論文
4-10-5

2계층 유사관계행렬 구축을 통한 질의 처리 (Fuzzy Query Processing through Two-level Similarity Relation Matrices Construction)

이기영(Ki-Young Lee)*

요 약

본 연구에서는 학술논문을 대상으로 하여 표제와 초록에 대한 2단계 색인어 유사관계행렬을 구축하였다. 동시출현빈도 기반으로 구축된 색인어 유사관계행렬은 호환관계에 따른 질의 확장으로 재현률을 유지하면서 2단계 내용기반 검색으로 정확률을 향상시키기 위한 색인구조이다.

따라서, 주제 분석을 통해 영역지식을 추출하고 이용자의 정보 요구와 영역지식을 퍼지논리 기반으로 추론하였다. 본 연구는 질의에 본질적으로 가지고 있는 용어 불일치 및 정보표현을 향상시키기 위한 연구이다.

Abstract

This paper construct two-level word similarity relation matrices about title and abstract to scientific treatise. As guide keyword similarity relation matrices which is constructed to co-occurrence frequency base same time keeps recall rater by query expansion by tolerance relation, it is index structure to improve the precision rate by two-level contents base retrieval.

Therefore, draw area knowledge through subject analysis and reasoned user's information request and area knowledge to fuzzy logic base.

This research is research to improve vocabulary mismatch problem and information expression having essentially on query.

1. 서 론

정보검색시스템은 시스템의 이용자가 필요로 하는 정보를 수집하여 내용을 분석한 뒤, 찾기 쉬운 형태로 조직하여 두었다가 정보에 대한 요구가 발생할 때 적합한 정보를 검색하여 제공하는 시스템이다[1]. 즉, 이용자의 정보습득을 효율적으로 할 수 있도록 도와주는 것이다. 대부분의 불리안 논리 모델을 기반으로 하는 정보검색시스템은 문헌이 색인 어휘에 의해서 정확하게 표현되며 이용자의 검색요구를 불리안 탐색요구에 의해서 정확하게 표현된다는 가정을 기반으로 한다[2,3,6]. 그러나 질의어의 엄격한 해석, 검색 결과 우선순위 제공, 색인 결정시 존재하는 불확실성에 대한 대비책 부재와 같은 검색 효율에 한계성을 드러낸다. 따라서, 검색효율의 한계를 극복하기 위한 문헌정보검색시스템은 대상문헌의 저자의 의도를 정확하게 추정하는 주제분석과 이용자의 정보요구를 정확하게 파악하는 요구분

*정회원 원광보건대학 컴퓨터응용개발과 조교수

논문접수 : 2003. 9. 15.
심사완료 : 2003.10. 1.

석을 통하여 양자간의 커뮤니케이션을 증진시킴으로서 검색효율성을 향상시킬 수 있다. 첫째, 주제분석과정은 색인이나 초록과 같은 문헌을 대표하는 색인처리 과정이며, 둘째, 요구분석과정은 이용자의 질문내용의 의미를 파악하여 관련된 문헌을 효과적으로 검색할 수 있도록 하는 탐색모형을 수립하는 과정이다. 그러나 주제분석과 요구분석은 과정상 불확실성과 애매성을 내포하게 된다.

따라서 본 논문은 주제분석 과정에서 발생하는 애매한 정보를 효율적으로 처리하고 해당 분야의 전문가의 지적 판단에 의존하는 전역적인 지식이 아닌 각 용어의 발생빈도를 이용하여 특정문헌에 대한 용어의 종속성을 퍼지화하고 각 문헌의 동시발생빈도를 이용하여 유사관계행렬을 구축한다. 영역지식을 표현하는 유사관계행렬은 퍼지 집합이론, 퍼지 합성을 통해 지식을 표현하고 추론한다. 대표적인 연구로는 퍼지 개념 네트워크를 기반으로 하는 검색[7,12,13]등이 있다.

또한, 요구분석 과정은 질의어의 형태로 주어지지만 이를 해석할 구체적인 방법은 정해져 있지 않다. 단지 검색요구에서 발생하는 애매한 정보를 처리하기 위해 사용자의 검색요구가 표현된 질의어를 입력받아 질의용어와 색인어들의 유사정도를 파악하는데 그치고 있다.

따라서, 질의어가 내포하고 있는 의미를 파악하여 질의를 확장하고 내용기반 검색을 통해 정확률을 향상시킬 수 있는 탐색모형을 수립한다. 대표적인 연구로는 유사관계 시소러스 기반 자동질의 확장[1,8], 지식베이스기반 퍼지 정보검색 기법[7,12,13] 등이 있다.

본 논문에서는 유사관계행렬을 통한 주제분석과 사용자 요구를 분석하기 위한 탐색모형을 수립한다. 주제분석은 동시 출현빈도 기반이며 탐색과정은 유사관계행렬의 호환클래스를 통한 질의 확장으로 용어불일치 문제를 해결하고자한다. 또한 재현률을 보장하면서 기존 퍼지 검색시스템의 단점인 정확률을 향상시키기 위한 2단계 내용기반 탐색모형을 수립하게 된다. 이를 설명하기 위해 본 논문을 다음과 같이 구성하였다. 2장에서는 관련연구로서 개념네트워크(concept network)을 통한 주제분석 및 탐색모형에 대하여 설명하고 3장에서는 학술논문을 대상으로 2단계 퍼지 유사관계행렬을 제안하고 구축한다. 4장에서는 퍼지 유사관계행렬을 통한 문헌검색 메커니즘을 자세하게 설명한다. 마지막으로 5장에서는 결론 및 향후 연구방향에 대하여 기술한다.

