韓國 컴퓨터情報學會 論文誌
第10卷 第4號, 2005. 9.

2005-10-4-1-1

2단계 퍼지 지식베이스를 이용한 질의 처리 모델

이 기 영*, 김 영 운**

Query Processing Model Using Two-Level Fuzzy Knowledge Base

Ki-Young Lee*, Young-Un 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. 따라서 본 논문에서는 문서 지식 구조를 파악하여 사용자 질의 용어와 색인어 사이의 내용 기반 유사도를 반영한 순위 재조정 모델을 제안한다. 이를 위해 전자는 시소러스 및 유사관계 행렬을 구축하여 주제 분석 메커니즘을 제공하고, 후자는 사용자 요구를 분석하기 위해 질의 확장 등의 탐색 모형을 수립하는 알고리즘을 제안한다. 따라서 본 논문에서 제안한 알고리즘은 검색 시스템의 정보 구조를 활용한 검색으로 재현율을 유지하면서 동시에 기존 퍼지 검색 모델의 단점인 정확률을 향상시키는 2단계 탐색모형을 수립하는 내용 기반 검색 기법이라 할 수 있다.

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. Accordingly, this study suggests the re-ranking retrieval model which reflects the content based similarity between user's inquiry terms and index words by grasping the document knowledge structure. In order to accomplish this, the former constructs a thesaurus and similarity relation matrix to provide the subject analysis mechanism and the latter propose the algorithm which establishes a search model such as query expansion in order to analyze the user's demands. Therefore, the algorithm that this study suggests as retrieval utilizing the information structure of a retrieval system can be content-based retrieval mechanism to establish a 2-step search model for the preservation of recall and improvement of accuracy which was a weak point of the previous fuzzy retrieval model.

▸ Keyword : Reduction Term, Fuzzy Compatibility Relation, Similarity Relation Matrices, Cluster Retrieval, Re-ranking Retrieval Model

• 제1저자 : 이기영
• 접수일 : 2005.06.11, 심사완료일 : 2005.07.30
* 원광보건대학 정보컨텐츠과 교수, ** 원광보건대학 정보컨텐츠과 겸임교수
※ 이 논문은 원광보건대학 교내연구비 지원에 의하여 연구되었음.

Ⅰ. 서 론

현재 웹 기반의 학술분야 전문(이하, 문서라고 기술함.) 검색 시스템은 사용자의 관심도를 자연어 형태의 질의로 표현하여 문서에서의 색인어(index) 발생 여부에 따라 검색 결과를 제공하는 키워드(keyword) 매칭 기법을 이용한다. 그러나 이러한 기법은 동의어 및 다의어 처리의 문제로 인하여 선별되지 않은 정보를 검색 결과로 제시함에 따라 반복적인 피드백(feedback) 필터링 과정을 거쳐야 하는 번거로운 작업을 필요로 한다[1,2,3,4]. 이러한 결과를 초래하는 주된 원인 중 하나는 대부분의 웹 기반 검색 시스템(web-based information retrieval system)에서 공통적으로 요구하는 질의 형식이 사용자의 관심도를 표현한 키워드나 색인어의 집합 형태로 구성된다는 점이다[5,6].

따라서 이러한 키워드나 색인어의 집합 형태의 질의를 이용하여 검색할 경우에 사용자에게 유용한 정보와는 전혀 관계없이 단순 질의에 해당하는 단어를 포함한 다량의 문서만을 제공하는 경우가 많다[2,7,8,9].

이를테면 사용자의 관심도가 반영되고 주어진 질의에 만족하는 정보뿐만 아니라 문서 주제에 연관된 정보를 보다 정확하게 검색할 수 있는 내용 기반의 검색 기법이 반영되지 못하고 있다. 이에 따라 본 연구에서는 사용자 관심을 표현하는 질의와 문서의 구조화 지식인 표제, 요약, 키워드에 대한 의미적 논리 구조를 이용하여 질의어 확장에 대한 연구를 수행한다. 또한 퍼지 관계 개념을 적용하여 색인어를 동일 공간상으로 표현하며 용어의 상대적인 개념 정도의 표현과 색인어 사이의 종속성 문제를 해결하고자 한다. 그리고 이를 기반으로 사용자 질의 용어와 색인어 사이의 내용 기반 검색 기법을 제안하여 사용자 관심도를 반영한 유연한 검색환경을 제공한다. 이는 대상 문서에서 저자의 의도를 추정하는 주제 분석 단계와 사용자 관심도를 정확하게 파악하여 관련문서를 효율적으로 검색하는 요구 분석 단계로 구성된다. 더불어 이와 같은 분석단계를 적용하면 문서의 내용을 구조화하거나 문서 집단으로 대표되는 특정 주제 영역의 지식구조를 파악하는 판단 기준으로 활용될 수 있다.

본 논문의 구성은 2장에서 질의어 확장에 대한 기존 연구를 설명하고, 3장에서는 본 논문의 기반이 되는 퍼지이론과 퍼지 검색모델을 설명한다. 4장에서는 도메인 지식을 표현하는 축소용어 집합, 시소러스 그리고 유사 관계 행렬을 생성하는 주제 분석 메커니즘과 퍼지 시소러스, 유사관계 행렬을 통한 문서 순위 재조정 알고리즘을 제안하고 5장에서는 제안된 알고리즘을 설계 및 구현한다. 그리고 6장에서는 시스템을 평가 검증하기 위해서 실험과 평가를 수행하며 마지막으로 7장에서는 결론 및 향후 연구과제에 대하여 기술한다.

Ⅱ. 관련 연구

이 장에서는 본 논문과 관련이 있는 사용자 검색 환경 및 문서구조화에 대한 내용에 관해서 기술한다.

2.1 질의어 확장 기법

대부분의 사용자는 문서 집합의 구성이나 검색환경에 대한 자세한 지식이 없기 때문에 사용자 자신의 정보 요구 목적에 적합한 질의를 구성하는 것이 매우 어렵다. 사실, 현재의 웹 검색의 인터페이스에서도 나타났듯이 효율적인 검색과 사용자 위주의 맞춤형 검색을 위해 질의를 재 작성하는데 사용자들은 많은 시간을 사용한다. 이에 따라 본 절에서는 본 연구와 관련하여 처음 사용자가 작성한 질의를 개선하도록 질의어를 확장에 대한 내용을 기술한다.

최근에 많이 연구되고 있는 질의 확장에 대한 연구는 이미 작성된 개인, 그룹의 프로파일을 이용하여 사용자 개인의 관심도와 선호도를 반영된 프로파일을 재구성 하는 기법과 의미적으로 유사한 용어들을 연결하여 시소러스를 구축하는 기법이 있다. 프로파일을 재구성하여 질의를 확장하는 기법은 사용자의 관심 분야에 관련된 키워드들로 작성된 개인이나 그룹 프로파일을 비교·참조하여 새로운 용어들로 확장시키는 기법이다[1,5,6,8].

