2단계 유사관계 행렬을 기반으로 한 순위 재조정 검색 모델

(A Re-Ranking Retrieval Model based on Two-Level Similarity Relation Matrices)

이 기 영 † 은 희 주 †† 김 용 성 †††
(Ki-Young Lee) (Hye-Ju Eun) (Yong-Sung Kim)

요 약 웹 기반의 학술분야 전문 검색 시스템은 사용자의 정보 요구 표현을 극히 제한적으로 허용함으로써 검색된 정보의 내용 분석과 정보 습득의 과정이 일관되지 못해 무분별한 정보 제공이 이루어진다. 본 논문에서는 용어의 상대적인 중요 정도를 축소용어 집합으로 구성하여 검색 시스템의 높은 시간 복잡도를 해결할 수 있도록 퍼지 검색 모델을 적용하였다. 또한 퍼지 호환관계의 특성을 만족하는 유사관계 행렬을 통해 사용자 질의를 정확하게 반영할 수 있도록 클러스터 검색을 수행하였다. 본 논문에서 제안한 퍼지 검색과 문서 클러스터 검색의 유사도 결합을 통한 순위 재조정 검색 모델은 검색 성능을 표현하는 정확률과 재현율 척도에서 향상됨을 입증하였다.

키워드 : 축소용어, 퍼지 호환관계, 유사관계 행렬, 클러스터 검색, 순위 재조정 모델

Abstract When Web-based special retrieval systems for scientific field extremely restrict the expression of user's information request, the process of the information content analysis and that of the information acquisition become inconsistent.

In this paper, we apply the fuzzy retrieval model to solve the high time complexity of the retrieval system by constructing a reduced term set for the term's relatively importance degree. Furthermore, we perform a cluster retrieval to reflect the user's query exactly through the similarity relation matrix satisfying the characteristics of the fuzzy compatibility relation. We have proven the performance of a proposed re-ranking model based on the similarity union of the fuzzy retrieval model and the document cluster retrieval model.

Key words : Reduction Term, Fuzzy Compatibility Relation, Similarity Relation Matrices, Cluster Retrieval, Re-ranking Retrieval Model

1. 서 론

웹 서비스 응용 기술의 발달과 인터넷 보급으로 인하여 정보는 기하급수적으로 생산·저장되며 사용자는 전 세계에 산재된 수많은 정보를 손쉽게 획득할 수 있게 되었다. 현재 웹 기반의 학술분야 전문(이하, 문서라고 기술함.) 검색 시스템은 사용자의 관심도를 자연어 형태의 질의로 표현하여 문서에서의 발생 여부에 따라 검색 결과를 제공하는 키워드 매칭 기법을 적용한다. 그러나 이러한 기법은 동의어 및 다의어 처리의 문제로 인하여 선별되지 않은 정보를 검색 결과로 제시함에 따라 반복적인 피드백 필터링 과정을 거쳐야 하는 번거로운 작업을 요구한다[1-4].

정보검색 모델은 다양한 주제들에 대하여 방대한 정보를 찾기 쉬운 형태로 조직화함으로써 수많은 사용자가 원하는 정보에 대하여 빠른 접근이 가능해야 한다. 이러한 관점에서 사용자 위주의 검색 시스템은 사용자의 관심도 및 정보 요구를 정확하게 표현해야 하며 검색된 결과 또한 사용자의 만족도를 충족시킬 수 있도록 문서의 내용 분석 과정과 정보 요구에 대한 정보 습득 과정이 일관된 메커니즘이 되어야 한다.

사용자의 관심도 및 정보 요구의 정확한 표현을 위해 지금까지 개인 또는 그룹 프로파일(profile) 이나 시소러

† 정 회 원 : 원광보건대학 컴퓨터응용개발과 교수
lky@wkhc.ac.kr
†† 비 회 원 : 전북대학교 전산통계학과
hjeun@chonbuk.ac.kr
††† 종신회원 : 전북대학교 전자정보공학부 교수
yskim@chonbuk.ac.kr
논문접수 : 2004년 5월 19일
심사완료 : 2004년 9월 6일

스(thesaurus)를 이용하여 질의어를 확장하는 방법이 많이 연구 진행되고 있다. 그러나 프로파일이나 시소러스 구축에 많은 시간이 소요될 뿐만 아니라 구축된 시소러스에 대해 구조화 및 적합성에 대한 문제가 발생한다. 즉, 용어 사이의 관계성 불일치 문제로 인하여 검색 시간의 증가, 거리 알고리즘에서 NOT 연산자 사용, min-max 함수를 이용한 OR 연산의 비효율성 문제 등이 발생한다[5].

따라서 본 논문에서는 시소러스에 대한 문제점을 해결하기 위해서 개념 정보(concept information)를 포함하고 있는 용어들의 유사도(similarity) 및 개념 거리(concept distance)를 이용하고자 한다. 즉, 사용자 관심을 표현하는 질의와 문서의 구조화 지식인 표제, 요약, 키워드에 대한 의미적 논리 구조를 기반으로 추출한 색인어 사이의 유사 정도를 퍼지 값으로 사상시켜 질의어를 확장하고자 한다. 또한 퍼지 관계 개념을 적용하여 용어의 상대적인 개념 정도의 표현과 색인어 사이의 종속성 문제를 해결하며, 시스템의 높은 시간 복잡도(time complexity)는 축소용어 집합(reduced term set)을 이용하여 처리하고자 한다. 그리고 사용자 질의 용어와 문헌의 색인어 사이의 내용 기반 유사도를 반영한 순위 재조정 모델을 제안한다. 다시 말해, 순위 재조정은 대상 문서에서 저자의 의도를 추정하는 색인어를 추출하는 주제 분석 단계와 사용자 관심도를 정확하게 파악하여 관련문서를 효율적으로 검색하는 요구 분석 과정으로 구성된다. 이를 위해 전자는 시소러스 및 유사관계 행렬(similarity relation matrix)을 구축하여 주제 분석 메커니즘을 제공하고, 후자는 사용자 요구를 분석하기 위해 질의 확장 등의 탐색 모형을 수립하는 알고리즘을 도입한다. 따라서 본 논문에서 제안한 알고리즘은 검색 시스템의 재현율을 유지하면서 동시에 기존 퍼지 검색 시스템의 단점인 정확률을 향상시키는 2단계 내용 기반 검색 기법이라 할 수 있다.

본 논문의 구성은 2장에서 주제 분석 과정 및 탐색모형에 대한 기존 연구를 설명하고, 3장에서는 도메인 지식을 표현하는 축소용어 집합, 시소러스 그리고 유사 관계 행렬을 생성하는 주제 분석 메커니즘과 퍼지 시소러스, 유사관계 행렬을 통한 검색 알고리즘을 제안하고 4장에서는 제안된 알고리즘을 평가 검증하기 위해서 실험과 평가를 한다. 마지막으로 5장에서는 결론 및 향후 연구과제에 대하여 기술한다.

2. 관련 연구

이 장에서는 본 논문과 관련이 있는 연구 현황 및 개념 네트워크, 퍼지 이론에 관해서 기술한다.

2.1 정보검색 모델의 연구 현황

대부분의 정보검색 모델은 불리언(boolean) 검색 및 벡터(vector) 기반 모델을 기반으로 하고 있다. 불리언 검색은 사용자 질의를 구성하기에는 편리하나 질의로 표현되는 각 개념의 상대적인 중요도를 나타내지 못하는 단점이 있다. 반면에 벡터 기반 모델은 사용자 질의 조건에 근접한 문서 검색이 가능하고 적합 순위를 부여할 수 있는 장점은 있으나 용어간의 가중치 부여에 의한 종속성이 단점으로 지적된다[3,5-7].

이러한 문제점을 해결하기 위하여 [8]의 퍼지 정보검색 모델은 불리언 정보검색 모델을 확장한 것으로 시소러스 자동구축 기능, 퍼지화된 우선순위 제공, 직접 관련 없는 제3의 개체 유추 검색 등을 제공하고 있다.

또한 잠재되어 있는 높은 시간 복잡도의 검색 연산을 낮추기 위한 확장된 연구로는 축소용어 집합을 적용한 연구 결과가 있다[8,9]. 그리고 대상 문서가 포함하는 저자의 의도를 정확하게 추론하여 파악함으로써 영역지식을 표현하는 자동 키워드 색인에 대한 연구는 광범위하게 연구되고 있다. 한편, 기존의 정보검색 시스템이 채택하고 있는 주제 분석 추론 과정은 일반적으로 통계적 속성에 기초한 것으로 문서를 대표하는 색인어 관계에 의해 문서집합을 정의하므로 색인어 문서 빈도나 색인어 사이에 존재하는 통계적 의존성을 주제 영역에서의 지식으로 이용한다.

그러나 사용자가 원하는 정보를 손쉽게 검색하기 위해 색인 정보는 특정한 체계나 규칙에 의해 적절하게 구조화되어야 한다[10]. 즉, 질의를 구성하는 탐색어 집합을 검색하기 전에 문서의 의미적 지식구조와 형태적 논리구조를 주제 분석에 반영하여 문서 클러스터를 수행함으로서 탐색어와 색인어 사이의 용어 불일치 문제를 해결할 수 있다[10,11]. 그러나 문서 클러스터를 구축하기 위해서는 문서를 정확하게 표현할 수 있는 특성(feature)을 결정해야 하고, 클러스터 사이의 구분이 명확하게 표현될 수 있도록 특성을 찾아야하는 문제점이 있다.

따라서 본 논문에서는 클러스터 검색의 이러한 문제점을 해결하고자 [12,13]을 응용하여 주제 분석 과정인 개념 네트워크를 정의하고, 이를 통한 탐색 모형을 제안하였다.

2.2 개념 네트워크

개념 네트워크는 노드와 링크로 구성되며 각 노드는 개념이나 문서를 표현하고 링크는 개념들 사이의 의미적인 관계 또는 문서와 개념 사이의 관계를 표현한다. 예를 들어 문서집합 $D=\{d_1, d_2, d_3, d_4\}$와 색인어로 구성된 개념 집합 $C=\{c_1, c_2, \cdots, c_7\}$을 개념 네트워크

로 나타내면 그림 1]과 같이 표현할 수 있으며 이것은 영역 전문가에 의해서 구성된 개념들 사이의 의미적인 연결을 나타낸다.

