시소러스의 연관성 정보를 이용한 문서의 순위 결정 방법

Document Ranking Methods Using Term Dependencies from a Thesaurus

이 준 호(Joon Ho Lee)*

□ 목 차 □

초 록

최근 시소러스를 기반으로 하는 불리안 검색 시스템에서 문서의 순위 결정에 사용될 수 있는 Relevance, R-Distance, K-Distance와 같은 방법들이 개발되었다. 이러한 방법들은 색인어들 사이의 연관성 정보를 이용하여 문서들의 순위를 결정함으로써 많은 경우에 높은 검색 효율을 제공할 지라도, 불리안 연산자 AND, OR, NOT에 대한 연산 방법이 문제점으로 지적되어 왔다. 본 논문에서는 개선된 퍼지 집합 모델과 확장된 불리안 모델을 시소러스가 제공하는 색인어들 사이의 연관성 정보를 효율적으로 이용할 수 있도록 확장함으로써, 기존 방법들의 문제점을 극복하는 새로운 순위 결정 방법 KB-FSM과 KB-EBM을 제안한다. 또한 KB-FSM과 KB-EBM이 Relevance, R-Distance, K-Distance보다 문서들의 순위를 보다 정확하게 결정함을 성능 비교를 통하여 입증한다.

ABSTRACT

In recent years various document ranking methods such as Relevance, R-Distance and K-Distance have been developed which can be used in thesaurus-based boolean retrieval systems. They give high quality document rankings in many cases by using term dependence information from a thesaurus. However, they suffer from several problems resulting from inefficient and ineffective evaluation of boolean operators AND, OR and NOT. In this paper we propose new thesaurus-based document ranking methods called KB-FSM and KB-EBM by exploiting the enhanced fuzzy set model and the extended boolean model. The proposed methods overcome the problems of the previous methods and use term dependencies from a thesaurs effectively. We also show through performance comparison that KB-FSM and KB-EBM provide higher retrieval effectiveness than Relevance, R-Distance and K-Distance.

* 한국과학기술원 인공지능연구센터
논문접수일:1993년 9월 5일

Ⅰ. 서 론

지난 30년 동안 과학과 기술 분야에서의 급속한 발전은 수 많은 주제들에 대해 방대한 양의 정보가 생성되는 정보화 사회를 탄생시켰다. 원하는 정보에 대한 정확하고 빠른 접근은 정보화 사회를 살아가는 현대인들에게 성공의 여부를 결정짓는 중요한 요소가 되었다. 그러나 대용량의 데이타로부터 주어진 시간 내에 원하는 정보를 발견하는 것은 매우 어려운 일이다. 이러한 문제점을 해결하기 위해 1960년도 초에 컴퓨터를 이용하여 지정된 정보를 검색하는 정보 검색(Information Retrieval)이라는 연구 분야가 확립되었다[1].

정보 검색 시스템의 중요한 역할 중의 하나는 검색된 각각의 문서에 대하여 순위 결정 방법을 적용하는 것이다. 순위 결정 방법은 문서가 질의를 만족하는 정도를 나타내는 문서값(Document Value)을 계산하고, 계산된 문서값에 따라 문서들에 순위를 부여한다. 높은 순위를 갖는 문서일수록 질의에 대한 만족도가 크며, 사용자는 높은 순위를 갖는 문서를 우선적으로 검토함으로써 필요한 정보를 얻는데 소모되는 시간을 최소화할 수 있다.

불리안 검색 시스템에서 문서는 색인어들의 집합으로 표현되고, 질의는 색인어들을 불리안 연산자 AND, OR, NOT으로 연결한 불리안 수식이며, 검색되는 문서는 질의로서 주어진 불리안 수식을 만족하는 문서들이다. 불리안 검색 시스템은 시스템의 구현이 용이하고 구현된 시스템이 짧은 검색 시간을 제공하기 때문에, 오늘날 정보 검색 분야에서 가장 널리 사용되고 있다. 한편, 많은 정보 검색 시스템들이 문서들을 색인하기 위해 시소러스를 사용하고 있다[2]. 불리안 검색 시스템이 문서의 색인을 위해 시소러스를 사용한다면, 색인어들 사이의 연관성 정보를 이용하여 문서들의 순위를 보다 정확하게 결정할 수 있다[3].

최근 시소러스를 기반으로 하는 불리안 검색 시스템에서 문서의 순위 결정에 사용될 수 있는 Relevance, R-Distance, K-Distance와 같은 방법들이 개발되었다[4-10]. 이러한 방법들은 색인어들 사이의 연관성 정보를 이용하여 문서들의 순위를 결정함으로써 많은 경우에 높은 검색 효율을 제공할 지라도, 다음과 같은 문제점을 지니고 있다[11-13]. 첫째, 질의가 항상 최소 논리합 정규형(Minimal Disjunctive Normal Form)으로 변환되어야 하며, 이는 검색 시간을 증가시키는 요인이 된다. 둘째, NOT 연산이 비효율적이다. 셋째, MIN 또는 MAX 함수를 이용한 OR 연산이 검색 효율을 저하시킨다.

본 논문에서는 개선된 퍼지 집합 모델(Fuzzy Set Model)과 확장된 불리안 모델(Extended Boolean Model)을 시소러스의 연관성 정보를 이용할 수 있도록 확장한 Knowledge-Based Fuzzy Set Model (KB-FSM)과 Knowledge-Based Extended Boolean Model (KB-EBM)을 제안한다. KB-FSM과 KB-EBM은 문서와 질의에 포함된 다양한 가중치를 효율적으로 처리하고, 기존의 방법들 Relevance, R-Distance, K-Distance가 지니고 있는 문제점들을 발생시키지 않는다. 또한 색인어들 사이의 연관성 정보를 효율적으로 이용함으로써 문서들의 순위를 보다 정확하게 결정한다.

본 논문의 구성은 다음과 같다. 2장에서 시소러스에 대하여 기술하고, 시소러스를 기반으로 하는 불리안 검색 시스템에서 사용될 수 있는 기존의 순위 결정 방법들 Relevance, R-Distance, K-Distance에 대하여 설명한다. 3장에서 개선된 퍼지

집합 모델과 확장된 불리안 모델을 설명하고, 이들을 시소러스의 연관성 정보를 이용할 수 있도록 확장한 KB-FSM과 KB-EBM을 제안한다. 4장에서 KB-FSM과 KB-EBM이 기존의 방법들보다 사람과 유사하게 문서들의 순위를 결정함을 성능 비교를 통하여 입증한다. 마지막으로 5장에서 결론 및 앞으로의 연구 방향을 제시한다.

Ⅱ. 시소러스를 기반으로 하는 순위 결정 방법

2.1 시소러스

시소러스는 문서에서 사용된 용어나 단어에 관계없이 문서의 주제를 기술하는 수단을 제공한다[2]. 시소러스는 분류 구조(Classification Structure), 제한된 어휘사전(Controlled Vocabulary), 순서화 체계(Ordering System)라고도 불리며, 개념을 표시하는 노드와 개념들 사이의 관계를 표시하는 링크로 구성된다. 개념들 사이의 관계로는 광의어(Broader-Term), 협의어(Narrower-Term)와 같은 'is-a' 관계와 동의어(Synonym), 관련어(Related-Term)등이 있다. 노드와 링크에 포함된 지식은 문서를 색인하고, 검색하는데 사용된다.

