한국 퍼지 및 지능시스템학회 논문지 2000, Vol. 10, No. 6 (2000)605~615

정보검색에서 시소러스를 이용한 효율적이고 효과적인 질의 평가 방법

Efficient and Effective Query Evaluation Method based on Thesaurus in Information Retrieval

최명복 · 김민구*

Myeongbok Choi and Minkoo Kim*

원주대학 행정전산과 , * 아주대학교 컴퓨터공학과

요 약

본 논문에서는 정보검색에서 시소러스를 이용한 효율적이고 효과적인 질의 평가 기법을 제안한다. 제안된 방법에서 시소러스 내부 용어들 간의 관계와 관련도가 용어 매트릭스로 표현되며, 용어들 간의 관계는 동의, 계층, 그리고 연관관계의 세 가지 관계가 제공된다. 시소러스 내부 용어들 간의 무시된 관련도가 퍼지 이론에 근거한 용어 매트릭스의 전이폐쇄 알고리즘에 의해 추론된다. 따라서 다양한 관계에 따른 시소러스에 표현된 지식을 이용할 수 있다. 또한 질의 평가시 용어 매트릭스를 이용하기 때문에 논문[3-7]에서 사용되는 방법보다 시간적으로 효율적이다. 그리고 정의된 용어 매트릭스는 논문[8]에서 발생되는 문제점을 제거하여 검색 효과를 높이기 위해 논문[6]에서 제안된 질의 평가 함수와 용이하게 통합시킨다.

ABSTRACT

In this paper, we propose an efficient and effective query evaluation method based on the thesaurus in information retrieval. In the proposed method the thesaurus is represented by a term matrix, where the elements in the term matrix consist of a pair representing a relation and a relational degree between terms. In the term matrix, there are three kinds of fuzzy relationships between terms, i.e., similar relation, hierarchical relation, and associative relation. The implicit fuzzy relationships between terms are inferred by the transitive closure of the term matrix based on fuzzy theory. Therefore the various knowledge of the thesaurus can be used by information retrieval system. Moreover our method is more efficient than the ones presented in the papers[3-7] because we use a term matrix in query evaluation, and the term matrix is easily integrated with query evaluation function in the paper[6] to overcome problems in the paper[8].

1. 서 론

정보검색 시스템의 주요한 목적을 극대화하기 위해서는 첫째 사용자의 정보요구와 문서의 색인인 검색형태(Search Patterns)를 명확하게 표현하는 것이고[1], 둘째는 정보요구를 만족시키는 적절한 정보들만을 탐색하고 탐색된 정보들에 대해 정보요구의 만족도에 따른 적합성 순위를 부여하는 것이다.

정확한 검색형태의 표현과 검색된 정보들의 정보요구에 부합하는 만족 정도를 제공하여 검색 효율을 높이기 위한 모델로 시소러스와 같은 지식베이스를 이용하는 정보검색 방법들이 제안되었다. 시소러스는 문서의 색인시 용어를 통제하는 목적 외에도 질의의 탐색어와 관련되어 있는 정보항목의 검색을 위해 사용되는데, 관련된 정보항목이 질의의 탐색어와 어느 정도 관련 있는지를 평가하는 색인어와 탐색어간의 질의평가에 반영함으로써 검색 효율을 크게 증가시킬 수 있다[2]. 그러나 기존의 시소러스 기반의 정보검색 모델들[3-6]은 용어들 간의 관련도를 질의 평가에 반영할 때 위상적 거리를 구하기 위해 그래프 형태의 시소러스를 탐색해야 한다.

시소러스 내부 용어들 간의 관련 정도를 [0.0, 1.0] 사이의 실수 값으로 부여한 모델[7]이 제안되었다. 이 모델[7]에서도 시소러스 내부 용어들 간의 개념적 거리는 그래프 형태의 개념적 네트워크에서 퍼지 기반의 Max-Min 방법에 의해 경로를 찾게 된다. 이상의 모델들[3-7]에서와 같이 그래프 형태의 시소러스에서 경로를 탐색하는 것은 실제 검색 시스템에 시소러스를 이용할 때, 시간적으로 매우 느릴 수 있다.

논문[8]은 시소러스 내부 용어들 간의 관련도를 [0.0, 1.0] 사이의 두 구간 값으로 부여할 수 있도록 허용하였으며, 시소러스 내부 용어들 간의 관계를 개념 매트릭스로 표현하였다. 개념 매트릭스는 그래프 형태의 시소러스에 비해 시간적 효율성을 제공한다.

그러나 논문[8]에서 제안된 질의 평가 기법은 시소러스 내부 용어들 간의 관련도를 적절히 반영하지 않기 때문에 검색 효과가 낮을 수 있으며, 검색된 문서들을 사용자의 질의에 부합하는 정도에 따라 순위(Ordering)를 부여하는데 어려울 수 있다.

이상에서 살펴본 모델들[3-8]의 또 다른 공통된 제한점 중의 하나는 시소러스 내부 용어들 간의 관계를 한가지 관계(주로 계층관계)로 제안하고 있다는 점이다. 그러나 실제 정보검색에서 많은 사용자들은 계층관계에 있는 용어뿐만 아니라, 동의관계 또는 연관관계에 있는 용어들을 이용하여 검색하게 된다. 따라서 용어들 간의 다양한 관계를 시소러스라는 지식베이스로 구축하여 검색에 이용하는 것은 좀 더 유용한 검색 시스템의 역할을 제공한다고 볼 수 있다.

본 논문에서는 시소러스를 이용한 효율적이고 효과적인 질의 평가 기법을 제안한다. 제안된 방법은 용어들 간의 관계의 성질과 표기에 있어서 논문[8]에서 사용된 개념 매트릭스와 다른 용어 매트릭스를 정의한다. 용어 매트릭스의 이용은 제안된 모델들[3-7] 보다 시간적 효율성을 제공한다. 그리고 정의된 용어 매트릭스는 논문[8]에서 발생되는 문제점을 제거하여 검색 효과를 높이기 위해 논문[6]에서 제안된 질의 평가함수와 용이하게 통합시킨다. 특히 용어 매트릭스에서 용어들 간의 관계는 계층, 동의, 그리고 연관관계의 세 가지 관계를 제공하여 다양한 지식을 반영할 수 있도록 하였다.

2. 지식 표현 방법

시소러스에 표현된 지식은 질의 평가시 평가 함수에 반영될 수 있기 때문에 시소러스의 구조는 평가 함수의 설계에 많은 영향을 준다. 이 절에서는 기존의 시소러스 구조와 질의 평가 방법을 살펴보고 몇 가지 문제점과 제한점을 제시한다.

2.1 의미 네트워크과 시소러스

Quillian[9]은 인간의 지식을 표현하는 방법으로 의미 네트워크(Semantic Network)을 제안하였다. Quillian의 제안 이후, 의미 네트워크는 지식베이스를 구축하기 위한 정보 구조로써 Cohen[10]과 같은 여러 문헌에서 사용되었다. 의미 네트워크는 아크(Arcs) 또는 화살표에 의해 명칭 노드(Labeled Node)들을 상호 연결하는 방향성 그래프로 광범위하게 묘사된다. 여기서 노드는 개념을 나타내고 링크는 개념들 사이를 여러 가지 종류의 관계로 연결시킨다.