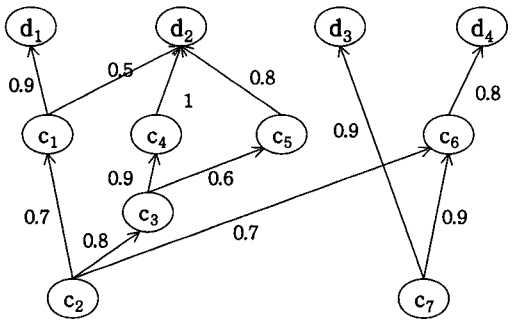


그림 1 개념 네트워크

이와 더불어 모든 도메인 개념들의 집합 $C=\{c_1, c_2, \cdots, c_7\}$에 대해 개념 네트워크 F(C)는 다음과 같은 관계 집합으로 표현될 수 있다.

$$F(C)=\{\langle c_1, c_2, w_{1.2}\rangle \mid c_1, c_2 \in C, w_{1.2} \in [0,1]\}$$

여기서, 관계$\langle c_1, c_2, w_{1.2}\rangle \in F(C)$은 c_1은 하의어로 c_2를 가지며, 두 개념들 사이의 퍼지 관련 정도가 $w_{1.2}$임을 의미한다. 이에 따라 그림 1에서 문서d_2는 $\langle d_2, c_1, 0.5\rangle$, $\langle d_2, c_4, 1\rangle$, $\langle d_2, c_5, 0.8\rangle \in F(C)$ 이다. 그리고 그림 1에서 이와 같은 관계들로부터 의미적으로 연결되는 하위 집합간의 관계는 전이적 성질(퍼지 이행관계)을 이용한다.

예를 들면$\langle c_1, c_2, w_{1.2}\rangle, \langle c_1, c_3, w_{1.3}\rangle, \langle c_2, c_4, w_{2.4}\rangle, \langle c_3, c_4, w_{3.4}\rangle \in F(C)$이면 $\langle c_1, c_4, w_{1.4}\rangle \in F(C)$가 성립한다. c_1와 c_4 사이의 퍼지 관련정도 $w_{1.4}$은 다음과 같은 자데(Zadeh) 퍼지 확장 논리(식 (1))를 이용하면 된다.

$$w_{i,l}=\max\{\min(w_{i,j}, w_{j,l}), \min(w_{i,k}, w_{k,l})\} \quad (1)$$

그림 1]에서 사용자 질의가 $Q=\{(c_2, 1.0)\}$일 경우, 문서 d_2에 대한 검색 상태 값(RSV)은 자데의 전이관계를 이용하여 다음과 같이 3가지의 절차에 의해서 평가된다 [12-14].

1) $c_2 \rightarrow c_1 \rightarrow d_2$ 이므로

$\min(w_{2.1}, w_{1.2})=\min(0.7, 0.5)=0.5$ 이다.

2) $c_2 \rightarrow c_3 \rightarrow c_4 \rightarrow d_2$ 이므로

$\min(w_{2.3}, w_{3.4}, w_{2.2})=\min(0.8, 0.9, 1)=0.8$ 이다.

3) $c_2 \rightarrow c_3 \rightarrow c_5 \rightarrow d_2$ 이므로

$\min(w_{2.3}, w_{3.5}, w_{5.2})=\min(0.8, 0.6, 0.8)=0.6$ 이다.

따라서 문서 d_2의 검색상태 값은 $\max(0.5, 0.8, 0.6)=0.8$로서 평가되어 상위의 상태 값을 가진 문서가 사용자에게 제시된다. 이와 같이 도메인 개념들 사이의 의미관계를 퍼지 정도로 표현한 개념 네트워크를 통해 사용자 질의와 개념적으로 서로 연관된 문서들을 검색하여 재현율을 향상시킬 수 있다. 그러나 이와 같은 방법은 영역 전문가에 의해서 개념 네트워크가 수동으로 유지되기 때문에 구축 및 유지 관리가 어렵고 다른 응용에 어려움이 있는 단점이 있다.

이를 해결하기 위해서는 개념 네트워크를 자동화하고, 탐색 과정은 다양한 탐색 요구를 지원하여 연관 검색이 가능하도록 지원하여야만 검색의 효율성을 높일 수가 있다. 또한 개념 네트워크와 더불어, 주제 분석과 탐색 모형에 대한 대표적인 연구로는 퍼지 관계를 이용한 검색 기법[5,7,9,15]과 유사관계 시소러스를 기반으로 자동 질의 확장[1,2,12,13] 등이 있다. 그러나 주제 분석 메커니즘과 사용자 요구를 분석하기 위한 탐색모형을 수립하는 메커니즘이 일관성이 없고 검색성능에서 정확률에 대한 한계를 나타낸다.

따라서 본 논문에서는 주제 분석 및 탐색 모형을 일관성 있도록 구현하여 재현율을 유지하면서 기존 퍼지 검색 시스템의 단점인 정확률을 향상시키기 위한 2단계 내용기반 탐색 모형을 제안하고자 한다.

2.3 퍼지 이론과 퍼지검색 모델

이 절에서는 본 논문에서 적용하는 퍼지 집합과 퍼지 집합 사이의 관계성, 그리고 퍼지검색모델에 관해서 기술한다.

2.3.1 퍼지 함수

퍼지 집합 A가 임의의 전체 집합 $X=x$에 대하여 [0, 1]값으로 표현되기 위해서는 $x=x_0$에 대해 집합 A의 소속 정도(membership degree)를 나타내는 소속 함수(membership function)는 다음과 같이 정의된다.

$$\mu_A: X \rightarrow [0,1] \quad (2)$$

이진 퍼지 관계(binary fuzzy relation)는 임의의 퍼지 집합 $x \in X, y \in Y$ 사이의 관계를 순서쌍(x, y)로 나타내고 (x, y)의 모임을 관계 R로 표시한다. 따라서, 주어진 이진 퍼지 관계 $R(x, y)$에 대해서, 각 $x \in X, y \in Y$에 대해 정의역 $dom\ R(x)=\max R(x, y)$ 이고, 치역 $ran\ R(y)=\max R(x, y)$이다.

이에 따라 $x \in X$, $y \in Y$에 대해서 이진관계 R의 소속 함수는 다음과 같이 정의된다.

$$\mu_R: X \times Y \rightarrow [0,1] \quad (3)$$

본 논문에서는 문서에서 발생한 색인어의 빈도를 문서에서의 소속 정도로 변환하기 위해 인공지능 시스템의 학습 알고리즘에 많이 적용되고 있는 시그모이드(sigmoid) 함수를 적용한다. 이 함수는 다음의 세 가지 특징을 만족한다.

1) $\sigma : R^{+} \rightarrow [0,1]$
2) $\sigma(F_1) > \sigma(F_2) \Leftrightarrow F_1 > F_2$ (4)
3) $\frac{d^2(\sigma)}{dF^2} \geq 0 \Leftrightarrow F \leq T_F$ and $\frac{d^2(\sigma)}{dF^2} \leq 0 \Leftrightarrow F \geq T_F$

식 (4) 중 첫 번째 조건식은 입력 값에 무관하게 항상 [0, 1]사이의 퍼지 값을 갖고, 두 번째 조건식은 S자 형태의 단조 증가 함수이며 마지막 조건식은 임계값(critical value)을 갖는 퍼지 소속 함수임을 의미한다[5,7]. 식 (4)를 만족하는 시그모이드 소속 함수에 대한 그래프는 그림 2와 같이 나타낼 수 있다.

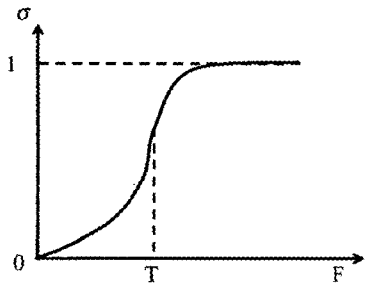


그림 2 시그모이드 함수

본 논문에서는 문서내의 색인어 발생 영역에 따른 빈도를 문서의 내용과 의미를 나타내는 중요도로 사상시키기 위해서 그림 3과 같이 각기 다른 임계 값을 갖는 시그모이드 소속 함수들을 정의하고, 이러한 각각의 임계 값은 질의어 확장과 문서 분류에 적용된 $\alpha-cut$의 알파(α) 값으로 이용한다.

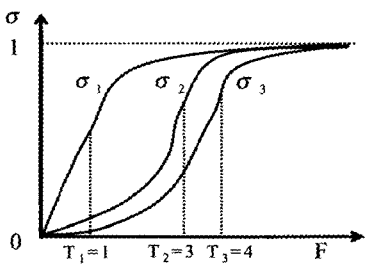


그림 3 정의한 시그모이드 함수

이를테면, T_1=1이고, T_2=3 그리고 T_3=4일 때, 임계 값을 갖으며 이 값은 문서에서 추출된 색인어의 발생 영역과 빈도에 따라 문서 의미의 중요도를 달리하는데 이용한다.

2.3.2 퍼지 유사관계 및 호환 관계

일반적인 집합에서의 이진관계 $R(x,x)$가 반사관계, 대칭관계, 전이관계를 만족하면 동치관계라 부른다. 임의의 집합 X에서 각각의 원소 x, y에 대하여 동치관계를 만족하는 집합 X의 모든 원소들을 포함하는 집합 A_x는 다음과 같이 관계로 표현할 수 있다.

$$A_x = \{y \mid (x,y) \in R(x,x)\} \quad (5)$$

집합 X가 퍼지 집합일 때, 임의의 $x, y, z \in X$에 대하여 정의된 퍼지 관계 $R \subseteq X \times X$이 다음과 같이 반사관계, 대칭관계, 전이관계가 정의될 때 퍼지 유사관계($\cong$)(similarity relation)라고 부른다.