다양한 시소러스들이 기존의 정보 검색 시스템에서 사용되어 왔다. Medical Subject Headings (MeSH)는 MEDLINE 시스템이 사용하는 시소

H Information Systems
H.3 Information Storage and Retrieval
H.3.0 General
H.3.1 Content Analysis and Indexing
H.3.1.1 Abstracting Methods
H.3.1.2 Dictionaries
H.3.1.3 Indexing Methods
H.3.1.4 Linguistic Processing
H.3.1.5 Thesauruses
H.3.2 Information Storage
H.3.2.1 Record Classification
H.3.2.2 File Organization
H.3.3 Information Search and Retrieval
H.3.3.1 Clustering
H.3.3.2 Query Formulation
H.3.3.3 Retrieval Models
H.3.3.4 Search Process
H.3.3.5 Selection Process

〈그림 1〉 CRCS 시소러스의 일부분

러스로서, 9단계의 계층적 트리 형태를 갖는 15,000여개의 색인어들로 구성되어 있다[14]. 동의어까지 고려한다면, MeSH는 전체적으로 100,000개 이상의 색인어들을 가지고 있다. 또다른 시소러스로서 Association of Computing Machinary (ACM)의 간행물들을 색인하기 위한 Computing Reviews Classification Scheme (CRCS)이 있다[15]. CRCS는 5단계의 계층적 트리 형태를 갖는 1,000여개의 색인어들로 구성되어 있으며, 색인어들 사이의 관계로서 'is-a' 관계만을 사용한다. 〈그림 1〉은 CRCS 시소러스의 일부분을 표시한 것으로, 각각의 색인코드와 색인어를 표현하고 있다. 색인코드에 따라 부모와 자식 노드를 구분할 수 있고, 각각의 부모 노드와 자식 노드사이에는 'is-a' 관계가 있다. 예를 들면, 색인어 'H Information Systems'는 'H.3 Information Storage and Retrieval'의 부모 노드이고, 'H.3 Information Storage and Retrieval'은 'H.3.1 Content Analysis and Indexing'의 부모 노드이다.

2.2 기존의 방법들

지금까지 시소러스의 색인어로 표현된 문서와 불리안 질의 사이의 개념적 근접성(Conceptual Closeness)[4, 5] 또는, 개념적 거리(Conceptual Distance)[6-10]를 측정하는 다음과 같은 순위 결정 방법들이 개발되었다.

- 적합성 알고리즘 (Relevance) [4, 5]
- 거리 알고리즘 (R-Distance) [6-9]
- 거리 알고리즘 (K-Distance) [10]

Relevance, R-Distance, K-Distance에서 입력된 불리안 질의는 Quine-McCluskey의 알고리즘[16]에 의해 항상 최소 논리합 정규형으로 변환된 후, 문서와 질의 사이의 적합성 또는 거리 계산에 사용된다. 즉, 불리안 질의 q는 다음과 같은 형태로 변환된다.

$$q = Con_1(q)\ OR\ Con_2(q)\ OR \cdots OR\ Con_p(q) = \bigvee_{i=1}^{p} Con_i(q)$$

$$Con_i(q) = L_{i1}\ AND\ L_{i2}\ AND \cdots AND\ L_{im_j} = \bigwedge_{j=1}^{m_j} L_{ij}$$

위의 식에서 L_{ij}는 i번째 논리곱(Conjunction)에서의 j번째 항목(Literal)으로 '탐색어$_{ij}$' 또는 'NOT 탐색어$_{ij}$'이고, m_i는 i번째 논리곱에서의 항목들의 수이며, p는 전체 질의를 구성하는 논리곱들의 수이다.

Relevance, R-Distance, K-Distance는 기본 거리 함수(Primitive Distance Function)를 기반으로 하여 질의와 문서 사이의 개념적 근접성 또는 개념적 거리를 계산한다. 기본 거리 함수 distance(t_i, t_j)는 시소러스에 포함되어 있는 임의의 두 색인어들 t_i와 t_j 사이의 개념적 거리를 나타내며, 다음과 같이 정의된다.

distance(t_i, t_j) = t_i와 t_j를 연결하는 최소의 'is-a' 링크 수

〈그림 2〉는 적합성 알고리즘 Relevance에서 문서와 질의 사이의 개념적 근접성을 계산하는 식을 보여준다. n개의 색인어를 갖는 문서 d = t_1 AND ... AND t_n을 가정하자. 먼저 질의를 최소 논리합 정규형으로 변환하고, 질의에 포함된 각각의 논리곱에 대하여 문서의 적합성을 계산한 후, 가장 큰 값이 최종적인 질의와 문서 사이의 적합성이 된다. 논리곱과 문서 사이의 적합성은 문서의 모든 색인어와 논리곱에 포함된 모든 탐색어 사이의 적합성을 더한 후, 정규화 값으로 나누어 준다. 정규화 값은 논리곱과 문서 사이의 적합성이 가질 수 있는 가장 큰 값이고, 이러한 정규화 과정은 문서나 논

$$\begin{aligned} RELEVANCE(q, d) &= RELEVANCE(Con_1 \text{ OR } \dots \text{ OR } Con_p, d) \\ &= \underset{i=1,\dots,p}{MAX}\ Relevance(Con_i, d) \end{aligned}$$

$$\begin{aligned} Relevance(Con_i, d) &= Relevance(L_{i1} \text{ AND } \dots \text{ AND } L_{im_i}, t_1 \text{ AND } \dots \text{ AND } t_n) \\ &= \frac{\sum_{j=1}^{m_i}\sum_{k=1}^{n} relevance(L_{ij}, t_k)}{MIN(m_i, n) + \frac{1}{2}(m_i \bullet n - MIN(m_i, n))} \end{aligned}$$

$$relevance(L_{ij}, t_k) = \begin{cases} \dfrac{1}{1 + distance(t_{ij}, t_k)} & \text{if } L_{ij} \text{ is } t_{ij} \\ \dfrac{-1}{1 + distance(t_{ij}, t_k)} & \text{if } L_{ij} \text{ is NOT } t_{ij} \end{cases}$$

〈그림 2〉 Relevance와 관련된 계산식들

리곱을 구성하는 어휘의 수에 따라 적합성이 계속 커지는 것을 방지한다. 이처럼 정규화된 적합성은 0과 1사이의 값을 갖게 된다. 색인어와 탐색어 사이의 적합성 계산 식은 기본 거리 함수 distance의 역함수로서, NOT 연산자를 포함하지 않는 경우에는 0과 1사이의 적합성을 갖고, NOT 연산자를 포함하는 경우에는 -1과 0사이의 적합성을 갖는다.

R-Distance는 질의와 문서 사이의 개념적 거리를 측정한다. R-Distance에서 주어진 질의는 우선 최소 논리합 정규형으로 변환되고, 각각의 논리곱 Con_i 와 문서 사이의 거리들이 계산된 후, 가장 작은 값이 최종적인 질의와 문서 사이의 거리가 된다. 논리곱과 문서 사이의 거리는 문서의 모든 색인어와 논리곱에 포함된 모든 탐색어 사이의 거리를 더한 후, 정규화 값으로 나누어 준다. NOT t_{ij} 와 t_k 사이의 개념적 거리 계산을 위하여 NOT t_{ij} 는 시소러스 내에서 t_{ij}와 가장 관련이 없는 색인어들의 집합 t_{ij-1}로 치환된다. 〈그림 3〉은 R-Distance와 관련된 계산식들을 보여준다.