시소러스를 이용한 질의 확장에서는 우선적으로 동의어와 하위어를 포함시키는데, 이는 하위어가 저 빈도 용어이면서 더 작은 개념을 나타내므로 검색 효과를 높여주기 때문이다. 하위어가 많을 때에는 하위어의 깊이에 제한을 두어서 제어하기도 한다.

수동 구축된 시소러스는 어떤 특정 영역에 대해서 구축되고, 그 영역의 문서들을 검색하기 위해 적용된 경우에는 아주 성공적이었으나 일반적인 영역에 대해 구축된 시소러스를 이용한 질의 확장에서는 그렇게 성공적인 성능 향상을 보이지 않았다. 또한 사람이 시소러스를 구축하는 과정은 많은 시간과 노력이 요구되며 일관성 유지가 어렵다.

2.2 클러스터링을 이용한 질의 확장

질의 확장의 목적은 사용자의 관심도를 반영한 문서를 더 많이 검색되도록 하는 것이므로, 연관 피드백 전략은 의미적으로 연결되는 문서들을 동일한 문서집단으로 클러스터링을 구성하는 개념을 기초로 하고 있다. 이 개념에 따르면 사용자의 관심과 관련된 문서로 판정된 문서는 더 큰 연관 문서 클러스터를 나타내는 용어를 포함하고 있으며, 이 경우 더 큰 연관 문서 클러스터를 나타내는 질의는 사용자의 도움을 받아서 점진적으로 질의가 확장된다[5,6,9,10].

이러한 연구 중 하나인 클러스터링을 통한 질의 확장 기법은 검색 문서를 대상으로 키워드의 동시 출현빈도를 기반으로 연관 행렬을 구성하고 이를 이용하여 질의를 확장하는 방법이다[2,3,10].

그러나 이러한 클러스터링을 적용한 질의 확장은 검색된 문서만을 대상으로 하기 때문에 질의를 확장하기 위해서는 검색된 문서의 내용에 접근해야 할 필요성이 빈번하게 발생하여 웹 환경에 적용하는 것은 비효율적이라 할 수 있다.

III. 퍼지 이론과 퍼지검색 모델

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

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.

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

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 $dom \quad R(x) = \max R(x, y)$ 이고, 치역 $ran \quad R(y) = \max R(x, y)$ 이다.

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

$$\mu_R : X \times Y \rightarrow [0,1] \quad \cdots\cdots (2)$$

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.

$$\begin{aligned} &1)\ \sigma : R^{+} \rightarrow [0,1] \\ &2)\ \sigma(F_1) > \sigma(F_2) \Leftrightarrow F_1 > F_2 \\ &3)\ \frac{d^2(\sigma)}{dF^2} \geq 0 \Leftrightarrow F \leq T_F \text{ and } \frac{d^2(\sigma)}{dF^2} \leq 0 \Leftrightarrow F \geq T_F \end{aligned} \quad \cdots\cdots (3)$$

식 3에서 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 3을 만족하는 시그모이드 소속 함수에 대한 그래프는 (그림 1)과 같이 나타낼 수 있다.

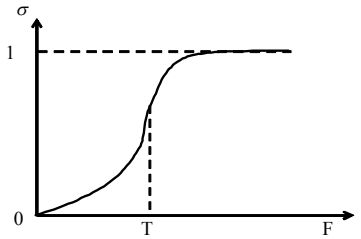

그림 1. 시그모이드 함수
Fig 1. A sigmoid function

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.

$$A_x = \{ y \mid (x,y) \in R(x,x) \} \quad \cdots\cdots (4)$$

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($\simeq$)(similarity relation)라고 부른다.

1) 반사관계: $\mu_{\simeq}(x,x) = 1$
2) 대칭관계: $\mu_{\simeq}(x,y) = \mu_{\simeq}(y,x)$ $\cdots\cdots (5)$
3) 전이관계: $\mu_{\simeq}(x,z) \geq \min\{\mu_{\simeq}(x,y), \mu_{\simeq}(y,z)\}$
   단, μ는 소속함수

퍼지 집합에서 유사관계를 만족하는 임의의 원소들은 퍼지 집합에 대한 소속정도의 값을 부여받고 일정 이상의 소속정도 값을 가진 유사클래스(similarity class)를 생성하여 유사관계를 만족하는 클래스를 분류하고 그룹화 할 수 있다. 또한, 퍼지 관계성의 하나인 호환관계(tolerance, compatibility relation)는 다음과 같이 반사와 대칭 성질을 만족하고 전이관계는 만족하지 않는다.

1) 반사관계: $\mu_{\cong}(x,x) = 1$ $\cdots\cdots (6)$
2) 대칭관계: $\mu_{\cong}(x,y) = \mu_{\cong}(y,x)$

일반적인 집합(crisp set)에서 호환관계 $R(X,X)$이 주어졌을 때 하나의 호환 클래스 집합 A는 A에 속하는 임의의 x, y에 대하여 x, y의 관계가 R에 속하면 X의 부분집합이다.

반면, 퍼지 집합에서도 퍼지 호환 관계 R이 임의의 집합 A에 주어지면, 호환관계를 만족하는 부분 집합들로 분할될 수 있는데 이와 같이 얻어진 부분집합들은 퍼지 호환 클래스라 한다. 퍼지 호환 관계에 $\alpha - cut$을 적용하여 생성된 α-호환 클래스 A_j는 다음과 같이 정의된다[4,11].

$$\mu_R(x, y) \geq \alpha, \ \forall x, y \in A_j \quad \cdots\cdots (7)$$

즉, 임의의 x, y 에 대하여 x, y의 관계가 특정 α값 이상이면 X의 부분집합으로 구성되고 이렇게 구성된 모든 호환 클래스들을 최대 호환 클래스 또는 완전 α-cover라고 한다[11].

Ⅳ. 문서 순위 재조정 알고리즘

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

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 확장된다.

4.1 시그모이드함수를 이용한 문서베이스 생성

원시 문서베이스를 구성하기 위해서는 문서 집합에서 용어를 추출하고 용어와 문서사이의 관련 정도를 가중치로 표현한다. 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. 첫째, 문서에서 추출한 색인어가 문서의 타이틀(T)이나 키워드 집합(K)에서 발생되었을 때 색인어 발생 빈도에 대한 문서에서의 중요 정도는 (그림 4-1)의 시그모이드 함수(σ_1)에 의해 <표 1>과 같이 구할 수 있다.

표 1. 타이틀, 키워드집합에서 소속 정도
Table 1. membership degrees of title and keyword set

F	0	1	2	3	4
σ_1	0	0.6	0.9	0.99	1

위의 <표 1>을 시그모이드 소속 함수를 적용하면 (그림 2)과 같다.

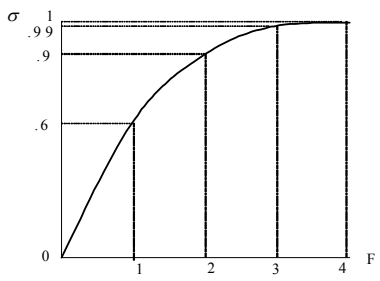

그림 2. 소속함수 (σ_1)
Fig 2. membership function(σ_1)

(그림 2)는 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 3)의 시그모이드 함수(σ_2)에 의해 <표 2>와 같이 구할 수 있다.