1) 반사관계: $\mu_{\cong}(x,x) = 1$
2) 대칭관계: $\mu_{\cong}(x,y) = \mu_{\cong}(y,x)$
3) 전이관계: $\mu_{\cong}(x,z) \geq \min\{\mu_{\cong}(x,y), \mu_{\cong}(y,z)\}$
단, μ는 소속함수 (6)

퍼지 집합에서 유사관계를 만족하는 임의의 원소들은 퍼지 집합에 대한 소속정도의 값을 부여받고 일정 이상의 소속정도 값을 가진 유사클래스(similarity class)를 생성하여 분류와 그룹화 할 수 있다. 또한 퍼지 관계 중 호환관계($\approx$) (compatibility relation)는 다음과 같이 반사와 대칭 성질을 만족하지만 전이관계는 만족하지 않는다.

1) 반사관계: $\mu_{\approx}(x,x) = 1$
2) 대칭관계: $\mu_{\approx}(x,y) = \mu_{\approx}(y,x)$ (7)

퍼지 집합에서는 퍼지 관계 R이 퍼지 호환 관계를 만족할 때 특정한 소속정도 α값을 선택하여 $\alpha-cut$을 적용하면 호환 클래스 집합 A_α을 생성할 수 있다.

$\alpha-cut$은 임의의 α ($0 \leq \alpha \leq 1$)값이 되는 함수 값에 대한 퍼지 상태 변수의 구간을 나타내며, 퍼지 집합의 원소들에 대해서 집합에 속할 기준을 정의할 때 사용된다. 임의의 X을 원소로 하는 퍼지 집합 A에 대해서 임의의 $\alpha \in [0,1]$ 값을 가진 $\alpha-cut$을 적용한 퍼지 집합 A_α는 다음과 같이 정의한다[13,14,17].

$$A_\alpha = \{x \mid A(x) \geq \alpha\} \quad (8)$$

따라서 퍼지 집합 A_α는 퍼지 집합에 속할 소속정도의 값이 α값 이상으로 이루어진 집합이다. α-호환 클래스는 임의의 x, y에 대하여 이들 관계가 α값 이상이면 X의 부분집합으로 구성된다. 이렇게 구성된 모든 호환 클래스들을 최대 호환 클래스 또는 완전 $\alpha-cover$라고 하고, 어떤 특정한 부류로 커버(*cover*)됨을 의미한다[17].

2.4 개선된 BK-퍼지 정보검색 모델(A-FIRM)

AFIRM(Advanced bandler-kohout Fuzzy information Retrieval Model)은 문서와 용어의 상대적인 의존도를 통계 확률적 기법에 의해 원시 문서베이스를 구성하고 퍼지 관계와 퍼지 관계곱을 이용하여 용어관계 시소러스를 자동화하였다[9,19].

용어관계 시소러스를 통해 용어 개념을 확장하는 관

계요구(R-request) 연산을 제공하며, 사용자 질의를 해석하고 검색하는 퍼지 검색요구(FS-request) 연산도 제공한다. 그림 4는 문서베이스 생성기와 시소러스 생성기를 통해 문서베이스, 시소러스를 자동으로 구축함으로서 영역 의존적인 지식을 추출하고 정보 제공자인 저자의 의도를 파악할 수 있다.

검색요구 과정은 시소러스와 질의 용어의 퍼지 합성을 통한 질의 확장 방법을 채택하였다. 또한 시간 복잡도의 검색 연산을 최소화하기 위해서 축소용어 집합을 추출하여 시소러스 및 문서베이스를 생성하는 것이 특징이다.

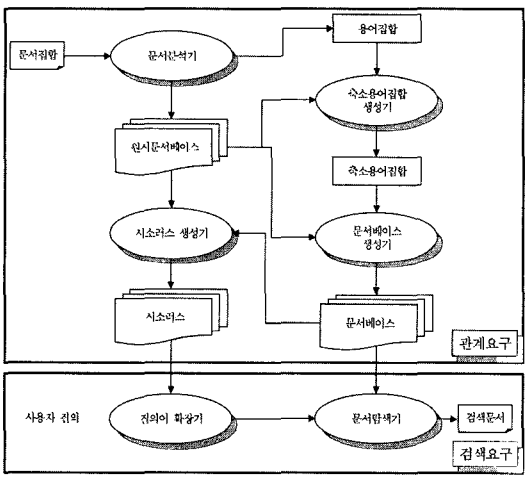


그림 4 A-FIRM의 시스템 구조

또한 A-FIRM은 높은 시간 복잡도를 해소하여 재현율을 향상시킬 수 있으며 문서 검색 결과의 우선순위를 제공하는 장점을 지니고 있다. 그러나 축소용어 집합에 대한 기준이 마련되어 있지 않고, 검색결과 우선 순위 조정에 대한 대비책이 단순하여 사용자가 느끼는 정확률에 대한 향상은 기대할 수가 없다.

따라서 본 논문에서는 그림 4의 시간 복잡도를 개선하면서 정확률을 향상시킬 수 있도록 퍼지 논리와 유사관계 행렬을 기반으로 한 순위 재조정 모델을 제안한다.

3. 문서 순위 재조정 알고리즘

본 논문에서는 학술분야 전문 정보검색을 위하여 표제와 초록에 대한 문서 구조적인 지식을 바탕으로 시소러스 및 유사관계 행렬을 구축하는 방법과 문서 순위 재조정 알고리즘을 제안한다.

3.1 시그모이드 함수를 이용한 문서베이스 생성

원시 문서베이스를 구성하기 위해서는 문서 집합에서 용어를 추출하고 용어와 문서사이의 관련 정도를 가중치로 표현한다. 본 논문에서는 용어 출현 빈도가 문서 내용을 대표한다는 가설에 근원으로 특정 문서에 대한 용어의 의미 관계를 퍼지화 하였다[5,8,19]. 이를 위해 퍼지 소속 함수로 대표적인 비선형 함수인 S자 형태의 시그모이드 함수를 이용하며, 표제 및 키워드에 발생한 색인어 빈도에 대한 임계 값은 1, 요약에 대한 색인어 임계 값은 2로 할당하였다. 시그모이드 함수 $\sigma(F)$는 다음의 조건을 만족한다(단, 도메인의 특성에 따라 임계 값은 유동적이다). 첫째, 문서에서 추출한 색인어가 문서의 타이틀(T)이나 키워드 집합(K)에서 발생되었을 때 색인어 발생 빈도에 대한 문서에서의 중요 정도는 그림 5의 시그모이드 함수(σ_1)에 의해 표 1과 같이 구할 수 있다.

표 1 타이틀, 키워드집합에서 소속 정도

F	0	1	2	3	4
σ_1	0	0.6	0.9	0.99	1

위의 표 1을 시그모이드 소속 함수를 적용하면 그림 5와 같다.

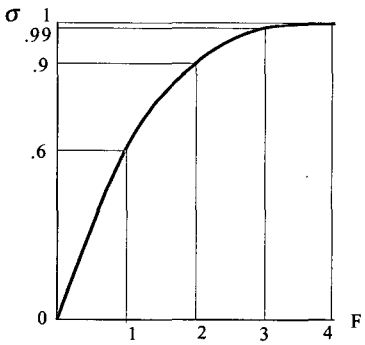


그림 5 소속함수 (σ_1)

그림 5는 타이틀이나 키워드 집합에서 빈도수가 2이상이면 매우 소속정도가 크며, 1인 경우도 소속정도가 0.6으로 대단히 크다.

둘째, 색인어가 문서의 요약부분(A)에 발생되었을 경우 빈도에 대한 소속 정도는 그림 6의 시그모이드 함수 (σ_2)에 의해 표 2와 같이 구할 수 있다.

표 2 요약에서의 소속 정도

F	0	1	2	3	4	5	6
σ_2	0	0.1	0.25	0.7	0.92	0.97	1

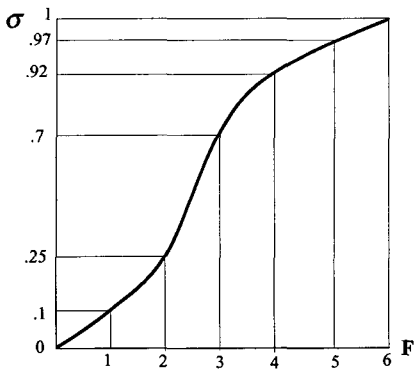


그림 6 소속함수 (σ_2)

시그모이드 함수로 나타내면 그림 6과 같다.

즉, 요약 부분에서 발생하는 색인어의 빈도수는 3회 정도가 되어야 소속정도가 중요하다는 것을 나타낸다. 이와 같이 시그모이드 함수는 상대적인 확률적 빈도(*probabilistic frequency*)를 절대적인 가능성(possibility) 값으로 사상시킨다[1,2,15]. 그리고 발생 영역별 빈도에 대한 중요도를 구분하여 생성하고, 문서 전체에 대한 중요도를 산출하기 위하여 본 논문에서는 [5]의 연구를 응용하여 문서를 대표하는 최종적인 색인어 가중치는 발생 영역의 가중치를 min-max 연산을 적용하여 생성하였다.

$$w_{ij} = \max\{\min(\mu_{ij}^{T}, \mu_{ij}^{A}), \min(\mu_{ij}^{T}, \mu_{ij}^{K}), \min(\mu_{ij}^{A}, \mu_{ij}^{K})\} \quad (9)$$

μ_{ij}^{T}: 문서 j에서 타이틀영역의 색인어 i에 대한 중요 정도

μ_{ij}^{A}: 문서 j에서 요약영역의 색인어 i에 대한 중요 정도

μ_{ij}^{K}: 문서 j에서 키워드영역의 색인어 i에 대한 중요 정도

식 (9)에 의하여 문서집합(D)와 색인어 집합(T)의 퍼지 관계인 원시 문서베이스는 다음과 같이 표현된다..

$$R = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_m \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{mn} \end{pmatrix} \end{array}$$

여기서, 문서 집합의 개수는 n이고 문서 집합에서 추출된 색인어는 m개이다. 본 논문에서 색인을 추출하는 과정으로 수동 색인을 채택한다. 채택한 주요 판점은 자동 색인보다는 추출된 색인어의 의미적인 종속관계를 나타내는 시소러스, 유사관계 행렬을 용이하게 구축하여 연관지식을 추출하고 내용기반 검색을 지원하기 위한 질의확장 모델을 제안하기 위함이다. 또한 실험 집합(KT-Set 2.0)에서 키워드 집합을 1차 색인으로 하고 간단한 수작업을 통해 색인용어를 최종적으로 결정한다.