K-Distance 알고리즘은 R-Distance와 같이 문서와 질의 사이의 개념적 거리를 측정한다. 그러나, R-Distance에서 사용하지 않았던 시소러스 링크에 대한 가중치, 색인어 가중치, 탐색어 가중치를 사용하였다. 또한 R-Distance에서 NOT 연산자의 처리를 위하여 치환되는 색인어의 수를 대폭 감소시켰다.

Ⅲ. 순위 결정 방법 KB-FSM과 KB-EBM

시소러스를 기반으로 하는 불리안 검색 시스템에서 문서의 순위 결정에 사용될 수 있는 순위 결정 방법 Relevance, R-Distance, K-Distance는 색인어들 사이의 연관성 정보를 이용하여 문서들의 순위를 결정함으로써 많은 경우에 높은 검색 효율을 제공할 지라도, 불리안 연산자 AND, OR, NOT에 대한 연산 방법이 문제점으로 지적되어 왔다[11-13]. 본 장에서는 개선된 퍼지 집합 모델과 확장된 불리안 모델을 시소러스의 연관성 정보를 이

$$\text{DISTANCE}(q, d) = \text{DISTANCE}(\text{Con}_1 \text{ OR } \ldots \text{ OR } \text{Con}_p, d)$$
$$= \min_{i=1,\ldots,p} \text{Distance}(\text{Con}_i, d)$$
$$\text{Distance}(\text{Con}_i, d) = \text{Distance}(L_{i1} \text{ AND } \ldots \text{ AND } L_{im_i}, t_1 \text{ AND } \ldots \text{ AND } t_n)$$
$$= \begin{cases} \dfrac{1}{m_i \bullet n} \sum_{j=1}^{m_i} \sum_{k=1}^{n} \text{distance}(L_{ij}, t_k) & \text{if } \text{Con}_i \neq d \\ 0 & \text{if } \text{Con}_i = d \end{cases}$$

$$\text{distance}(L_{ij}, t_k) = \begin{cases} \text{distance}(t_{ij}, t_k) & \text{if } L_{ij} \text{ is } t_{ij} \\ \dfrac{1}{|t_{ij}^{-1}|} \sum_{t \in t_{ij}^{-1}} \text{distance}(t, t_k) & \text{if } L_{ij} \text{ is NOT } t_{ij} \end{cases}$$

$$t_{ij}^{-1} = \left\{ X \in V \mid \text{distance}(t_{ij}, X) = \max_{Y \in V} \text{distance}(t_{ij}, Y) \right\},$$
where V is the set of terms of a thesaurus

〈그림 3〉 R-Distance와 관련된 계산식들

용할 수 있도록 확장함으로써, 기존 방법들의 문제점들을 극복할 수 있는 새로운 순위 결정 방법 KB-FSM과 KB-EBM을 제안한다.

3. 1 개선된 퍼지 집합 모델과 확장된 불리안 모델

MIN과 MAX 연산자는 검색 효율을 저하시키는 특성을 지니고 있기 때문에, 이들을 불리안 연산자 계산식으로 사용하는 기존의 퍼지 집합 모델[17-20]은 정보 검색 시스템을 위한 검색 모델로서 부적합한 것으로 알려져 왔다[21, 22]. 퍼지 집합 이론에 대한 연구가 시작된 이후로, MIN과 MAX를 대신할 수 있는 다양한 퍼지 연산자들이 개발되어 왔다[23]. 이러한 퍼지 연산자들이 검색 효율에 미치는 영향을 분석함으로써, 긍정적 보상 연산자를 기반으로 하는 개선된 퍼지 집합 모델이 제안되었다[24-27]. 개선된 퍼지 집합 모델을 기반으로 하는 정보 검색 시스템은 다음에서 설명되는 〈T, Q, D, F〉로 정의될 수 있다.

(1)T는 질의와 문서를 표현하기 위해 사용되는 색인어들의 집합이다.

(2)Q는 시스템이 인식할 수 있는 질의들의 집

Query = (LQ, w_{LQ}) AND (RQ, w_{RQ}) |
(LQ, w_{LQ}) OR (RQ, w_{RQ}) |
NOT (LQ, w_{LQ}) |
Term
LQ = Query
RQ = Query

〈그림 4〉 가중치를 갖는 불리안 질의에 대한 BNF 문법

합이다. Q에 속하는 각각의 질의 q는 〈그림 4〉의 BNF(Backus Normal Form) 문법에 의해 정의될 수 있다. 0부터 1사이의 값을 갖는 w_{LQ}와 w_{RQ}는 질의 가중치로서 질의 내에서 탐색어와 탐색절이 갖는 중요도를 나타낸다. 예를 들면, 질의 q가 (((t1, w1) OR (t2, w2)), w1,2) AND (t3, w3)일 때, wi는 탐색어 t의 가중치이고 w1,2는 탐색절 (t1 OR t2)의 가중치이다.

(3)D는 문서들의 집합이다. D에 속하는 각각의 문서 d는 wi가 색인어 ti의 가중치일 때, ((t1, w1), ... (tn, wn))와 같이 표현된다. 색인어 가중치 wi는 0부터 1사이의 값을 갖는다.

(4)F는 문서값을 계산하는 순위 결정 함수(Ranking Function)로서 다음과 같이 정의된다.

$$F: D \times Q \rightarrow [0, 1]$$

순위 결정 함수 F는 각 쌍의 (d, q)에 0부터 1 사이의 값을 지정한다. 이 값은 문서 d 와 질의 q 사이의 유사성을 의미하며, 질의 q에 대한 문서 d의 문서값이다. 함수 F(d, q)는 다음과 같은 2단계 과정을 거쳐서 계산된다.

i)질의에 나타난 각각의 색인어 ti에 대하여, 기본 소속 함수 F(d, ti)는 문서 d에서 색인어 ti의 가중치 wi로 정의된다.

ii)불리안 연산자 AND, OR, NOT은 〈그림 5〉에서 주어진 식을 이용하여 계산된다. 두개 이상의 불리안 연산자를 포함하는 불리안 질의는 가장 안쪽에 위치하는 절부터 순환적으로 계산된다. 〈그림 6〉은 퍼지 집합 모델에서 지금까지 개발된 질의 가중치를 연산하는 3가지 방법들을 보여준다[17, 18, 28, 29]. p는 주어진 질의에서의 탐색어 또는 탐색절이고, w는 해당 탐색어 또는 탐색절의 가중치이다.

$$F(d, t_1 \text{ AND } t_2) = \gamma \bullet \text{MIN}(F(d,t_1), F(d,t_2)) + \frac{(1-\gamma)(F(d,t_1)+F(d,t_2))}{2}, \quad 0 \le \gamma \le 1$$

$$F(d, t_1 \text{ OR } t_2) = \gamma \bullet \text{MAX}(F(d,t_1), F(d,t_2)) + \frac{(1-\gamma)(F(d,t_1)+F(d,t_2))}{2}, \quad 0 \le \gamma \le 1$$

$$F(d, \text{NOT } t_1) = 1 - F(d,t_1)$$

〈그림 5〉 개선된 퍼지 집합 모델의 불리안 연산자 계산식

$$\text{(a)} \quad f(F(d,p), w) = w \bullet F(d,p)$$

$$\text{(b)} \begin{cases} f(F(d,p), w) = w \bullet F(d,p) & \text{when p appears in an OR clause} \\ f(F(d,p), w) = \frac{1}{w} \bullet F(d,p) & \text{when p appears in an AND clause} \end{cases}$$

$$\text{(c)} \begin{cases} f(F(d,p), w) = F(d,p) & \text{if } F(d,p) \ge w \\ f(F(d,p), w) = 0 & \text{if } F(d,p) < w \end{cases}$$