2. 개념 네트워크

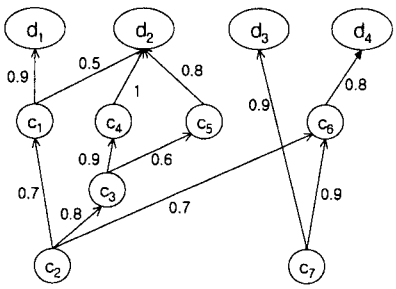
대상문헌이 포함하는 저자의 의도를 정확하게 추론하고 파악함으로서 영역지식을 표현하는 자동 키워드 색인에 대한 연구는 광범위하게 연구되고 있다.

기존의 정보검색시스템이 채택하고 있는 주제분석 추론과정은 일반적으로 통계적 속성에 기초한 것이다. 문헌을 대표하는 색인어 관계에 의해 문헌집합을 정의하며 색인어 문헌 빈도나 색인어 사이에 존재하는 통계적 의존성을 주제 영역에서의 지식으로 이용한다[2,9,10].

Shyi-Ming Chen, Jeng-Yih Wang은 퍼지 검색에서 주제분석과정인 개념네트워크를 정의하고 이를 통한 탐색모형을 제안하였다. 개념네트워크는 노드와 링크로 구성되며 각 노드는 개념이나 문서를 표현하며 링크는 개념들 사이의 의미적인 관계 또는 문서와 개념간의 관계를 표현한다. 다음과 같이 문서집합과 개념집합을 예를 들어 [그림 2-1]와 같이 구성되었다고 가정한다.

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

$$C = \{c_1, c_2, \cdots, c_7\}$$



[그림 2-1] 개념 네트워크
[Fig. 2-1] A Concept Network

[그림 2-1]은 영역 전문가에 의해서 구성된 개념들 사이의 의미적인 연결을 나타낸 것이다. 모든 도메인 개념들의 집합 C에 대해 개념네트워크 F(C)는 다음과 같은 관계집합으로 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_1와 c_2 개념들 사이의 퍼지 관련정도를 의미한다. [그림 2-1]에서 문서 d_2는 다음과 같은 개념집합으로 표현된다.

$$F(c_i, c_j) \in [0,1]$$

$$d_2 = \{(c_1, 0.5),\ (c_4, 1),\ (c_5, 0.8)\}$$

[그림 2-1]에서 문서 d_2의 개념집합과 하위 집합간의 관계는 전이적 성질(퍼지 이행관계)을 이용한다.

즉,

$\langle c_1, c_2, w_{1.2}\rangle$, $\langle c_2, c_3, w_{2.3}\rangle \in F(C)$ 이면 $\langle c_1, c_3, w_{1.3}\rangle \in F(C)$ 가 성립한다.

[그림 2-1]에 표현된 예제를 이용하여 설명하면 $\langle c_1, c_2, 0.7\rangle$이고 $\langle c_2, c_3, 0.8\rangle$이면 c_1와 c_3의 퍼지 관련정도 $w_{1.3}$은 자데(Zadeh) 퍼지 확장 논리를 통해

$$F(c_i, c_k) = \min(F(c_i, c_j), F(c_j, c_k))$$

$$w_{1.3} = \max(\min(c_1, c_2), \min(c_2, c_3))$$로 평가되어 c_1와 c_3의 퍼지관계는 0.7의 가중치를 가지게된다.

동일하게,

$$F(c_1, c_2), F(c_3, c_4), \cdot\ \cdot F(c_{n-1}, c_n)$$

에 대한 퍼지 관련정도에 의해

$$F(c_1, c_n) = \min(F(c_1, c_2), F(c_2, c_3) \cdot\ \cdot F(c_{n-1}, c_n)$$

평가할 수 있다. 개념 네트워크를 통한 문서 d_2에 대한 검색 상태 값(RSV)은 자데의 전이관계를 이용하여 다음과 같이 3가지의 루틴에 의해서 평가된다. 단, 유향링크(directed links)에 대한 고려는 배제하였다.

$$Q = \{(c_2, 1.0)\}$$

1) $c_2 \rightarrow c_1 \rightarrow d_2$

$$\min(F(c_2, c_1), F(c_1, d_2)) = \min(0.7, 0.5) = 0.5$$

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

$$\min(F(c_2, c_3), F(c_3, c_4), F(c_2, d_2)) = \min(0.8, 0.9, 1) = 0.8$$

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

$$(F(c_2, c_3), F(c_3, c_5), F(c_5, d_2)) = \min(0.8, 0.6, 0.8) = 0.6$$

따라서, d_2의 검색상태 값은

$\max(0.5, 0.8, 0.6) = 0.8$로서 평가되어 상위의 상태 값을 가진 문서가 사용자에게 제시된다. 따라서, 도메인 개념들 사이의 의미관계를 퍼지정도로 표현한 개념 네트워크를 통해 사용자 질의와 개념적으로 서로 연관된 유사한 의미의 개념을 가진 문서들을 검색하여 연관검색이 가능하며 재현률을 향상시킬 수 있다. 이와 같은 방법은 영역 전문가에 의해서 개념네트워크가 수동으로 유지되기 때문에 구축 및 유지 관리가 어렵고 다른 응용에 어려움이 있다.

따라서 이를 해결하기 위해 개념네트워크를 자동화하고 탐색과정은 다양한 탐색요구를 지원하여 연관검색이 가능하도록 지원하여야 한다.