많은 정보검색 시스템들은 전통적으로 문서에 독립적인 방법으로 특정 주제 문제를 설명하기 위해서 시소러스를 사용하고 있다[11]. 시소러스는 노드와 링크로 구성된다. 노드는 개념을, 그리고 링크는 광의어, 협의어, 동의어와 같은 관계에 의해 개념들 간의 상호 의존성을 표현한다. 그러므로 시소러스는 개념들을 연결하는 다소 제약된 의미 네트워크의 일종으로 볼 수 있다. 논문[5]에서는 시소러스의 구조를 트리(Tree), 가중치 트리(Weighted Tree), 유향 무순환 그래프(Directed Acyclic Graph; DGA), 가중치 DGA, 그리고 그래프(Graph)의 5가지로 분류하였다.

2.2 계층적 개념 그래프

계층적 개념 그래프(HCG; Hierarchical Conceptual Graph)[5]는 가중치가 부여된 계층적 시소러스로 단일의 루트 노드(Root Node)를 갖는다. HCG에서 노드들은 색인어들의 집합 T를 구성하며, 링크는 색인어

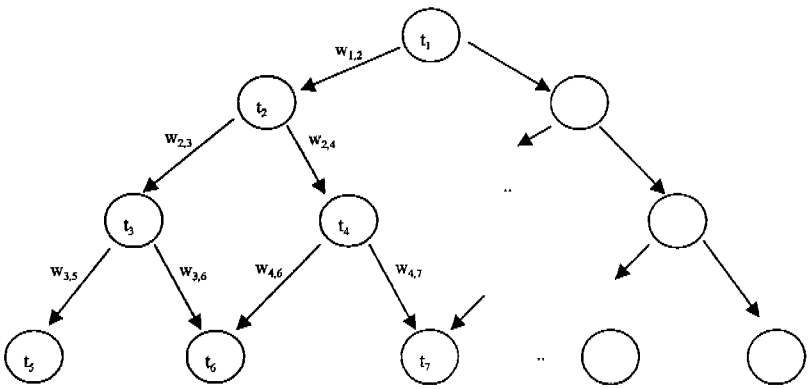

그림 1. 계층적 개념 그래프의 구조
Fig. 1. Structure of HCG

들 간의 "gencralization" 관계 G의 정도를 반영하는 가중치 w_{ij}를 갖는다. 기호 t_i G t_j에서, 색인어 t_i는 색인어 t_j보다 좀더 일반적인 광의적 의미의 용어이며, 역으로 t_j는 t_i보다 좀더 특정적인 협의적 의미를 갖는 용어이다. G는 비반사적(irreflexive), 비대칭적(asym-metric), 그리고 전이적(Transitive)이다. 앞의 그림 1은 계층적 개념 그래프의 구조를 나타낸다.

다음과 같이 HCG 구조와 같은 그래프로 구성된 시소러스 기반의 질의 평가 함수들이 제안되었다.

- Relevance Algorithm(Relevance) [3]
- Distance Algorithm(R-Distance) [4]
- Distance Algorithm(K-Distance) [5]
- Distance Algorithm(E-Relevance) [6]

이상의 평가 함수들은 불리언 질의와 문서 사이의 개념적 거리(Conceptual Distance)를 계산하기 위하여 시소러스에서 광의어 또는 협의어와 같은 계층관계만을 고려한다. 또한 시소러스 내부 용어들 간의 개념적 거리는 두 용어들 사이의 위상적 거리에 근거하여 계산한다. 두 용어들 간의 개념적 거리는 두 노드를 연결하는 최단 경로를 찾은 후, 경로상의 링크의 가중치를 합함으로써 계산된다. 찾아진 두 노드들 t_i와 t_j의 최단 경로상의 개념적 거리 $d(t_i, t_j)$는 다음과 같이 정의된다.

$$d(t_i, t_j) = w_{i,x1} + w_{x1,x2} + w_{x2,x3} + \cdots + w_{xn,j} \quad (1)$$

수식에서 $w_{i,x1}$, $w_{x1,x2}$, $w_{x2,x3}$, $w_{xn,j}$는 시소러스에서 노드 t_i와 t_j 사이의 최단 경로 상에 있는 노드들 간의 가중치를 나타낸다. 예를 들어, 그림 1에서 노드 t_4와 t_5 사이의 개념적 거리는 $w_{2,4}+w_{2,3}+w_{3,5}$와 $w_{4,6}+w_{3,6}+w_{3,5}$중에서 좀더 작은 값 중의 하나이다.

이상에서 살펴본 HCG와 같은 그래프 형태를 갖는 시소러스에서 내부 용어들 간의 개념적 거리를 질의 평가에 반영하기 위해서는 최단 경로를 찾는 그래프 탐색이 행해져야 한다.

2.3 개념 네트워크

논문[7]은 퍼지 정보검색을 위한 개념 네트워크를 제안하였다. 개념 네트워크는 노드와 유향 링크(Directed Link)로 구성된다. 노드는 개념 또는 문서를 표현하며, 각 링크는 두 개념들을 연결하든지 또는 특정한 개념 C_i와 문서 d_i를 연결한다. 또한 링크에는 [0, 1] 사이의 가중치가 부여된다. 링크에 부여된 가중치 w_{ij}는 개념 C_i와 C_j 사이 또는 개념 C_i와 문서 d_i 사이의 관련 정도를 의미하게 된다. 그림 2는 논문[7]로부터의 개념 네트워크를 보여준다. 그림 2로부터 문서 $d_2 = \{(C_1, 0.5), (C_2, 1.0), (C_5, 0.8)\}$와 같이 개념들의 퍼지 집합으로 표현할 수 있다.

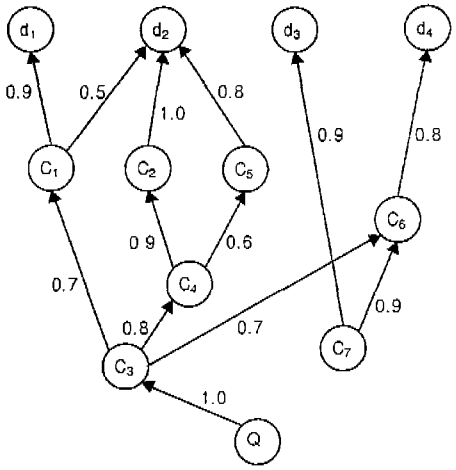

그림 2. 개념 네트워크
Fig. 2. Concept Network

개념 C_i로부터 개념 C_j로 링크가 존재할 때, 링크에 부여되는 가중치 w_{ij}는 $F(C_i, C_j)$로 표현할 수 있다. 여기서 F는 매핑 함수로 F: $C \times C \rightarrow [0, 1]$(단, $F(C_i, C_j) \in [0, 1]$)이다. 만약 개념 C_i로부터 C_j까지의 관련 값이 $F(C_i, C_j)$이고, 개념 C_j로부터 C_k까지의 관련 값이 $F(C_j, C_k)$이면, 개념 C_i로부터 개념 C_k까지의 관련 값은 다음과 같은 수식에 의해 얻어질 수 있다[7].

$$F(C_i, C_k) = \max_{C_j} \min[F(C_i, C_j), F(C_j, C_k)] \quad (2)$$

예로 그림 2에서와 같이 개념 네트워크가 4개의 문서 d_1, d_2, d_3, d_4와 7개의 개념 C_1, C_2, …, C_7로 구성되며 질의가 $Q = \{(C_3, 1.0)\}$으로 주어지면, 개념 C_3에 대한 문서 d_2의 관련 값을 구할 수 있다[7]. 그림 2에서와 같이 개념 C_3로부터 문서 d_2까지의 관련 값을 결정하는데 다음과 같은 3가지 경로들이 적용될 수 있다. 그러므로 논문[7]에 의해 개념 C_3에 대한 문서 d_2의 관련 값은 각 경로에 대한 값들의 최대 값으로 Max(0.5, 0.8, 0.6) = 0.8을 얻게 된다.