〈그림 6〉 퍼지 집합 모델에서 질의 가중치 연산 방법

확장된 불리안 모델은 불리안 모델, 퍼지 집합 모델, 벡터 공간 모델을 모두 포함하는 통합된 검색 모델로서, 매개변수 p의 값에 따라 서로 다른 문서값을 생성한다[30-32]. p의 값은 1부터 ∞까지의 값을 갖을 수 있으며, p의 값이 감소함에 따라 불리안 연산자의 해석이 완화된다. 즉 p의 값이 1일 때 불리안 연산자 AND와 OR의 의미가 완전히 사라지기 때문에 벡터 공간 모델의 결과와 일치하는 문서값을 생성하고, p의 값이 ∞일 때 불리안 모델과 퍼지 집합 모델의 결과와 일치하는 문서값을 생성한다.

확장된 불리안 모델은 질의 가중치와 문서 가중치의 연산 방법을 모두 제공한다. 확장된 불리안 모델에서 질의와 문서 사이의 유사도, 즉 문서값은 퍼지 집합 모델에서 사용된 2단계 과정을 이용하여 계산될 수 있다. 두 검색 모델 사이의 유일한 차이점은 불리안 연산자와 질의 가중치에 대한 계산식으로, 확장된 불리안 모델의 계산식은 〈그림 7〉과 같다.

3.2 KB-FSM과 KB-EBM의 설계

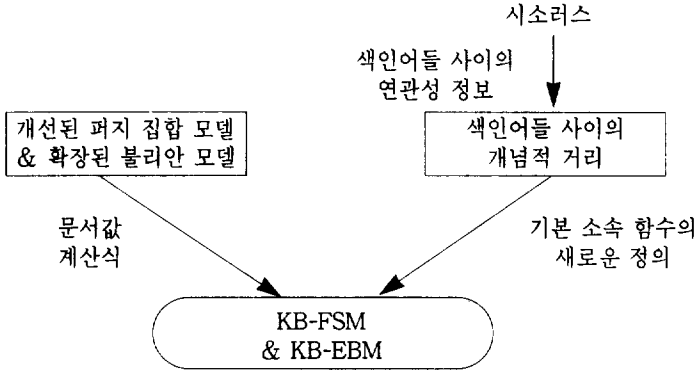
〈그림 8〉은 제안하는 방법의 개발에 사용된 기본적 개념을 표현하고 있다. KB-FSM과 KB-EBM은 시소러스의 연관성 정보를 이용하여 새롭게 정의된 기본 소속 함수를 사용하며, 불리안 연산자와 질의 가중치의 연산을 위하여 각각 개선된 퍼지 집합 모델과 확장된 불리안 모델의 문서값 계

$$F\left(d,\left(t_1,w_1\right)\ \text{AND}\ \left(t_2,w_2\right)\right)=1-\left[\frac{w_1^{\,p}\left(1-F\left(d,t_1\right)\right)^p+w_2^{\,p}\left(1-F\left(d,t_2\right)\right)^p}{w_1^{\,p}+w_2^{\,p}}\right]^{1/p}$$

$$F\left(d,\left(t_1,w_1\right)\ \text{OR}\ \left(t_2,w_2\right)\right)=\left[\frac{w_1^{\,p}F\left(d,t_1\right)^p+w_2^{\,p}F\left(d,t_2\right)^p}{w_1^{\,p}+w_2^{\,p}}\right]^{1/p}$$

$$F\left(d,\text{NOT}\ \left(t_1,w_1\right)\right)=1-w_1F\left(d,t_1\right)$$

〈그림 7〉 확장된 불리안 모델의 불리안 연산자와 질의 가중치에 대한 계산식



〈그림 8〉 KB-FSM과 KB-EBM의 개발에 사용된 기본적 개념

산식을 수정없이 이용한다.

3.3 새로운 기본 소속 함수의 정의

Relevance, R-Distance, K-Distance는 시소러스에 포함되어 있는 두 색인어들 사이의 개념적 거리를 계산하는 기본 거리 함수 distance(t_i, t_j)를 이용하여 질의와 문서 사이의 개념적 근접성 또는 개념적 거리를 측정하였다. 본 절에서는 기본 거리 함수 distance(t_i, t_j)를 이용하여 개선된 퍼지 집합 모델과 확장된 불리안 모델의 기본 소속 함수 F(d, t)를 새롭게 정의한다.

가장 간단한 경우로서, 문서 d가 단지 하나의 색인어 $\{t_k\}$만으로 표현되어 있는 경우를 고려하자. 이때 기본 소속 함수 F(d, t)는 F(t_k, t)로 간소화된다. 함수 F(t_k, t)는 두 색인어 t_k와 t 사이의 개념적 근접성을 의미하기 때문에, 기본 거리 함수 distance(t_k, t)에 반비례한다. 본 논문에서는 F(t_k, t)를 함수 distance^{-1}(t_k, t)로 다음과 같이 정의한다.

$$\text{distance}^{-1}(t_k,t)=\frac{\lambda}{\lambda+\text{distance}(t_k,t)},\quad 0<\lambda<\infty$$

distance^{-1}(t_k, t)의 값은 0부터 ∞ 사이의 임의의 λ 값에 대하여 0부터 1사이의 값을 생성한다. distance^{-1}(t_k, t)는 distance(t_k, t)가 0일 때 1로 계산되고, distance(t_k, t)가 무한대로 접근할 때 0에 접근한다. 매개변수 λ는 distance^{-1}(t_k, t)가 distance(t_k, t)에 반비례하는 정도를 조절한다. λ의 값은 문서값에 영향을 미치기 때문에, 실험을 통하여 가장 높은 검색 효율을 제공하는 λ의 값이 결정되어야 한다.

다음에서 두개 이상의 색인어로 표현되어 있는 문서 d=$\{t_1, ..., t_n\}$에 대한 기본 소속 함수 F(d, t)를 정의한다. 문서 d에 포함되어 있는 각각의 색인어 t_i에 대하여 distance^{-1}(t_i, t)를 계산한 후, 그 값들을 합산한 값을 unF(d, t)로 정의한다.

$$\text{unF}(d,t)=\sum_{i=1}^{n}\text{distance}^{-1}(t_i,t)$$

문서에 포함된 색인어의 수가 많을수록, unF(d, t)의 값이 크다. 따라서 문서 d가 두개 이상의 색인어로 표현되어 있을 경우, 기본 소속 함수 F(d, t)는 다음과 같이 정규화된 형태로 정의되어야 한다. 정규화 값은 unF(d, t)가 생성할 수 있는 가장 큰 값이다.

$$\text{F}(d,t)=\frac{\sum_{i=1}^{n}\text{distance}^{-1}(t_i,t)}{1+\frac{\lambda}{\lambda+1}(n-1)}$$

