한국 퍼지 및 지능시스템학회 논문지 2000, Vol. 10, No. 3 (2000)242~251

퍼지관계곱을 이용한 정보검색시스템의 성능 개선

Performance Improvement of Information Retrieval System by means of Fuzzy Relational Product

김창민 · 김용기

Chang-Min Kim and Yong-Gi Kim

경상대학교 컴퓨터과학과 및 전산개발연구소

요 약

퍼지관계 개념을 응용한 BK-퍼지정보검색기법은 형태론에 입각하는 기존의 정보검색기법과는 달리 문서와 용어의 상대적 의미에 근거한 정보검색 기법이다. 그러나 BK-퍼지정보검색기법은 높은 시간복잡도(time complexity)의 검색 연산을 내재하고 있어 실제 대용량의 정보 검색은 사실상 불가능하다. 본 논문에서는 BK-퍼지정보검색모델의 시간복잡도를 낮추기 위해, 축소용어집합(reduced term set)을 이용한 개선된 BK-퍼지정보검색모델(A-FIRM)을 제안한다. 개선된 BK-FIRM은 시스템 처리시간과 신뢰도 간 상충점(trade-off)을 제공한다. 축소용어집합은 용어집합의 부분집합으로서 검색결과의 신뢰도와 밀접한 관계를 가진다. 동일한 크기의 축소용어집합이 주어질 때, 보다 적절한 용어들로 구성된 축소용어집합이 보다 나은 검색 신뢰도를 이끈다. 따라서 보다 적절한 축소용어집합 구성을 위한 축소용어집합 추출방법이 요구된다. 본 논문에서는 축소용어집합 추출방법을 크게 무작위 추출, 규칙에 의한 추출, 인간에 의한 직관적 추출 방법으로 구분하고 검색결과의 신뢰도 변화 형태를 분석한다.

1. 소 개

과학과 기술 분야의 급속한 발전은 수많은 주제들에 대한 방대한 양의 정보를 생성하는 정보화 사회를 탄생시켰다. 원하는 정보에 대한 정확하고 빠른 접근은 정보화 사회를 살아가는 현대인들에게 성공 여부를 결정짓는 중요한 요소가 되었다. 그러나 대용량의 데이터로부터 원하는 정보를 한정된 시간 내에 검색하는 것은 쉬운 일이 아니다. 1960년대 초, 이러한 문제점을 해결하기 위하여 컴퓨터를 이용하여 정보를 검색하는 정보검색(information retrieval)[1] 분야가 확립되었다.

정보검색의 대표적인 검색모델은 불리언 식으로 표현된 질의어를 이용하여 정보를 검색하는 불리언 검색 모델이다. 불리언 검색 모델은 구현하기 쉽고 질의어의 처리 시간 면에서 효율적이기 때문에 가장 널리 쓰이고 있는 검색 모델이다. 불리언 검색 모델은 질의어가 적절히 입력되면 조회율과 정확도 면에서 좋은 성능을 보인다. 그러나 질의어의 엄격한 해석, 검색결과 우선 순위에 대한 대비책 부재, 색인 결정 시 존재하는 불확실성에 대한 대비책 부재와 같은 검색 효율의 한계성을 보인다. 따라서 불리언 검색 연산을 융통성 있게 해석하여 탐색결과의 정확도와 조회율을 향상시키기 위해 다양한 모델이 제안되었다. 대표적인 모델로서, 퍼지집합론에 근거한 Fox와 Sharat의 MMM(Max Min and Max) 모델[2], Paice의 Paice 모델[3], 정규화된 용어와 역문헌의 빈도수 통계치를 이용한 Fox의 P-norm 모델[4], 퍼지집합론과 퍼지관계곱을 이용한 Bandler와 Kohout의 퍼지정보검색모델[5] 등이 있다.

일찌기 Rijsbergen은 자료검색(data retrieval)과 정보검색(information retrieval)을 비교 분석하였다. 정보검색은 자료검색에 반해 항목의 부정확한 매칭(partial match, best match), 항목에 대한 적합도 명세 필요성 그리고 항목의 다형 분류법(polythetic classification)과 같은 특징을 가지고 있다. Rijsbergen은 이와 같은 정보검색의 모호한 특성으로부터 정보검색과 퍼지이론의 자연스러운 융화를 이끌어낼 수 있다고 말한다[6,7]. Bandler와 Kohout는 Rijsbergen의 통찰력 있는 관점에 근거하여 불리언 정보검색모델을 확장한 BK-퍼지정보검색모델(BK-FIRM : Bandler and Kohout's fuzzy information retrieval model)을 제안하였다. Bandler와 Kohout의 퍼지정보검색모델은 시소러스 자동 구축 기능, 검색 결과의 퍼지화된 우선 순위 제공,

본 연구는 1998년 과학재단의 핵심전문연구(과제번호 981-0919-102-2)에 의해 수행되었습니다.

직접 관련 없는 제 3의 개체 유추 검색 등과 같은 장점을 가지고 있다. 그러나 BK-퍼지정보검색모델은 높은 시간복잡도(time complexity)의 검색연산을 내재하고 있어 대용량의 정보 검색이 요구되는 분야에 적용하는 것은 사실상 불가능하다.

본 논문에서는 축소용어집합을 이용하여 BK-퍼지정보검색모델의 시간복잡도를 낮추는 개선된 BK-퍼지정보검색모델(A-FIRM : Advanced Bandler and Kohout's fuzzy information retrieval model)을 제안하고 평가한다.

2. BK- 퍼지정보검색모델

Bandler와 Kohout의 BK-퍼지정보검색모델은 형태론에 입각한 기존의 정보검색기법과는 달리 문서와 용어의 상대적 의미를 표현하는 퍼지관계와 퍼지관계곱을 이용하는 정보검색기법으로서 자동 시소러소(thesaurus) 구축기능과 검색결과의 퍼지화된 우선 순위 제공과 같은 기능을 기본적으로 가지고 있다. BK-퍼지정보검색모델은 우선 문서집합과 용어집합을 정의하고 문서와 용어의 상대적 의미를 문서와 용어의 퍼지관계행렬로 표현하며 이에 퍼지관계곱 연산을 적용하여 시소러스를 형성한다. 이후 사용자로부터 요구된 질의어를 해석하고 시소러스를 이용하여 확장한 후 문서와 용어의 퍼지관계를 이용하여 문서를 검색한다. BK-퍼지정보검색모델은 시소러스를 이용하여 주어진 용어의 의미를 확장하는 관계요구(R-request) 연산과 사용자로부터 주어진 질의어를 해석하여 적합한 문서를 검색하는 퍼지검색요구(FS-request) 연산을 제공한다[6-11].