<예 1> 문서영역이 표제, 키워드, 요약으로 구성된 문서집합인 KT-Set(2.0)을 대상으로 색인어 발생 빈도별 소속 값을 생성하고 문서 집합(D)에서 색인어(T) 의미를 표현하는 원시문서베이스(R)은 식 (9)를 적용하여 구성하면 다음과 같다.

$D=\{d_1, d_2, d_3, d_4, d_5\}$

$T=\{t_1, t_2, t_3, \cdots, t_9\}$

$$= \left\{ \begin{array}{l} \text{다층퍼셉트론, 가중치행렬, 경험적방법, 뉴런,} \\ \text{계층적구조, 관계모델, 관계데이타모델, 문자인식, 관계대수} \end{array} \right\}$$

$$R = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{ccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 \end{array} \\ \begin{pmatrix} 0.94 & 0.50 & 0.50 & 0.80 & 0.00 & 0.50 & 0.50 & 0.80 & 0.00 \\ 1.00 & 0.00 & 0.50 & 0.00 & 0.94 & 0.00 & 0.50 & 0.00 & 0.94 \\ 0.94 & 0.00 & 0.80 & 1.00 & 0.00 & 0.00 & 0.80 & 1.00 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix} \end{array}$$

3.2 시간 복잡도와 축소용어 행렬

퍼지 정보검색 시스템은 높은 시간 복잡도를 가지는 검색 연산을 필요로 하여 적용 분야에 많은 제약이 따른다[9]. 따라서 본 논문에서는 도메인에 의존적인 축소용어 행렬을 생성하며 이를 기반으로 유사관계 시소러스를 구성하는 방법을 제안한다.

(1) 원시 문서베이스와 축소용어 행렬

축소용어 집합은 원시 문서베이스를 이용해서 생성하며 용어 집합의 부분집합으로 높은 시간 복잡도 문제를 처리한다. 본 논문에서는 [1,2]에서 시소러스를 구축하기 위해 각 색인어의 관계 값을 도출하는 방법을 적용하였으며 문서 집합에서도 각 색인어의 관계 값을 생성한다. 이는 두개의 퍼지 집합 사이의 동치관계를 논리적 동치인 불리언 대수를 퍼지 집합에 적용하기 위해 다음과 같이 퍼지 소속 함수로 표현한다.

$$\left.\begin{array}{l} u_{A\cong B}(w) = \max\{\min\ (u_A(w), u_B(w)), \\ \qquad\qquad \min\ (1-u_A(w), 1-u_B(w))\} \\ t_i = u_{w_i \cong w_j} = \dfrac{1}{|d|} \sum_{k=1}^{|d|} u_{w_i \cong w_j}(D_k) \end{array}\right\} \quad (10)$$

$u_A(w)$: 임의의 원소 w가 퍼지 집합 A에 속할 정도

$u_B(w)$: 임의의 원소 w가 퍼지 집합 B에 속할 정도

t_i : 색인어 i가 문서집합(도메인)에서의 관계 정도

$|d|$: 전체 문서의 개수

$u_{w_i \cong w_j}(D_k)$: 문서 k에서 색인어 i, j 사이의 유사정도

또한 축소용어 집합은 각 용어의 퍼지 값에 대하여 $\alpha-cut$을 적용함으로써 도메인 영역에서 문서를 분류하기에 부적합한 색인어를 제거할 수 있는 장점을 갖고 있다. 본 논문에서는 식 (10)의 소속 함수를 이용하여 각 색인어가 도메인 전체 영역에서의 관계 정도를 평가하는 방법을 이용한다. 즉, $u_A(w) = u_B(w)$일 경우를 평가하는 방법[1,2]을 응용하여 도메인에서 색인어를 평가하고 $\alpha-cut$에 의한 축소용어 집합을 생성하였다. 문서

와 축소용어의 퍼지 관계를 표현하는 축소용어 집합 기반의 문서베이스는 원시 문서 베이스에서 문서 내용을 대표할 수 있는 축소용어만을 추출하여 구성한다.

여기서, 축소용어 집합 R_r은 다음과 같다.

$$R_r = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_r \end{array} \\ \begin{pmatrix} I_{11} & I_{12} & \cdots & I_{1r} \\ I_{21} & I_{22} & \cdots & I_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ I_{n1} & I_{n2} & \cdots & I_{nr} \end{pmatrix} \end{array}$$

단, r은 축소행렬의 첨자로 $r < m$이다.

<예 2> <예 1>에서 구한 원시 문서베이스(R)에서 식 (10)을 이용하고 $\alpha-cut$을 0.8로 적용했을 경우 축소행렬은 다음과 같이 얻을 수 있다.

원시베이스(R)에서 색인어 t_1의 계산 절차는 다음과 같다.

$$\begin{aligned} \mu_{t_1}(w) = & \max\{\min(0.94,\ 0.94), \min(1-0.94, 1-0.94)\} + \\ & \max\{\min(1.00,\ 1.00), \min(1-1.00, 1-1.00)\} + \\ & \max\{\min(0.94,\ 0.94), \min(1-0.94, 1-0.94)\} + \\ & \max\{\min(0.94,\ 0.94), \min(1-0.94, 1-0.94)\} + \\ & \max\{\min(0.94,\ 0.94), \min(1-0.94, 1-0.94)\} = 4.76 \end{aligned}$$

$t_1 = 4.76/5 = 0.95$이므로, 계산결과가 0.8이상인 색인어를 추출하여 다음과 같은 행렬을 얻을 수 있다.

$$R_r = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} \begin{array}{ccccc} r_1 & r_2 & r_3 & r_4 & r_5 \end{array} \\ \begin{pmatrix} 0.94 & 0.50 & 0.00 & 0.50 & 0.00 \\ 1.00 & 0.00 & 0.94 & 0.00 & 0.94 \\ 0.94 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix} \end{array}$$

(2) 유사관계 시소러스

축소용어로 구성된 문서베이스를 기반으로 색인어 사이의 퍼지 관련 정도를 나타내는 시소러스를 구성한다. 시소러스는 다음 식 (11)과 같이 문서베이스와 원시 문서베이스의 퍼지 관계곱 연산을 이용한다.

$$\left.\begin{aligned} & S_r = R^T \otimes R_r \\ & s_{ij} = \bigvee_{i,j=1..n} (\min(w_{in}, I_{nj}), (\min(1-w_{in}), (1-I_{nj}))) \end{aligned}\right] \quad (11)$$

s_{ij} : 색인어 i와 축소용어 j의 유사 정도

w_{in} : 문서 n에서 색인어 i의 중요 정도

I_{nj} : 축소행렬 문서 n에서 축소용어 j의 중요정도

퍼지 관계곱 연산은 특정 문서에서 색인어간 동시 출현 빈도가 많을수록 색인어 유사성이 높다는 가정 하에 식 (11)의 동시 출현 빈도를 고려하였다.

여기서, 유사관계 시소러스 S_r은 다음과 같다.

$$S_r = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} r_1 & r_2 & \cdots & r_r \end{array} \\ \begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1r} \\ S_{21} & S_{22} & \cdots & S_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ S_{m1} & S_{m2} & \cdots & S_{mr} \end{pmatrix} \end{array}$$

축소용어 집합 기반의 시소러스는 색인어의 의미를 정의하기 위한 관점에서 구축하였으므로 개념 행렬(concept matrices)이라 할 수 있다[1,2,5,9].

<예 3> <예 2>에서 구한 축소 행렬과 원시 문서베이스의 퍼지 관계를 식 (11)을 이용하면 유사관계 시소러스를 다음과 같이 구할 수 있다.

유사관계 시소러스(S_r)의 원소 s_{23}의 계산 절차는

$$\begin{aligned} s_{23} = & \max(\min(0.50, 0.00),\ \min(1-0.50,\ 1-0.00)) + \\ & \max(\min(0.00,\ 0.94), \min(1-0.00,\ 1-0.94)) + \\ & \max(\min(0.00,\ 0.00), \min(1-0.00,\ 1-0.00)) + \\ & \max(\min(0.00,\ 0.00), \min(1-0.00,\ 1-0.00)) + \\ & \max(\min(0.00,\ 0.00), \min(1-0.00,\ 1-0.00)) \\ = & \ 3.56 \end{aligned}$$

이므로 $s_{23} = 3.56/5 = 0.71$되며, 다른 원소도 같은 방법으로 계산하면 다음과 같은 행렬을 구할 수 있다.

$$S_r = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{array} \begin{array}{c} \begin{array}{ccccc} r_1 & r_2 & r_3 & r_4 & r_5 \end{array} \\ \begin{pmatrix} 0.95 & 0.14 & 0.24 & 0.14 & 0.24 \\ 0.14 & 0.90 & 0.71 & 0.90 & 0.71 \\ 0.38 & 0.64 & 0.64 & 0.64 & 0.64 \\ 0.46 & 0.60 & 0.35 & 0.60 & 0.35 \\ 0.24 & 0.71 & 0.99 & 0.71 & 0.99 \\ 0.14 & 0.90 & 0.71 & 0.90 & 0.71 \\ 0.38 & 0.64 & 0.64 & 0.64 & 0.64 \\ 0.46 & 0.60 & 0.35 & 0.60 & 0.35 \\ 0.24 & 0.71 & 0.99 & 0.71 & 0.99 \end{pmatrix} \end{array}$$

(3) 유사관계 행렬

본 논문에서는 검색 순위를 재조정하기 위한 방안으로 다음 식 (12)와 같이 퍼지 호환관계(tolerance relation)의 특성을 만족하는 유사관계 행렬을 정의한다. 이는 문서 베이스에서 누락된 색인어 정보를 고려하기 위한 방안으로 원시 문서 베이스를 기반으로 퍼지 호환관계를 만족하는 행렬이며 동시 출현 빈도를 기반으로 하였다.

$$\left.\begin{aligned} & S = R^T \otimes R \\ & \bar{s}_{ij} = \bigvee_{i=1,..m} (\min(w_{im}, w_{mi}), (\min(1-w_{im}, 1-w_{mi})) \end{aligned}\right] \quad (12)$$

$\bar{s}_{ij}$: 색인어 i와 j의 유사 정도

w_{im} : 원시 문서베이스의 문서 m에서 색인어 i의 중요 정도

여기서, 유사관계 행렬 S는 다음과 같다.

$$S = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_m \end{array} \\ \begin{pmatrix} \bar{S}_{11} & \bar{S}_{12} & \cdots & \bar{S}_{1m} \\ \bar{S}_{21} & \bar{S}_{22} & \cdots & \bar{S}_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ \bar{S}_{m1} & \bar{S}_{m2} & \cdots & \bar{S}_{mn} \end{pmatrix} \end{array}$$

<예 4> <예 3>의 축소용어 집합 기반의 호환 관계를 식 (12)를 이용하여 유사관계 행렬을 구축하면 다음과 같다.