지금까지 가중치를 갖지 않는 문서들만을 고려하였다. 다음에서는 가장 일반적인 경우로서 색인어 가중치를 갖는 문서들에 대하여 기본 소속 함수 F(d, t)를 정의한다. 문서 d가 색인어와 가중치의 쌍들로 $\{(t_1, w_1), ..., (t_n, w_n)\}$와 같이 표현되어 있다고 하자. 가중치를 갖지 않는 문서는 색인어들에 대한 가중치를 1로 가정하고 있기 때문에, 두 색인어 t_i와 t 사이의 유사도는 함수 distance^{-1}(t_i, t)와 색인어 t_i의 가중치 wi의 곱으로 정의되어야 한다. 따라서 본 논문에서 제안하는 기본 소속 함수 F(d, t)는 다음과 같다.

$$\text{F}(d,t)=\frac{\sum_{i=1}^{n}\left(\text{distance}^{-1}(t_i,t)\bullet w_i\right)}{1+\frac{\lambda}{\lambda+1}(n-1)}$$

결론적으로 순위 결정 알고리즘 KB-FSM과 KB-EBM은 다음과 같이 요약될 수 있다. 첫째,

문서는 시소러스의 색인어와 가중치의 쌍들 {(t_1, w_1), ..., (t_n, w_n)}로 표현된다. 둘째, 질의는 탐색어와 탐색절 가중치를 포함하는 불리안 질의이다. 마지막으로 문서와 질의 사이의 유사도는 개선된 퍼지 집합 모델과 확장된 불리안 모델에서 도입한 식들과 본 절에서 새롭게 정의한 기본 소속 함수를 사용하여 계산된다.

3.4 다양한 특성의 기본 소속 함수들

0부터 1사이의 값만을 생성하도록 기본 소속 함수를 정의한다면, 개선된 퍼지 집합 모델과 확장된 불리안 모델의 문서값 계산식을 변경없이 사용할 수 있다. 본 절에서는 3.3절에서 제시한 기본 소속 함수와 다른 특성을 갖는 4개의 기본 소속 함수를 추가로 정의한다. 이들은 모두 색인어들 사이의 개념적 거리를 측정하는 기본 거리 함수 distance(t_i, t_j)를 기반으로 하고 있다.

$$F_{closest}(d,t) = \max_{t_i \in d}\left(\text{distance}^{-1}(t_i,t) \bullet w_i\right)$$

$$F_{average}(d,t) = \frac{1}{2}\left\{\max_{t_i \in d}\left(\text{distance}^{-1}(t_i,t) \bullet w_i\right) + \frac{\sum_{i=1}^{n}\left(\text{distance}^{-1}(t_i,t) \bullet w_i\right)}{1+\frac{\lambda}{\lambda+1}(n-1)}\right\}$$

$$F_{square}(d,t) = \frac{\sum_{i=1}^{n}\left(\left(\text{distance}^{-1}(t_i,t)\right)^2 \bullet w_i\right)}{1+\frac{\lambda}{\lambda+1}(n-1)}$$

$$F_{square\text{-}closest}(d,t) = \max_{t_i \in d}\left(\left(\text{distance}^{-1}(t_i,t)\right)^2 \bullet w_i\right)$$

Ⅳ. 성능 비교

4.1 성능 평가 자료와 방법

순위 결정 방법의 성능 평가는 순위 결정 방법이 정한 문서의 순위와 사용자가 정한 문서의 순위를 비교하는 것이 가장 바람직하다. 일반적으로 두가지 서로 다른 순위가 얼마나 유사한가를 계산하기 위하여 순위 연관 방법(Rank Correlation Method)이 사용된다[4, 6, 7, 10]. k개의 객체 $e_1,...,e_k$가 있을때, 이에 대한 서로 다른 순위 $r_1,...,r_k$와 $r_1',...,r_k'$ 사이의 Spearman 연관 계수 는 다음과 같다[33].

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

Spearman 연관 계수 는 두 순위가 일치할 때 1이고, 서로 관련이 없는 순위일 때 0이며, 서로 정반대의 순위일 때 -1의 값을 갖는다.

순위 결정 방법의 성능 평가를 위해서는 시소러스의 색인어들로 표현된 문서들과 불리안 질의, 그리고 불리안 질의와 문서 쌍에 대해 사용자가 정한 순위가 필요하다. 본 논문에서는 〈그림 9〉에서 기술된 4개의 성능 평가 자료들을 사용하였다. 두 개의 성능 평가 자료는 기존의 순위 결정 방법들 R-Distance와 K-Distance의 성능 평가를 위해 사용되었던 것이며, 나머지 두개의 성능 평가 자료는 보다 정확한 성능 비교를 위하여 새롭게 생성되었다.

성능 평가 자료 1은 NOT 연산자를 포함하지 않는 4개의 불리안 질의와 9개의 문서들, 그리고 15

자료번호	문서수	질의수	참여인원수	주제영역
1	9	4	15	Information Storage and Retrieval
2	6	5	20	Artificial Intelligence
3	7	5	15	Distributed Database System
4	7	5	15	

〈그림 9〉 성능 평가 자료

Query 1: Retrieval Models (h.3.3.3)
Query 2: Retrieval Models (h.3.3.3) AND
Search Process (h.3.3.4)
Query 3: Information Search and Retrieval (h.3.3)
Query 4: Information Search and Retrieval (h.3.3) AND
Retrieval Models (h.3.3)

〈그림 10〉 성능 평가 자료 1에 포함된 불리안 질의

Query 5: Artificial Intelligence (i.2) AND
NOT Knowledge Representation Formalisms and Methods (i.2.4)
Query 6: Artificial Intelligence (i.2) AND
Speech Recognition and Understanding (i.2.7.5) AND
NOT Deduction and Theorem Proving (i.2.3)
Query 7: Artificial Intelligence (i.2) AND
NOT Deduction and Theorem Proving (i.2.3)
Query 8: Artificial Intelligence (i.2) AND
Frames and Scripts (i.2.4.1) AND
NOT Programming Languages (d.3)
Query 9: Artificial Intelligence (i.2) AND
Deduction and Theorem Proving (i.2.3) AND
NOT Applications and Expert Systems (i.2.1)

〈그림 11〉 성능 평가 자료 2에 포함된 불리안 질의

명의 학생들이 정한 문서들의 순위들로 구성되어 있다. 질의와 문서는 CRCS의 색인어들로 표현되어 있으며, 색인어들은 Information Storage and Retrieval 분야의 내용이다[6]. 성능 평가 자료 2는 NOT 연산자를 포함하는 5개의 불리안 질의와 6개의 문서들, 그리고 20명의 학생들이 정한 문서들의 순위들로 구성되어 있다[10]. 질의와 문서는 CRCS의 색인어들로 표현되어 있고, 문서는 Communications of the ACM에서 임의로 추출되었으며, 문서의 내용은 Artificial Intelligence 분야이다. 〈그림 10〉과 〈그림 11〉은 사용된 성능 평가 자료에 포함된 질의들이다.

기존의 성능 평가 자료에 포함되어 있는 불리안 질의들은 AND와 NOT 연산자만을 지니고 있다. OR 연산자에 대한 성능 평가와 AND와 NOT 연산자에 대한 보다 정확한 성능 평가를 위하여 두 개의 성능 평가 자료를 생성하였다. 각각의 성능 평가 자료는 〈그림 12〉와 〈그림 13〉에서 제시된 5개의 불리안 질의와 7개의 문서들, 그리고 15명의 학생들이 정한 문서들의 순위로 구성되어 있다. 질의와 문서는 CRCS의 색인어들로 표현되어 있고, 문서의 내용은 Distributed Database System 분야이다[13, 34].