2.1 BK-퍼지정보검색모델과 시소러스

BK-퍼지정보검색모델에서 퍼지관계요구(R-request)는 시소러스를 이용하여 주어진 용어에 관한 다른 용어들의 연관성을 제공하는 연산이다. 시소러스는 용어의 상호 연관성을 유지하는 구조체라 말할 수 있는데, 전통적 정보검색 기법에서는 용어간 유사어를 유지하는 동의어 사전을 의미한다. BK-퍼지정보검색모델에서 시소러스 구축은 4단계의 처리절차로 구성된다. 첫 번째 단계에서는 수식 (1)-(3)과 같이 문서집합 D, 용어집합 T, 문서와 용어의 퍼지관계 $\tilde{R}$을 설정하여 초기화 한다. 두 번째 단계에서는 수식 (4)과 같이 퍼지관계 $\tilde{R}$의 전치행렬 $\tilde{R}^T$와 퍼지관계 $\tilde{R}$에 퍼지관계곱 연산 @(◁, ▷ 또는 □)을 적용하여 결과 퍼지관계 $\tilde{B}$를 산출한다. 세 번째 단계에서는 수식 (5)과 같이 퍼지관계 $\tilde{B}$를 이진행렬로 변환하기 위하여 $\tilde{B}$에 α-cut을 적용하여 O_α를 산출한다. 네 번째 단계에서는 수식 (6)과 같이 O_α에 핫세 다이어그램(Hasse diagram)을 적용하여 용어에 관한 계층구조를 이끌어낸다[6].

$$D = \{d_1, d_2, \cdots d_n\} \tag{1}$$

$$T = \{t_1, t_2, \cdots t_n\} \tag{2}$$

$$\tilde{R} = D \times T = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kn} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{matrix} \quad \begin{matrix} t_1 & t_2 & \cdots & t_n \end{matrix} \tag{3}$$

$$\tilde{B} = \tilde{R}^T \times \tilde{R}$$

$$= \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nk} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_k \end{matrix} \quad \begin{matrix} d_1 & d_2 & \cdots & d_n \end{matrix} \quad @ \quad \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{kn} & v_{k2} & \cdots & v_{kn} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{matrix} \quad \begin{matrix} t_1 & t_2 & \cdots & t_n \end{matrix}$$

$$= \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{matrix} \quad \begin{matrix} t_1 & t_2 & \cdots & t_n \end{matrix} \tag{4}$$

$$O_\alpha = \alpha\_\text{cut}(\tilde{B}, \alpha)$$

$$= \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{matrix} \quad \begin{matrix} t_1 & t_2 & \cdots & t_n \end{matrix} \tag{5}$$

$$H(O_\alpha) \tag{6}$$

2.2 퍼지관계곱

Bandler와 Kohout는 이진 관계곱 연산을 확장하여 퍼지정보검색모델의 시소러스 형성에 이론적 바탕이 되는 퍼지관계곱(fuzzy relational product) 연산을 제안하였다.

퍼지이론에서 '집합 $\tilde{A}$가 집합 $\tilde{B}$의 부분집합이다'는 수식 (7)과 같은 의미를 가진다.

$$\tilde{A} \subset \tilde{B} \;=\; \mu_{\tilde{A}}(x) \le \mu_{\tilde{B}}(x),\ {}^{\forall}x \in U \tag{7}$$

'퍼지집합 $\tilde{A}$가 퍼지집합 $\tilde{B}$의 부분집합이다'에 대한 정도는 수식 (8)(9)와 같이 정의할 수 있다.

$$\frac{1}{|U|}\Sigma(\mu_{\tilde{A}}(x) \rightarrow \mu_{\tilde{B}}(x)) \tag{8}$$

$$MIN(\mu_{\tilde{A}}(x) \rightarrow \mu_{\tilde{B}}(x),\ x \in U) \tag{9}$$

수식 (8)은 '퍼지집합 $\tilde{A}$가 퍼지집합 $\tilde{B}$에 내포된다'의 평균등급을 나타낸다. 수식 (9)는 '퍼지집합 $\tilde{A}$가 퍼지집합 $\tilde{B}$에 내포된다'의 최소등급을 나타낸다. 일반적으로 수식 (8)은 수식 (9)보다 융통성이 커서 정보검색 분야에 보다 적절하다.

집합 A, B, C와 퍼지관계 $\tilde{R}: A \times B$과 $\tilde{S}: B \times C$가 주어지고 $a_i \in A$, $c_i \in C$라 할 때, $\tilde{R}$과 $\tilde{S}$의 퍼지관계곱 $(\tilde{R}@\tilde{S})_{ik}$는 A의 원소 a_i와의 원소 c_k의 의미상 포함관계를 나타내는 것으로서 수식 (10)-(12)과 같이 세 가지 퍼지관계곱 연산 $\triangleleft$, $\triangleright$ 또는 $\square$를 이용하여 표현될 수 있다.

$$(R \triangleleft S)_{ik} \;=\; \frac{1}{|B|}\Sigma(R_{ij} \rightarrow S_{jk}) \tag{10}$$

$$(R \triangleright S)_{ik} \;=\; \frac{1}{|B|}\Sigma(R_{ij} \leftarrow S_{jk}) \tag{11}$$

$$(R \square S)_{ik} \;=\; \frac{1}{|B|}\Sigma(R_{ij} \leftrightarrow S_{jk}) \tag{12}$$

수식 (10)의 $\triangleleft$ 연산자는 퍼지삼각서브논리곱(fuzzy triangle sub-product)이라고 하고 수식 (9)는 a_i가 c_k에 포함되는 정도를 의미한다. 수식 (11)의 $\triangleright$ 연산자는 퍼지삼각슈퍼논리곱(fuzzy triangle super product)이라고 하고 수식 (11)은 a_i가 c_k를 포함하는 정도를 의미한다. 수식 (12)의 $\square$ 연산자는 퍼지사각논리곱이라고 하고 수식 (12)은 a_i와 c_k가 유사한 정도를 의미한다[6-12].