유사관계 행렬(S)의 원소 $\bar{s}_{23}$의 계산 절차는 다음과 같다.

$$\begin{aligned} \bar{s}_{23} = & \max(\min(0.50, 0.50), \min(1-0.50, 1-0.50)) + \\ & \max\{\min(0.00,\ 0.50), \min(1-0.00, 1-0.50)\} + \end{aligned}$$

$max\{min(0.00,\ 0.80), min(1-0.00, 1-0.80)\}+$
$max\{min(0.00,\ 0.00), min(1-0.00, 1-0.00)\}+$
$max\{min(0.00,\ 0.00), min(1-0.00, 1-0.00)\}=3.20$

이고 $\bar{s}_{23}=3.20/5=0.64$이므로 다른 원소도 같은 방식으로 계산하면 된다.

$$S=\begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{array}\begin{array}{c} \begin{array}{ccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 \end{array} \\ \begin{pmatrix} 1.00 & 0.14 & 0.38 & 0.46 & 0.24 & 0.14 & 0.38 & 0.46 & 0.24 \\ 0.14 & 1.00 & 0.64 & 0.60 & 0.71 & 0.90 & 0.64 & 0.60 & 0.71 \\ 0.38 & 0.64 & 1.00 & 0.66 & 0.64 & 0.64 & 0.76 & 0.66 & 0.64 \\ 0.46 & 0.60 & 0.66 & 1.00 & 0.35 & 0.60 & 0.66 & 0.86 & 0.35 \\ 0.24 & 0.71 & 0.64 & 0.35 & 1.00 & 0.71 & 0.64 & 0.35 & 0.99 \\ 0.14 & 0.90 & 0.64 & 0.60 & 0.71 & 1.00 & 0.64 & 0.60 & 0.71 \\ 0.38 & 0.64 & 0.76 & 0.66 & 0.64 & 0.64 & 1.00 & 0.66 & 0.64 \\ 0.46 & 0.60 & 0.66 & 0.86 & 0.35 & 0.60 & 0.66 & 1.00 & 0.35 \\ 0.24 & 0.71 & 0.64 & 0.35 & 0.99 & 0.71 & 0.64 & 0.35 & 1.00 \end{pmatrix} \end{array}$$

3.3 퍼지관계를 이용한 질의어 확장

3.2장의 주제 분석 과정을 통해 축소용어 집합을 생성하고 이를 기반으로 퍼지 관계곱을 통한 시소러스, 유사관계 행렬을 구성하였다. 본 논문에서는 이를 기반으로 탐색 모형을 수립하는 과정을 제안한다.

(1) 사용자 질의 표현

질의 용어에 대해 도메인 지식을 확장하기 위해서 유사 관계 행렬을 활용하여 질의를 확장하며 다음과 같이 질의 연산자를 정의한다.

$$\left.\begin{array}{l} 1)\ x_i OR x_j=\mu(x_i)\vee\mu(x_j) \\ 2)\ x_i AND x_j=\mu(x_i)\wedge\mu(x_j) \\ 3) NOT\ x_i=\neg\mu(x_i)=1-\mu(x_i) \\ 4) VERY x_i=q_{very}(\mu(x_i))=\mu(x_i)^2 \\ 5) FAIRY x_i=q_{fairy}(\mu(x_i))=\mu(x_i)^{1/2} \end{array}\right] \quad (13)$$

여기서, $x_i, x_j\in[0,1],\ 1\le i\le n,\ 1\le j\le n$

질의 벡터(q)는 식 (13)을 만족하며 사용자 질의는 다음 <예 5>와 같이 사용자 질의 벡터로 표현된다.

$Q=\{(c_1/x_1),(c_2/x_2),\cdots(c_n/x_n), q=(x_1,x_2,\cdots,x_n)$일 경우에 질의 벡터에 관한 연산은 다음과 같다.

<예 5> 사용자 질의가 $q=\{t_6/0.8\}$이면 질의벡터는 다음과 같이 표현된다.

$$Q=q\times T=\{t_1/0, t_2/0, t_3/0, t_4/0, t_5/0, t_6/0.8, t_7/0, t_8/0, t_9/0\}$$

(2) 유사관계 시소러스 기반 질의 확장

사용자 정보 요구에 대한 질의는 도메인 지식을 확장하기 위해 시소러스와 퍼지 합성을 통해 확장된 질의베이스로 구성된다. 질의베이스(Q_r)는 다음 식 (14)와 같이 사용자에 의해서 표현된 질의 Q와 축소용어 집합과 문서와의 퍼지 관계를 나타내는 문서베이스(S_r) 사이의 퍼지 합성에 의해 생성한다.

$$\left.\begin{array}{l} Q_r=Q\circ S_r \\ \mu_{Q\cdot S_r}(x,z)=\underset{y\in Y}{Max}\{Min(\mu_Q(x,y),\mu_{S_r}(y,z))\} \end{array}\right] \quad (14)$$

$\mu_Q(x,y)$: 사용자 질의와 색인어와의 관계

$\mu_{S_r}(y,z)$: 문서와 축소용어 집합과의 관계 정도(시소러스)

$\mu_{Q\cdot S_r}(x,z)$: 사용자 질의와 축소용어 집합과의 관계 정도

여기서, 질의 확장집합 Q_r은 다음과 같다.

$$Q_r=\{qr_1, qr_2,\cdots, qr_r\}$$

<예 6> 질의 확장집합 Q_r은 식 (14)를 이용하여 시소러스와 퍼지 합성을 통해 확장된다.

질의베이스(Q_r)의 계산 절차는 다음과 같다.

$$\begin{aligned} qr_1=max\{&min(0.00,0.95),\ min(0.00,0.14),\ min(0.00,0.38)\\ &min(0.00,0.46),\ min(0.00,0.24),\ min(0.80,0.14)\\ &min(0.00,0.38),\ min(0.00,0.46),\ min(0.00,0.24)\\ \}& \end{aligned}$$

이므로, 같은 방식으로 계산하면

$Q_r=Q\cdot S_r=\{r_1/0.14, r_2/0.80, r_3/0.71, r_4/0.80, r_5/0.71\}$

이다.

(3) 유사성 평가

이와 같이 구성된 질의베이스와 문서베이스에 대한 유사도를 평가함으로써 문서 검색상태 값(RSV)을 파악할 수 있고, 유사성 척도 방법은 다음 식 (14)와 같이 정의한다. 즉, 각 문서를 평가하기 위하여 식 (15)에 명시된 유사도 척도 방법을 이용하여 시소러스 기반의 1단계 문서 검색을 수행할 수 있다.

$$\left.\begin{array}{l} O_r=Q_r\otimes R_r^T \\ \quad=\{RSV(d_1), RSV(d_2),\cdots, RSV(d_n)\} \\ RSV(d_i)=\dfrac{\sum_{\vec{q(j)}\neq'-'and\ j=1,\cdots,n} T(1-|I_{ij}-Qr_j|)}{k} \end{array}\right] \quad (15)$$

I_{ij} : 축소용어 집합(R_r)에서 문서 i와 축소용어 j의 관계 정도

Qr_j : 사용자 질의와 축소용어 j와의 관계 정도

d_i : 문서 i의 검색 상태 값(RSV)

<예 7> 문서 베이스와 질의 베이스와의 유사도 값을 식 (15)에 의하여 평가하여 1단계 문서 검색(퍼지 검색) 절차 및 결과는 다음과 같다.

$$\begin{aligned} d_1=&(1-|0.94-0.14|)+(1-|0.50-0.80|)+\\ &(1-|0.00-0.71|)+(1-|0.50-0.80|)+\\ &(1-|0.00-0.71|)=2.18 \end{aligned}$$

$RSV(d_1)=2.18/5=0.44$이므로, 같은 방식으로 계산하면

$O_r=Q_r\otimes R_r^T$
$=\{d_1/0.44, d_2/0.41, d_3/0.24, d_4/0.24, d_5/0.24\}$ 이다.

3.4 문서 관련성 평가를 통한 클러스터 검색

기존의 시소러스의 사용 방법은 검색 영역을 확장하는 것으로 문서 검색의 정확률 및 재현율을 향상시킬 수 있다[8,15,16,20]. 그러나 정확한 정보의 검색에는 완전하지 못함으로 본 논문에서는 그림 7과 같이 영역 지

식을 확장하는 방법으로 퍼지 유사관계 행렬을 활용하여 질의 개념을 확장하는 방법을 제안한다.

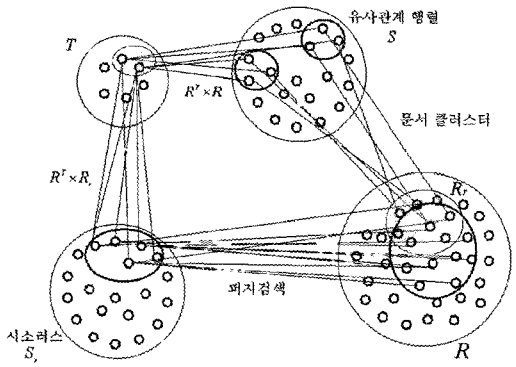
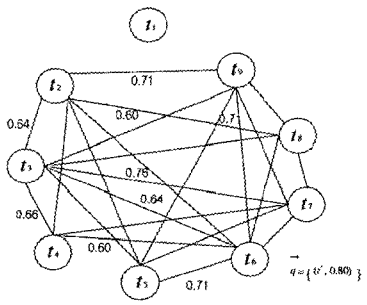


그림 7 퍼지 색인어 유사관계에 의한 검색

문서 검색 시스템에서 문서는 문서 구조를 고려한 색인어의 집합에 의해 표현되며, 색인어 연관관계 정도에 의해 문서의 내용을 구조화하거나 문서집단으로 대표되는 특정 주제 영역의 지식 구조를 파악할 수 있다. 따라서 사용자에 의해서 구성되는 질의는 질의 용어의 관계에 따라 사용자 요구를 파악할 수 있기 때문에 질의 용어에 대한 지식 또한 포괄적인 지식에서 세부적인 지식으로 확장되어진다. 또한 사용자들은 특정 도메인에 대한 상당한 지식을 가지고 있으므로 포괄적인 의미 검색과 다양한 사용자 요구를 정확하게 표현할 수 있도록 설계되어야 한다[14]. 원시 문서베이스를 기반으로 생성한 유사관계 행렬에서 호환관계를 만족하는 호환 클래스를 분류하고 원시 질의에 추가함으로서 재현율을 향상시키고자 한다. 즉, 도메인의 영역 지식을 반영한 원시 질의를 확장하며 추가된 질의 가중치는 전이관계에 의한 퍼지 확장 원리를 이용하여 부여한다.