3. 계층적 유사관계행렬 구축

문서를 기술하는 저자는 표제를 만들기 위해서 때때로 이해하기 어렵고, 환상적이고, 강한 인상을 주는 단어들을 사용하는 경향이 있다. 표제가 문서를 대표하는 주제를 적절하게 표현 못하는 경우가 있기 때문에 문서의 주제분석을 위해 표제만을 이용하는 것은 문서 초록이나 문서 전문을 이용하는 것보다 효과적이지 못하다는 연구가 있다[3,15]. 또한, 문서 전문을 이용하는 것은 부적합하기 때문에 본 논문에서는 문서검색을 위하여 표제와 초록에 대한 문헌 구조적인 지식을 바탕으로 2단계의 색인어관계행렬을 구축하여 문서를 구조적으로 분석하며 의미지식정도를 파악하고자 한다. 이절에서는 2단계 색인어관계행렬을 구축하는 과정을 설명한다.

수식 (1)(2)(3)와 같이 문서집합(D), 표제에서 추출한 색인집합(T_T) 그리고 요약에서 추출한 색인집합(T_A)을 정의한다.

$$D = \{d_1, d_2, d_3, \cdots, d_k\} \quad (1)$$

$$T_T = \{t_1, t_2, t_3, \cdots, t_n\} \quad (2)$$

$$T_A = \{t_1, t_2, t_3, \cdots, t_m\} \quad (3)$$

이를 기반으로 (4)(5)와 같이 표제 색인어 및 요약 색인에 대한 퍼지 관계를 표현하는 원시문서베이스를 생성할 수 있다.

(D^T)은 문서 표제에 출현한 색인집합에 대한 상대적 관련성을 간접적으로 추정한 원시문서베이스이다.

색인어의 중요정도를 표현하는 가중치는 용어빈도를 기반으로 하였다.

$$D^T = D \times T_T = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ w_{k1} & w_{k2} & \cdots & w_{kn} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ d_k \end{matrix} \quad (4$$

$$t_1 \quad t_2 \quad \cdots \quad t_n$$

(D^A)은 문서 요약에 출현한 색인집합에 관한 상대적 관련성을 표현한다. 이는 표제를 통해 우선 1차 검색된 문서를 재평가하기 위해 본 논문에서는 정의하였다.

$$D^A = D \times T_A = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ v_{k1} & v_{k2} & \cdots & w_{km} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ d_k \end{matrix} \quad 5$$

$$t_1 \quad t_2 \quad \cdots \quad t_m$$

(D^T), (D^A)와 같이 색인어 집합에 의해 문헌을 표현하는 원시문헌베이스를 구성하기 위해서는 원시문서에서 용어를 추출하고 용어와 문서사이의 관련정도를 가중치로 표현(w_{ik} v_{ik})하는데 본 논문에서는 용어출현빈도가 문서내용을 대표한다는 가정[5,11]을 기초로 하여 특정 문서에 대한 용어의 종속성을 퍼지화하였다.

용어의미관계를 퍼지화하기 위한 가정은 다음과 같다. 첫째, 용어빈도수를 입력 값으로 하며 입력 값에 무관하게 [0,1]사이의 퍼지 값으로 사상한다. 둘째, 문서의 특성상 용어 발생빈도가 많으면 중요 용어일 가능성이 높으므로 멤버쉽 함수는 단조증가형태를 이룬다. 셋째, 도메인의 특성에 따른 임계 값(critical value)을 가지며 임계값은 도메인 의존적으로 적용된다.

본 논문에서는 신경망에서 많이 사용되는 활성화함수 중에서 대표적인 비 선형함수인 S자 형태의 시그모이드(Sigmoid) 함수를 이용

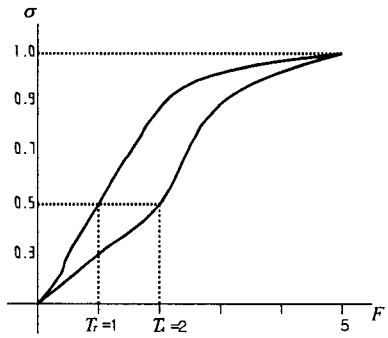
하며, 표제에서 발생하는 색인어에 대한 임계값은 1로 설정하였고 요약에 대한 색인어 임계값은 2로 할당하였다. 멤버쉽(membership) 함수($\sigma(F)$는 다음의 조건을 만족한다.

$$\sigma : R^{+} \rightarrow [0,1]$$

$$\sigma(F_1) > \sigma(F_2) \Leftrightarrow F_1 > F_2$$

$$\frac{d^2(\sigma)}{dF^2} \geq 0 \Leftrightarrow F \leq T_F \text{ and } \frac{d^2(\sigma)}{dF^2} \leq 0 \Leftrightarrow F \geq T_F$$

위의 조건을 만족하고 본 논문에서의 사용되는 임계값(T_T, T_A)을 [그림 3-1]와 같이 나타낼 수 있다.



[그림 3-1] 시그모이드 멤버쉽 함수
[Fig. 3-1] sigmiod membership function

이는 상대적(relative)인 확률적 빈도(probabilistic frequency)를 절대적(absolute)인 가능성(possibility) 퍼지 소속 값으로 사상시키는 기능을 하는 멤버쉽 함수이다.

본 논문에서 색인을 추출하는 과정은 수동색인을 채택하였다. 자동으로 하는 색인 방법은 자연어문장에서 색인어를 추출하는 형태분석기가 요구되는데 대용량의 사전데이터가 구축되어야하고 본 논문의 주요관점이 자동색인보다는 추출된 색인어의 의미적인 종속관계를 나타내는 유사관계행렬(similarity relation matrix)을 구축하여 연상지식을 추출하고 내용기반 검색을 지원하기 위한 질의확장 모델이므로 키워드 집합을 1차 색인으로 하고 간단한 수작업을 통해 색인용어를 최종적으로 결정하였다.