4.2 성능 평가 결과

4개의 성능 평가 자료에 본 논문에서 제안한 KB-FSM, KB-EBM과 기존의 방법들 Relevance, R-Distance, K-Distance를 적용하고, 순위 결정 방법들이 결정한 순위와 학생들이 정한 순위 사이의 Spearman 연관 계수를 계산한 후, 비교하였다.

KB-FSM은 문서값을 조절할 수 있는 2개의 매개변수 γ와 λ를 가지고 있다. 4개의 γ값 0, 0.3, 0.6, 0.9에 대하여, λ값을 0.5부터 2까지 증가시키면서 KB-FSM의 성능을 평가하였다. 본 논문에서는 기본 소속 함수로 사용될 수 있는 5개의 새로운 함수를 정의하였다. 따라서 이들 기본 소속 함수들이 KB-FSM의 검색 효율에 미치는 영향을 우선적으로 비교한다. 〈그림 14〉는 성능 비교의 결과를 보여준다. 기본 소속 함수 F가 사용되고, γ와 λ의 값들이 각각 0.3과 1.4일때 가장 높은 검색 효율을 제공하였다.

KB-EBM은 문서값을 조절할 수 있는 2개의 매개변수 p와 λ를 가지고 있다. 4개의 p값 1, 2, 5, 9에 대하여, λ값을 0.5부터 2까지 증가시키면서 KB-EBM의 성능을 평가하였다. 〈그림 15〉는 5개의 기본 소속 함수들들에 대한 KB-EBM의 성능을 보여준다. 기본 소속 함수 F가 사용되고, p와 λ의 값들이 각각 2와 1.4일때 가장 높은 검색 효율을 제공하였다.

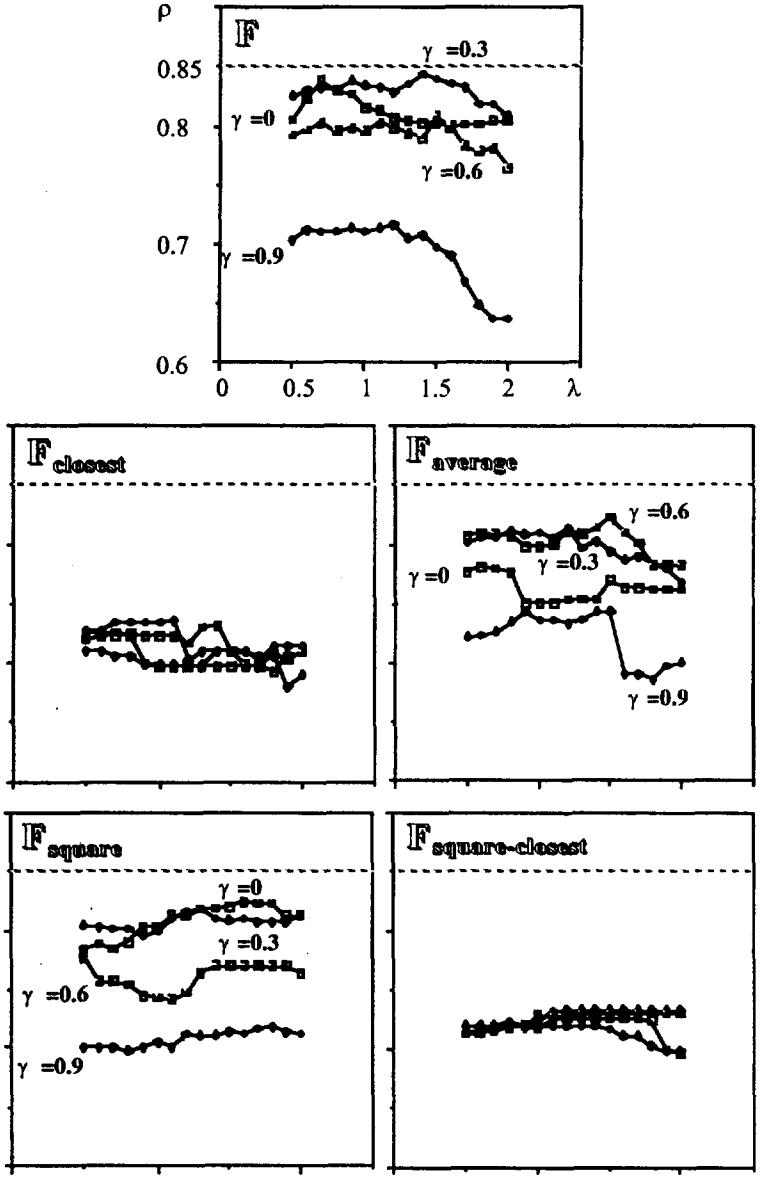
〈표 1〉부터 〈표 4〉는 성능 평가 자료 1부터 4에 순위 결정 방법 Relevance, K-Distance, R-Distance, KB-FSM, KB-EBM을 적용한 결과이다. 이들로부터 KB-FSM과 KB-EBM이 실험에서 사용된 다양한 형태의 질의에 관계없이 높은 검색 효율을 제공함을 알 수 있다. 〈표 5〉는 순위 결정 방법들에 대한 성능 평가의 결과를 요약한 것이다. Spearman 연관 계수의 평균, 최저치, 최고치 모두에서 KB-FSM과 KB-EBM이 가장 좋은 성능을 보여주고 있다. 또한 KB-FSM과 KB-EBM은 가장 작은 분산값을 가지고 있으며, 이는 제안하는 방법들이 가장 안정되어 있음을 의미한다.

Query 10: Distributed Systems (h.2.4.1) AND
Distributed Systems (c.2.4) AND
NOT Trees (g.2.2.4)
Query 11: Query Processing (h.2.4.2) AND
Data Models (h.2.1.1) AND
NOT Distributed Systems (c.2.4)
Query 12: Distributed Systems (h.2.4.1) AND
Trees (g.2.2.4) AND
Combinatorial Algorithms(g.2.1.1) AND
NOT Distributed Databases (c.2.4.2)
Query 13: Logical Design (h.2.1) AND
Distributed Systems (h.2.4.1) AND
Query Processing (h.2.4.2) AND
NOT Graph Theory (g.2.2)
Query 14: Systems (h.2.4) AND
Languages (h.2.3) AND
NOT Distributed Systems (c.2.4)

