자동으로 생성한 다음, 사용자 질의 검색과 같은 방법으로 관련된 문제를 검색하게 된다. 예를 들어, 사용자가 현재 학습하고 있는 106번 문제는 "E-Mail"에 관련된 문제이며, IDX(p_{106})={E-Mail : 0.9, Telnet : 0.5, FRP : 0.5, SMTP : 0.95, NNTP : 0.7, Internet : 0.7, Protocol : 0.7}과 같은 색인 정보를 가지고 있다. 따라서, 이 색인 정보와 불리언 연산자 OR를 이용하여 Q="E-Mail : 0.9 OR Telnet : 0.5 OR FRP : 0.5 OR SMTP : 0.95 OR NNTP : 0.7 OR Internet : 0.7 OR Protocol : 0.7"과 같은 초기 질의를 자동으로 구성할 수 있다. 퍼지 문제 검색 시스템은 사용자 질의 검색과 같은 방법으로 자동 구성된 질의를 처리하여 관련된 문제들을 검색할 수 있다. 이 방법으로 검색된 37번 문제는 실제 "E-Mail"과 의미적으로 상당히 관련된 문제이며, IDX(p_{37})={TCP : 0.85, IP : 0.85, Protocol : 0.8, Host : 0.75, E-Mail : 0.85, SNMP : 0.6, SMTP : 0.9, UDP : 0.6, TFTP : 0.5}와 같은 색인 정보를 가지고 있다.

5. 평 가

일반적으로 정보 검색 시스템에 대한 성능은 정확률(Precision)과 재현율(Recall)로 평가된다. 퍼지 문제 검색 시스템에서 정확률은 검색된 모든 문제들 중 사용자 질의로부터 실제 의미적으로 연상될 수 있는 문제에 대한 비율이며, 재현율은 사용자 질의로부터 실제 연상될 수 있는 모든 문제들 중 검색된 문제에 대한 비율이다. 이들은 아래 식에 의해 정량화될 수 있다.

$$\text{Precision} = \frac{n(R \cap A)}{n(A)}, \quad \text{Recall} = \frac{n(R \cap A)}{n(R)}$$

여기서, n(R)은 사용자 질의에서 연상될 수 있는 모든 문제들의 개수이며, n(A)는 사용자 질의에 의해 검색된 문제의 개수를 나타낸다. 또한, n(R∩A)은 검색된 문제들 중 실제 사용자 질의에서 연상되는 문제들에 대한 개수이다.

본 논문의 퍼지 문제 검색 시스템은 연상학습을 지원한다는 관점에서 일정한 정확률과 높은 재현율의 검색 성능을 나타내야 한다. 이 검색 성능을 다음과 같은 환경에서 평가하였다. 먼저, 문제 데이타베이스는 "네트워크관리사" 자격증 시험 문제들 중에서 225문제를 임의적으로 선택하여 구성하였다. 또한, 문제

질의 검색에 대한 성능은 사용자 질의 검색에 대한 성능으로부터 유추할 수 있기 때문에 사용자 질의 검색만을 평가하였다. 이때, 사용자 질의는 단일 질의, 이접 질의 그리고 연접 질의 각각 1개씩을 사용하였으며, 질의 확장 정도는 0.5로 설정하였다.

사용자 질의로부터 연상될 수 있는 문제를 평가하기 위한 전문가들은 3명의 네트워크 전공 대학원 학생과 2명의 전산학 전공 대학원 학생으로 구성하였다. 이 평가 전문가들은 문제 은행 데이타베이스에서 각각의 사용자 질의로부터 연상될 수 있는 문제를 0과 1사이의 값으로 평가한 다음, 5명의 평균이 0.5 이상인 문제는 사용자 질의로부터 연상될 수 있다고 결정하였다. 이 0.5는 응용 도메인이나 선정된 평가 전문가에 따라 다르게 결정될 수도 있다. 또한, 퍼지 시소러스를 사용하지 않는 단순 검색과 이를 사용하는 퍼지 검색에 대한 각각의 성능을 동시에 평가하여 연상학습에 대한 적합성을 비교하였다.