2.3 BK-퍼지정보검색모델과 검색요구

BK-퍼지정보검색모델은 사용자로부터 주어진 질의어를 해석하여 적합한 문서를 검색하는 검색요구 연산을 제공한다. BK-퍼지정보검색모델에서 검색요구는 다음과 같은 절차를 따른다. 우선 수식(13)(14)과 같이 문서집합 D와 용어집합 T을 정의하고 수식(15)과 같이 문서와 용어와의 퍼지관계 $\tilde{R}$이 존재한다고 가정한다. 이때 임의의 검색식 S가 주어지면 수식(16)과 같이 퍼지관계 $\tilde{R}$을 검색식 S에 적용하여 검색식 S에 대한 적합도를 표현하는 문서의 퍼지집합 D를 얻고 수식(17)과 같이 S에 α-cut을 적용, α-레벨 집합화하여 최종 k개의 결과 문서를 가지는 집합 D'_α를 구한다[5,7].

$$T = \{t_1, t_2, \cdots, t_n\} \tag{13}$$

$$D = \{d_1, d_2, \cdots, d_n\} \tag{14}$$

$$\tilde{R} = D \times T = \begin{matrix} \\ d_1 \\ d_2 \\ \vdots \\ d_k \end{matrix} \begin{matrix} \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kn} \end{bmatrix} \\ t_1 \quad t_2 \quad \cdots \quad t_n \end{matrix} \tag{15}$$

$$\begin{aligned} \tilde{D} &= S(\tilde{R}) \\ &= S\left(\begin{matrix} \\ d_1 \\ d_2 \\ \vdots \\ d_k \end{matrix} \begin{matrix} \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kn} \end{bmatrix} \\ t_1 \quad t_2 \quad \cdots \quad t_n \end{matrix} \right) \\ &= \{d_1/s_1,\ d_2/s_2,\ \cdots,\ d_k/s_k\} \end{aligned} \tag{16}$$

$$D'_\alpha = \{d'_1, d'_2, \ldots, d'_{k'}\} \tag{17}$$

2.4 BK-퍼지정보검색모델 검색요구 예

BK-퍼지정보검색모델의 검색요구를 적용례를 통하여 살펴보자. 우선 문서집합과 용어집합이 수식 (18)(19)과 같이 주어지고 문서와 용어의 퍼지관계가 수식 (20)과 같이 주어진다. 이때 사용자로부터 주어진 검색식이 수식 (21)과 같다고 할 때 검색식에 문서와 용어의 퍼지관계를 대입하여 수식 (22)과 같은 검색결과를 얻고 수식 (23)과 같이 검색결과에 $\alpha - \text{cut} = 0.9$를 적용하여 최종 검색결과 d_4를 얻는다.

$$D = \{d_1, d_2, d_3, d_4\} \tag{18}$$

$$T = \{\text{비행기}(t_1), \text{항공기}(t_2), \text{전투기}(t_3), F_{16}(t_4), F_{14}(t_5), \text{바다}(t_6)\} \tag{19}$$

$$\tilde{R} = D \times T = \begin{matrix} \\ d_1 \\ d_2 \\ d_3 \\ d_4 \end{matrix} \begin{matrix} \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 0.0 & 0.1 \\ 0.3 & 0.3 & 0.3 & 0.0 & 0.9 & 1.0 \\ 1.0 & 1.0 & 0.1 & 0.1 & 0.8 & 0.1 \\ 0.8 & 0.8 & 1.0 & 1.0 & 1.0 & 0.0 \end{bmatrix} \\ t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5 \quad t_6 \end{matrix} \tag{20}$$

$$S = \text{very(바다) and rather}(F14) \text{ and not}(F16) \tag{21}$$

$$\tilde{D}=S(\tilde{R})=\{d_1/0.0, d_2/0.95, d_3/0.01, d_4/0.0\} \quad (22)$$

$$D'_{0.9} = \{d_2\} \quad (23)$$

3. 개선된 BK-퍼지정보검색모델

Bandler와 Kohout가 제안한 BK-퍼지정보검색모델은 확장된 불리언 정보검색모델로서 자동 시소러스 구축, 검색결과의 퍼지화된 우선 순위 제공, 직접 관련 없는 제 3의 개체 유추 검색 등과 같은 장점을 제공한다. 그러나 BK-퍼지정보검색모델에는 높은 시간복잡도(time complexity)의 검색연산이 내재되어 있어 다양한 분야 적용을 어렵게 한다. 본 논문에서는 개선된 축소용어집합(reduced term set)을 이용하여 BK-퍼지정보검색모델 시간복잡도를 개선하는 개선된 BK-퍼지정보검색모델을 제안한다.

3.1 개선된 BK-퍼지정보검색모델과 시소러스

BK-퍼지정보검색모델은 자동으로 시소러스를 구축하는 기능을 가지고 있다. 개선된 BK-퍼지정보검색모델 역시 자동으로 시소러스를 구축하는 기능을 가지고 있는데 다음과 같은 4단계의 처리절차를 따른다. 첫 번째 단계에서는 수식 (24)-(26)과 같이 문서집합 D, 용어집합 T, 문서와 용어의 퍼지관계 $\tilde{R}$을 설정한다. 두 번째 단계에서는 수식 (27)과 같이 Ω연산을 이용하여 용어집합 T로부터 축소용어집합 T_r를 산출한다. 세 번째 단계에서는 수식 (28)과 같이 퍼지관계 R과 축소용어집합 T_r과의 투영(projection)에 의하여 문서집합과 축소용어집합과의 퍼지관계 $\tilde{R}_r$를 산출한다. 네 번째 단계에서는 수식 (29)과 같이 퍼지관계 $\tilde{R}$의 전치행렬 $\tilde{R}^T$와 퍼지관계 $\tilde{R}_r$에 퍼지관계곱 연산 @(◁, ▷또는 □)을 적용하여 결과 퍼지관계 $\tilde{B}_r$를 산출한다.