이와 같이 구축된 원시문서베이스를 이용하여 색인어간의 퍼지 관련정도를 나타내는 퍼지 색인어관계행렬을 구성한다.

색인어 유사관계행렬(S^T)은 특정 문서에서 색인어간 동시 출현빈도가 많을수록 색인어 유사성이 높다는 가정 하에 동시 출현빈도를 고려하였다.

본 논문에서 제안하는 퍼지 유사관계행렬은 퍼지 호환관계(tolerance relation)의 특성을 만족한다.

퍼지 유사관계행렬의 각 요소간의 관계는

$$\mu(t_i, t_j) = \mu(t_j, t_i), \qquad 0 \leq \mu(t_i, t_j) \leq 1,$$

$\mu(t_i, t_i) = 1$을 만족한다.

아래의 (6)(7)의 수식에 의하여 구성된 퍼지 유사행렬은 색인들의 의미를 정의하기 위한 관점에서 구축하였으므로 관련연구에서 표현한 개념행렬(Concept Matrices)이며 S^T는 축소 용어집합이다 할 수 있다 [1,9,12,13]. 이는 모든 문서에서 표현된 색인어를 모두 고려할 수 있다는 장점을 가진다.

$$u_{A \cong B}(w) = Max\{Min\ (u_A(w), (u_B(w)),$$

$$Min\ (1 - u_A(w), 1 - u_B(w))\} \qquad (6)$$

$$u_{ij} = u_{w_i \cong w_j} = \frac{1}{d} \sum_{k=1}^{d} u_{w_i \cong w_j}(d_k) \qquad (7)$$

색인어 유사관계행렬(S^T)은 원시문서베이스(D^T)에서 (6)(7)의 수식을 적용하여 구성한다. 이는 1차 검색을 수행하기 위해 구성된 문서 표제에 존재하는 색인어간의 관련정도를 의미한다. 색인어관계행렬에서 퍼지 호환관계에 따라 질의 확장 및 가중치를 조정함으로서

재현률을 향상시킬 수 있다.

$$S^T = D_{ik}^T \times D_{ik}^T$$

$$\begin{bmatrix} \bigvee_{i=1..m}(\min(\mu_{i1},\mu_{i1}),\min(1-\mu_{i1},1-\mu_{i1})) & \bigvee_{i=1..m}(\min(\mu_{i1},\mu_{i2}),\min(1-\mu_{i1},1-\mu_{i2})) & \cdots & \bigvee_{i=1..m}\min(\mu_{i1},\mu_{im}),\min(1-\mu_{i1},1-\mu_{im})) \\ \bigvee_{i=1..m}\min(\mu_{i2},\mu_{i1}),\min(1-\mu_{i2},1-\mu_{i1})) & \bigvee_{i=1..m}\min(\mu_{i2},\mu_{i2}),\min(1-\mu_{i2},1-\mu_{i2})) & \cdots & \bigvee_{i=1..m}\min(\mu_{i2},\mu_{im}),\min(1-\mu_{i2},1-\mu_{im})) \\ \cdots & \cdots & \cdots & \cdots \\ \bigvee_{i=1..m}\min(\mu_{im},\mu_{i1}),\min(1-\mu_{im},1-\mu_{i1})) & \bigvee_{i=1..m}\min(\mu_{im},\mu_{i2}),\min(1-\mu_{im},1-\mu_{i2})) & \cdots & \bigvee_{i=1..m}(\mu_{im},\mu_{im}),\min(1-\mu_{im},1-\mu_{im})) \end{bmatrix}$$

(8)

동일한 방법으로 (S^A)은 원시문서베이스 (D^A)에서 동시발생빈도 기반으로 구성된 유사관계행렬이다. 1차 검색된 문서집합에서 검색순위를 재조정하여 정확률을 향상시키기 위해 구성된 문서 요약에 존재하는 색인어간의 관련정도 의미한다.

$$S^A = D_{ik}^A \times D_{ik}^A$$

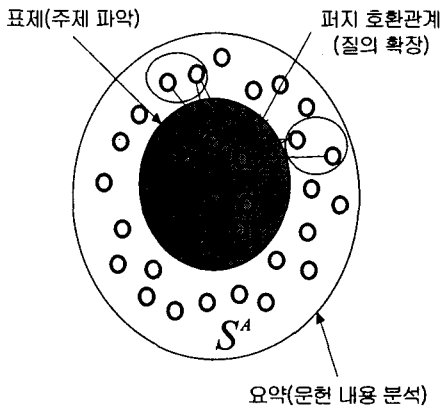
$$\begin{bmatrix} \bigvee_{i=1..m}(\min(\mu_{i1},\mu_{i1}),\min(1-\mu_{i1},1-\mu_{i1})) & \bigvee_{i=1..m}(\min(\mu_{i1},\mu_{i2}),\min(1-\mu_{i1},1-\mu_{i2})) & \cdots & \bigvee_{i=1..m}\min(\mu_{i1},\mu_{im}),\min(1-\mu_{i1},1-\mu_{im})) \\ \bigvee_{i=1..m}\min(\mu_{i2},\mu_{i1}),\min(1-\mu_{i2},1-\mu_{i1})) & \bigvee_{i=1..m}\min(\mu_{i2},\mu_{i2}),\min(1-\mu_{i2},1-\mu_{i2})) & \cdots & \bigvee_{i=1..m}\min(\mu_{i2},\mu_{im}),\min(1-\mu_{i2},1-\mu_{im})) \\ \cdots & \cdots & \cdots & \cdots \\ \bigvee_{i=1..m}\min(\mu_{im},\mu_{i1}),\min(1-\mu_{im},1-\mu_{i1})) & \bigvee_{i=1..m}\min(\mu_{im},\mu_{i2}),\min(1-\mu_{im},1-\mu_{i2})) & \cdots & \bigvee_{i=1..m}(\mu_{im},\mu_{im}),\min(1-\mu_{im},1-\mu_{im})) \end{bmatrix}$$

(9)

(8)(9)와 같이 동시 발생빈도를 기반으로 유사관계행렬을 구성하였다. 이는 시소러스와 같은 문서집합에 의존적인 색인이라 할 수 있다. 사용자 요구 분석과정에서는 질의를 확장하기 위해 퍼지 호환관계의 $\alpha - Cut$을 기반으로 질의 용어를 추가하며, 문서베이스와 동일한 형태의 질의베이스 벡터를 생성함으로서 전체적인 사용자 의도를 파악한다.