1) 첫 번째 경로: $C_3 \rightarrow C_1 \rightarrow d_2$:
   Min(0.7, 0.5) = 0.5.
2) 두 번째 경로: $C_3 \rightarrow C_4 \rightarrow C_2 \rightarrow d_2$:
   Min(0.8, 0.9, 1.0) = 0.8.
3) 세 번째 경로: $C_3 \rightarrow C_4 \rightarrow C_5 \rightarrow d_2$:
   Min(0.8, 0.6, 0.8) = 0.6.

이러한 추론 절차는 n개의 문서가 존재할 때 n번 반복되어야 한다. 이상의 2.2절과 2.3절에서 살펴본 HCG와 같은 그래프 형태의 시소러스와 개념 네트워크로 표현된 시소러스 구조에서 용어들 간의 관련도를 추론하기 위해서는 그래프를 탐색해야 한다. 특히 복합 질의(AND 또는 OR 연결자로 구성된 질의)가 주어질 때, 이러한 추론 절차는 매우 느릴 수 있고 비효율적이다. 이러한 제한점은 결과적으로 실제적인 정보검색 시스템이 구현되었을 때, 대부분의 사용자들에게 만족할 만한 검색 속도를 제공하지 못할 것이다.

2.4 개념 매트릭스

논문[8]은 논문[7]에서 제시한 그래프 형태의 개념 네트워크(예로 그림 2)를 모델링하기 위하여 다음과 같이 개념 매트릭스를 정의하였다.

[정의 2.1] 개념들의 집합 $C = \{C_1, C_2, C_3, \cdots, C_n\}$일 때 개념 매트릭스 M은 퍼지 매트 릭스[12]로 다음의 특성을 갖는다. 여기서 $M(C_i, C_j) \in [0, 1]$는 개념 C_i로부터 C_j까지의 관련도를 나타낸다.

1) 반사관계
$M(C_i, C_j) = 1, \forall C_i \in C.$
2) M은 대칭관계가 아닐 수 있다.
$M(C_i, C_j) \neq M(C_j, C_i).$
3) 전이관계
$M(C_i, C_k) \geq \underset{C_j \in C}{\text{Max Min}}[M(C_i, C_j), M(C_j, C_k)].$

[정의 2.2] 개념 매트릭스 M은 다음과 같이 표현된다.

$$M = \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{pmatrix}$$

여기서 n은 개념들의 개수이며, f_{ij}는 개념 C_i로부터 개념 C_j까지의 관련 값(단, $f_{ij} \in [0, 1]$, $1 \leq i \leq n$, $1 \leq j \leq n$)이다.

논문[8]에서 질의 디스크립터 Q는 질의 디스크립터 벡터 q에 의해 다음과 같이 표현할 수 있다. 여기서 x_i는 탐색어 t_i의 가중치이다.

$$Q = \{(t_1, x_1), (t_2, x_2), \cdots, (t_n, x_n)\},$$
$$q = \langle x_1, x_2, \cdots, x_n \rangle, \quad (3)$$
(단, $x_i \in [0, 1]$, $1 \leq i \leq n$)

질의 디스크립터 Q와 비슷하게 문서들을 표현하기 위한 d는 다음과 같이 문서 디스크립터로 표현할 수 있다. 여기서 n은 개념집합, T의 원소 수, 그리고 w_i는 그 문서에 대한 색인어 t_i의 가중치를 나타낸다.

$$d = \langle (t_1, w_1), (t_2, w_2), \cdots, (t_n, w_n) \rangle$$
(단, $t_i \in T$, $w_i \in [0, 1]$, $1 \leq i \leq n$) (4)

문서 집합 D는 가중치 w에 대해 d와 t간의 문서 디스크립터 벡터로 다음과 같이 표현할 수 있다. 여기서 m은 문서들의 수이고 n은 개념들의 수이다. w_{ij} (단, $w_{ij} \in [0, 1]$, $1 \leq i \leq m$, $1 \leq j \leq n$)는 문서 d_i에 대한 개념 t_j의 가중치를 나타낸다.

$$\begin{array}{c} \\ d_1 \\ d_2 \\ \cdot \\ d_m \end{array} \begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & t_3 & \cdots & t_n \end{array} \\ \begin{pmatrix} w_{11} & w_{12} & w_{13} & \cdots & w_{1n} \\ w_{21} & w_{22} & w_{23} & \cdots & w_{2n} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ w_{m1} & w_{m2} & w_{m3} & \cdots & w_{mn} \end{pmatrix} \end{array}$$

문서 디스크립터 매트릭스 D에서 문서 d_i에 대한 용어 t_j의 관련성 정도는 전문가에 의해 결정될 수 있다. 그러나 전문가는 어떤 개념들에 대해서는 특정 문서로의 관련성을 무시할 수 있다[8]. 무시된 관련성은 개념과 개념 사이의 관련성을 나타내는 개념 매트릭스의 모든 연결 관계를 찾아봄으로써 구할 수 있다. 이러한 개념 매트릭스에서 개념들 간의 모든 연결 관계는 개념 매트릭스 M의 전이폐쇄(Transitive Closure)에 의해 쉽게 파악할 수 있다. 개념 매트릭스 M의 전이폐쇄 T는 다음과 같은 간단한 알고리즘[13]에 의해 구할 수 있다.

[정의 2.3] 관계 $R(X, X)$가 있을 때, 관계 $R(X, X)$의 전이폐쇄 $R_T(X, X)$는 다음과 같은 3단계로 구성되는 간단한 알고리즘에 의해 결정될 수 있다.

1. $R' = R \cup (R \circ R)$.
2. If $R' \neq R$, make $R = R'$ and go to Step 1.
3. Stop: $R' = R_T$.

디스크립터 매트릭스 D와 개념 매트릭스 M의 전이폐쇄 T에 의해 새로운 매트릭스 D^*가 두 매트릭스 간의 Max-Min 합성연산[13]에 의해 구해진다. 구해진 D^*의 문서 디스크립터 벡터 d_i와 질의 디스크립터 벡터 q가 주어질 때, d_i와 q 사이의 유사 정도는 수식 (5)과 같이 RSV(Retrieval Status Values)로 계산된다. 또한 수식에서 $q(j)$는 질의 디스크립터 벡터 q의 j번째 요소를 의미하며 k는 질의 디스크립터 벡터에 명시된 개념들의 개수이다. 기호 "-"는 사용자에 의해서

무시된 개념을 위해 사용될 수 있다. w_{ij}는 D^*의 개체 디스크립터 벡터 d_i에 대한 각 개념들 간의 가중치이고, x_i는 질의에 있는 탐색어 t_i의 가중치이다(단, $0 \leq \mathrm{RSV}(d_i) \leq 1$)[8].

$$RSV(d_i) = \frac{\sum S(w_{ij}, x_j)\ (\text{단}, q(j) \neq \text{"-" and } j = 1, 2, \ldots, n)}{k} \tag{5}$$

$$S(A, B) = 1 - |A - B| \tag{6}$$

또한 논문[8]에서 질의는 AND 또는 OR로 구성되는 두 가지 형태로 분류하였다. AND로 연결된 질의는 탐색어들의 AND 연결을 나타내는 질의 디스크립터 벡터에 의해 쉽게 표현된다.

$$q = \langle x_1, x_2, \cdots, x_n \rangle \ (\text{단}, x_j\ [0, 1],\ 1 \leq j \leq n) \tag{7}$$