표 2. 요약에서의 소속 정도
Table 2. membership degrees of abstract

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

시그모이드 함수로 나타내면 (그림 3)과 같다.

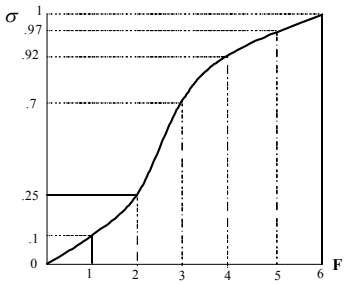

그림 3. 소속함수(σ_2)
Fig 3. membership function(σ_2)

즉, 요약 부분에서 발생하는 색인어의 빈도수는 3회 정도가 되어야 소속정도가 중요하다는 것을 나타낸다. 이와 같이 시그모이드 함수는 상대적인 확률적 빈도를 절대적인 가능성 값으로 사상시킨다. 그리고 발생 영역별 빈도에 대한 중요도를 구분하여 생성하고, 문서 전체에 대한 중요도를 산출하기 위하여 본 논문에서는 문서를 대표하는 최종적인 색인어 가중치는 발생 영역의 가중치를 min-max 연산을 적용하여 생성하였다.

$$w_{ij} = \max\{\min(\mu^{T}_{ij}\mu^{A}_{ij}), \min(\mu^{T}_{ij}\mu^{K}_{ij}), \min(\mu^{A}_{ij}\mu^{K}_{ij})\} \quad \cdots\cdots (8)$$

단, μ^{T}_{ij}: 문서 j에서 타이틀영역의 색인어 i에 대한 중요 정도

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

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

식 8에 의하여 문서집합(D)와 색인어 집합(T)의 퍼지 관계인 원시 문서베이스는 다음과 같이 표현된다.

$$\mathbf{R} = \begin{array}{c} \\ \mathbf{d}_1 \\ \mathbf{d}_2 \\ \vdots \\ \mathbf{d}_n \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{t}_1 & \mathbf{t}_2 & \cdots & \mathbf{t}_m \end{array} \\ \begin{bmatrix} \mathbf{w}_{11} & \mathbf{w}_{12} & \cdots & \mathbf{w}_{1m} \\ \mathbf{w}_{21} & \mathbf{w}_{22} & \cdots & \mathbf{w}_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{w}_{n1} & \mathbf{w}_{n2} & \cdots & \mathbf{w}_{nm} \end{bmatrix} \end{array}$$

여기서, 문서 집합의 개수는 n이고 문서 집합에서 추출된 색인어는 m개이다. 본 논문에서 색인을 추출하는 과정으로 수동 색인을 채택한다. 채택한 주요 관점은 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 문서 집합인 KT-Set(2.0)을 대상으로 색인어 발생 빈도별 소속값을 생성하고 문서 집합(D)에서 색인어(T) 의미를 표현하는 원시문서베이스(R)은 식 8을 적용하여 구성하면 다음과 같다.

$$D=\{d_1,d_2,d_3,d_4,d_5\},\quad T=\{t_1,t_2,t_3,\cdots,t_9\}$$

$$=\{\text{다층퍼셉트론, 가중치행렬, 경험적방법, 뉴런, 계층적구조, 관계모델, 관계데이터모델, 문자인식, 관계대수}\}$$

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

4.2 시간 복잡도와 축소용어 행렬

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

4.2.1 원시 문서베이스와 축소용어 행렬

축소용어 집합은 원시 문서베이스를 이용해서 생성하며 용어 집합의 부분집합으로 높은 시간 복잡도 문제를 처리한다. 본 논문에서는 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.

$$u_{A\geq B}(w)=\max\{\min(u_A(w),u_B(w)),\ \min(1-u_A(w),1-u_B(w))\} \quad \cdots\cdots (9)$$

$$t_i=u_{w_i\equiv w_j}=\frac{1}{|d|}\sum_{k=1}^{d}u_{w_i\equiv w_j}(D_k)$$

단, $u_A(w)$: 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

$u_{w_i\equiv w_j}(D_k)$: 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 9의 소속 함수를 이용하여 각 색인어가 도메인 전체 영역에서의 관계 정도를 는 각 색인어의 평균소속정도를 평가하는 방법을 이용한다[2,3]. 즉, $u_A(w)=u_B(w)$일 경우를 평가하는 방법을 응용하여 도메인에서 색인어를 평가하고 $\alpha-cut$에 의한 축소용어 집합을 생성하였다. 문서와 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.

$$R_r=\begin{array}{c} \\ d_1\\ d_2\\ \vdots\\ d_n \end{array}\begin{array}{c} \begin{array}{cccc} r_1 & r_2 & \cdots & r_r \end{array}\\ \begin{bmatrix} I_{11} & I_{12} & \cdots & I_{1r}\\ I_{21} & I_{22} & \cdots & I_{2r}\\ \vdots & \vdots & \cdots & \vdots\\ I_{n1} & I_{n2} & \cdots & I_{nr} \end{bmatrix} \end{array}$$

단, r은 축소행렬의 첨자로 $r<m$이다.

<예 2> 예 1에서 구한 원시 문서베이스(R)에서 식9를 이용하고 $\alpha-cut$을 0.95로 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, 전체 색인어의 결과 값은 다음과 같이 얻을 수 있다.

$$t_1 = 0.95,\quad t_2 = 0.90,\quad t_3 = 0.76,\quad t_4 = 0.96,$$
$$t_5 = 0.95,\quad t_6 = 0.90,\quad t_7 = 0.76,\quad t_8 = 0.96$$
$$t_9 = 0.99$$

따라서, 계산 결과가 $\alpha - cut$을 만족하는 색인어를 추출함으로서 다음과 같은 축소 용어로 구성된 행렬을 얻을 수 있다.

$$\mathbf{R}_r = \begin{array}{c} \\ \mathbf{d_1} \\ \mathbf{d_2} \\ \mathbf{d_3} \\ \mathbf{d_4} \\ \mathbf{d_5} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{r_1} & \mathbf{r_2} & \mathbf{r_3} & \mathbf{r_4} & \mathbf{r_5} \end{array} \\ \begin{bmatrix} \mathbf{0.94} & \mathbf{0.80} & \mathbf{0.00} & \mathbf{0.80} & \mathbf{0.00} \\ \mathbf{1.00} & \mathbf{0.00} & \mathbf{0.94} & \mathbf{0.00} & \mathbf{0.94} \\ \mathbf{0.94} & \mathbf{1.00} & \mathbf{0.00} & \mathbf{1.00} & \mathbf{0.00} \\ \mathbf{0.94} & \mathbf{0.00} & \mathbf{0.00} & \mathbf{0.00} & \mathbf{0.00} \\ \mathbf{0.94} & \mathbf{0.00} & \mathbf{0.00} & \mathbf{0.00} & \mathbf{0.00} \end{bmatrix} \end{array}$$