<예 8> 유사관계 행렬(S)에서 사용자 질의($q=\{t_6/0.8\}$)에 대해서, 분류정도($0.6-cut$)와 식 (6)에 따라 호환클래스 생성하고 질의를 확장하면 다음과 같다(유사한 문서집합에 대한하여 변별력을 향상시키기 위하여 $a-cut$은 재조정한다).



위 그림은 유사관계 행렬 S을 개념 네트워크로 나타낸 것이다. 사용자 질의 $q=\{t_6/0.8\}$에 대하여 호환클래스는 유사관계에 따라 사용자 질의에 확장되는 용어(Q_e)는 $t_2, t_3, t_5, t_6, t_7, t_9$이다.

따라서, 색인어 유사관계 행렬 기반의 호환클래스는

$$C/R_{0.6} = \{C_1 = \{t_3, t_5, t_6, t_7\}, C_2 = \{t_2, t_5, t_6, t_9\}, C_3 = \{t_2, t_5, t_6, t_7\}\}$$

이며 확장 질의의 가중치는 전이적 성질(퍼지 이행관계)을 이용하여 계산되며

$$Q_e = \{(t_2, 0.90), (t_3, 0.64), (t_5, 0.71), (t_6, 0.80), (t_7, 0.64), (t_9, 0.71)\}$$

가 된다.

확장된 용어의 가중치는 관련연구에서 개념 네트워크 기반 검색 방법의 전이 관계를 이용하였다. 원시 문서베이스가 9개의 색인어로 구성되었다고 가정할 경우 이를 기반으로 유사성 척도에 의하여 문서 검색 상태 값(RSV)을 생성할 수 있다(단, 확장된 용어를 포함한 용어에 대하여만 유사성을 파악한다).

$$\left.\begin{aligned} O &= Q_e \otimes R \\ &= \{RSV(d_1), RSV(d_2), \cdots, RSV(d_n)\} \\ RSV(d_j) &= \frac{\sum_{q(i) \neq 0 \text{ and } j=1,\cdots,n} T(1-|w_{ij}-Qe_j|)}{k} \end{aligned}\right] \quad (16)$$

Q_e: 유사관계 행렬(S)에 의해 확장된 질의베이스

w_{ij}: 원시 문서베이스(R)의 문서와 색인어 관계정도

<예 9> 질의확장에 따른 문서 클러스터 절차 및 결과는 다음과 같다.

$$Q_e = \{t_1/-, t_2/0.90, t_3/0.64, t_4/-, t_5/0.71, t_6/0.80, t_7/0.64, t_8/-, t_9/0.71\}$$

$$R = \begin{array}{c} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \\ d_5 \end{array} \begin{array}{c} t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5 \quad t_6 \quad t_7 \quad t_8 \quad t_9 \\ \begin{pmatrix} 0.94 & 0.50 & 0.50 & 0.80 & 0.00 & 0.50 & 0.50 & 0.80 & 0.00 \\ 1.00 & 0.00 & 0.50 & 0.00 & 0.94 & 0.00 & 0.50 & 0.00 & 0.94 \\ 0.94 & 0.00 & 0.80 & 1.00 & 0.00 & 0.00 & 0.80 & 1.00 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix} \end{array}$$

에 대하여 식 (16)을 적용하면

$$d_1 = 0.60 + 0.86 + 0.29 + 0.7 + 0.86 + 0.29 = 3.6$$
$$RSV(d_1) = 3.6/6 = 0.60$$

$$d_2 = 0.10 + 0.86 + 0.77 + 0.20 + 0.86 + 0.77 = 3.56$$
$$RSV(d_2) = 3.56/6 = 0.59$$

$O=\{d_1/0.60, d_2/0.59, d_3/0.43, d_4/0.27, d_5/0.27\}$을 얻을 수 있다.

이와 같이 문서 클러스터를 통해 문서 상태 값에 따라 문서를 평가할 수 있다. 여기서, 원시 문서베이스(R)의 반전된 부분만이 유사성을 평가하는데 활용된다.

3.5 유사도결합을 통한 문서 순위 재조정 알고리즘

본 논문에서는 축소용어 집합을 기반으로 작성된 시소러스를 통한 각 문서의 검색상태 값(O_r)과 원시 문서베이스를 기반으로 작성된 각 문서의 검색 상태 값(O)은 1차 질의 확장과 2차 내용 질의에 대한 검색 상태 값을 의미한다. 따라서 1단계 질의 확장으로 재현율을 유지하고, 정확률을 높이기 위한 2단계 유사관계 행렬 기반의 클러스터 검색을 수행하였다. 본 논문에서는 검색 순위를 재조정(Re-ranking)을 통해 적합한 문서가 상위에 검색될 수 있도록 검색 상태 값을 재조정한다.

$$Sim_{combined} = \alpha O_r + \beta O = \{RSV(d_1), RSV(d_2), \cdots, RSV(d_n)\} \quad (17)$$

O_r: 축소용어 행렬 기반 퍼지검색(문서상태 값)

O: 유사관계 행렬 기반 클러스터 검색(문서상태 값)

α, β : 1로 설정

$Sim_{combined}$: 문서 상태 값의 순위 재조정 결과

<예 11> 식 (17)에서 도메인의 특성에 따라 α, β을 적용하여 재조정한다.

최종적인 문서 상태 값은

$O_r = \{d_1/0.44, d_2/0.41, d_3/0.24, d_4/0.24, d_5/0.24\}$이고,

$O = \{d_1/0.60, d_2/0.59, d_3/0.43, d_4/0.27, d_5/0.27\}$, 여기서 α와 β 값을 1:1로 설정하면

$Sim_{combined} = \{d_1/0.52, d_2/0.50, d_3/0.34, d_4/0.26, d_5/0.26\}$이다.

따라서, 검색상태 값이 0.5 이상을 선택하면 $Sim_{0.5} = \{d_1/0.52, d_2/0.50\}$이다.

위와 같이 각 단계별로 계산하는 개괄적 흐름 절차는 그림 8과 같으며, 이에 대한 구체적인 알고리즘은 다음과 같으며, 구성하는 각 개별적 알고리즘은 부록에 제시되어 있다.

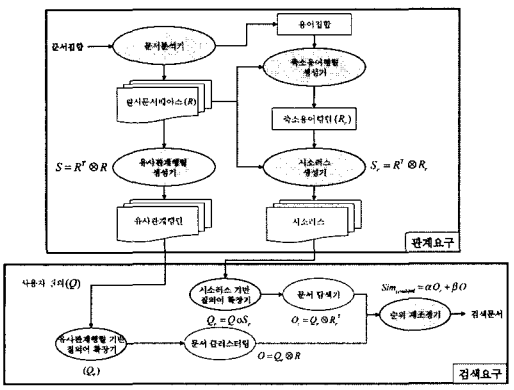


그림 8 문서 순위 재조정 검색 모델

Algorithm 3.0 문서 순위 재조정 알고리즘

입력 : 각 문서
출력 : 문서 순위가 재조정 된 문서 집합

```
function Re_ranking-document( )
 {
   get document_set( ) ;  //수집된 문서 입력
   occurrence-frequency( );
    //형태소 분석기를 이용한 색인어의 위치 정보와 빈도수 계산
   source_base( );
   // 문서분석기(원시 문서베이스 생성)
   reduction_term_set(); // 축소용어 행렬 생성기
   thesaurus_creation(); // 시소러스 생성기
   similarity_relation_matrix(); //유사관계 행렬 생성기
   get_user_query( );
   query_expansion( ); //시소러스기반 질의어 확장기
   fuzzy_search( ); //문서탐색기(퍼지 검색)
   cluster_search( );
   //유사관계 행렬기반 질의어 확장기 및 문서 클러스터
   similarity_combine_rerank();
   //유사도 결합을 통한 순위 재조정기
   put-document_set( );
 }
```

4. 실험 및 평가

본 논문에서는 시간 복잡도를 해소하기 위한 축소행렬을 구성하였고 이를 기반으로 퍼지 검색을 수행하였다. 또한 정확률을 향상시키기 위해 유사관계 행렬을 기반으로 문서 클러스터 검색을 제안하였으며 퍼지 검색과 문서 클러스터 검색의 유사도 결합을 통한 순위 재조정 모델을 설계하고 구현하였다. 검색 순위 재조정 모델의 성능 평가의 절차는 첫째, 문서집합에서 색인어와 출현 빈도를 추출하고, 시그모이드 함수를 적용하여 원시 문서베이스를 구성한다. 둘째, 동시 출현 빈도를 기반으로 문서 구조 특성을 고려하여 축소용어 집합, 시소러스를 구성하여 이를 기반으로 질의확장을 통한 퍼지 검색을 수행한다. 셋째, 색인어 군집에 의한 확장 검색 기법인 문서 클러스터 검색은 원시 문서베이스를 기반으로 유사관계 행렬을 구축하였고 분류정도에 따른 검색을 수행한다. 넷째, 퍼지 검색과 클러스터 검색의 유사도 결합을 통하여 순위를 재조정 모델의 재현율(recall)과 정확률(precision)을 평가한다.