또한 OR로 연결된 질의는 다음의 수식 (8)과 같이 AND로 연결된 질의 디스크립터들을 OR로 연결한 질의를 의미한다. 또한 OR로 연결된 질의와 개체 사이의 유사 정도는 다음과 같은 수식 (9)와 같이 계산한다[8].

$$q_1 \text{ OR } q_2 \tag{8}$$

$$RSV^*(o_i) = \mathrm{Max}[RSV_1(o_i), RSV_2(o_i)] \quad (\text{단}, 0 \leq RSV^*(o_i) \leq 1,\ 1 \leq i \leq m) \tag{9}$$

이상의 질의 평가 수식 (5)는 개념 매트릭스를 이용하기 때문에 시간적 효율성을 제공한다. 그러나 수식 (5)는 질의 q에 대한 문서 d_i의 유사 정도를 계산하기 위해 q에 주어진 각 개념들에 대응되는 d_i에 대한 개념만을 고려한 유사 측도 방법이다. 다시 말해서 문서 d_i가 몇몇 개념들에 의해 색인되어 있을 때, 수식 (5)는 질의에 있는 임의의 탐색어와 d_i를 설명하는 모든 개념들 간의 의존 정보를 반영하지 않기 때문에 효과적으로 문서들을 순서화하기 어려울 수 있다. 또한 수식 (9)와 같이 OR로 연결된 복합 질의인 경우, OR는 최대 값(Max)으로 연산하고 있다. 이와 같이 OR 연산자를 Max로 해석하면 검색된 문서들을 순서화 하기 어려울 수 있으며, 원하지 않는 결과가 발생될 수 있다. 특히, 사용자들이 질의를 줄 때 탐색어의 개수를 많이 사용하지 않는 경우 또는 문서에 대한 색인어의 가중치가 0 또는 1로 제한된다면, 수식 (5)와 같은 질의 평가 방법, 그리고 OR 연산자를 Max로 해석하는 것은 질의 평가 결과가 몇 가지 값들로만 계산될 수 있기 때문에 검색된 문서들을 순서화 하는 것은 어려울 수 있다. 예를 들어 논문[8]의 다음과 같은 7개의 개념과 7개의 문서로 되어 있는 문서 디스크립터 매트릭스 D와 개념 매트릭스 M이 주어질 때를 고려해 보자.

$$D = \begin{pmatrix} 1.0 & 1.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.4 & 0.0 & 0.0 & 0.0 & 0.8 \\ 0.0 & 0.4 & 1.0 & 0.0 & 0.0 & 0.0 & 0.5 \\ 0.0 & 0.0 & 0.0 & 1.0 & 1.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.9 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.7 \\ 0.0 & 0.8 & 0.5 & 0.0 & 0.9 & 0.7 & 1.0 \end{pmatrix}$$

$$M = \begin{pmatrix} 1.0 & 1.0 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.4 & 0.0 & 0.0 & 0.0 & 0.8 \\ 0.0 & 0.4 & 1.0 & 0.0 & 0.0 & 0.0 & 0.5 \\ 0.0 & 0.0 & 0.0 & 1.0 & 1.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.9 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.7 \\ 0.0 & 0.8 & 0.5 & 0.0 & 0.9 & 0.7 & 1.0 \end{pmatrix}$$

만약, $q = \langle 0.6, \_, \_, \_, \_, \_, \_ \rangle$ OR $\langle _, _, _, _, _, _, 0.8 \rangle$라면, 다음과 같은 결과를 얻는다.

$$RSV^*(d_1) = \mathrm{Max}[RSV_1(d_1) = 0.9,\ RSV_2(d_1) = 0.9] = 0.9$$
$$RSV^*(d_2) = \mathrm{Max}[RSV_1(d_2) = 0.6,\ RSV_2(d_2) = 0.9] = 0.9$$
$$RSV^*(d_3) = \mathrm{Max}[RSV_1(d_3) = 0.4,\ RSV_2(d_3) = 0.8] = 0.8$$
$$RSV^*(d_4) = \mathrm{Max}[RSV_1(d_4) = 1.0,\ RSV_2(d_4) = 0.9] = 1.0$$
$$RSV^*(d_5) = \mathrm{Max}[RSV_1(d_5) = 0.6,\ RSV_2(d_5) = 0.9] = 0.9$$
$$RSV^*(d_6) = \mathrm{Max}[RSV_1(d_6) = 0.8,\ RSV_2(d_6) = 0.9] = 0.9$$
$$RSV^*(d_7) = \mathrm{Max}[RSV_1(d_7) = 0.4,\ RSV_2(d_7) = 0.9] = 0.9$$

이상의 결과에서와 같이 d_1, d_2, d_5, d_6, d_7은 같은 순위를 갖게 된다. 그러나, 예로 d_1과 d_2를 비교해 보면 d_1, d_2는 각각 $[RSV_1(d_1) = 0.9,\ RSV_2(d_1) = 0.9]$와 $[RSV_1(d_2) = 0.6,\ RSV_2(d_2) = 0.9]$로 d_2 보다 d_1이 좀더 질의에 관련되어 있다고 볼 수 있다. 이러한 단점은 우선적으로 수식 (5), (9)에서 질의 q에 대한 문서 d_i의 관련정도를 계산할 때 q에 주어진 각 개념들에 대응되는 d_i에 대한 개념만을 고려하기 때문이며, OR 연산자를 최대 값(Max)으로 연산하기 때문이다.

3. 용어들간 추론 규칙과 추론 기법

시소러스 내부 용어들 간의 관계는 여러 가지가 있을 수 있다. ISO 2788[14]에서는 이에 대한 권장 사항을 규정하고 있다. 권장된 대표적인 관계는 동의, 계층, 연관관계이다. 본 논문에서는 사람들이 사용하는 용어들 간의 관계가 동의, 계층, 연관관계로 분류

될 수 있다고 가정할 때 용어들 간의 여러 가지 관계에 따른 다양한 지식을 검색에 이용할 수 있도록 이들 세 가지 관계들 간의 추론 규칙과 추론 기법을 설명한다. 제안된 추론 규칙과 추론 기법은 다분히 인식론적이다.

3.1 추론 규칙

먼저 두 용어들이 같은 관계로만 구성된 추론 규칙과 예를 살펴본다. 다음과 같은 세 가지 규칙이 존재한다.

[규칙1]

t_1	H	t_2	$(D_{1,2})$
t_2	H	t_3	$(D_{2,3})$
t_1	H	t_3	$(D_{1,3}(H))$

[규칙1의 예]

생물	H	식물	$(D_{1,2})$
식물	H	꽃	$(D_{2,3})$
생물	H	꽃	$(D_{1,3}(H))$

[규칙2]

t_1	E	t_2	$(D_{1,2})$
t_2	E	t_3	$(D_{2,3})$
t_1	E	t_3	$(D_{1,3}(E))$

[규칙 2의 예]

아빠	E	아버지	$(D_{1,2})$
아버지	E	부친	$(D_{2,3})$
아빠	E	부친	$(D_{1,3}(E))$

[규칙3]

t_1	R	t_2	$(D_{1,2})$
t_2	R	t_3	$(D_{2,3})$
t_1	R	t_3	$(D_{1,3}(R))$

[규칙3의 예]