계층적 색인어관계행렬(S^T , S^A)을 2-Level(1: 표제, 2: 요약)로 정의함으로서 2단계 질의를 확장하고 문서 내용지식을 고려하여 정확률을 향상시키기 위함이다.

[그림 3-2]에 문서의 구조를 고려한 2단계 주제분석과정은 다음과 같다.



[그림 3-2] 문서의 시맨틱 구조

[Fig. 3-2] A Semantic Structure of Document

이와 같이 주제분석과정은 색인어 유사관계행렬(S^T, S^A)과 원시문서베이스(D)의 퍼지합성을 통해 영역 의존적인 문서베이스(D^*)을 구성할 수 있으며, 요구분석 과정은 1단계 확장된 질의베이스와 문서베이스의 유사도를 측정하여 문서를 우선 검색하며, 문서베이스 (D^*)의 구조는 다음과 같다.

$$D^* = D \times S^A = \begin{matrix} \\ d_1 \\ d_2 \\ \cdot \\ \cdot \\ d_k \end{matrix} \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1m} \\ \mu_{21} & \mu_{22} & \cdots & \mu_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \mu_{k1} & \mu_{k2} & \cdots & \mu_{km} \end{bmatrix} \begin{matrix} \\ d_1 \\ d_2 \\ \cdot \\ \cdot \\ d_k \end{matrix}$$

$$\begin{matrix} t_1 & t_2 & \cdots & t_m \end{matrix}$$

문서베이스(D^*)에 대한 퍼지합성 규칙은 다음과 같이 표현된다.

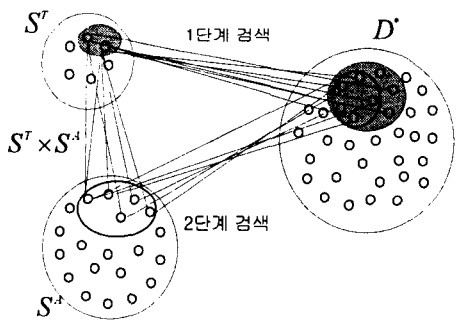
$$D^* = D \times S^A =$$

$$\begin{bmatrix} \bigvee_{i=1..m}(\min(\mu_{i1},\mu_{i1})) & \bigvee_{i=1..m}(\min(\mu_{i1},\mu_{i2})) & \cdots & \bigvee_{i=1..m}\min(\mu_{i1},\mu_{im})) \\ \bigvee_{i=1..m}\min(\mu_{i2},\mu_{i1})) & \bigvee_{i=1..m}\min(\mu_{i2},\mu_{i2})) & \cdots & \bigvee_{i=1..m}\min(\mu_{i2},\mu_{im})) \\ \cdots & \cdots & \cdots & \cdots \\ \bigvee_{i=1..m}\min(\mu_{ik},\mu_{i1})) & \bigvee_{i=1..m}\min(\mu_{ik},\mu_{i2})) & \cdots & \bigvee_{i=1..m}(\mu_{ik},\mu_{im})) \end{bmatrix}$$

(10)

우선, 문서 표제에 출현한 색인어를 통해 질의과정을 수행함으로서 탐색과정을 단순화할 수 있으며 재현률을 향상시키기 위해 표제 유사관계행렬(S^T)에서 호환관계를 만족하는 색인어를 질의어에 추가하여 검색을 수행한다. 1단계에서의 호환관계에 따라 검색된 문서를 대상으로 의미확장을 통해 문서가 연상하는 함축 지식을 추론하기 위해 유사관계행렬(S^A)의 퍼지 호환관계에 따라 색인어를 재분류하고 질의 확장을 통한 검색상태에 따라 문서 순위를 재조정한다.

사용자 질의에 따라 문서를 검색하는 탐색과정은 다음의 [그림 3-3]와 같이 표현할 수 있다.



[그림 3-3] 퍼지 키워드 유사관계에 의한 검색
[Fig. 3-3] Search by Fuzzy Similarity Relation

기존의 시소러스의 사용은 검색영역확장으로 많은 문서의 검색은 가능하게 했지만 정확한 정보의 검색에는 완전하지 못하므로 [그림 3-3]와 같이 1단계 문서의 주제영역을 설정하고 2단계를 통해 영역지식을 확장하는 방법으로 문서의 정확률을 향상시키고자 한다.

4. 퍼지 유사관계행렬을 이용한 2단계 검색

문서검색시스템에서 문서는 문헌구조를 고려한 색인어의 집합에 의해 표현되며 각 색인어의 연관관계 정도에 따라서 문서의 내용을 분류할 수 있다. 동일하게 사용자에 의해서 구성되는 질의도 질의 용어의 관계에 따라 사용자 요구를 파악할 수 있다. 따라서, 질의 용어에 대한 지식 또한 포괄적인 지식에서 세부적인 지식으로 확장되어진다.