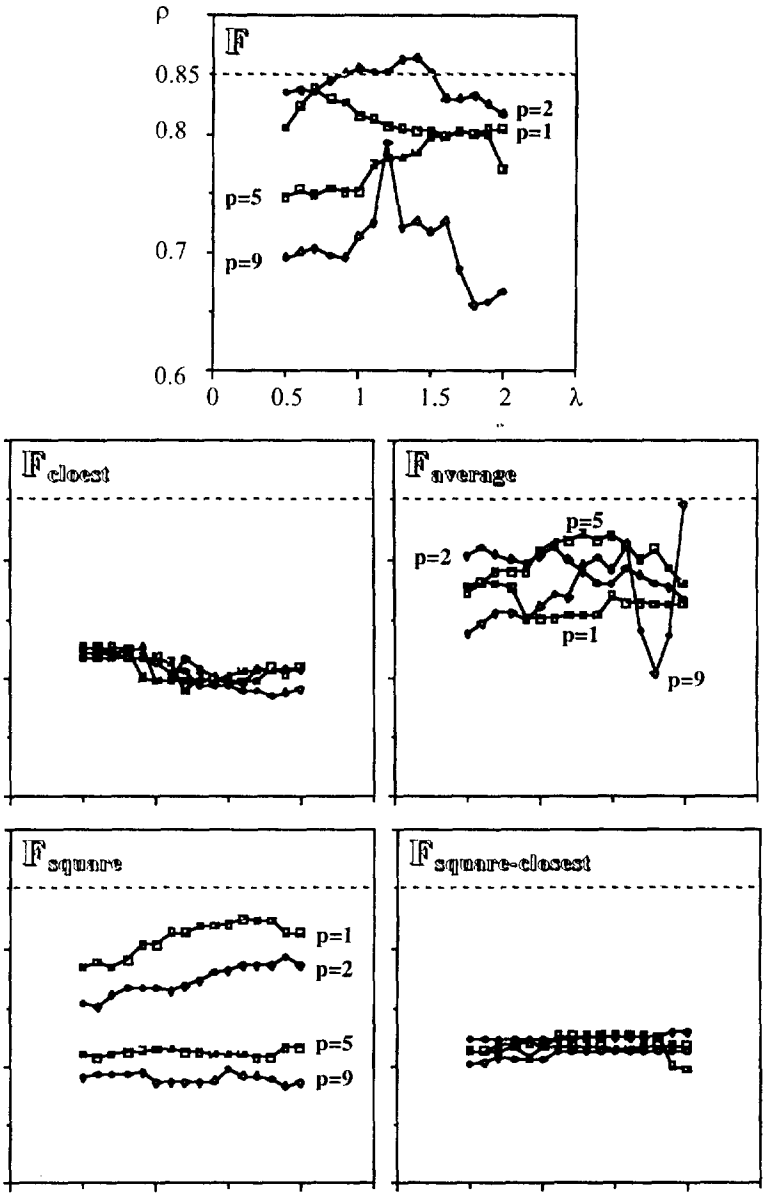
〈그림 12〉 성능 평가 자료 3에 포함된 불리안 질의

Query 15: Distributed Systems (h.2.4.1) OR
Distributed Systems (c.2.4) OR
Trees (g.2.2.4)
Query 16: Query Processing (h.2.4.2) OR
Data Models (h.2.1.1) OR
Distributed Databases (c.2.4.2)
Query 17: Distributed Systems (h.2.4.1) OR
Trees (g.2.2.4) OR
Combinatorial Algorithms(g.2.1.1)
Query 18: Distributed Systems (h.2.4.1) OR
Logical Design (h.2.1) OR
Distributed Databases (c.2.4.2)
Query 19: Systems (h.2.4) OR
Languages (h.2.3)

〈그림 13〉 성능 평가 자료 4에 포함된 불리안 질의



〈그림 14〉 KB-FSM에서 5개 기본 소속 함수들의 성능 비교



〈그림 15〉 KB-EBM에서 5개 기본 소속 함수들의 성능 비교

〈표 1〉 성능 평가 자료 1을 사용한 검색 효율 비교

	Query1	Query2	Query3	Query4	Average
Relevance	0.867	0.783	0.933	0.800	0.846
R-Distance	0.879	0.742	0.842	0.783	0.812
K-Distance	0.867	0.833	0.904	0.817	0.855
KB-FSM	0.867	0.867	0.917	0.817	0.863
KB-EBM	0.867	0.867	0.917	0.800	0.863

〈표 2〉 성능 평가 자료 2를 사용한 검색 효율 비교

	Query5	Query6	Query7	Query8	Query9	Average
Relevance	0.943	0.829	0.943	0.314	0.257	0.657
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
KB-FSM	0.943	0.943	1.000	0.657	0.714	0.851
KB-EBM	0.943	0.943	0.943	0.657	0.714	0.840

〈표 3〉 성능 평가 자료 3를 사용한 검색 효율 비교

	Query10	Query11	Query12	Query13	Query14	Average
Relevance	0.857	0.786	0.750	0.884	0.723	0.800
R-Distance	0.786	0.357	0.679	0.179	0.670	0.534
K-Distance	0.929	0.902	0.679	0.750	0.688	0.790
KB-FSM	0.893	0.500	0.964	0.929	0.634	0.784
KB-EBM	0.893	0.857	1.000	0.857	0.777	0.877

〈표 4〉 성능 평가 자료 4를 사용한 검색 효율 비교

	Query15	Query16	Query17	Query18	Query19	Average
Relevance	0.902	0.268	0.705	0.607	0.714	0.639
R-Distance	0.268	0.929	0.420	0.616	0.991	0.645
K-Distance	0.554	0.884	0.420	0.027	0.786	0.534
KB-FSM	0.857	0.964	0.750	0.848	0.964	0.877
KB-EBM	0.893	0.929	0.750	0.848	0.964	0.877

〈표 5〉 성능 평가 결과에 대한 요약

	Average	Variance	Worst	Best
Relevance	0.730	0.048	0.257	0.943
R-Distance	0.552	0.185	-0.771	0.991
K-Distance	0.755	0.053	0.027	0.986
KB-FSM	0.844	0.018	0.500	1.000
KB-EBM	0.864	0.008	0.657	1.000

Ⅴ. 결론 및 앞으로의 연구

문서의 순위 결정 방법은 정보 검색 시스템의 중요한 구성 요소 중의 하나이다. 정보 검색 시스템은 검색된 문서에 대하여 순위 결정 방법을 적용함으로써 문서가 질의를 만족하는 정도를 나타내는 문서값을 계산하고, 계산된 문서값에 따라 문서들에 순위를 부여한다. 높은 순위를 갖는 문서일수록 질의에 대한 만족도가 크며, 사용자는 높은 순위를 갖는 문서를 우선적으로 검토함으로써 필요한 정보를 얻는데 소모되는 시간을 최소화할 수 있다.

불리안 검색 시스템은 짧은 검색 시간을 제공하고 질의를 비교적 쉽게 표현할 수 있기 때문에, 정보 검색 분야에서 가장 널리 사용되어 왔다. 불리안 검색 시스템이 문서의 색인을 위해 시소러스를 사용한다면, 다음과 같은 장점을 추가로 얻을 수 있다. 첫째, 색인어가 시소러스로부터 선택되기 때문에, 문서에서 사용된 특정한 용어에 관계없이 같은 내용을 갖는 문서는 같은 색인어에 의해 검색될 수 있다. 둘째, 색인어들 사이의 연관성 정보를 이용하여 문서값을 보다 정확하게 계산할 수 있다. 본 논문에서는 시소러스를 기반으로 하는 불리안 검색 시스템에서 사용될 수 있는 문서의 순위 결정 방법에 대하여 고찰하였다.

지금까지 Relevance, R-Distance, K-Distance와 같은 방법들이 시소러스를 기반으로 하는 불리안 검색 시스템에서 문서의 순위 결정을 위하여 개발되었다. 이러한 방법들은 문서값 계산에 시소러스의 'is-a' 관계를 이용함으로써 많은 경우에 사람과 유사하게 문서의 순위를 결정할 지라도, 가중치와 불리안 연산자에 대한 효율적인 연산 방법을 지원하지 않는다. 본 논문에서는 개선된 퍼지 집합 모델과 확장된 불리안 모델의 문서값 계산식을 이용하여 새로운 순위 결정 방법 KB-FSM과 KB-EBM을 제안하였다. 이들 방법은 기존 방법들의 문제점들을 극복하고, 또한 문서들의 순위를 보다 정확하게 결정한다.