사용자 질의	n(R)	단순 검색				퍼지 검색			
		n(A)	n(R∩A)	재현율	정확률	n(A)	n(R∩A)	재현율	정확률
단일 질의	25	17	15	0.600	0.882	24	21	0.840	0.875
이접 질의	15	8	6	0.400	0.750	20	14	0.933	0.700
연접 질의	34	28	25	0.735	0.892	30	28	0.823	0.933
계				0.57	0.84			0.86	0.83

그림 10 검색 성능 평가에 대한 비교

그림 10은 퍼지 시소러스를 이용하지 않은 단순 검색과 이를 이용한 퍼지 검색에 대한 성능 평가 결과이다. 여기서, 단일 질의는 프로토콜 중에서 'TCP'로부터 연상되는 문제들이라는 의미에서 Q="TCP", 이접 질의는 'Web' 또는 'E-Mail'과 관련된 프로토콜로부터 연상되는 문제들이라는 의미에서 Q="Web OR E-Mail" 그리고 연접 질의는 인터넷에서 일반적으로 사용되는 프로토콜에서 연상되는 문제들이라는 의미에서 Q="Internet AND Protocol"을 각각 사용하였다. 이 결과에서 퍼지 검색이 단순 검색에 비해 일정한 정확률과 높은 재현율을 요구하는 연상학습의 특성을 잘 반영하고 있음을 알 수 있다. 즉, 단순 검색에 비해 정확률은 단지 1%의 감소를 보였지만 재현율이 29% 향상되었다.

6. 결론 및 향후 연구 과제

본 논문에서는 문제 은행에서 사용자들의 연상학습을 효과적으로 지원할 수 있는 퍼지 검색 시스템을 설계하고 구현하였다. 또한, 정확률과 재현율 평가를 통해 이 검색 시스템이 사용자 연상학습에 매우 적합한 검색 성능을 나타냄을 검증하였다. 퍼지 검색 시스템은 크게 문제 관리 컴포넌트와 문제 검색 컴포넌트로 구성된다. 문제 관리 컴포넌트는 문제들을 데이타베이스에 체계적으로 관리할 수 있도록 지원하며, 문제 검색 컴포넌트는 효과적인 연상학습을 위해 사용자 질의를 통한 검색과 문제 질의를 통한 검색을 동시에 지원한다. 전자는 사용자 질의로부터 개념적으로 연상되는 문제들을 검색할 수 있게 하며, 후자는 사용자가 현재 학습하고 있는 문제로부터 연상되는 다른 문제들을 검색할 수 있게 한다. 여기서, 퍼지 시소러스는 사용자 질의의 탐색어들과 문제의 색인어들 사이의 용어 불일치 문제를 해결할 수 있게 하여, 일정한 정확률과 높은 재현율의 검색 성능을 유지시킴으로써 문제 컨텐츠를 통한 학습자의 연상학습을 효과적으로 지원할 수 있게 한다. 따라서, 현재 인터넷에서 운영되는 여러 교육 지원 사이트의 문제 은행 검색 서비스에 본 논문에서 제안하는 퍼지 검색 방법을 적용한다면, 사용자의 연상학습에 대한 욕구를 상당히 만족시킬 수 있을 것이다.

본 논문의 향후 연구과제로 이 퍼지 문제 검색 방법을 실제 응용 도메인에 적용하기 위해서는 자동 색인 방법을 이용하여 문제 은행 데이타베이스를 쉽게 구축할 수 있도록 해야 한다. 또한, 해당 응용 도메인에 적합한 퍼지 시소러스를 구축하고 인터넷 환경에 적합할 수 있도록 본 시스템을 JAVA나 ASP로 변환해야 한다.

참 고 문 헌

[1] 이기호, 최윤희, "웹 그룹웨어 원격 교육 시스템의 설계 및 구현", 정보과학회 논문지(C), Vol. 4, No. 1, pp. 126-134, 1998.

[2] Daolsoft Inc, "e-bussiness 문제은행 Solution", http://www.daolsoft.com, 2000.

[3] Icedu Inc, "자격증 문제은행", http://www.icedu.com, 2000.

[4] Oedae Language Institute, "인터넷 외국어 강좌", http://www.toptop.co.kr, 1999.

[5] W. B. Croft, "*What Do People Want from Information Retrieval?*," D-Lib Magazine, 1995.

[6] H. J. Peat and P. Willett, "*The Limitation of Term Co-occurrence Data for Query Expansion in Document Retrieval System,*" Journal of the American Society for Information Science, Vol. 42, No. 5, pp. 378-383, 1991.