자동차	R	도로	$(D_{1,2})$
도로	R	신호등	$(D_{2,3})$
자동차	R	신호등	$(D_{1,3}(R))$

여기서 t_1, t_2, t_3는 용어이며, $D_{1,2}$, $D_{2,3}$은 용어들 간의 관련도이다(단, $D_{i,j} \in [0, 1]$). H는 계층관계, E는 동의관계, 그리고 R은 연관관계를 의미한다. $D_{1,3}(H)$, $D_{1,3}(E)$, $D_{1,3}(R)$등은 $D_{1,2}$, $D_{2,3}$을 사용한 t_1과 t_3간의 추론된 관련도와 관계가 된다. 예로 $D_{1,3}(H)$는 용어 t_1과 t_3 사이의 관련도의 계산 결과가 $D_{1,3}$이고 H(계층관계)로 추론됨을 의미한다. 제안된 추론규칙 오른쪽은 그 추론규칙을 적용한 하나의 예를 보여준다. [규칙1의 예]는 "식물은 생물이고 꽃은 식물이므로 꽃은 생물이다"라고 말할 수 있다. 이러한 추론규칙들은 다양한 예제를 통한 관찰에 바탕을 두고 있으며 인식론적으로 비교적 잘 받아들여지는 규칙들이다. 비슷한 방법으로 다음의 [규칙4]에서 [규칙9]까지를 생각할 수 있다. [규칙4]부터 [규칙9]까지에 각각 표기되어 있는 $(D_{1,3}({}_H\mathrm{E}))$, $(D_{1,3}({}_E\mathrm{H}))$, $(D_{1,3}({}_HR))$, $(D_{1,3}(R_H))$, $(D_{1,3}({}_ER))$, 그리고 $(D_{1,3}(R_E)$는 각각의 규칙에서 추론된 관련도와 관계가 된다. 예를 들어 [규칙4]에서 $(D_{1,3}(H_E))$는 계층관계와 동의관계로 구성되어 있는 경우 용어 t_1과 t_3 사이의 관련도는 $D_{1,3}$으로 계산되며 대문자 H인 계층관계로 추론됨을 의미한다.

[규칙4]

t_1	H	t_2	$(D_{1,2})$
t_2	E	t_3	$(D_{2,3})$
t_1	H	t_3	$(D_{1,3}(H_E))$

[규칙5]

t_1	E	t_2	$(D_{1,2})$
t_2	H	t_3	$(D_{2,3})$
t_1	H	t_3	$(D_{1,3}({}_EH))$

[규칙6]

t_1	E	t_2	$(D_{1,2})$
t_2	R	t_3	$(D_{2,3})$
t_1	R	t_3	$(D_{1,3}({}_ER))$

[규칙7]

t_1	H	t_2	$(D_{1,2})$
t_2	R	t_3	$(D_{2,3})$
t_1	R	t_3	$(D_{1,3}({}_HR))$

[규칙8]

t_1	R	t_2	$(D_{1,2})$
t_2	H	t_3	$(D_{2,3})$
t_1	R	t_3	$(D_{1,3}(R_H))$

[규칙9]

t_1	R	t_2	$(D_{1,2})$
t_2	E	t_3	$(D_{2,3})$
t_1	R	t_3	$(D_{1,3}(R_E))$

3.2 퍼지이론에 기반한 추론 기법

이 절에서는 앞에서 언급한 추론 규칙을 이용하여 실제적으로 용어들 간의 관련도를 계산하는 인식론적 추론 기법을 제안하려고 한다. 3.1절에서 정의한 추론 규칙들을 살펴보면 몇 가지 흥미 있는 사실을 발견할 수 있다. 계층관계들로 구성된 [규칙1]과 연관관계들로 구성된 [규칙3]의 추론 결과는 동의관계들만으로 구성된 [규칙2]의 추론 결과보다 더 많이 전이에 의하여 영향을 받는다. 왜냐하면 계층관계와 연관관계는 동의관계 보다는 의미상의 거리가 일반적으로 더 멀기 때문이다. 예를 들면, [규칙1의 예]에서 꽃과 생물 간의 추론 결과는 [규칙2의 예]에서 아빠와 부친간의 추론 결과보다 전이의 영향을 더 많이 받기 때문에 의미적 밀접도는 상대적으로 약해진다. 따라서 다음과 같은 조건 식 (11), (10)가 유도된다. 여기서 $D_{i,k}(E)$, $D_{i,k}(H)$, $D_{i,k}(R)$은 용어들 간의 의미적 관련도가 $D_{i,j}$, $D_{j,k}$일 때 각각 동의, 계층, 연관관계에 대한 추론 결과이다.

$$D_{i,k}(E) \geq D_{i,k}(H) \tag{10}$$

$$D_{i,k}(E) \geq D_{i,k}(R) \tag{11}$$

또한 계층관계, 동의관계, 연관관계에서 서로 같은 관계들 간의 추론 결과는 링크의 개수에 의존한다. 즉, 링크의 개수가 많아질수록 추론 결과의 의미적 밀접도는 점점 더 약해진다. 이러한 가정을 계층관계에 적용하면 다음과 같이 표현할 수 있다. 여기서 $D_{i,l}(H)$는 $D_{i,j}$, $D_{j,k}$, $D_{k,l}$에 대한 추론 결과이다.

$$D_{i,k}(H) \geq D_{i,l}(H) \tag{12}$$

식 (12)은 링크의 개수에 의존하므로 관련도들 중 작은 값 보다 크지 않다는 의미를 내포하고 있다. 즉, 관련도들 중 작은 값에서 H_α 만큼 보상한 값과 같다. 그러나 H_α는 추론 결과를 관련도들 중 작은 값에서 크게 벗어나게 하지는 않는다. 각 관계에 적용하면 다음과 같은 수식을 얻게 된다.

$$D_{i,k}(H) \leq \min[D_{i,j}, D_{j,k}] = \min[D_{i,j}, D_{j,k}] - H_\alpha \quad (13)$$

$$D_{i,k}(E) \leq \min[D_{i,j}, D_{j,k}] = \min[D_{i,j}, D_{j,k}] - E_\alpha \quad (14)$$

$$D_{i,k}(R) \leq \min[D_{i,j}, D_{j,k}] = \min[D_{i,j}, D_{j,k}] - R_\alpha \quad (15)$$

또한 식 (10), (11)에 의해 H_α, E_α, R_α 간에 다음의 조건 식이 성립한다.

$$E_\alpha \leq H_\alpha \quad (16)$$

$$E_\alpha \leq R_\alpha \quad (17)$$

이상의 조건 식을 만족시키는 연산식은 관계들 각각에 대해 여러 가지가 존재할 수 있다. 본 연구에서는 이상의 조건 식을 만족시키는 서로 같은 관계들 간의 추론에 다음과 같은 연산 식을 사용한다.

$t_i = t_k$ 경우,

$$D_{i,k}(E) = D_{i,k}(H) = D_{i,k}(R) = 1 \quad (18)$$

$t_i \neq t_k$ 경우,

$$D_{i,k}(E) = \mathrm{Min}[D_{i,j}, D_{j,k}] \quad (19)$$

$$D_{i,k}(H) = (D_{i,j} \cdot D_{j,k}) \quad (20)$$

$$D_{i,k}(R) = (D_{i,j} \cdot D_{j,k}) \quad (21)$$

또한, 이상의 조건식을 바탕으로 서로 다른 관계들 간의 추론를 위한 연산식을 이끌어 낼 수 있다. 서로 다른 관계들 간의 추론은 3.1절의 [규칙4], [규칙5], [규칙6], [규칙7], [규칙8], [규칙9]에 해당한다. 본 연구에서 이러한 추론 규칙들의 추론은 앞에서 정의한 조건식 (10), (11)에 근거한다. 예를 들어 [규칙4]와 [규칙5]를 생각해보면 계층관계와 동의관계로 구성된 [규칙4] 또는 이와 역 관계일 때의 [규칙5]에 대한 추론 결과는 계층관계들 만으로 구성된 추론 결과 보다 의미적 밀접도가 강하다. 또한 동의관계들 만으로 구성된 추론 결과 보다는 약하다. 따라서 다음과 같은 조건이 성립한다.