앞으로의 연구는 다음과 같다. 첫째, KB-FSM과 KB-EBM은 문서들의 순위를 결정하기 위해 'is-a' 관계만을 이용한다. 시소러스는 'is-a' 관계 이외에 동의어, 관계어 등의 유용한 관계들을 가지고 있다. 이러한 관계들을 문서값 계산에 효율적으로 이용하는 방법이 연구되어야 한다. 둘째, 기본 거리 함수 distance의 계산에 있어서 하나의 'is-a' 링크로 연결된 색인어들 사이의 개념적 거리를 1로 가정하고 있다. 일반적으로 시소러스에서 인접한 색인어들 사이의 개념적 거리는 서로 다르게 평가되는 것이 바람직하다. 따라서 시소러스의 링크에 색인어들 사이의 개념적 거리에 따른 가

중치를 부여하고, 이를 효율적으로 연산하는 방법이 연구되어야 한다.

참고문헌

[1] G. Salton, "Historical Note: The Past Thirty Years in Information Retrieval," *Journal of the American Society for Information Science*, Vol. 38, No. 5, pp. 375-380, 1987.

[2] E. Svenonius, "Unanswered Questions in the Design of Controlled Vocabularies," *Journal of American Society for Information Science*, Vol. 37, No. 5, pp. 331-340, 1986.

[3] H.P. Giger, "Concept Based Retrieval in Classical IR Systems," *Proceedings of the 11th International ACM SIGIR Conference on Research and Development in Information and Retrieval*, 1988, pp. 275-289.

[4] R. Rada, S. Humphrey, and C. Coccia, "A Knowledge-Base for Retrieval Evaluation," *ACM Annual Conference*, pp. 360-366, 1985.

[5] R. Rada, S. Humphrey, et al., "Relevance on a Biomedical Classification Structure," *The Expert Systems in Government Symposium*, pp. 532-537, 1985.

[6] C.F. McMath, R.S. Tamaru, and R. Rada, "A Graphical Thesaurus-Based Information Retrieval System," *International Journal of Man-Machine Studies*, Vol. 31, No. 2, pp. 121-147, 1989.

[7] H. Mili and R. Rada, "Merging Thesauri: Principles and Evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 2, pp. 204-220, 1988.

[8] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and Application of a Metric on Semantic Nets," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 19, No. 1, pp. 17-30, 1989.

[9] R. Rada and E. Bicknell, "Ranking Documents with a Thesaurus," *Journal of the American Society for Information Science*, Vol. 40, No. 5, pp. 304-310, 1989.

[10] Y.W. Kim and J.H. Kim, "A Model of Knowledge Based Information Retrieval with Hierarchical Concept Graph," *Journal of Documentation*, Vol. 46, No. 2, pp. 113-136, 1990.

[11] J.H. Lee, M.H. Kim and Y.J. Lee, "A Knowledge-Based Approach to Rank Documents for Boolean Queries," *Far-East Workshop on Future Database Systems*, Kyoto, Japan, pp. 315-322, 1992.

[12] J.H. Lee, M.H. Kim and Y.J. Lee, "Ranking Documents in Thesaurus-Based Boolean Retrieval Systems," *Information Processing & Management*, Vol. 30, No. 1, pp. 79-1, 1994.

[13] J.H. Lee, M.H. Kim and Y.J. Lee, "Information Retrieval Based on Conceptual Distance in Is-a Hierarchies," *Journal of Documentation*, Vol. 49, No. 2, pp. 188-207, 1993.

[14] D.B. McCarn, "MEDLINE: An Introduction to On-Line Searching," *Journal of the American Society for Information Science*, Vol. 31, No. 3, pp. 181-192, 1980.

[15] J.E. Sammet and A. Ralston, "The

New (1982) Computing Reviews Classification System - Final Version," *Communications of the ACM*, Vol. 25, No. 1, pp. 13-25, 1982.

[16] E.J. McCluskey, "Minimization of Boolean Functions," *Bell System Technical Journal*, Vol. 35, No. 6, pp. 1417-1444, 1956.

[17] D.A. Buell, "A General Model of Query Processing in Information Retrieval System," *Information Processing & Management*, Vol. 17, No. 5, pp. 249-262, 1981.

[18] T. Radecki, "Fuzzy Set Theoretical Approach to Document Retrieval," *Information Processing & Management*, Vol. 15, No. 5, pp. 247-259, 1979.

[19] W.M. Sachs, "An Approach to Associative Retrieval through the Theory of Fuzzy Sets," *Journal of the American Society for Information Science*, Vol. 27, pp. 85-87, 1976.

[20] W.G. Waller and D.H. Kraft, "A Mathematical Model of a Weighted Boolean Retrieval System," *Information Processing & Management*, Vol. 15, pp. 235-245, 1979.

[21] A. Bookstein, "Fuzzy Requests: An Approach to Weighted Boolean Searches," *Journal of the American Society for Information Science*, Vol. 31, No. 4, pp. 240-247, 1980.

[22] S.E. Robertson, "On the Nature of Fuzz: A Diatribe," *Journal of the American Society for Information Science*, Vol. 29, No. 6, pp. 304-307, 1978.

[23] H.J. Zimmermann, *Fuzzy Set Theory and Its Applications*, 2nd ed., Kluwer Academic Publishers, 1991.

[24] J.H. Lee, M.H. Kim and Y.J. Lee, "Enhancing the Fuzzy Set Model for High Quality Document Rankings," *Microprocessing and Multiprogramming - The Euromicro Journal*, Vol. 35, No. 5, pp. 337-344, 1992.

[25] J.H. Lee, W.Y. Kim, M.H. Kim and Y.J. Lee, "Enhancing the Fuzzy Set Model with Positively Compensatory Operators," *International Symposium on Database Systems on Advanced Applications*, Taejon, Korea, pp. 368-375, 1993.

[26] J.H. Lee, W.Y. Kim, M.H. Kim and Y.J. Lee, "On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework," *International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA USA, pp. 291-297, 1993.

[27] M.H. Kim and J.H. Lee and Y.J Lee, "Analysis of Fuzzy Operators for High Quality Information Retrieval," *Information Processing Letters*, Vol. 46, No. 5, pp. 251-256, 1993.

[28] A. Bookstein, "A Comparison of Two Weighting Schemes for Boolean Retrieval," *Journal of the American Society for Information Science*, Vol. 32, No. 4, pp. 275-279, 1981.

[29] D.A. Buell and D.H. Kraft, "Threshold Values and Boolean Retrieval System," *Information Processing & Management*, Vol. 17, No. 3, pp. 127-136, 1981.

[30] G. Salton, E.A. Fox, and H. Wu, "Extended Boolean Information Retrieval," *Communications of the ACM*, Vol. 26, No. 11,

pp. 1022-1036, 1983.

[31] G. Salton, E.A. Fox, and E. Voorhees, "Advanced Feedback Methods in Information Retrieval," *Journal of the American Society for Information Science*, Vol. 36, No. 3, pp. 200-210, 1985.

[32] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley, 1989.

[33] M. Kendall, *Rank Correlation Methods*, 4th ed., Charles Griffin & Company Ltd., 1975.

[34] J.H. Lee, M.H. Kim and Y.J. Lee, *The Extended Boolean Model Using Term Dependencies from a Thesaurus*, Department of Computer Science, Korea Advanced Institute of Science and Technology, 1992. (Technical Report CS-TR-92-76)