4.1 실험분석 방법

본 논문에서 제안한 순위 재조정 모델의 검색 효율(retrieval effectiveness)을 측정하기 위해 다음 식 (18)의 순위 정확률과 식 (19)의 순위 재현율을 이용하였다[4].

$$order\_precision = \frac{Match_{Doc}}{Rank_{Doc|n}} \quad (18)$$

$Match_{Doc}$: 테스트 질의의 결과와 $Rank_{Doc|n}$이 일치하는 문서의 수

$Rank_{Doc|n}$: 문서순위결정 결과 중에서 상위 n개의 문서

$$order\_recall = \frac{KT-Set_{match}}{Rank_{match}} \quad (19)$$

$KT-Set_{match}$: KT-Set에서 테스트 질의에 대한 결과 문서의 수

$Rank_{match}$: 문서 순위 결정 결과가 KT-Set의 테스트 질의에 대한 결과 문서를 모두 포함하는 순간의 순위

$n = KT-Set_{match}$

4.2 실험 대상

본 논문에서 제안한 문서 순위 재조정 알고리즘의 비교·평가하기 위하여 이용하는 실험 대상(Domain)은 다음과 같다. 실험 문서는 KT-Set(2.0)의 전체 문서 중에서 표제, 요약, 키워드 집합으로 구조화된 문서 1000개(문서번호11000번)를 문서집합을 구성한다. 사실, 실험 문서 KT-Set(2.0)은 총 4414개의 문서로 구성되어 있으나 3414개의 문서는 신문사설, 보고서 등으로 문서 구조가 정형화되어 있지 않다. 따라서 본 논문에서 제안한 알고리즘의 실험에는 적절하지 않아 대상에서 제외한다. 그리고 KT-Set(2.0)의 전체 질의 집합은 표 3과 같이 총 50개의 질의로 구성되어 있으나 실험 결과의 신뢰성을 높이기 위하여 질의에 대한 적합 문서수가 표 4와같이 5개 이상인 테스트 질의를 추출하여 실험 질의 데이타로 이용한다.

따라서 본 논문의 실험 집합에 대한 통계 정보는 표 4와 같다.

표 4 실험집합의 통계정보

질의 번호	질의어 수(질의 내용)	적합문서 수 ($KT-Set_{match}$)	cluster 검색의 분류 기준값 (α)
8	2(멀티미디어&데이타베이스)	22	0.6
16	2(자동번역||기계번역)	7	0.6
21	2(문자인식||필기체인식)	19	0.6
22	2(지능형&정보검색)	5	0.6
25	1(지리정보)	8	0.6
26	2(음성인식||음성생성)	13	0.6
33	3(초고속&정보&통신망)	77	0.6
39	2(신경망&퍼지제어)	40	0.6
43	2(시소러스||형태소)	22	0.6
50	2(샘플||샘플링)	6	0.6

4.3 실험 비교 및 평가

본 실험에서는 제안된 모델의 성능을 비교·평가하기 위하여 일반적인 검색 모델 중 구조화된 문서 정보를 이용한 3개의 순위 검색하는 모델을 선정하여 평가한다. 첫 번째 모델은 「Persin」의 문서 가중치 순위 결정방법[21]이고 두 번째 모델은 「은희주」의 퍼지 맴버쉽 함수를 이용한 퍼지 검색 기법[5]이며 세 번째 모델은 「Koczy」의 계층적 용어 클러스터 기법[2]이다. 또한 각각의 검색 모델에 대한 검색 효율 측정 도구는 평균 순위 재현율과 평균 순위 정확률을 적용한다.

(1) 분류기준 값 실험

문서 클러스터 기법을 수행하는 과정에서의 분류 기준 값(α)의 변화에 따라 순위 정확률을 평가한 내용이다. 분류 적합도는 유사한 문서 집합에 대한 변별력을 향상시키기 위하여 색인어 내용 지식을 동적으로 분류하기 위하여 사용하였다.

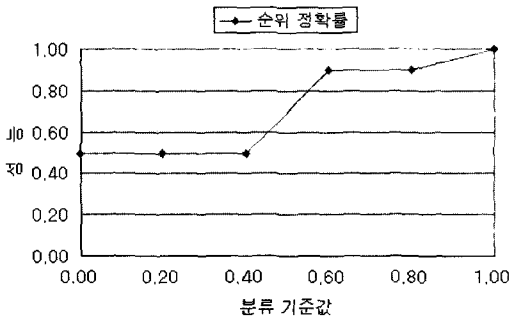


그림 9 분류 기준 값(α)에 따른 순위 정확률

평가 결과에서 문서 클러스터 기법의 분류 기준 값(α)이 0.1~0.5일 경우 순위 정확률은 0.5이며 0.6~0.7일 경우에 가장 우수한 성능을 보였다. 그리고 0.8이상일 경우에 정확률은 1이었으나, 재현율은 0.4로 평가 되었다. 따라서 본 논문에서는 분류 기준 값이 0.6일 경우를 적용하여 실험을 수행하였다.

(2) 비교 평가 결과

표 3 KT-Set 질의 집합의 통계정보

질의번호	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
검색문서	0	1	0	1	0	0	0	22	1	1	0	1	1	0	0	7	2
질의번호	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
검색문서	0	0	0	19	5	2	0	8	13	3	0	0	0	0	0	77	0
질의번호	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	
검색문서	0	0	0	0	40	1	2	1	22	0	0	0	0	0	0	6	

본 실험에의 평가 결과는 다음 그림 10과 같이 「Persin」의 문서 순위 결정 기법의 평가에서 평균 순위 재현율이 0.81과 평균 순위 정확률은 0.86의 성능을 보였다. 그리고 「은희주」의 퍼지 멤버쉽 함수를 이용한 퍼지 검색은 평균 순위 재현율이 0.89를 유지하는데 비해 평균 순위 정확률은 0.82의 성능을 보였고 「Koczy」의 계층적 용어 클러스터 기법은 평균 순위 재현율과 평균 순위 정확률은 각각 0.85와 0.87의 성능이 측정되었다. 그리고 본 논문에서 제안하는 순위 재조정 모델에서는 퍼지 검색과 문서 클러스터 기법의 단점인 순위 정확률과 순위 재현율이 0.9이상으로 나타남으로써 퍼지 검색의 재현율을 유지면서 정확률이 향상되었음을 알 수 있다.

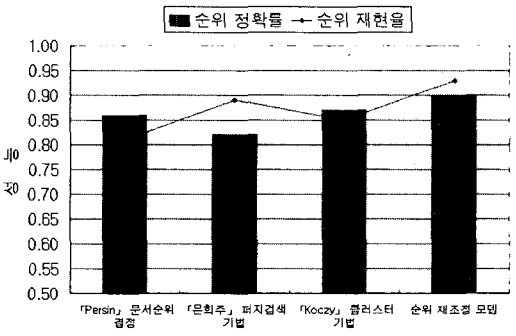


그림 10 평균 순위 재현율과 평균 순위 정확률

(3) T-test 분석

마지막으로 본 논문에서는 그림 10의 실험 결과인 평균 순위 재현율과 평균 순위 정확률에 대한 신뢰성을 검증하기 위하여 재현율과 정확률의 평균값의 차이를 SPSS 8.0을 통해 ANOVA와 T-test를 실시하였다. ANOVA 분석에 의한 결과, 재현율은 F=17.324, Sig.=0.000이고, 정확률은 F=13.269, Sig.=0.000로 통계적으로 유의한 차이를 나타내었다. 그리고 Persin의 알고리즘과 본 논문에서 제안한 순위 재조정 모델의 T-test 분석 결과 재현율은 t=-7.104, Sig.=0.000이고 정확률이 t=-3.689, Sig.=0.002로 나타나 통계적으로 유의한 차이를 보여주고 있다.

평가 결과를 종합적으로 살펴보면 퍼지 검색기법은 재현율이 상대적으로 우수하였고 문서 클러스터 기법은 정확률이 우수함을 보인다. 그리고 본 논문에서 제안한 순위 재조정 기법은 재현율을 보장하면서 정확률이 향상되었음을 알 수 있다.

5. 결론 및 향후 연구과제

현재 이용하고 있는 웹 기반의 학술분야 전문 검색 시스템은 실제 얻어진 정보들 중에서 사용자의 관심도가 많이 반영된 문서를 선별하는데 많은 문제점이 있다. 이에 따라 본 논문에서는 검색 시스템이 사용자의 정보 요구를 충족시킬 수 있도록 문서 내용 분석 과정과 정보 요구에 대한 정보 습득 과정의 일관된 메커니즘을 제공하고자 한다. 또한 개념 정보들이 속해있는 용어들의 유사도 및 개념 거리를 이용하여 개념 정보를 사용자의 관심을 표현하는 질의어와 문서에서 추출한 색인어간의 유사 정도를 퍼지 값으로 사상시켜 질의어를 확장하고자 한다.

본 논문은 퍼지 검색, 문서 클러스터 기법 및 유사성 결합을 통한 재순위화 모델로 구성된다. 퍼지 검색에서 축소용어 집합은 높은 시간 복잡도를 처리하고자 구성하였으며 색인어 자신의 중요 정도를 min-max 연산을 통하여 $\alpha-cut$을 만족하는 용어만을 대상으로 하였다. 그리고 사용자 요구가 반영된 문서들을 검색하기 위하여 탐색어 집합을 검색 전에 확장하는 방법인 문서 클러스터 기법으로 검색 속도와 정확도를 높일 수 있도록 하였다. 문서 클러스터 기법에서는 사용자 질의에 대하여 유사관계 행렬을 기반으로 분류 기준 값(α) 이상의 호환관계를 만족하는 용어들로 확장하여 문서를 검색하였다. 마지막으로 퍼지 검색의 재현율의 특성과 의미적으로 연결된 문서들을 클러스터링하는 문서 클러스터 기법의 유사도를 결합함으로써 사용자의 정보 요구를 충족시킬 수 있도록 문서 검색 순위를 재조정한다.

본 논문의 실험적 평가를 위해 순위 정확률과 순위 재현율을 이용하였고 실험 집합으로는 KT-Set의 총 50개의 테스트 질의들 중에서 10개 질의를 기준으로 실험하였다. 평가결과를 요약해보면 퍼지 검색에서 재현율이 상대적으로 우수하였고 문서 클러스터 기법에서는 분류 기준 값(α)이 0.6~0.7일 경우에 문서집합에 대한 변별력이 가장 우수하였다. 본 논문에서 제안한 유사성을 결합을 통한 재순위화 모델의 평가에서는 퍼지 검색과 문서 클러스터 기법의 단점인 정확률과 재현율이 0.9 이상으로 나타남으로서 퍼지 검색의 재현율을 유지하면서 정확률이 향상되었음을 알 수 있다.