4.2.2 유사관계 시소러스

축소용어로 구성된 문서베이스를 기반으로 색인어 사이의 퍼지 관련 정도를 나타내는 시소러스를 구성한다. 시소러스는 다음 식 10과 같이 문서베이스와 원시 문서베이스의 퍼지 관계곱 연산을 이용한다.

$$S_r = R^T \otimes R_r$$
$$s_{ij} = \underset{i,j=1..n}{\vee} (\min(w_{in} I_{rj}), (\min(1-w_{in}), (1-I_{rj}))) \quad \cdots\cdots\cdots\cdots (10)$$

단, s_{ij} : 색인어 i와 축소용어 j의 유사 정도

w_{in} : 문서 n에서 색인어 i의 중요 정도

I_{rj} : 축소행렬 문서 n에서 축소용어 j의 중요정도

퍼지 관계곱 연산은 특정 문서에서 색인어간 동시 출현 빈도가 많을수록 색인어 유사성이 높다는 가정 하에 식 10의 동시 출현 빈도를 고려하였다.

여기서, 유사관계 시소러스 S_r은 다음과 같다.

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

축소용어 집합 기반의 시소러스는 색인어의 의미를 정의하기 위한 관점에서 구축하였으므로 개념 행렬(concept matrices)이라 할 수 있다.

<예 3> 예 2에서 구한 축소 행렬과 원시 문서베이스의 퍼지 관계를 식 10을 이용한 유사관계 시소러스를 구성하면 다음과 같다.

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.

$$\mathbf{S}_r = \begin{array}{c} \\ \mathbf{t_1} \\ \mathbf{t_2} \\ \mathbf{t_3} \\ \mathbf{t_4} \\ \mathbf{t_5} \\ \mathbf{t_6} \\ \mathbf{t_7} \\ \mathbf{t_8} \\ \mathbf{t_9} \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbf{r_1} & \mathbf{r_2} & \mathbf{r_3} & \mathbf{r_4} & \mathbf{r_5} \end{array} \\ \begin{bmatrix} \mathbf{0.95} & \mathbf{0.37} & \mathbf{0.24} & \mathbf{0.37} & \mathbf{0.24} \\ \mathbf{0.14} & \mathbf{0.70} & \mathbf{0.71} & \mathbf{0.70} & \mathbf{0.71} \\ \mathbf{0.38} & \mathbf{0.76} & \mathbf{0.64} & \mathbf{0.76} & \mathbf{0.64} \\ \mathbf{0.37} & \mathbf{0.96} & \mathbf{0.45} & \mathbf{0.96} & \mathbf{0.45} \\ \mathbf{0.24} & \mathbf{0.45} & \mathbf{0.99} & \mathbf{0.45} & \mathbf{0.99} \\ \mathbf{0.22} & \mathbf{0.60} & \mathbf{0.61} & \mathbf{0.60} & \mathbf{0.61} \\ \mathbf{0.38} & \mathbf{0.76} & \mathbf{0.64} & \mathbf{0.76} & \mathbf{0.64} \\ \mathbf{0.37} & \mathbf{0.96} & \mathbf{0.45} & \mathbf{0.96} & \mathbf{0.45} \\ \mathbf{0.24} & \mathbf{0.45} & \mathbf{0.99} & \mathbf{0.45} & \mathbf{0.99} \end{bmatrix} \end{array}$$

4.2.3 유사관계 행렬

본 논문에서는 검색 순위를 재조정하기 위한 방안으로 다음 식 11과 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.

$$S = R^T \otimes R$$

$$\overline{s_{ij}} = \bigvee_{i=1,m} (\min(w_{im}, w_{mi}), (\min(1-w_{im}, 1-w_{mi})) \quad \cdots\cdots (11)$$

단, $\overline{s_{ij}}$: 색인어 i와 j의 유사 정도

w_{im}: 원시 문서베이스의 문서 m에서 색인어 i의 중요 정도

여기서, 유사관계 행렬 (S)는 다음과 같다.

$$\mathbf{S} = \begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_m \end{array} \\ \begin{bmatrix} \overline{S}_{11} & \overline{S}_{12} & \cdots & \overline{S}_{1m} \\ \overline{S}_{21} & \overline{S}_{22} & \cdots & \overline{S}_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ \overline{S}_{m1} & \overline{S}_{m2} & \cdots & \overline{S}_{mm} \end{bmatrix} \end{array}$$

<예 4> 예 3의 축소용어 집합 기반의 호환 관계를 식 11을 이용하여 유사관계 행렬을 구축하면 다음과 같다.

유사관계 행렬(S)의 원소 $\overline{s_{23}}$의 계산 절차는 다음과 같다.

$$\begin{aligned} \overline{s_{23}} = & \max(\min(0.50, 0.50), \min(1-0.50, 1-0.50)) + \\ & \max(\min(0.00, 0.50), \min(1-0.00, 1-0.50)) + \\ & \max(\min(0.00, 0.80), \min(1-0.00, 1-0.80)) + \\ & \max(\min(0.00, 0.00), \min(1-0.00, 1-0.00)) + \\ & \max(\min(0.00, 0.00), \min(1-0.00, 1-0.00)) = 3.20 \end{aligned}$$

이고 $\overline{s_{23}} = 3.20/5 = 0.64$이므로 다른 원소도 같은 방식으로 계산하면 된다.

$$\mathbf{S} = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{array} \begin{array}{c} \begin{array}{ccccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 & t_9 \end{array} \\ \begin{bmatrix} 1.00 & 0.14 & 0.38 & 0.37 & 0.24 & 0.22 & 0.38 & 0.37 & 0.24 \\ 0.14 & 1.00 & 0.64 & 0.70 & 0.71 & 0.80 & 0.64 & 0.70 & 0.71 \\ 0.38 & 0.64 & 1.00 & 0.76 & 0.64 & 0.54 & 0.76 & 0.76 & 0.64 \\ 0.37 & 0.70 & 0.76 & 1.00 & 0.45 & 0.60 & 0.76 & 0.96 & 0.45 \\ 0.24 & 0.71 & 0.64 & 0.45 & 1.00 & 0.61 & 0.64 & 0.45 & 0.99 \\ 0.22 & 0.80 & 0.54 & 0.60 & 0.61 & 1.00 & 0.54 & 0.60 & 0.61 \\ 0.38 & 0.64 & 0.76 & 0.76 & 0.64 & 0.54 & 1.00 & 0.76 & 0.64 \\ 0.37 & 0.70 & 0.76 & 0.96 & 0.45 & 0.60 & 0.76 & 1.00 & 0.45 \\ 0.24 & 0.71 & 0.64 & 0.45 & 0.99 & 0.61 & 0.64 & 0.45 & 1.00 \end{bmatrix} \end{array}$$

4.3 퍼지관계를 이용한 질의어 확장

4.2장의 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.