$$D = \{d_1, d_2, \cdots, d_n\} \quad (24)$$

$$T = \{t_1, t_2, \cdots, t_n\} \quad (25)$$

$$\tilde{R} = D \times T = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{21} & v_{22} & \cdots & v_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kn} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{matrix} \quad (26)$$
$$t_1 \quad t_2 \quad \cdots \quad t_n$$

$$T_r = \Omega(T, n', \chi, \gamma\}$$
$$= \{r_1, r_2, \cdots, r_{n_r}\} \quad (27)$$

$$\tilde{R}_r = D \times T_r$$
$$= \text{projection}(R, T_r)$$
$$= \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n_r} \\ v_{21} & v_{22} & \cdots & v_{2n_r} \\ \vdots & \vdots & \ddots & \vdots \\ v_{k1} & v_{k2} & \cdots & v_{kn_r} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{matrix} \quad (28)$$
$$r_1 \quad r_2 \quad \cdots \quad r_{n_r}$$

$$\tilde{B}_r = \tilde{R}^T \times \tilde{R}_r$$
$$= \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nk} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{matrix} @ \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n_r} \\ v_{21} & v_{22} & \cdots & v_{2n_r} \\ \vdots & \vdots & \ddots & \vdots \\ v_{kn} & v_{k2} & \cdots & v_{kn_r} \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{matrix}$$
$$d_1 \quad d_2 \quad \cdots \quad d_k \qquad r_1 \quad r_2 \quad \cdots \quad r_{n_r}$$
$$= \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n_r} \\ v_{21} & v_{22} & \cdots & v_{2n_r} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn_r} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{matrix} \quad (29)$$
$$r_1 \quad r_2 \quad \cdots \quad r_{n_r}$$

여기서 유의할 점은 개선된 BK-퍼지정보검색모델은 생성된 시소러스를 계층구조화하지 않는다는 것이다. 개선된 BK-퍼지정보검색모델에서는 퍼지관계 행렬 내에 존재하는 관계성을 그대로 이용한다. 따라서 퍼지관계행렬을 계층구조화된 트리로 표현할 필요가 없다.

3.2 개선된 BK-퍼지정보검색모델과 검색

개선된 BK-퍼지정보검색모델은 사용자로부터 주어진 질의어를 해석하고 시소러스를 이용하여 확장하며 적합한 문서를 검색하는 검색요구 연산을 제공한다. 개선된 BK-퍼지정보검색모델은 사용자로부터 주어진 검색요구를 처리하기 위해 우선 수식 (26)-(30)과 같이 문서집합 D, 용어집합 T, 축소용어집합 T_r, 문서집합과 축소용어집합과의 퍼지관계 $\tilde{R}_r$, 시소러스 $\tilde{B}_r$가 정의되어야 한다. 이후 사용자로부터 질의어 S가 입력되면 수식 (30)과 같이 질의어를 해석하는 연산 Φ를 이용하여 $\tilde{Q}$를 산출한다. 수식 (31)에서는 처리된 질의어 $\tilde{Q}$와 시소러스 $\tilde{B}_r$을 합성하여 확장된 질의어 $\tilde{Q}_r$을 산출한다. 수식 (32)에서는 퍼지관계곱을 이용하여 $\tilde{R}_r$과 확장된 질의어 $\tilde{Q}_r$을 합성하여 검색결과 $\tilde{O}$를

얻고 수식 (33)에서는 $\tilde{O}$에 α-cut을 적용하여 최종 문서 검색결과 $\tilde{Q}_\alpha$를 구한다. 이때 연산 ∘, @는 각각 퍼지관계합성연산, 퍼지관계곱연산을 의미한다.

$$\tilde{Q}=Q\times T=\Phi(S)=\begin{bmatrix} v_1 & v_2 & \cdots & v_n \\ t_1 & t_2 & \cdots & t_n \end{bmatrix} \tag{30}$$

$$\tilde{Q}_r = \tilde{Q}^T \circ \tilde{B}_r \tag{31}$$

$$= \begin{bmatrix} v_1 & v_2 & \cdots & v_n \\ t_1 & t_2 & \cdots & t_n \end{bmatrix} \circ \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n_r} \\ v_{21} & v_{22} & \cdots & v_{2n_r} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn_r} \\ r_1 & r_2 & \cdots & r_{n_r} \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{matrix}$$

$$\begin{bmatrix} v_1 & v_2 & \cdots & v_{n_r} \\ t_1 & t_2 & \cdots & t_{n_r} \end{bmatrix} \tag{31}$$

$$\tilde{O} = \tilde{Q} \mathbin{@} \tilde{R}_r^T$$

$$= \begin{bmatrix} v_1 & v_2 & \cdots & v_{n_r} \\ t_1 & t_2 & \cdots & t_{n_r} \end{bmatrix} @ \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{n_rk} \\ d_1 & d_2 & \cdots & d_k \end{bmatrix} \begin{matrix} r_1 \\ r_2 \\ \vdots \\ r_{n_r} \end{matrix}$$

$$= \{d_1/v_1,\ d_2/v_2,\ \cdots,\ d_k/v_k\} \tag{32}$$

$$\tilde{O}'_\alpha = \alpha\_\text{cut}(\tilde{O},\ \text{level})$$
$$= \{d_1,\ d_2,\ \cdots,\ d_{k\alpha}\} \tag{33}$$

3.3 개선된 BK-퍼지정보검색모델에서의 검색 예

개선된 BK-퍼지정보검색모델에서의 검색요구를 실제 예를 통하여 살펴보자. 우선 문서집합, 용어집합, 문서집합과 용어집합과의 퍼지관계가 수식 (34)-(36)과 같이 주어지고 축소용어집합과 시소러스, 문서집합과 축소용어집합과의 퍼지관계가 수식 (37)-(39)과 같이 주어진다. 이때 수식 (40)과 같이 사용자로부터 질의어가 주어지면 이를 수식 (41)과 같이 용어의 퍼지집합으로 변환하고 수식 (42)과 같이 시소러스를 이용하여 확장하는데 본 예에서는 Max-Min 합성[16]을 이용한다. 수식 (43)은 퍼지삼각서브논리곱을 이용하여 질의어와 문서와의 관련성을 문서의 퍼지집합으로 나타내며 이에 α-cut을 적용하여 수식 (44)과 같은 최종 검색결과를 산출한다.