앞으로의 향후 연구로는 주제 분석 과정에서 도메인의 특성에 따라 매개변수를 자동화하는 연구와 축소용어 집합에 대한 신뢰도를 향상시킬 수 있도록 문서 요약과 문서의 내용, 결론을 포함한 전체 문서에서의 용어 추출 방법에 대한 연구를 계속할 것이다. 또한 탐색 모형에서의 사용자 관심이 반영될 수 있도록 프로파일을 이용한 개념 질의 확장에 대한 연구를 계속 진행할 계획이다.

참 고 문 헌

[1] Laszlo T. Koczy, T. D. Gedeon, "Information retrieval by fuzzy relations and hierarchical co-occurrence," Part I. TR97-01, Dept. of Info. Eng., School of Comp. Sci. & Eng., UNSW, 1997.

[2] Laszlo T. Koczy, T. D. Gedeon, "Information retrieval by fuzzy relations and hierarchical co-occurrence," Part II. TR97-03, Dept. of Info. Eng., School of Comp. Sci. & Eng., UNSW, 1997.

[3] 문성빈, "적합성 피이드백을 이용한 전문검색시스템의 검색 효율성 증진을 위한 연구", 정보관리학회지, 제10권 2호, 1993.

[4] 우선미, "사용자 프로파일과 잠재적 구조분석을 이용한 검색된 문서의 순위 결정 방법", 전북대학교 대학원 박사학위논문, 2001.8.

[5] 은희주, "퍼지함수와 관계성을 적용한 질의 확장 및 문서 분류 시스템", 전북대학교 대학원 박사학위논문, 2003.8.

[6] 정영미, "정보검색론", 구미무역, 1988.

[7] 김철, 이승채, 김병기, "색인어 퍼지 관계와 서열 기법을 이용한 정보 검색 방법론", 한국정보처리학회 논문지 제3권 제5호, 1996.9.

[8] Kim, Chang-Min, Kim, Yong-Gi, "An Improvement of Bandler-Kohout Fuzzy Information Retrieval Model using Reduced Set," IEEE International Fuzzy Systems Conference Proceedings, August, 1999.

[9] 김창민, 김용기, "퍼지 관계곱 기반 퍼지정보 검색 시스템 구현", 정보처리학회 논문지, 제8-B권 제2호, 2001.4, pp. 115-122.

[10] 유영준, "문헌정보학에서 지식 구조에 관한 연구", 연세대학교 대학원 박사학위논문, 2003.8.

[11] 남궁황, "문단의 의미구조에 의한 전문검색시스템의 설계 및 평가에 관한 연구", 중앙대학교 대학원 박사학위논문, 1999.

[12] Shyi-Ming Chen, Yih-Jen Horng, "Fuzzy Query Processing for Document Retrieval Based on Extended Fuzzy Concept Networks," IEEE Transactions on Systems, MAN. and CyberNetics-Part B: CyberNetics, Vol. 29, No. 1, February, 1999.

[13] Shyi-Ming Chen, Jeng-Yih Wang, "Document Retrieval Using Knowledge-Based Fuzzy Information Retrieval Techniques," IEEE Transactions on Systems, MAN. and CyberNetics, Vol. 25, No. 5, May, 1995.

[14] 최재훈, 김지숙, 조기환, "문제은행에서 연상학습을 지원하는 퍼지 검색 시스템", 정보과학회지, 제29권 제4호, 2002.4.

[15] Bandler W. and Kohout L. J., "The Identification Operators and Fuzzy Relational Products," International Journal of Man-Machine Studies 12(1980) 89-116. Reprinted in: Mamdani E. H. and Gaines B. R., eds., Fuzzy Reasoning and its Applications(Academic press London, 1981)

[16] 이종득, "시소러스 기반의 정보검색 시스템 구축을 위한 개념 그룹화 방법", 전북대학교 대학원 박사학위논문, 1998.3.

[17] 이광형, 오길록, "퍼지 이론 및 응용", 홍릉 과학 출판사, 1991.

[18] Chia-Hui Chang, Ching-chi Hsu, "Enabling Concept-Based Relevance Feedback for Information Retrieval on the WWW," IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 4, July/August, 1999.

[19] Ogawa, Y. et al. "A Fuzzy Document Retrival System Using the Keyword Connection Matrix and a Learning Method," Fuzzy Sets and Systems Vol 39, pp.163-179, 1991.

[20] Takagi, T., Tajima, M., "Query expansion using conceptual fuzzy sets for search engine," Proceedings of the 10th IEEE International Conference on Fuzzy Systems - Vol. 3, 2002.12.

[21] Michael Persin, "Document Filtering for Fast Ranking," ACM-SIGIR, pp.339-348, 1994.

부 록

Algorithm 3.1 색인어 발생 영역 정보와 빈도 계산

입력 : 각 문서
출력 : 각 색인어 위치정보와 빈도

```
function occurrence_frequency( )
{
  for (i=1 ; ; i++ )
  {
    preprocessing( Di );
     // 형태소 분석기를 사용하여 용어 분리
    for(j=1 ;  ; j++)
    {
      extract_keyword( Kij)
      // 문서 i에서 색인어 Kj 을 추출
      search_keyword_location( Ki )
      // 추출한 색인어의 위치 정보
      occurrence_frequent ( Kj,  Di )
    // 문서 Di에서 추출된 색인어 Kj의 빈도 계산
    }
  }
}
```

Algorithm 3.2 색인어의 발생 위치에 따른 소속 정도 계산

입력: 각 색인어의 위치정보 및 빈도
출력: 원시문서베이스 생성

```
function source_base( )
{
  for(i=1; ; i++)
  {
  source_base_process(i, read_occurrence_frequency_value);
   // 색인어별 발생위치에 따른 소속정도로 사상
  source_min_max(i, read_occurrence_frequency_value);
    //타이틀, 키워드, 요약의 중요정도 min_max 연산
  }
}
```

Algorithm 3.3 축소용어 행렬 생성기

입력 : 원시 문서 베이스
출력 : 축소 용어 행렬

```
function reduction_term_set()
{
  for(i=1; ; i++)
  {
    reduction_min_max(i, read_source_value);
    // 색인어 자신의 중요정도 min_max 연산
  }
  for(j=1; ; j++)
  {
    if(reduction_term_value( t_j >= α))
   //중요정도가 α이상인 색인어로 축소용어행렬 구성
     reduction_term_matrix(j, reduction_max_value);
  }
}
```

Algorithm 3.4 시소러스 생성기

입력 : 원시 문서베이스, 축소용어 집합
출력 : 유사관계 시소러스

```
function thesaurus_creation( )
{
  for(i=1; ; i++)  //  S_r = R^T ⊗ R_r
  {
    thesaurus_min_max(i, read_source_value,
                      reduction_term_value);
  }
}
```

Algorithm 3.5 유사관계 행렬 생성기

입력 : 원시 문서 베이스
출력 : 유사관계행렬

```
function similarity_relation_matrix()
{
  for(i=1; ; i++)  //  S = R^T ⊗ R
  {
    similarity_relation_matrix_min_max
                        (i, read_source_value);
  }
}
```

Algorithm 3.6 시소러스 기반 질의어 확장기

입력: 사용자 질의 벡터, 시소러스
출력: 질의 확장 벡터

```
function query_expansion()
{
  question_base_process();  // 질의벡터 생성( Q)
  for(i=1;  ; i++) //  Q_r = Q ∘ S_r
  {
    if(read_thesaurus_val(i) > question_base[i])
    {
      if(query_expansion_val[i]< question_base[i])
        query_expansion_val[i]= question_base[i];
    }
    else
    {
   if(query_expansion_val[i]< read_thesaurus_val(i))
    query_expansion_val[i]=read_thesaurus_val(i);
    }
  }
}
```

Algorithm 3.7 문서탐색기(퍼지검색)

입력: 확장 질의 벡터, 축소용어 집합
출력: 검색된 문서집합

```
function fuzzy_search()
{
  for(i=1; ; i++)  //  O_r = Q_r ⊗ R_r^T
  {
  fuzzy_search_process(read_query_expansion_value,
                       read_reduction_term_value);
  }
}
```

Algorithm 3.8 유사관계 행렬 기반 질의어 확장기 및 문서 클러스터

입력: 질의 벡터, 유사관계 행렬
출력: 호환클래스에 따른 문서 클러스터

```
function cluster_search()
{
  for(i=1; ; i++) // 호환클래스 생성( Qe)
  {
search_node_expansion(read_similarity_relation_matrix_value,
                      get_query_base);
  }
  for(i=1; ; i++) //  O= Qe ⊗ R
  {
    query_expansion_search(node_value);
  }
}
```

Algorithm 3.9 순위재조정기(2단계 순위 재조정)

입력: 퍼지 검색, 클러스터 검색의 검색문서 유사도
출력: 순위 재설정된 문서집합

```
function similarity_combine_rerank()
{
  for(i=1; ; i++)  //  Simcombind= α Or+ β O
  {
    similarity_reranking(fuzzy_search_value,
                         cluster_search_value);
  }
}
```


이 기 영

1992년 2월 광주대학교 컴퓨터공학과 졸업(공학사). 1994년 2월 전북대학교 전산통계학과 졸업(이학석사). 1997년 2월 전북대학교 전산통계학과(박사수료). 1998년 3월~원광보건대학 컴퓨터응용개발과 부교수. 관심분야는 퍼지 정보검색, 멀티미디어시스템 등

은 희 주

1998년 2월 전북대학교 컴퓨터과학과 졸업(이학사). 2000년 2월 전북대학교 전산통계학과 졸업(이학석사). 2003년 8월 전북대학교 전산통계학과 졸업(이학박사) 관심분야는 퍼지 클러스터링, 퍼지 유전자 알고리즘, 퍼지 정보검색, 퍼지 데이타 마이닝 등

김 용 성
정보과학회논문지 : 소프트웨어 및 응용
제 31 권 제 10 호 참조