4.3.1 사용자 질의 표현

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

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) $VERY x_i = q_{very}(\mu(x_i)) = \mu(x_i)^2$ ····· (12)
5) $FAIRY x_i = q_{fairy}(\mu(x_i)) = \mu(x_i)^{1/2}$

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

질의 벡터(q)는 식 12를 만족하며 사용자 질의는 다음 예 5와 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 $q = \{t_4/0.8\}$이면 질의벡터는 다음과 같이 표현된다.

$$Q = q \times T = \{ \begin{array}{ccccc} t_1/0.0, & t_2/0.0, & t_3/0.0, & t_4/0.80, & t_5/0.0, \\ t_6/0.0, & t_7/0.0, & t_8/0.0, & t_9/0.0 & \end{array} \}$$

4.3.2 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 13과 같이 사용자에 의해서 표현된 질의(Q)와 축소용어 집합과 문서와의 퍼지 관계를 나타내는 문서베이스(S_r) 사이의 퍼지 합성에 의해 생성한다.

$$\mu_{Q \circ S_r}(x,z) = \max_{y \in Y} \{ Min\{ \mu_Q(x,y), \mu_{S_r}(y,z) \} \} \quad \cdots\cdots (13)$$

단, $\mu_Q(x,y)$: 사용자 질의와 색인어와의 관계

$\mu_{S_r}(y,z)$: 문서와 축소용어 집합과의 관계정도

$\mu_{Q \circ S_r}(x,z)$: 질의와 축소용어 집합과의 관계정도

여기서, 질의 확장집합 Q_r은 다음과 같다.

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

<예 6> 질의 확장집합 Q_r은 식 13을 이용하여 시소러스와 퍼지 합성을 통해 확장된다.

질의베이스(Q_r)의 계산 절차는 다음과 같다.

$qr_1 = max\{ min(0.00, 0.95), min(0.00, 0.14), min(0.00, 0.38)$
$min(0.80, 0.37), min(0.00, 0.24), min(0.00, 0.22)$
$min(0.00, 0.38), min(0.00, 0.37), min(0.00, 0.24)$
$\}$

이므로, 같은 방식으로 계산하면

$Q_r = Q \circ S_r = \{ r_1/0.37, r_2/0.80, r_3/0.45, r_4/0.80, r_5/0.45 \}$이다.

4.3.3 유사성 평가

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

$$P_r = Q_r \otimes R_r^T = \{ RSV(d_1), RSV(d_2), \cdots, RSV(d_n) \} \quad \cdots\cdots\cdots\cdots (14)$$

$$RSV(d_i) = \frac{\sum_{\vec{q(j)} \neq 0 \text{ and } j=1,\cdots,n} T(1 - |I_{ij} - Qr_j|)}{k}$$

단, 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 14에 의하여 평가하여 1단계 문서 검색(퍼지 검색) 절차 및 결과는 다음과 같다.

$d_1 = (1 - |0.94 - 0.37|) + (1 - |0.80 - 0.80|) +$
$(1 - |0.00 - 0.45|) + (1 - |0.00 - 0.45|) +$
$(1 - |0.00 - 0.71|) = 3.53$

$RSV(d_1) = 3.53/5 = 0.71$ 이므로, 같은 방식으로 계산하면

$P_r = Q_r \otimes R_r^T$

$= \{ d_1/0.71, d_2/0.36, d_3/0.63, d_4/0.39, d_5/0.39 \}$이다.

4.4 문서 관련성 평가를 통한 클러스터 검색

기존의 시소러스의 사용 방법은 검색 영역을 확장하는 것으로 문서 검색의 정확률 및 재현율을 향상시킬 수 있다.

그러나 정확한 정보의 검색에는 완전하지 못함으로 본 논문에서는 영역 지식을 확장하는 방법으로 퍼지 유사관계 행렬을 활용하여 질의 개념을 확장하는 방법을 제안한다.

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($q = \{ t_4/0.8 \}$)에 대해서, 분류정도($0.50 - cut$)와 식13에 따라 호환클래스 생성하고 질의를 확장하면 다음과 같다(유사한 문서집합에 대한하여 변별력을 향상시키기 위하여 $\alpha - cut$은 도메인에 의존적으로 재조정된다).

사용자 질의 $q = \{ t_4/0.8 \}$에 대하여 호환클래스는 유사관계에 따라 사용자 질의에 확장되는 용어(Q_e)는

$t_2, t_3, t_4, t_6, t_7, t_8$이다.

따라서, 색인어 유사관계 행렬 기반의 호환클래스는

$C/R_{0.50} = \{ C_1 = \{ t_2, t_3, t_4 \}, C_2 = \{ t_3, t_4, t_6, t_7 \}, C_3 = \{ t_2, t_4, t_6, t_8 \}$

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

$Q_e = \{ (t_2, 0.70), (t_3, 0.76), (t_4, 0.80),$
$(t_6, 0.60), (t_7, 0.76), (t_8, 0.96) \}$

가 된다.

확장된 용어의 가중치는 관련연구에서 개념 네트워크 기반 검색 방법의 전이 관계를 이용하였다. 원시 문서베이스

가 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).

$$P = Q_e \otimes R = \{RSV(d_1), RSV(d_2), \cdots, RSV(d_n)\} \quad \cdots\cdots (15)$$

$$RSV(d_i) = \frac{\sum_{q(j) = r' \text{ and } j=1,\cdots,n} T(1 - |w_{ij} - Qe_j|)}{k}$$

단, 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_2/0.70, t_3/0.76, t_4/0.80, t_5/−, t_6/0.60, t_7/0.76, t_8/0.96, t_9/−}

R	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9
d_1	0.94	0.50	0.50	0.80	0.00	0.50	0.50	0.80	0.00
d_2	1.00	0.00	0.50	0.00	0.94	0.00	0.50	0.00	0.94
d_3	0.94	0.00	0.80	1.00	0.00	0.00	0.80	1.00	0.00
d_4	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
d_5	0.94	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00

에 대하여 식 15를 적용하면

P = { d_1/0.84, d_2/0.40, d_3/0.73, d_4/0.24, d_5/0.32 }

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.

4.5 유사도결합을 통한 순위 재조정 알고리즘

본 논문에서는 축소용어 집합을 기반으로 작성된 시소러스를 통한 각 문서의 검색상태 값(P_r)과 원시 문서베이스를 기반으로 작성된 각 문서의 검색 상태 값 (P)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.

$$Sim_{combined} = \alpha P_r + \beta P = \{RSV(d_1), RSV(d_2), \cdots, RSV(d_n)\} \quad \cdots\cdots (16)$$

단, P_r: 축소용어 행렬 기반 퍼지검색(문서상태 값)

P: 유사관계 행렬 기반 클러스터 검색(문서상태 값)

α, β : 1로 설정

$Sim_{combined}$: 문서 상태 값의 순위 재조정 결과

<예 10> 식 16에서 도메인의 특성에 따라 α, β을 적용하여 재조정한다.