$$D = \{d_1,\ d_2,\ d_3,\ d_4\} \tag{34}$$

T={비행기(t_1), 항공기(t_2), 전투기(t_3), $F16$(t_4), $F14$(t_5), 바다(t_6)} (35)

$$\tilde{R} = D\times T = \begin{bmatrix} 1.0 & 1.0 & 1.0 & 1.0 & 0 & 0.1 \\ 0.3 & 0.3 & 0.3 & 0 & 0.9 & 1.0 \\ 1.0 & 1.0 & 0.1 & 0.1 & 0.8 & 0.1 \\ 0.8 & 0.8 & 1.0 & 1.0 & 1.0 & 0 \\ t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{matrix} \tag{36}$$

$$T_r = \{F16(r_1),\ F14(r_2)\} \tag{37}$$

$$\tilde{R}_r=D\times T_r=projection(\tilde{R},\ T_r)=\begin{bmatrix} 1.0 & 0 \\ 0 & 0.9 \\ 0.1 & 0.8 \\ 1.0 & 1.0 \\ r_1 & r_2 \end{bmatrix} \begin{matrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{matrix} \tag{38}$$

$$\tilde{B}_r=T\times T_r=\begin{bmatrix} 0.7 & 0.7 \\ 0.7 & 0.7 \\ 0.92 & 0.75 \\ 1.0 & 0.75 \\ 0.6 & 1.0 \\ 0.75 & 0.95 \\ r_1 & r_2 \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{matrix} \tag{39}$$

S = very(바다) and rather($F14$) and not($F16$) (40)

$$\tilde{Q} = \Phi(S)$$
$$= \{t_1/0,\ t_2/0,\ t_3/0,\ t_4/0,\ t_5/0.64,\ t_6/0.89\} \tag{41}$$

$$\tilde{Q} = \tilde{Q}^T \circ \tilde{B}_r$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0.2 & 0.64 & 0.89 \\ t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \end{bmatrix} \circ \begin{bmatrix} 0.7 & 0.7 \\ 0.7 & 0.7 \\ 0.92 & 0.75 \\ 1.0 & 0.75 \\ 0.6 & 1.0 \\ 0.75 & 0.95 \\ r_1 & r_2 \end{bmatrix} \begin{matrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \end{matrix} \tag{42}$$

$$= \{r_1/0.75,\ r_2/0.89\}$$

$$\tilde{O} = Q_r \circ R_r^T$$

$$= \begin{bmatrix} 0.75 & 0.89 \\ r_1 & r_2 \end{bmatrix} \triangleleft \begin{bmatrix} 1.0 & 0 & 0.1 & 1.0 \\ 0 & 0.9 & 0.8 & 1.0 \\ d_1 & d_2 & d_3 & d_4 \end{bmatrix} \begin{matrix} r_1 \\ r_2 \end{matrix} \tag{43}$$

$$= \{d_1/0.55,\ d_2/0.63,\ d_3/0.65,\ d_4/1.0\}$$

$$O_{0.9} = \{d_4\} \tag{44}$$

3.4 축소용어집합

개선된 BK-퍼지정보검색모델에서는 BK-퍼지정보검색모델의 시간복잡도를 개선하기 위하여 축소용어집합을 이용한다. 축소용어집합은 용어집합의 부분집합으로서 상수개의 원소로 구성된다. 축소용어집합 추출은 n개의 원소로 구성된 임의의 용어집합 T에서 n_r개의 원소로 구성된 축소용어집합(reduced term set) T_r을 선별하는 것이다. 이때 축소용어집합을 추출하는 연산 Ω는 수식 (45)와 같은 형태로 정의될 수 있다.

$$T_r = \Omega(T, n_r) \tag{45}$$

수식 (45)에서 이때 상수 n_r를 산정하는 것은 매우 중요하다. 왜냐하면 n_r은 검색시스템의 처리능력 및 검색결과의 신뢰도와 직접 관련되어 있기 때문이다. 적절한 n_r의 산정은 검색시스템의 처리능력 χ과 요구되는 검색결과의 최소 신뢰도 γ를 고려하여야 한다. 적정한 n_r은 χ와 γ에 비례하므로 적절한 상수 k를 이용하여 수식 (46)을 정의할 수 있다.

$$n_r = k\gamma\chi \tag{46}$$

개선된 BK-퍼지정보검색모델에서 검색신뢰도는 축소용어집합의 적절성 정도와 연관이 있다. 주어진 축소용어집합의 적합도 s_r가 주어지면 수식 (46)은 수식 (47)과 같이 변환할 수 있다. 시스템 성능 χ와 검색신뢰도 γ이 정해질 때 축소용어집합의 적합도가 높아질수록 보다 작은 축소용어집합을 구성할 수 있고 동일한 환경에서 보다 나은 검색 결과를 이끌어낼 수 있다.

$$n_r = \frac{k'}{s_r}\gamma\chi \tag{47}$$

축소용어집합 추출 방법은 크게 무작위 추출, 규칙에 의한 자동 추출 혹은 인간에 의한 수동 추출과 같은 3가지로 분류할 수 있다. 무작위 추출은 난수 생성기를 이용하여 임의의 축소용어집합을 생성하는 방법이고, 규칙에 의한 자동 추출은 규칙을 정하고 이를 이용하여 축소용어집합을 형성하는 방법이다. 인간에 의한 수동 추출은 인간이 직접 참여하여 축소용어집합을 직접 추출하는 방법이다. 인간에 의한 수동 추출은 높은 적합도를 이끌어 낼 수 있으나 본 논문에서는 무작위 추출과 규칙에 의한 자동 추출을 비교 분석한다.

3.5 적합 빈도 추출법

규칙에 의한 자동 추출법은 미리 규칙을 정해놓고 이를 이용하여 용어집합으로부터 축소용어집합를 선정하는 방법이다. 본 논문에서는 적합 빈도 추출을 제안한다. 이는 용어와 문서와의 관계에 기초하여 가장 많은 문서와 관련성을 가지는 용어를 추출하는 방법으로 수식 (48)과 같다. 적합 빈도 추출 함수 Ω는 모든 용어에 대하여 문서집합과 용어집합의 퍼지관계 $\tilde{R}$로부터 임계값 ρ보다 크거나 같은 값을 가지는 문서들의 빈도를 추출하여 그 빈도가 높은 용어를 축소용어로 선별한다.

$$\begin{aligned} T_r &= \Omega(\tilde{R}, \rho, n_r) \\ &= \mathrm{Sort}(\mathrm{Freq}(\alpha\_\mathrm{cut}(\tilde{R}, \rho)), n_r) \end{aligned} \tag{48}$$

4. 두 모델 간 시간복잡도 비교

BK-퍼지정보검색모델은 내재되어 있는 검색연산의 높은 시간복잡도 때문에 적용분야에 큰 제약을 받는다. 문서의 개수가 n_D이고 용어의 개수를 n_T일 때, BK-퍼지정보검색모델의 퍼지검색요구의 시간복잡도는 $\Theta(n_D \times n_T)$이고 시소러스 구축의 시간복잡도는 $\Theta(n_D \times n_T^2)$이다.