$$D_{i,k}(E) \geq D_{i,k}(H_E) \geq D_{i,k}(H) \quad (22)$$

$$D_{i,k}(E) \geq D_{i,k}({}_E H) \geq D_{i,k}(H) \quad (23)$$

본 연구에서는 이러한 조건을 만족시키는 동의관계와 계층관계 간의 추론 결과는 계층관계들 만으로 구성된 추론 결과와 동의관계들 만으로 구성된 추론 결과의 합을 평균하여 구한다. 따라서 다음과 같은 식 (24), (25)가 성립하게 된다.

$$D_{i,k}(H_E) = (D_{i,k}(H) + D_{i,k}(E)) / 2 \quad (24)$$

$$D_{i,k}({}_E H) = (D_{i,k}(E) + D_{i,k}(H)) / 2 \quad (25)$$

비슷한 방법으로 동의관계와 연관관계들 간의 추론 결과, 계층관계와 연관관계들 간의 추론 결과, 그리고 이들 각각의 역 관계들에 대한 추론 결과도 다음과 같이 얻을 수 있겠다.

$$D_{i,k}({}_E R) = (D_{i,k}(E) + D_{i,k}(R)) / 2 \quad (26)$$

$$D_{i,k}(R_E) = (D_{i,k}(R) + D_{i,k}(E)) / 2 \quad (27)$$

$$D_{i,k}({}_H R) = (D_{i,k}(H) + D_{i,k}(R)) / 2 \quad (28)$$

$$D_{i,k}(R_H) = (D_{i,k}(R) + D_{i,k}(H)) / 2 \quad (29)$$

또한 두 용어 t_i와 t_k가 다중 링크에 의해 연결된 경우에는 다중 링크의 각 경로에 대한 추론 결과들 중 가장 큰 값을 취할 수 있겠다. 따라서 수식 (18)-(21)는 다음과 같이 표현할 수 있다.

$t_i = t_k$ 경우,

$$D_{i,k}(E) = D_{i,k}(H) = D_{i,k}(R) = 1 \quad (30)$$

$t_i \neq t_k$ 경우,

$$D_{i,k}(E) = \max_j \mathrm{Min}[D_{i,j}, D_{j,k}] \quad (31)$$

$$D_{i,k}(H) = \max_j (D_{i,j} \cdot D_{j,k}) \quad (32)$$

$$D_{i,k}(R) = \max_j (D_{i,j} \cdot D_{j,k}) \quad (33)$$

4. 제안된 용어 매트릭스와 질의 평가 기법

이 부분에서는 2장에서 설명된 그래프 형태의 시소러스 탐색에 따른 시간적 제한점을 극복하기 위하여 용어 매트릭스를 정의한다. 정의된 용어 매트릭스는 논문[8]에서 사용된 개념 매트릭스와 매트릭스의 성질과 표기에 있어서 다르다. 정의된 용어 매트릭스는 용이하게 3장에서 제안된 동의, 계층, 연관관계들 간의

추론 기법을 용이하게 모델링한다. 그리고 정의된 용어 매트릭스는 2장의 논문[8]에서 발생되는 단점들을 제거하여 검색 효과를 높이기 위해 논문[6]에서 제안된 평가함수와 용이하게 통합시킨다. 제안하는 용어 매트릭스는 다음과 같이 정의할 수 있다.

[정의4.1] 용어들의 집합 $T_{er}=\{t_1, t_2, t_3, \cdots, t_n\}$일 때 용어 매트릭스 MT는 첫 번째 구성원소가 관련도이고 두 번째 구성원소가 관계인 쌍으로 표현되는 다음과 같은 특성을 갖는 퍼지 매트릭스[12]이다. 여기서 $MT(t_i, t_j)$는 인접한 용어 t_i로부터 t_j까지의 관련도를 나타낸다. 관련도는 [0, 1] 사이의 값이다.

1) 반사관계
   $MT(t_i, t_i)=1, \forall t_i \in T_{er}$
2) 대칭관계
   $MT(t_i, t_j)=MT(t_j, t_i)$.
3) 전이관계가 아닐 수 있다.
   $MT(t_i, t_k) \geq \underset{t_j \in T_{er}}{\text{Max Min}}[MT(t_i, t_j), MT(t_j, t_k)]$.

[정의4.2] 용어 매트릭스 MT은 다음과 같이 표현된다.

$$MT=\begin{pmatrix} (f_{11}, R_{el}) & (f_{12}, R_{el}) & \cdots & (f_{1n}, R_{el}) \\ (f_{21}, R_{el}) & (f_{22}, R_{el}) & \cdots & (f_{2n}, R_{el}) \\ \cdot & \cdot & \cdots & \cdot \\ (f_{n1}, R_{el}) & (f_{n2}, R_{el}) & \cdots & (f_{nn}, R_{el}) \end{pmatrix}$$

여기서 n은 용어들의 개수이며, f_{ij}는 용어 t_i로부터 용어 t_j까지의 관련도(단, $f_{ij} \in [0, 1]$, $1 \leq i \leq n$, $1 \leq j \leq n$)이다. R_{el}은 관계 H, E, R 중 하나일 수 있다.

2장에 있는 [정의2.3]의 전이폐쇄 알고리즘은 용어 매트릭스 MT에서 무시된 용어들 간의 관련도 계산을 위해 사용할 수 있다. 전이폐쇄 알고리즘의 단계 1에 있는 합집합 기호∪와 합성 기호 ∘는 여러 가지의 연산자가 적용될 수 있다. 본 논문에서는 앞에서 제안한 시소러스 내부 용어들 간의 관련도의 추론 기법인 수식 (30)-(33)을 모델링 하기 위해 합집합 기호 ∪는 퍼지 집합의 Max 연산자[13], 그리고 합성 기호 ∘는 다음과 같은 두 가지의 연산자를 전이폐쇄 알고리즘에 적용할 것이다. 즉, 두 개의 퍼지 관계 R과 S가 각각 $R \subseteq X \times Y$, $S \subseteq Y \times Z$일 때, 다음과 같이 계산한다.

동의관계의 경우,

$$\mu_{R \circ S}(x, z)=\underset{y \in Y}{\text{Max Min}}[\mu_R(x, y), \mu_S(y, z)]. \quad (34)$$

(단, $x \in X$, $z \in Z$)

계층, 연관관계의 경우,

$$\mu_{R \circ S}(x, z)=\underset{y \in Y}{\text{Max}}[\mu_R(x, y) \cdot \mu_S(y, z)]. \quad (35)$$

(단, $x \in X$, $z \in Z$)

수식 (34), (35)는 각각 최대-최소 합성, 최대-곱 합성이라고 부른다[13]. 이러한 합성 연산들을 전이폐쇄 알고리즘에 적용하면 3장에서 제안한 인식론적 관점에 근거한 추론 기법(수식 (30)-(33))의 의미를 그대로 반영하여 용어들 사이의 모든 관련도를 추론해 내게 된다. 그러므로 전이폐쇄 알고리즘에 의해 생성된 새로운 용어 매트릭스 MT*는 시소러스 내부 용어들 간의 모든 관련도가 추론되어 있기 때문에 사용자의 질의와 문서 사이의 유사 정도를 계산하는 다양한 질의 평가 함수(예로 논문들[3-6]에서 제안된 함수)에 용이하게 사용될 수 있다. 예를 들어 논문[6]은 다음과 같은 질의 Q가 반복 적용될 때, 문서 d에 대해 수식 (36)-(41)과 같은 질의 평가 함수를 제안하였다. 수식에서 T_L, T_R은 탐색어, x_1, x_2는 탐색어의 가중치, t_i는 문서 d에 대한 색인어의 집합, w_i는 색인어의 가중치이다.