최종적인 문서 상태 값은

P_r = { d_1/0.71, d_2/0.36, d_3/0.63, d_4/0.39, d_5/0.39 }이고,

P = { d_1/0.84, d_2/0.40, d_3/0.73, d_4/0.24, d_5/0.32 },

여기서 α와 β 값을 1:1로 설정하면

$Sim_{combined}$ = { d_1/0.78, d_2/0.38, d_3/0.68, d_4/0.32, d_5/0.36 }

이다. 따라서, 검색상태 값이 0.5 이상을 선택하면

$Sim_{0.5}$ = { d_1/0.78, d_3/0.68 }이다.

V. 시스템 설계 및 구현

본 논문에서는 학술분야 전문 정보검색을 위하여 표제, 요약 그리고 키워드에 대한 문서 구조적인 지식을 기반으로 색인어 쌍의 동시 출현 빈도를 이용하여 문헌 지식구조를 파악한다. 또한 문서내에서의 상호빈도에 대한 퍼지 관계를 이용하여 유사관계 시소러스를 생성한다. 이는 색인어를 개념공간상의 한 개념으로 취급함으로서 문서기반 색인이라 할 수 있다. 또한 문서베이스에서 누락된 색인어 정보를 고려하기 위한 방안으로 동시 출현빈도를 기반으로 퍼지 호환관계를 만족하는 유사관계 행렬을 구성한다. 전체시스템의 흐름은 (그림 4)과 같으며 검색시스템은 사용자 위주의 검색 확장을 통해 1차 퍼지검색을 수행한다.

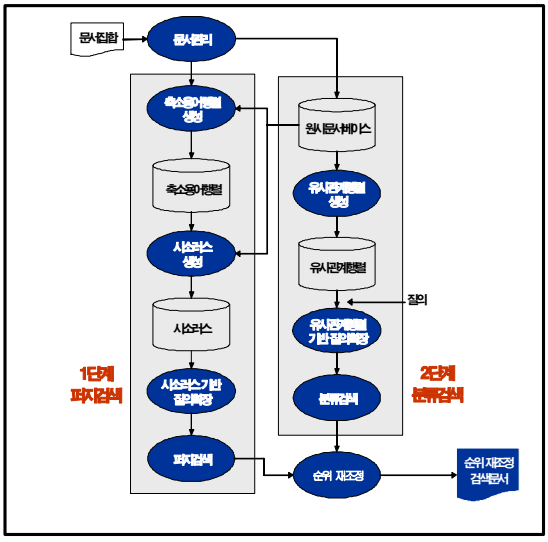

그림 4. 2단계 질의 처리 모델
Fig 4. Two-Level Query Processing Model

5.1 검색시스템 설계

본 논문에서 제안한 검색 시스템은 (그림 5)과 같이 입력된 문서들을 대상으로 문서에 대한 정보를 추출하여 색인을 생성하고 관리하는 부분인 문서분석기, 원시 문서베이스를 기반으로 키워드간의 상호 의존관계를 통해 문서내용을 분석하기 위한 내용분석기, 그리고 확장된 질의를 이용하여 문서를 검색하고 유사도 결합을 통한 문서 순위를 재조정하는 검색처리기로 구성된다.

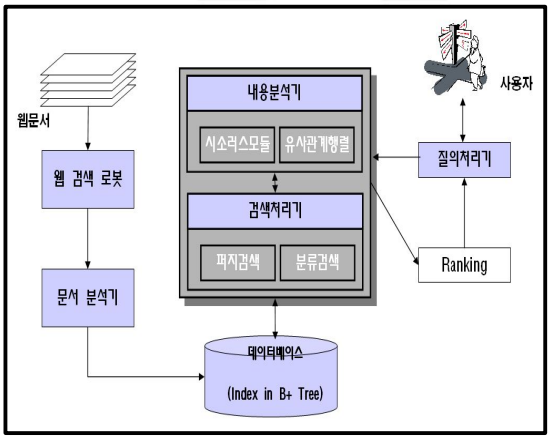

그림 5. 검색시스템 구조
fig 5. retrieval system structure

5.2 시스템 구현

본 시스템의 구현 환경은 IBM-PC 호환 기종에서 MS사의 Windows XP 운영체제에서 MSSQL 2000의 데이터베이스를 기반으로 C# 프로그램을 이용하여 구현하였다. 구현된 문서 순위 재조정 시스템은 문서 등록기, 내용분석기 그리고 문서 검색기로 구성하였다.

본 논문에서 형태소 분석은 문서의 문장 단위나 텍스트 파일 형태로 입력받아 형태소 분석 및 태깅, 구문분석을 통하여 1000개의 문서를 대상으로 KT-Set(2.0)의 키워드와 색인어 발생빈도 수를 고려하여 색인어를 추출하였다.

5.2.1 문서 등록 화면

본 논문에서 구현한 시스템의 문서 등록 부분은 다음 (그림 6)과 같이 실험 대상의 데이터를 문서 등록 구조에 따라 입력할 수 있도록 설계하였다.

구현 과정을 기술하기 위하여 실험 결과의 예로 KT-Set(2.0) 질의번호 25와 26의 21개의 적합문서에서 추출한 색인어 140개를 기준으로 하였다.

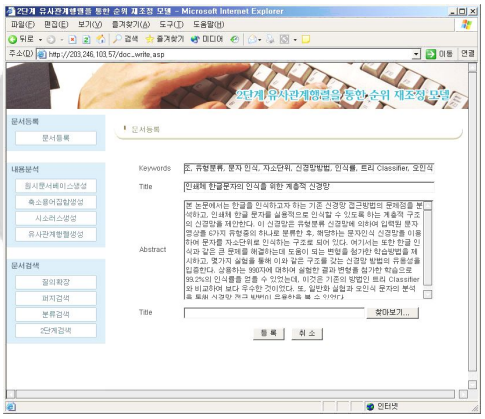

그림 6. 문서 등록
Fig 6. document registration

5.2.2 각 색인어의 발생 영역별 소속 정도 계산

본 시스템의 문서 등록 구조에 따라 저장되며 키워드, 타이틀 그리고 요약 색인어 집합에서 발생된 빈도수에 대한 가중치는 소속 함수에 따라 구분되어 계산된다. 그리고 문서를 대표하는 최종적인 색인어 가중치는 식 8의 발생영역의 가중치 연산을 통해 계산된다. 문서집합과 색인어 집합의 퍼지관계를 표현하는 원시 문서베이스는 다음 (그림 7)과 같다.