개선된 퍼지정보검색모델은 BK-퍼지정보검색모델에 비해 낮은 시간복잡도를 갖는다. 문서의 개수가 n_D, 용어의 개수를 n_T이고 축소용어가 원소 개수가 c_{n_r}이라고 할 때 개선된 BK-퍼지정보검색모델에서의 시간복잡도를 산출하면, 문헌 검색의 시간복잡도는 $n_D \times c_{n_r}$에 비례하고 시소러스를 구축의 시간복잡도는 $n_D \times n^T \times c_{n_r}$에 비례한다. 이때 c_{n_r}가 검색시스템과 검색결과의 신뢰도를 고려하여 산출된 상수임을 고려하면 문헌검색의 시간복잡도는 $\Theta(n_D)$이고 시소러스 구축의 시간복잡도는 $\Theta(n_D \times n_T)$임을 알 수 있다. 따라서 개선된 퍼지정복검색모델은 문헌 검색의 시간복잡도를 $\Theta(n_D \times n_T)$에서 $\Theta(n_D)$로 낮추고 시소러스 형성 시간복잡도를 $\Theta(n_D \times n_T^2)$에서 $\Theta(n_D \times n_T)$로 낮춘다.

5. 개선된 BK-퍼지정보검색모델의 신뢰도 분석

문서검색시스템에서 검색결과의 신뢰도란 사용자의 질의어에 대한 검색된 문서의 적합도를 의미한다. 그러나 검색 문서의 적합도는 주관적인 것이고 비교 대상 역시 불확실하여 이를 정량적인 값으로 산정하는 것은 매우 어려운 일이다. 본 논문에서 제안하는 개선된 BK-퍼지정보검색모델의 검색 신뢰도 분석은 축소용어집합을 적용하지 않는 BK-퍼지정보검색모델의 검색결과와의 유사도 산출함으로써 산정한다.

5.1 검색결과 신뢰도 측정

개선된 BK-퍼지정보검색모델의 검색 신뢰도는 BK-퍼지정보검색모델의 검색결과에 대한 유사도로써 산출하며 그 대상은 두 모델에 의한 검색결과에서 상위에 랭크되는 30개의 문서를 추출하여 구성한 두 개의 문서집합으로 한다.

본 논문에서는 두 개의 문서집합의 유사도를 산출하기 위해 Dice 상관계수(Dice coefficient)를 이용한다. Dice 상관계수는 벡터의 거리 및 유사도를 측정하는 방법 중의 하나로서 간소화와 정규화 기능을 가지고 있어 정보검색분야에 자주 이용된다[17]. 두 개의 벡터 D_i, D_j가 주어질 때 Dice 상관계수 $S_{D_iD_j}$는 수식 (46)과 같다.

$$S_{D_iD_j} = \frac{2\sum_k weight_{ik} \cdot weight_{jk}}{\sum_k weight_{ik}^2 + \sum_k weight_{jk}^2} \tag{46}$$

이때 문서집합은 이진 용어 가중치로 구성된 벡터로 볼 수 있다. D_i와 D_j는 두 개의 집합 A와 B로 표현가능하고 수식 (46)은 수식 (47)으로 변환할 수 있다.

$$S_{AB} = \frac{2|A \cap B|}{|A| + |B|} \tag{47}$$

BK-퍼지정보검색모델의 상위 30개의 검색결과를 집합 α, 개선된 BK-퍼지정보검색모델의 상위 30개의 검색결과를 집합 β라 할 때 검색 결과 α와 β의 유사도는 수식 (48)을 이용하여 산출 가능하다.

$$S_{\alpha\beta} = \frac{2|(\alpha \cap \beta)|}{30+30} = \frac{|(\alpha \cap \beta)|}{30} \tag{48}$$

이때, BK-FIRM에 의한 검색결과가 B이고, 임의의 축소용어집합 추출 방법에 의한 A-FIRM의 검색결과가 Φ라 할 때, Φ의 신뢰도 $R_{B\Phi}$는 수식 (49)과 같이 Φ와 B의 유사도 산출 $S_{B\Phi}$에 의한다.

$$R_{B\Phi} = S_{B\Phi} \tag{49}$$

5.2 축소용어집합 생성 방법에 따른 신뢰도 분석

3.4절에서는 축소용어와 축소용어집합의 생성방법에 관하여 살펴보았다. 본 절에서는 난수를 이용하여 용어집합 중 일부를 축소용어집합으로 산출하는 무작위 추출과 비교적 많은 문서들과 높은 관련성을 가지는 용어를 선별하는 적합 빈도 추출에 의해 산출된 검색결과의 신뢰도를 비교한다.

5.1.2 무작위 추출법과 적합 빈도 추출법의 비교

난수를 이용하여 축소용어집합을 추출하는 무작위 추출은 확률적으로 용어집합의 특성을 잘 나타내지만 문서집합과의 연관성을 전혀 고려하지 않아 낮은 검