사용자의 정보 요구에 대한 포괄적인 의미를 정확하게 파악하기 위해서 전문가에 의해서 구성된 시소러스를 활용하는 방법과 영역지식을 표현하는 퍼지 유사관계행렬을 활용하여 질의 개념을 확장하여 파악하는 방법이 있는데 본 논문에서는 후자의 방법이며 이 절에서는 유사관계행렬을 이용하여 질의를 확장하고 처리하는 방법을 논의한다.

4.1 호환관계에 따른 질의 확장

일반적인 정보검색시스템에서 사용자는 정형화된 형태의 질의를 통해 자신이 요구하는 정보를 검색한다. 특히, 사용자들은 특정 도메인에 대한 상당한 지식을 가지고 있다고 할 수 있기 때문에 포괄적인 의미검색이나 다양한 사용자 요구를 정확하게 표현할 수 있도록 설계되어야 한다[6]. 그러나 일반적인 정보검색의 사용자 질의는 모두 동일한 의미적 중요성을 가지는 탐색어들에 대한 불리안형태로만 표현된다. 따라서, 퍼지 불리안 형태의 질의를 사용하여 탐색어들에 대한 의미적 중요성을 표현할 수 있도록 하였다.

여기에서 원시질의 간의 질의 형태는 다음과 같다.

$$Q = \{(t_1, v_1), (t_2, v_2), \cdots, (t_n, v_n)\}$$
$$\vec{q} = [v_1, v_2, \cdots, v_n] \quad (11)$$

본 논문에서는 사용자 요구분석 과정이 주제분석과정과 일관된 메커니즘을 유지하기 위해 주제 개념을 파악하여 질의를 확장하는 방

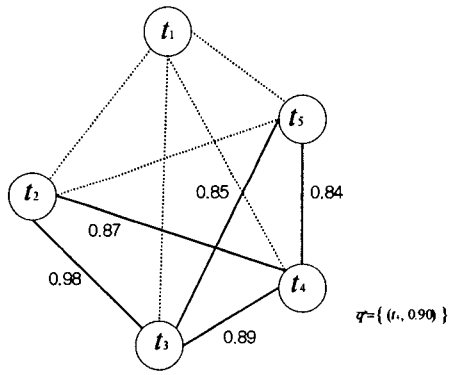
법을 이용한다.

유사관계행렬에서 호환관계를 만족하는 호환클래스를 분류하고 원시질의에 추가함으로서 재현률을 향상시키고자 한다.

즉, 도메인의 영역지식을 이용하여 사용자의 원시질의를 정확하게 표현하기 위하여 호환클래스를 적용하며 추가된 질의 가중치는 전이관계에 의한 퍼지확장 원리를 이용하여 부여한다.

사용자 원시질의는 다음과 같이 용어벡터로 구성되었다고 가정한다.

색인어 유사관계행렬(S^T) 및 사용자 질의가 [그림 4-1]와 같이 부여되었다고 가정한다. 단, 여기에서는 $0.8-cut$을 적용한 색인어 관계이며, 호환클래스는 실선으로 표현하였다.



[그림 4-1] 색인어 네트워크에서 호환관계
[Fig. 4-1] the tolerance relations of Keyword Network

[그림 4-1]에서와 같은 사용자 질의에서 호환클래스는 (12)와 같이 분류되며 원시 질의에 확장되는 용어($\overrightarrow{q_e}$)는 t_2, t_3, t_5로 가정할 수 있다.

$$C/R_{0.9}=\{C_1=\{t_1\},\ C_2=\{t_2, t_3, t_4\},\ C_3=\{t_3, t_4, t_5\}\} \quad (12)$$

$$\overrightarrow{q_e} = [(t_2, 0.89), (t_3, 0.89), (t_4, 0.90), (t_5, 0.80)]$$

확장된 용어의 가중치는 관련연구에서 개념네트워크 기반 검색방법의 전이 관계를 이용하였다.

4.2 2단계 내용기반 확장을 통한 검색순위 재조정

문서는 색인어 조합의 형태에 의해서 주제를 표현한다. 따라서, 특정 주제를 표현하기 위해 색인어들은 매우 밀접한 의존관계를 가진다. 주제분석과정에서 색인어집합을 초기화하고 <그림 3-1>의 시그모이드 멤버쉽 함수를 적용하여 (16)(17)와 같이 원시문서베이스(D^T, D^A)를 구성한다.

예를 들어 문서, 색인어집합을 정의하고 질의를 확장하고 검색하는 단계를 고려해본다.

$$D=\{d_1, d_2, d_3, d_4, d_5\} \quad (13)$$

$$T_T=\{t_1, t_2, t_3, t_4, t_5\} \quad (14)$$

$$T_A=\{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9\} \quad (15)$$

원시문헌에서 용어빈도를 기반으로 상대적 관련성을 퍼지화하여 도메인 지식을 표현하는 원시문서베이스(D^T, D^A)를 구성한다.

(D^T)은 문헌들의 표제에 출현한 색인어들 간의 관계를 반영한 원시문서베이스이다.