[7] J. Y. Nie and M. Brisebois, "*An Inferential Approach to Information Retrieval and its Implementation using a Manual Thesaurus,*" Artificial Intelligence Review, Vol. 10, No. 5, pp. 409-439, 1996.
[8] MediArc Inc, "원격가상교육", http://dasan.sejong.ac.kr/~inlee/class/99-2comp/edudpt /edu2/won4.htm, 1999.
[9] J. H. Lee, M. H. Kim and Y. J. Lee, "*Ranking Documents in Thesaurus-Based Boolean Retrieval Systems,*" Information Processing and Management, Vol. 30, No. 1, pp. 79-91, 1994.
[10] J. H. Lee, M. H. Kim and Y. H. Cho, "*Using Term Dependencies of a Thesaurus in the Fuzzy Set Model,*" Microprocessing and Microprogramming, Vol. 39, pp. 105-108, 1993.
[11] S. K. M. Wong and Y. Y. Yao, "*A Generalized Binary Probabilistic Independence Model,*" Journal of the American Society for Information Science, Vol. 41, No. 5, pp. 342-329, 1990.
[12] B. Y. Ricardo and R. N. Berthier, *Modern Information Retrieval, Addison-Wesley*, 2000.
[13] Y. Qiu, "*Automatic Query Expansion Based on a Similarity Thesaurus,*" Ph. D. Thesis, ETH Zurich, Institute of Computer Systems, 1995.
[14] N. Fuhr, "*Probabilistic Models in Information Retrieval,*" The Computer Journal, Vol. 35, No. 3, pp. 243-255, 1992.
[15] H. L. Larsen and R. R. Yager, "*The Use of Fuzzy Relational Thesauri for Classificatory Problem Solving in Information Retrieval and Expert Systems,*" IEEE Transactions on Systems, Man and Cybernetics, Vol. 23, No. 1, pp. 31-41, 1993.
[16] 최재훈, 김기헌, 양재동, "객체기반 시소러스 시스템의 설계 및 구현: 반자동화 방식의 구축, 추상화 방식의 개념 브라우징 및 질의기반 참조", 정보과학회논문지(데이타베이스), Vol. 27, No. 1, 2000.
[17] C. Buckley, G. Salton and J. Allan, "*The Effect of Adding Relevance Information in a Relevance Feedback Environment,*" In Proceedings of the 17th Annual International ACM/SIGIR Conference, pp. 292-300, Dublin, Ireland, 1994.
[18] M. Mitra, A. Singhal and C. Buckley, "*Improving Automatic Query Expansion,*" In Proceedings of the 21th Annual International ACM/SIGIR Conference, pp. 206-214, Melbourne. Australia, 1998.
[19] H. Schuetze and J. O. Pedersen, "*A Co-occurrence Based Thesaurus and Two Applications to Information Retrieval,*" Information Processing and Management, Vol. 33, No. 3, pp. 307-318, 1997.
[20] 신정훈, 안윤애, 류근호, 박현주, "문헌검색을 위한 지식기반 질의 처리기 구현", 정보과학회논문지(C), Vol. 3, No. 5, pp. 522-532, 1997.
[21] M. M. Gupta and J. Qi, "*Theory of t-norms and Fuzzy Inference Methods,*" Fuzzy Sets and Systems, Vol. 40, No. 3, pp. 431-450, 1991.

최 재 훈
1994년 전북대학교 전자계산학 학사 졸업. 1996년 전북대학교 전산통계학과 석사 졸업. 2000년 전북대학교 전산통계학과 박사 졸업. 2000 ~ 현재 한국전자통신연구원 선임연구원. 관심분야는 Bio-Informatics, DBMS, Software Engineering, System Modeling, Fuzzy Theory, Information Retrieval, Multimedia Processing

김 지 숙
1998년 전북대학교 컴퓨터과학과 학사 졸업. 2001년 전북대학교 교육대학원 교육학석사 졸업. 관심분야는 ICT 교육 시스템, 정보검색, 데이터베이스, 멀티미디어 처리, 퍼지 시스템

조 기 환
1985년 전남대학교 계산통계학과 학사 졸업. 1987년 서울대학교 계산통계학과 석사 졸업. 1996년 영국 Newcastle 대학교 전산학과 박사 졸업. 1987~97년 한국전자통신연구원 선임연구원. 1997 ~99년 목포대학교 컴퓨터과학과 전임강사. 1999년 ~ 현재 전북대학교 전자정보공학부 조교수. 관심분야는 이동컴퓨팅, 무선인터넷, 컴퓨터통신.