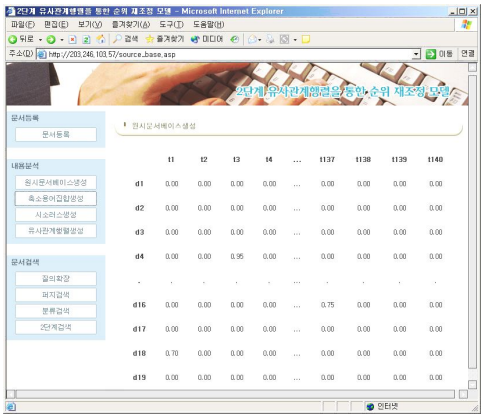

그림 7. 원시 문서베이스
fig 7. original document base

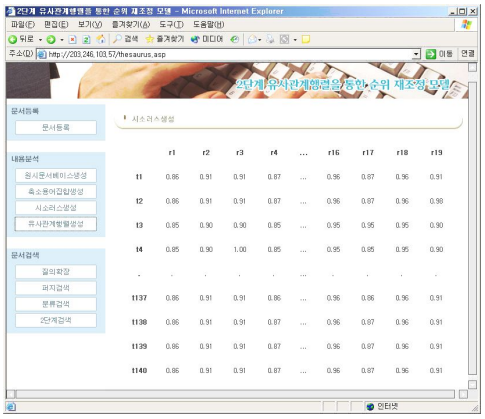

그림 9. 유사관계 시소러스
fig 9. similarity relation thesaurus

5.2.3 축소용어행렬 생성기

퍼지 정보검색 시스템의 높은 시간 복잡도를 해결하고 도메인 의존적인 축소 용어집합을 생성하기 위하여 각 색인어 관계 값을 도출하는 방법인 식 9의 소속 함수를 적용하였으며 (그림 8)는 원시 문서와 축소 용어와의 관계행렬을 표현한다.

그림 8. 축소용어 집합
fig 8. reduced term set

5.2.4 시소러스 생성기

문서를 대표하는 축소용어로 구성되어 문서 상태를 표현하고 문서베이스를 기반으로 개념 공간상의 색인어 사이의 관련정도를 구현하기 위해, 먼저 앞장에서 정의한 식 10을 적용하여 (그림 9)와 같이 유사관계 시소러스를 생성한다.

5.2.5 유사관계행렬 생성기

축소 용어로 구성된 문서베이스에서 누락된 색인어 정보를 고려하고 클러스터링을 이용한 군집생성을 통한 검색을 지원하기 위하여 식 11을 적용하여 (그림 10)와 같이 유사관계 행렬을 생성한다.

그림 10. 유사관계 행렬
fig 10. similarity relation matrix

5.2.6 퍼지검색

본 시스템에서 퍼지 검색 모듈은 도메인 지식을 확장하기 위해 시소러스와 퍼지 합성을 통해 질의벡터를 질의베이스로 관련 용어들로 확장하여 식 12, 식 13을 적용하여 확장 질의베이스를 구성하며 확장된 질의베이스와 문서베이스의 유사성 척도를 통한 퍼지검색을 수행한다. 유사성 척도는 식 14를 적용하며 (그림 11)와 같이 1차 퍼지검색을 수행한다.

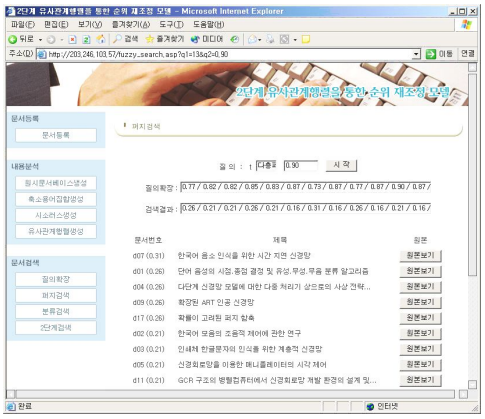

그림 11. 1단계 퍼지검색
fig 11. 1-step fuzzy retrieval

5.2.7 분류검색

퍼지 검색은 검색 영역을 확장하여 검색성능을 향상시킬 수 있는 장점은 있으나 문서간의 변별력이 부족하여 사용자 질의와 의미적으로 연결된 검색에는 완전하지 못하다. 따라서 본 논문에서는 퍼지 호환 관계를 만족하는 색인어 유사 정도의 값에 알파-컷을 적용하여 일정 이상의 유사도 값을 가진 색인어 클래스를 생성하여 질의어를 확장할 수 있게 한다. 본 시스템에서 분류검색 모듈은 식 15를 적용하여 (그림 12)와 같이 분류 검색을 지원한다.

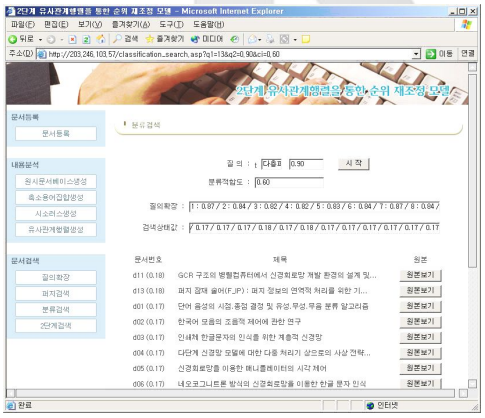

그림 12. 2단계 분류 검색
fig 12. 2-step classification retrieval

5.2.8 문서순위 재조정기

본 시스템의 문서 순위 재조정 모듈은 1단계 퍼지검색에서 검색된 문서의 검색 상태 값 그리고 2단계 내용 기반 검색을 지원하는 분류 검색을 통하여 검색된 문서 상태 값을 식 16을 적용하여 검색 순위를 재조정함으로서 사용자에게 적합한 문서가 상위에 검색될 수 있도록 한다.

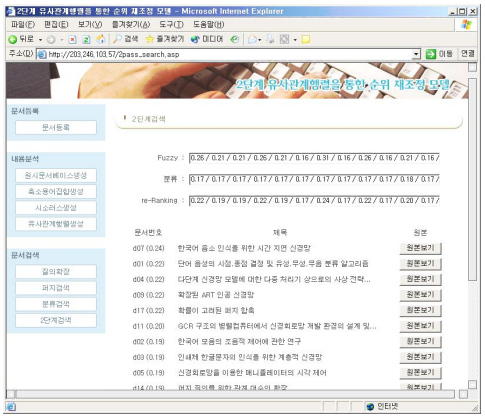

그림 13. 순위 재조정 검색 모듈
fig 13. Re-ranking retrieval module

사용자 요구가 반영된 순위 재조정 검색 환경은 (그림 13)과 같다.

VI. 실험 및 평가

본 논문에서 제안한 문서 순위 재조정 알고리즘의 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 구축하여 분류정도에 따른 검색을 수행하였다. 넷째, 퍼지 검색과 클러스터 검색의 유사도 결합을 통하여 순위를 재조정 모델의 재현율(recall)과 정확률(precision)을 평가하였다.

본 실험에서는 제안된 모델의 성능을 비교·평가하기 위하여 3개의 순위 검색 모델을 선정하여 평가한다. 첫 번째 모델은 [12,13]의 문서 가중치 순위 결정방법이고, 두 번째 모델은 [6]의 퍼지 소속 함수를 이용한 퍼지 검색 기법이며, 세 번째 모델은 [3]의 계층적 용어 클러스터 기법이다.