아래의 예제는 KT-SET 1.0의 질의 39번의 검색 결과 문헌중에서 5개의 문헌을 임으로 추출하여 표현하였다. 추출된 용어들은 빈도를 기반으로 시그모이드 멤버쉽 함수를 이용하여 가능성 퍼지소속 값으로 사상(mapping) 하였다. 여기에서 표제에 대한 임계 값은 1이고 요약에 대한 임계값은 3으로 설정하였다.

$$D^T = \begin{bmatrix} 0.94 & 0.50 & 0.50 & 0.80 & 0.00 \\ 1.00 & 0.00 & 0.50 & 0.00 & 0.94 \\ 0.94 & 0.00 & 0.80 & 1.00 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.50 & 0.00 \end{bmatrix}$$

(16)

(D^A)은 문헌들의 요약에서 출현한 색인어들간의 관계를 반영한 원시문서베이스이다.

$$D^A = \begin{bmatrix} 0.94 & 0.50 & 0.50 & 0.80 & 0.00 & 0.00 & 0.93 & 0.00 & 0.00 \\ 1.00 & 0.00 & 0.50 & 0.00 & 0.94 & 0.80 & 0.00 & 0.00 & 0.30 \\ 0.94 & 0.00 & 0.80 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.93 & 0.00 \\ 0.94 & 0.00 & 0.00 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 & 0.30 \end{bmatrix}$$

(17)

동시 출현빈도를 기반으로 색인어간의 관련정도는 (6)(7)의 수식을 이용하여 구성한다. 표제에서의 색인어간의 관계에 따라 사용자 질의에 대한 탐색요구가 해석된다.

$$S^T = \begin{bmatrix} 1.00 & 0.14 & 0.38 & 0.45 & 0.24 \\ 0.14 & 1.00 & 0.64 & 0.60 & 0.71 \\ 0.38 & 0.64 & 1.00 & 0.66 & 0.64 \\ 0.45 & 0.60 & 0.66 & 1.00 & 0.35 \\ 0.24 & 0.71 & 0.64 & 0.35 & 1.00 \end{bmatrix}$$

(18)

(S^A)은 표제 유사관계행렬에서 무시되는 탐색요구를 내용기반으로 확장하기 위한 요약부분의 유사관계행렬이다.

$$S^A = \begin{bmatrix} 1.00 & 0.14 & 0.38 & 0.45 & 0.24 & 0.20 & 0.22 & 0.22 & 0.16 \\ 0.14 & 1.00 & 0.64 & 0.60 & 0.71 & 0.74 & 0.90 & 0.71 & 0.78 \\ 0.38 & 0.64 & 1.00 & 0.66 & 0.64 & 0.64 & 0.64 & 0.45 & 0.58 \\ 0.45 & 0.60 & 0.66 & 1.00 & 0.35 & 0.38 & 0.66 & 0.35 & 0.48 \\ 0.24 & 0.71 & 0.64 & 0.35 & 1.00 & 0.96 & 0.63 & 0.63 & 0.80 \\ 0.20 & 0.74 & 0.64 & 0.38 & 0.96 & 1.00 & 0.65 & 0.65 & 0.80 \\ 0.22 & 0.90 & 0.64 & 0.66 & 0.63 & 0.65 & 1.00 & 0.63 & 0.69 \\ 0.22 & 0.71 & 0.45 & 0.35 & 0.63 & 0.65 & 0.63 & 1.00 & 0.69 \\ 0.16 & 0.78 & 0.58 & 0.48 & 0.80 & 0.80 & 0.69 & 0.69 & 1.00 \end{bmatrix}$$

(19)

또한, 영역지식 관계를 가진 색인어 유사관계행렬(S^A)와 원시문서베이스(D^A)의 퍼지 합성을 통해 문서베이스(D^*)를 구성한다.

$$D^* = \begin{bmatrix} 0.94 & 0.90 & 0.64 & 0.80 & 0.63 & 0.65 & 0.93 & 0.63 & 0.69 \\ 1.00 & 0.74 & 0.64 & 0.50 & 0.94 & 0.94 & 0.65 & 0.65 & 0.80 \\ 0.94 & 0.64 & 0.80 & 1.00 & 0.64 & 0.64 & 0.66 & 0.45 & 0.58 \\ 0.94 & 0.71 & 0.45 & 0.45 & 0.63 & 0.65 & 0.63 & 0.93 & 0.69 \\ 0.94 & 0.50 & 0.50 & 0.50 & 0.35 & 0.38 & 0.50 & 0.35 & 0.48 \end{bmatrix}$$

(20)

문서베이스(D^*)는 확장된 질의베이스($\vec{q}_e^T$)와의 유사성을 측정하여 각 문서에 대한 검색상태 값(RSV)을 표현하게 된다.

사용자 질의는 표현형식에는 제약이 없지만 본 논문에서는 검색과정을 표현하기 위해 간단히 질의 벡터로 표현하였다. 각 질의 용어는 독립된 색인 행렬로 간주한다.

$$Q = \{(t_1, x_1), (t_2, x_2), \cdots, (t_n, x_n)\ ,$$

$$\vec{q} = (x_1, x_2, \cdots, x_n)$$

(21)

여기에서 $x_i \in [0, 1]$로 소속되며 사용자의 탐색요구에 대한 관계정도를 의미한다. 사용자 질의가 다음과 같다고 가정하고 탐색과정을 설명한다.

$$Q = \{(t_1, 0.8), (t_3, 0.8))$$

$$\vec{q} = (0.8, -, 0.8, -, -) \qquad (22)$$

(22)의 표현에서 "-"는 사용자에 의해서 정의되지 않는 질의 용어이다. 따라서 기존의 검색에서는 무시되는 용어이다. 재현률을 유지하고 연상검색이 가능하도록 질의를 확장하는 방법은 (18)(19)의 퍼지 유사관계행렬에서의 호환클래스를 통해 수행된다.