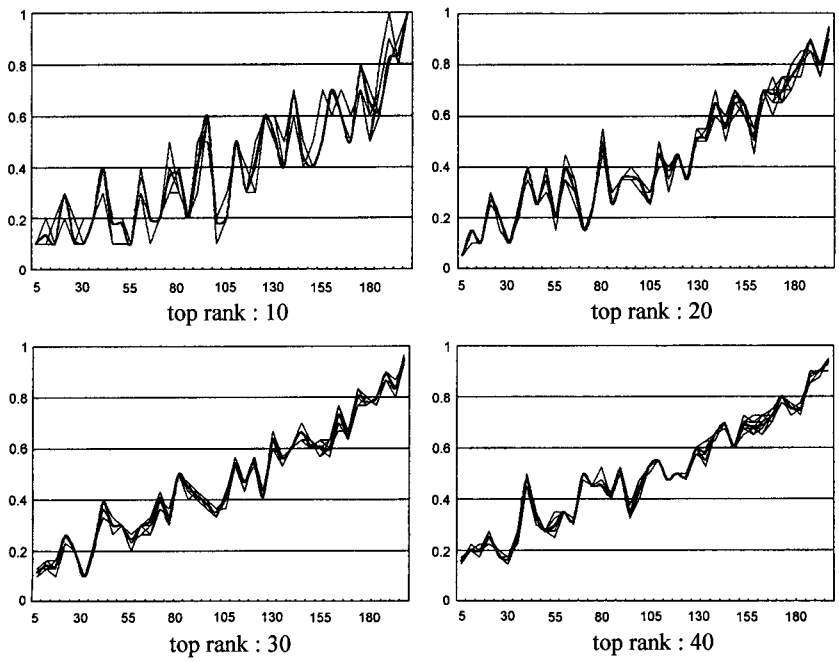


그림 1. 무작위 추출에 의한 검색 신뢰도

색 신뢰도를 가진다. 주어진 임계값을 상회하는 용어들을 추출 후 그 빈도가 높은 용어를 축소용어로 선택하는 적합 빈도 추출은 용어와 문서집합간의 연관성을 인정하고 이에 근거하여 축소용어를 추출하므로 보다 나은 검색 신뢰도를 보인다.

그림 1은 무작위 추출에 의한 검색 결과의 신뢰도

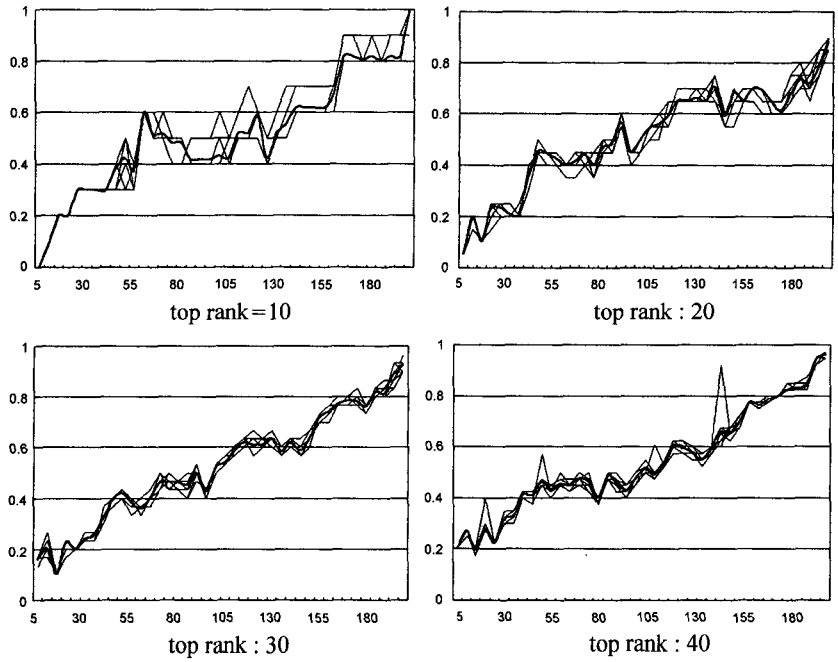


그림 2. 적합 용어 빈도 추출법의 의한 검색 신뢰도

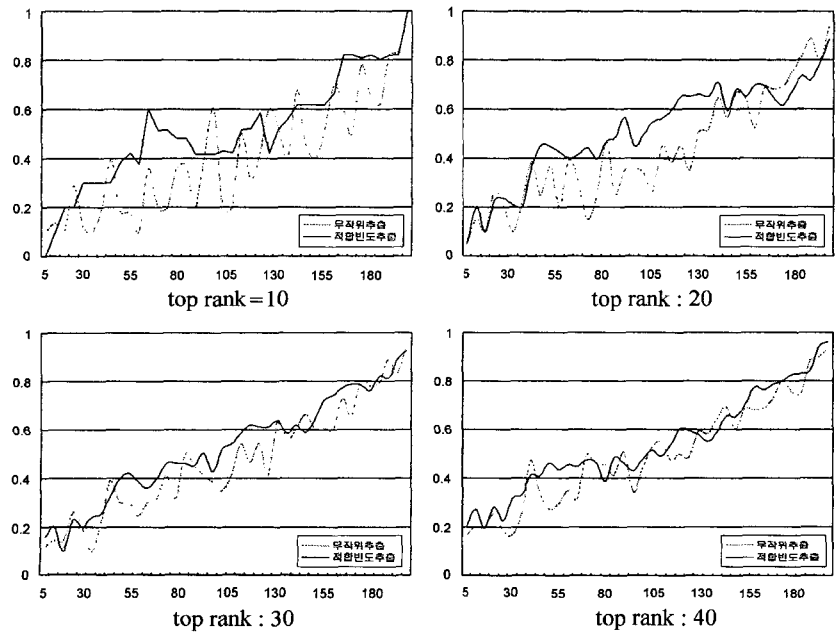


그림 3. 무작위 추출법과 적합 용어 빈도 추출법의 신뢰도 비교

이다. 각 축소용어집합에 대하여 50회의 질의어를 입력한 실험 결과이다. 그림 1에서 진한 곡선은 50개의 질의어에 대한 평균 신뢰도의 변화 유형이다.

그림 2은 적합 빈도 추출에 의한 검색 결과의 신뢰도이다. 각 축소용어집합에 대하여 50회의 질의어를 입력하였으며, 진한 곡선은 50개의 질의어에 대한 평균 신뢰도의 변화 유형이다.

그림 3는 무작위 추출과 적합 빈도 추출에 의한 검색결과의 평균 신뢰도의 변화 유형을 비교하여 보여준다. 그림 3에서 적합 빈도 추출에 의한 결과가 무작위 추출에 의한 결과보다 우수하다. 따라서 축소용어집합 생성시 특정 처리를 가하여 보다 나은 축소용어집합의 선택하는 가능하다.

6. 결론 및 향후과제

퍼지정보검색은 의료진단(medical diagnosis), 정보검색, 수기분류(handwriting classification) 등 수많은 문제 해결에 응용되고 있다. 특히 의료진단 분야에서는 진단 자료를 이용한 환자 관리 처리(diagnostic data and patient management processes), 의학적 증상 증후 비교(medical sign and symptom comparison)와 같은 부분에 퍼지 정보 검색 기법이 이용되고 있다[6]. 뿐만 아니라 퍼지정보검색은 용어, 질의, 문헌과 같은 개체들의 관계를 퍼지관계행렬로 표현하고 개체연결과 같은 1차원적 단순검색 뿐만 아니라 개체간의 연관성에 근거하여 직접 관련이 없는 제3의 개체도 검색 가능하므로 단순 매칭(matching)으로 해결하기 힘든 화상, 동영상, 소프트웨어 재사용 등과 같은 분야에 특히 유용한 기법이다.