Q = <(T_L, x_1)> AND <(T_R, x_2)> |
<(T_L, x_1)> OR <(T_R, x_2)> |
<(T_L, x_1)> AND NOT <(T_R, x_2)>,
(단, x_1, x_2[0, 1])

d = <(t_1, w_1), (t_2, w_2), ⋯, (t_n, w_n)>
(단, $t_i \in T_{er}$, $w_i \in [0, 1]$, $1 \leq i \leq n$)

$$\text{Sim}((T_L, x_1) \text{ AND } (T_R, x_2), d) = 1-\left[\frac{x_1^p(1-SI(T_L, d))^p+x_2^p(1-SI(T_R, d))^{1/p}}{x_1^p+x_2^p}\right], \quad (36)$$

$$\text{Sim}((T_L, x_1) \text{ OR } (T_R, x_2), d) = \left[\frac{x_1^p(SI(T_L, d))^p+x_2^p(SI(T_R, d))^{1/p}}{x_1^p+x_2^p}\right], \quad (37)$$

$$\text{Sim}((T_L, x_1) \text{ AND NOT } (T_R, x_2), d) = x_1(SI(TL, d)) \cdot (1-x_2(SI(TR, d)), \quad (38)$$

$$SI(T, d)=\frac{\sum_{i=1}^{n} Rev(T, t_i) \cdot w_i}{1+\frac{1}{2}(n-1)}, \quad (39)$$

$$Rev(T,t_i)=\frac{1}{1+distance(T,t_i)}, \quad (40)$$

distance(T, t_i) = T와 t_i를 연결하는 최소 링크 수. (41)

수식 (41)은 용어들 간의 개념적 거리를 계산하기 위해서 그래프 형태의 시소러스 구조(예로 그림 1)를 탐색해야 한다. 또한 개념적 거리는 수식 (1)과 같이 위상적 거리에 기반을 두고 있으며, 용어들 간의 한가지(계층관계)관계 만을 고려하고 있다. 그러나 시소러스에서 용어들 간의 다양한 관련성은 정보검색 행동에서 인간의 지식을 반영하는데 매우 중요할 수 있다. 이러한 제한점들은 본 논문에서 제안된 용어 매트리스를 이용하면 해결된다. 다시 말해서 수식 (41)의 *distance*(T, t_i)는 단순히 최소 링크 수로 그래프 형태의 시소러스를 탐색하는 대신 용어 매트릭스 MT에 전이폐쇄 알고리즘을 이용하여 생성된 메트릭스 MT*의 값을 이용하면 된다. 이유는 용어 매트릭스가 시소러스 내부 용어들 간에 여러 가지 관계들 간의 개념적 거리를 계산하기 위해 본 논문에서 제안된 수식(34), (35)를 이용하고 있기 때문이다. 이와 같이 MT*의 값을 이용하여 계산된 각 문서의 결과 값은 작은 값일수록 사용자의 질의에 좀더 관련되어 있다는 의미가 된다.

또한 이상의 질의 평가 수식들은 2.4절에서 설명된 논문[8]의 단점들을 해결할 수 있다. 우선 질의 평가 알고리즘의 수식 (39)에서 $SI(T, d)$는 질의의 탐색어 T와 문서 d를 설명하는 모든 색인어들과의 용어 의존성 정보를 반영하기 때문에 검색 효과가 증진될 수 있다. 또한, 수식에서 변수 $p(1\le p\le\infty)$는 AND, OR에 대한 다양한 해석을 제공한다[15]. p의 값이 감소할수록 AND와 OR 연산자들 사이의 구별이 점점 없어지며, p의 값이 ∞이면 퍼지 집합 모델에서와 같이 Min과 Max로 해석된다. 따라서 p를 조절하면 논문[8]에서 발생되는 OR 연산자의 문제점을 해결할 수 있다.

5. 실험 및 결론

시소러스와 같은 지식베이스를 이용한 문서의 순위는 인간 전문가를 시뮬레이션하는 것으로 질의 평가 함수에 의한 순위와 인간 전문가의 순위를 비교하는 것은 바람직하다[3-6]. 이전의 질의 평가 함수들[3-6]은 두 순위 결과들이 얼마나 일치하는지를 보기 위하여 스피어맨 순위 상관 계수(Spearman Rank Correlation Method)[16] ρ를 사용하였다. k개의 개체 e_1, e_2, …, e_k에 대한 두 순위 r_1, r_2, …, r_k와 s_1, s_2, …, s_k 사이의 스피어맨 순위 상관 계수는 다음의 수식 (42)과 같다. 만약 두 순위가 동등하다면 상관 계수 ρ는 1이 된다. 또한 ρ가 0으로 계산되면 두 순위가 관련이 없음을 나타내며, 역으로 두 순위가 되어 있는 경우 ρ는 -1이 된다.

$$\rho = 1-6\times\left(\frac{\sum_{i=1}^{k}(r_i-s_i)^2}{k(k^2-1)}\right) \quad (42)$$

논문[4]는 R-Distance를 평가하기 위하여 4개의 불리언 질의와 "Information Storage and Retrieval"의 한 부분을 설명하는 CRCS 용어들을 가지고 색인된 9개의 문서들, 그리고 15명의 학생 전문가에 의해 판단된 순위 결과를 이용하였다. CRCS(Computing Reviews Classification Structure)는 계층적 시소러스로 현재 CCS(ACM Computing Classification System)[17]으로 이름이 변경되었다. 여기서 불리언 질의에 있는 용어들의 가중치는 모두 0 또는 1이며, 질의는 부정어(Negated Term)를 포함하지 않는다. 또한 논문[5]는 부정어를 포함하는 5개의 불리언 질의와 "Communications of the ACM"에 있는 6개 논문을 사용하여 K-Distance를 평가하였다. 평가를 위해 20명

Query1: Retrieval Models (h.3.3.3)
Query2: Retrieval Models (h.3.3.3) AND
Search Process (h.3.3.3)
Query3: Information Search and Retrieval (h.3.3)
Query4: Information Search and Retrieval (h.3.3) AND
Retrieval Models (h.3.3.3)

그림 3. 논문[4]에서 사용된 4개의 불리언 질의
Fig. 3. Four boolean queries from the test data used in paper[4]

Query5: Artificial Intelligence (i.2) AND
NOT Knowledge Representation Formalisms and Methods (i.2.4)
Query6: Artificial Intelligence (i.2) AND
Speech Recognition and Understanding (i.2.7.5) AND
NOT Deduction and Theorem Proving (i.2.3)
Query7: Artificial Intelligence (i.2) AND
NOT Deduction and Theorem Proving (i.2.3)
Query8: Artificial Intelligence (i.2) AND
Frames and Scripts (i.2.4.1) AND
NOT Programming Languages (d.3)
Query9: Artificial Intelligence (i.2) AND
Deduction and Theorem Proving (i.2.3) AND
NOT Applications and Expert Systems (i.2.1)

그림 4. 논문[5]에서 사용된 5개의 불리언 질의
Fig 4. Five boolean queries from the test data used in paper[5]

의 학생 전문가에 의해 판단된 순위 결과가 사용되었으며, 문서는 "Artificial Intelligence"의 한 부분을 설명하는 CRCS 용어들을 가지고 색인되었다. 논문[6]은 논문[4-5]의 테스트 데이터를 이용하여 E-Relevance를 실험하였다. 그림 3과 4는 이들 실험에서 사용된 질의를 보여준다. 각 질의에서 괄호 안의 기호 (예로 h.3.3.3)는 계층적 시소러스 내에서 그 용어의 위치 정보를 나타낸다.