6.1 실험 분석 방법

본 논문의 검색 효율(retrieval effectiveness)을 측정하기 위해 재현율(recall)과 정확률(precision) 척도를 이용한다. 검색된 문서가 적합한 문서를 어느 정도 포함하는가를 나타내는 정확률과 전체 문서 중에서 적합한 문서를 어느 정도 찾아내는가를 나타내는 재현율의 의미를 본 실험의 성능평가에 적용하기 위하여 순위 정확률과 순위 재현율을 이용하였다[1,2]. 일반적인 검색 효율 척도를 응용한 순위 정확률 및 순위 재현율의 공식은 식 17, 식 18과 같다.

$$order\_precision = \frac{Match_{Doc}}{Rank_{Doc|n}} \quad \cdots\cdots (17)$$

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

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

그리고 일반적인 재현율의 의미를 적용하기 위해 순위 재현율의 공식은 식 18과 같다.

$$order\_recall = \frac{KT-Set_{match}}{Rank_{match}} \quad \cdots\cdots (18)$$

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

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

$$n = KT-Set_{match}$$

6.2 실험 대상

본 논문에서 제안한 문서 순위 재조정 알고리즘의 비교·평가를 위하여 이용하는 실험 대상은 다음과 같다. 실험 문서는 KT-Set(2.0)의 전체 문서 중에서 표제, 요약, 키워드 집합으로 구조화된 문서 1000개(문서번호1~1000번)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 총 50개의 질의로 구성되어 있으나 실험 결과의 신뢰성을 높이기 위하여 질의에 대한 적합 문서수가 같이 5개 이상인 10개의 테스트 질의를 추출하여 실험 질의 데이터로 이용한다.

테스트 질의에 대한 적합 문서수가 <표 3>과 같이 적합문서가 5개 이상인 테스트 질의를 추출하여 실험 질의 데이터로 이용한다.

표 3. 실험집합의 통계정보
Table 3. statistical group of test set

질의 번호	질의어 수(질의 내용)	적합문서 수	분류검색의 분류 기준값
8	2(멀티미디어&데이타베이스)	22	0.70
16	2(자동번역|기계번역)	7	0.65
21	2(문자인식|필기체인식)	19	0.65
22	2(지능형&정보검색)	5	0.60
25	1(지리정보)	8	0.60
26	2(음성인식|음성생성)	13	0.75
33	3(초고속&정보&통신망)	77	0.70
39	2(신경망&퍼지제어)	40	0.75
43	2(시소러스|형태소)	22	0.70
50	2(샘플|샘플링)	6	0.75

6.3 비교 평가

본 실험에의 평가 결과는 다음 (그림 14)과 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.

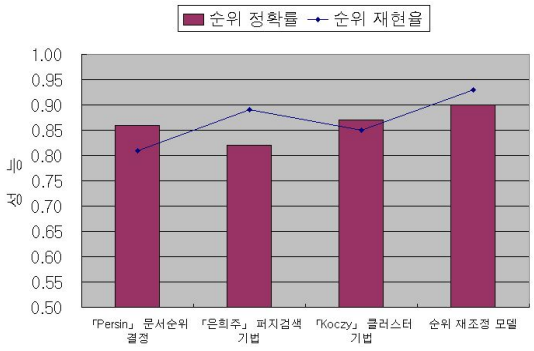

그림 14. 순위 재현율과 순위 정확률 평가
fig 14. evaluate the order by recall and precision

보 요구를 충족시킬 수 있도록 문서 검색 순위를 재조정하였다.

본 논문의 실험적 평가를 위해 순위 정확률과 순위 재현율을 이용하였고 실험 집합으로는 KT-Set의 총 50개의 테스트 질의들 중에서 10개 질의를 기준으로 실험하였다. 본 논문에서 제안한 재순위화 모델의 평가에서는 (그림 14)와 같이 퍼지 검색의 단점인 순위 정확률과 문서 클러스터 기법의 단점인 순위 재현율을 상호 보안할 수 있는 방법을 연구 하였으며 순위재조정 모델의 순위 정확률과 순위 재현율이 0.9이상으로 나타남으로서 퍼지 검색의 재현율을 유지하면서 정확률이 향상되었음을 알 수가 있었다. 그러나, 퍼지 검색은 시간복잡도 측면에서는 그 성능을 개선하였지만 축소용어집합이 작아질수록 검색결과의 신뢰도가 감소할 수 있다. 따라서 신뢰도 감소를 최소화 할 수 있도록 도메인에 의존적인 자동화 방안이 요구된다.

Ⅶ. 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 제안하였다. 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 확장시킨 방안도 제안하였다. 이를테면, 본 논문은 퍼지 검색, 문서 클러스터 기법 및 유사성 결합을 통한 재순위화 모델이라고 할 수가 있다. 먼저 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 정

참고문헌

[1] 우선미, “사용자 프로파일과 잠재적 구조분석을 이용한 검색된 문서의 순위 결정 방법”, 전북대학교 대학원 박사학위논문, 2001.8

[2] 이기영, 은희주, 김용성, “2단계 유사관계행렬을 기반으로 한 순위 재조정 검색 모델”, 정보과학회논문지, 제31권 제11호, 2004.11

[3] 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

[4] 은희주, “퍼지함수와 관계성을 적용한 질의 확장 및 문서 분류 시스템”, 전북대학교 대학원 박사학위논문, 2003.8

[5] 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.

[6] J.C.Lamirel, M.B. Ahmed, "A System for information Retrieval based on classification Components: SARCI," The 5th world Multi-conference on Informatics SCI, IIIS, IEEE, 2001
[7] 김창민, 김용기, "퍼지 관계곱 기반 퍼지정보 검색 시스템 구현", 정보처리학회 논문지, 제8-B권 제2호, 2001.4, pp. 115-122.
[8] R. Hoch "Using Information Retrieval techniques for text classification in document analysis," SIGIR' 98, 1998
[9] Yang, Y., "An evaluation of statistical approaches to text categorization," Journal of Information Retrieval, Vol 1, No.1 1999
[10] 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.
[11] 오길록, "퍼지이론 및 응용", 홍릉출판사, 1997
[12] Michael Persin, "Document Filtering for Fast Ranking," ACM-SIGIR, pp. 339-348, 1994
[13] 정희진, 정충영, "퍼지이론을 이용한 정보시스템", 한국컴퓨터정보학회 논문지, 제9권 제4호, 2004.9

저 자 소 개

이 기 영

1992년 2월 광주대학교 컴퓨터과학과 졸업(이학사)
1994년 2월 전북대학교 전산통계학과 졸업(이학석사)
2005년 2월 전북대학교 전산통계학과(이학박사)
1998년 3월~현재 원광보건대학 정보컨텐츠과 부교수
<관심분야> 퍼지 클러스터링, 정보검색, 데이터 마이닝

김 영 운

2003년 원광대학교 컴퓨터정보통신공학부졸업(공학사)
2005년 원광대학교 컴퓨터공학과 졸업(공학석사)
1997년~1999년 영원한친구 개발팀장
2000년~현재 파라(PARA) 대표
2003년~현재 원광보건대학 정보컨텐츠과 겸임교수
<관심분야> 컴퓨터그래픽스, 영상처리, EAI, XML