기존의 검색이론과는 달리 BK-퍼지정보검색모델만이 가지고 있는 특성에도 불구하고 실제 대용량의 문서나 용어를 다루는 검색시스템 적용이 어려웠던 것은 BK-퍼지정보검색모델 자체 검색연산의 높은 시간복잡도 때문이었다. 본 연구에서는 축소용어집합을 이용하여 BK-퍼지정보검색모델의 시간복잡도를 낮춘 개선된 퍼지정보검색모델을 제안한다.

축소용어집합은 개선된 퍼지정보검색모델의 핵심요소이다. 축소용어집합의 크기는 검색모델의 시간복잡도와 검색 결과의 신뢰도에 직접 연관되고, 검색 결과의 신뢰도는 축소용어집합추출방법과도 연관된다. 적절한 축소용어추출방법은 제한된 처리시간이 주어질 때 검색결과의 신뢰도를 높일 수 있다. 본 논문에서는 축소용어집합추출방법을 크게 무작위 추출, 규칙에 의한 자동 추출, 인간에 의한 수동 추출로 구분하고 무작위 추출 방법과 규칙에 의한 자동 추출 방법을 실험 비교하였다.

결론적으로, 개선된 BK-퍼지정보검색모델은 BK-퍼지정보검색모델 보다 향상된 시간복잡도를 가진다. 축소용어집합의 크기가 작아질수록 처리시간은 줄어들지만 검색 결과 신뢰도는 감소함을 알 수 있다. 따라서 A-FIRM에서 축소용어집합은 시스템 처리시간과 신뢰도 간 상충점(trade-off)을 제공한다. 동일한 크기의 축소용어집합이 이용되더라도 선정되는 축소용어집합의 적합성에 따라 검색결과의 신뢰도가 영향을 받는다.

참고문헌

[1] Rijsbergen, C. J. van, *Information Retrieval,* 2nd edition, butterworths, 1979.

[2] Fox, E. A., and Sharat, S., "A Comparison of Two Methods for Soft Boolean Interpretation in Information Retrieval," Technical Report TR-86-1, Virginia Tech, Department of Computer Science, 1986.

[3] Paice, C. P., "Soft Evaluation of Boolean Search Queryies in Information Retrieval Systems," Information Technology, Res. Dev. Application, 1984.

[4] Fox, E. A., "Extending the Boolean and Vector Space Models of Information Retrieval with P-Norm Queryies and Multiple Concept Types," Cornell University, 1983.

[5] Bollmann, P., and Konrad, E., "Fuzzy Document Retrieval," in: Trappl R., Klir G. J. and Ricciardi L., eds., Progress in Cybernetics and Systems Research, vol. 3 (Hemisphere Publ. Comp., and John Wiley, Washington and New York) 355-363, 1976.

[6] Kohout, L. J., and Harris, M., "Computer Representation of Fuzzy and Crisp Relations by Means of Threaded Trees Using Foresets and Aftersets," *Journal of Fuzzy Logic and Intelligent Systems,* vol. 3, no.1, 1993.

[7] Kohout, L. J., Keravnou E. and Bandler W., "Automatic Documentary Information Retrieval by means of Fuzzy Relational Products," In Gaines, B. R., Zadeh L. A. and Zimmermann, H. J., editors Fuzzy Sets in Decision Analysis, pages 308-404, North-Holland, Amsterdam, 1984.

[8] Bandler, W., and Kohout L. J., "Fuzzy Power Sets and Fuzzy Implication Operator," *Fuzzy Sets and Systems* 4, 13-30, 1980

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

[10] Keravnou, E. "Fuzzy Relational Products in Information Retrieval Systems," B. Tech. Dissertation, Dept. of Computer Science, Brunel University, 1982.

[11] Kohout, L. K., Bandler, W., "Fuzzy Relational

Products as a Tool for Analysis and Synthesis of the Behaviour of Complex Natural and Artificial Systems," in: Wang S. K. and Chang P. P. eds., Fuzzy Sets: Theory and Application to Policy Analysis and Information Systems, Plenum Press, New York, 341-367, 1980.

[12] Kim, Yong-Gi and Kohout, L. J., "Use of Fuzzy Relational Products and Algorithms for generating Control strategies in resolution based Automated Reasoning," Proceedings of the fourth International Fuzzy System Association (IFSA) world congress, (Brussels, Belgium), July 7-12, 1991.

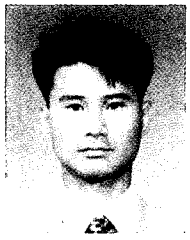
[13] Kim, Yong-Gi and Kohout, L. J., "Comparison of Fuzzy Implication Operators by means of Weighting Strategy in on Applied Computing (SAC'92)," Kansas City, March 1-3, 1992.

[14] Keravnou, E., "System for Experimental Verification of Deviance of Fuzzy Connectives in Information Retrieval Application," Second World Conference on Mathematics at the Service of Man. Topic 7, Measuring "Deviance in Non-Classical Logics and Modelling, Las Palmas (Canary Islands), June-July, 1982.

[15] Bandler, W., and Kohout, J. "Semantics of Implication operators and fuzzy relational products," Intl. Journal of Man-Machine Studies, 1980.

[16] Zimmermann, H. J. Fuzzy Set Theory and Its Application, Kluwer Academic Publishers, 1991.

[17] Salton, G., Automatic Text Processing, Addison-Wesley, 1989.



김 창 민 (Chang-Min Kim)

1997년 : 경상대학교 컴퓨터과학과 (이학사)
1999년 : 경상대학교 컴퓨터과학과 (공학석사)
1999년~현재 : 경상대학교 컴퓨터과학과 박사과정
관심분야 : 인공지능, 지식기반시스템, 자율무인잠수정, 지능항해시스템

김 용 기 (Yong-Gi Kim)

1978년 : 서울대학교 공과대학(공학사)
1987년 : University of Montana (전산학석사)
1992년 : Florida State University (전산학박사)
1982년~1984 : KIST시스템공학연구소 연구원
1992년~현재 : 경상대학교 컴퓨터과학과 교수
관심분야 : 인공지능, 지식기반시스템, 자율무인잠수정, 지능항해시스템