본 논문에서도 논문[4-6]에서 사용된 테스트 자료를 가지고 용어 매트릭스를 계산한 후 이를 수식 (41)에 적용하여 실험하였다. 본 논문에서 이를 B-Relevance라고 명명하였다. 실험을 위해 계층구조의 시소러스에서 용어들 간의 거리는 초기 값이 모두 0.85로 주어졌다. 이상의 실험 결과는 다음의 표 1, 표 2와 같다.

표 1, 2에서 각 평가 함수에 대해 평균값들을 합하면 논문[6]의 E-Relevance는 논문[4]의 R-Distance, 논문[5]의 K-Distance보다 평균 검색 효과가 높다. B-Relevance는 논문[6]의 E-Relevance에서 제안된 질의 평가 s함수를 이용하기 때문에 그 결과는 E-Relevance와 같다. 그러나 B-Relevance에서는 논문[6]의 E-Relevance에서 사용되는 위상적 거리(수식 (41))에 근거한 그래프 탐색 대신에 동의, 계층, 연관관계의 세 가지 관계를 표현하는 용어 매트릭스을 이용하여 추론된 값을 이용하기 때문에 논문[6]을 비롯한 논문들[3-7]보다 시간적으로 효율적이며 다양한 지식을 사용할 수 있다.

비슷한 방법으로 본 논문에서는 논문[8]에서 제안된 질의 평가 함수를 이상에서 설명된 테스트 자료를 가지고 실험하였다. 실험을 위해 시소러스에서 용어들 간의 거리는 초기 값이 모두 0.85로 주어졌다. 실험 결과 질의에 대해 문서들을 순위화 하는데 어려웠다. 예를 들어 그림 1의 질의 (Query1: Retrieval Models)에 대한 결과는 표 3과 같이 2가지 순위만으로 표시된다.

이러한 결과는 2.4절에서 설명된 논문[8]의 단점 때문이다. 그러나 본 논문에서의 질의 평가 방법은 용어 매트릭스에 기반한 용어들간 의존도를 적절히 반영하기 위하여 수식 (39)을 이용하기 때문에 이러한 단점이 제거된다.

앞으로의 연구 방향은 다양한 관계를 갖는 시소러스를 구축하고 좀더 많은 질의를 이용하여 검색 성능을 평가해야 할 것이다.

표 1. Negated term이 포함되지 않은 질의에 대한 스피어맨 상관 계수
Table 1. Spearman correlation coefficients for queries without negated terms

평가 함수	Query1	Query2	Query3	Query4	평 균
R-Distance	0.879	0.742	0.842	0.783	0.812
K-Distance	0.867	0.833	0.903	0.817	0.855
E-Relevance	0.867	0.817	0.933	0.867	0.871
B-Relevance	0.867	0.817	0.933	0.867	0.871

표 2. Negated term이 포함된 질의에 대한 스피어맨 상관 계수
Table 2. Spearman correlation coefficients for queries with negated terms

평가 함수	Query5	Query6	Query7	Query8	Query9	평 균
R-Distance	0.486	0.886	0.829	-0.086	-0.771	0.269
K-Distance	0.986	0.943	0.943	0.600	0.829	0.860
E-Relevance	0.943	1.000	0.943	0.657	0.714	0.851
B-Relevance	0.943	1.000	0.943	0.657	0.714	0.851

표 3. 논문[8]의 질의 평가 함수에 의한 RSV
Table 3. RSV of query evaluation function from paper[8]

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8	d_9
RSV(di)	1.00	0.85	0.85	1.00	1.00	0.85	1.00	1.00	1.00

참고문헌

[1] Radecki, T. Fuzzy set theoretical approach to document retrieval, Information Processing & Management, Vol. 15, pp. 247-259, 1979.

[2] Salton, G. & McGill, M. J. Introduction to modern information retrieval New York: McGraw-Hill, 1983.

[3] Rada, R., Mili, H., Bicknell, E., & Blettner, M. Development and application of a metric on semantic nets, IEEE Transactions on systems, man, cybernetics, Vol. 19, No. 1, pp. 17-36, January, 1989.

[4] McMath, C. F., Tamaru, R. S., Rada, R. A., graphical thesaurus-based information retrieval system, International Journal of Man-Machinc Studies, 31(2), pp. 121-147, 1989.

[5] Kim. Y. W., Kim, J. H., A model of knowledge based information retrieval with hierarchical concept graph, Journal of Documentation, 46(2). pp. 113-136, 1990.

[6] Lee, J. H., Kim, M. H., & Lee, J. H., Ranking documents in thesaurus-based boolean retrieval systems, Inforamtion Processing & System, 30(1), pp. 79-91, 1994.

[7] Lucarella, D., Morara, R. FIRST, Fuzzy information retrieval system, Journal of Information Sciene, Vol. 17, pp. 81-91, 1991.

[8] Chen, S. M., Wang, J. Y, Document retrieval using knowledge-based fuzzy information retrieval techniques, IEEE Transactions on systems, man, cybernetics, Vol. 25, No. 5, pp. 793-803, 1995.

[9] M. R. Quillian, "Semantic memory," in Semantic information processing, M. Minsky ed., MIT Press,

Cambridge Massachusetts, pp. 227-270, 1968.
[10] P. R. Cohen and R. Kjeldsen, "Information retrieval by constrained spreading activation in semantic networks," Information Processing & Management, Vol. 23, No. 2, pp. 255-268, 1987.
[11] G. Salton, "A New Comparison between Conventional Indexing(MEDLARS) and Automatic Text Processing(SMART)," Journal of the American Society for Information Science, Vol. 23, No. 2, pp. 75-84, 1972.
[12] Kandel, A. Fuzzy mathematical techniques with applications. CA: Addison-Wesley, 1986.
[13] George J. K.; Tina A. F. Fuzzy Sets, Uncertainty, and Information, Prentice-Hall International, Inc., 1988.
[14] ISO 2788, International Organization for Standardization: Guidelines for the Establishment and Development of Monolingual Thesauri, 2nd ed., Geneva, ISO, 1986.
[15] Salton, G., Fox, E. A. & Wu, H. Extended boolean information retrieval, Communications of the ACM, 26(11), pp. 1022-1036.
[16] M. Kendall, Rank Correlation Methods(4th ed), London and High Wycombe: Charles Griffin & Company LTD, 1975.
[17] Http://www.acm.org/class/. The ACM Computing Classification System [1998 Version], Valid in 2000, 2000.

최 명 복 (Myeong-Bok Choi)

1992년 : 호서대학교 전자계산학과(학사)
1994년 : 아주대학교 컴퓨터공학과(석사)
1994년~현재 : 아주대학교 컴퓨터공학과(박사과정)
1997년~현재 : 원주대학 행정전산과 (조교수)
관심분야 : 지능형 정보검색, 퍼지 응용, 지식표현, 의사결정 시스템

김 민 구 (Min-Koo Kim)

서울대학교 계산통계학과(이학사)
한국과학기술원 전산학과(공학석사)
1989년 : Pennsylvania 주립대(박사)
1981년~현재 : 아주대학교 컴퓨터공학과 (교수)
관심분야 : 지식표현, 상식추론, 지능형 교수 시스템, 지능형 정보검색 시스템