각 표제에 대한 유사관계행렬(S^T)와 요약부분의 유사관계행렬(S^A)에서 $\alpha - cut_{0.6}$의 호환클래스를 질의에 추가하였으며 가중치는 자데의 퍼지 확장논리를 통해 수행하였다. 예제에서 질의 확장 결과는 다음과 같다. 즉, 문서베이스와 동일한 벡터를 생성하였고 유사성 척도에 따라 각 문서를 평가한다.

$$\vec{q}_e^T = (0.80, 0.60, 0.80, 0.60, 0.64)$$

(23)

$$\vec{q}_e^A = (-, -, -, -, -, 0.64, 0.64, -, -)$$

문서를 검색하는 단계는 확장된 질의($\vec{q}_e^T$)와 문서베이스(D^*)의 유사성 정도를 파악하여 문서를 1차적으로 검색하고, 2단계로 1차 검색문서를 기반으로 질의어 ($\vec{q}_e^A$)와 문서베이스(D^*)의 유사성을 파악한다.

즉, 1단계에서 표제 색인어에 의해 검색을 수행하고 2단계에서 내용분석을 통해 문서검색 순위를 재조정한다. 유사성 척도 방법은 다음과 같다.

$$S(x,y) = 1 - |x - y|$$

$$RSV(d_i) = \frac{\sum_{\vec{q}(j) \neq '-' \text{and } j=1,\cdots,n} T(u_{ij}, v_j)}{k} \quad (24)$$

전체적인 검색상태 값(RSV)의 조정은 다음과 같이 정의할 수 있다.

$$\vec{q}_{expand} = \alpha \vec{q}_e^T + \beta \vec{q}_e^A \quad (25)$$

예제에서 각 문서에 대한 검색 상태 값은 다음과 같으며 1단계 검색을 통해 재현률을 유지하고, 정확률을 높이기 위한 2단계 검색을 통하여 검색순위를 재조정한다.

$\vec{q}_e^T$ =

$d_1 = 0.84,\ d_2 = 0.82,\ d_3 = 0.88,\ d_4 = 0.85,\ d_5 = 0.81$

$\vec{q}_e^A$ =

$d_1 = 0.85,\ d_2 = 0.84,\ d_3 = 0.99,\ d_4 = 0.99,\ d_5 = 0.80$

각 문서의 최종 검색상태 값(RSV)는 다음과 같다.

$\vec{q}_{expand}$ =

$d_1 = 0.85,\ d_2 = 0.84,\ d_3 = 0.94,\ d_4 = 0.82,\ d_5 = 0.81$

여기에서 ($\vec{q}_{expand}$)은 1차 질의 확장($\vec{q}_e^T$)와 2차 내용 질의($\vec{q}_e^A$)에 대한 검색상태 값을 의미하며 전체질의에 대한 (25)의 α, β은 도메인의 특성에 따라 적용할 수 있으나 본 논문에서는 1로 설정하였다.

간단한 예제를 통하여 검색상태 값의 조정결과를 살펴보았다. 예제에서 1차 검색과 2차 검색에서의 결과 값에 대한 흐름이 거의 유사함으로 간단한 2차 검색만으로 문헌의 내용을 평가할 수 있음을 보인다. 그러나 상위 5개 적합문서를 대상으로 실험하였기 때문에 1단계와 2단계의 검색상태 값에 대한 변화는 미세함을 알 수 있다. 따라서, 내용기반 2단계 검색을 통하여 사용자 우선순위 및 2차 검색만으로 검색을 수행할 수 있으므로 시스템을 사용자 검색 요구에 유동적으로 대응할 수 있도록 구성할 수 있다.

5. 결 론

본 연구에서는 주제분석을 통해 영역지식을 추출하며, 이용자의 요구 분석과정과 영역지식을 퍼지논리 기반으로 추론하여 이용자의 질의에 본질적으로 가지고 있는 용어불일치 및 정보표현을 향상시키기 위한 연구이다. 따라서, 주제분석을 통해 키워드의 동시발생빈도를 기반으로 문서구조 특성을 고려하여 2계층 유사관계행렬을 구축하였고 탐색모형과의 일관된 메커니즘을 구현하기 위하여 유사행렬을 기반으로 도메인 의존적인 문서베이스를 구성하였다.

또한 기존 퍼지 논리기반 검색시스템은 재현률은 향상되나 정확률을 처리하는데 문제가 있으므로 문서구조의 특성을 반영하여 표제에 대해서 우선 질의 용어를 확장하여 검색을 수행하고 결과문서를 대상으로 세부적인 지식을 표현하는 유사관계행렬(S^A)을 통해 2차 검색을 수행하였다. 향후 연구로는 키워드 유사행렬을 구성함에 있어 좀더 체계적인 색인어간

의 유사성 측정방법 및 질의 확장에 있어 전체질의의 포괄적인 의미를 포함하는 개념질의를 통해 질의를 확장할 수 있는 방법에 대한 연구가 필요하다.

참고 문헌

[1] 김창민, 김용기, 퍼지관계곱 기반 퍼지정보검색시스템 구현, 정보처리학회논문지, 제8-B권 제2호, 2001.4, pp. 115-122.

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

[3] 문성빈, 적합성 피이드백을 이용한 전문검색시스템의 검색 효율성 증진을 위한 연구, 정보관리학회지, 제
10권 2호, 1993

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

[5] 정영미, "정보검색론", 구미무역, 1988

[6] 최재훈, 김지숙, 조기환, 문제은행에서 연상학습을 지원하는 퍼지검색시스템, 정보과학회지, 제29권 제4호, 2002.4

[7] 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.

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

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

[11] 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

[12] 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.

[13] 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.

[14] 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.

이기영

1992년 광주대학교 전산통계학과(공학사)
1994년 전북대학교 대학원 전산통계학과(이학석사)
1997년 전북대학교 대학원 전산통계학과(박사수료)
1998년 ~ 현재 원광보건대학 컴퓨터응용개발과 조교수
관심분야 : 인공지능, 정보검색, 멀티미디어시스템