

높은 검색 효과를 위한 다른 특성을 갖는 가중치 기법의 이용

Using Different Properties of Weighting Schemes for High Retrieval Effectiveness

이준호

한국과학기술연구원 연구개발정보센터

Joon Ho Lee

Korean Research and Development Information Center
Korea Institute of Science and Technology

요약

질의 또는 문서에 대한 상이한 표현 방법 또는 상이한 검색 기법은 서로 다른 집합의 문서들을 검색함이 알려져 왔다. 최근 이러한 특성을 이용하여 다양한 표현 방법 또는 검색 기법을 결합함으로써 보다 높은 검색 효과를 얻을 수 있음이 입증되었다. 본 논문에서는 질의와 문서에 대한 하나의 표현과 하나의 검색 기법하에서 서로 다른 특성을 갖는 가중치 기법을 결합함으로써 보다 높은 검색 효과를 얻을 수 있음을 기술한다. 문서의 형태를 분류하고 가중치 기법의 특성을 기술한 후, 이를 기반으로 하여 서로 다른 특성을 갖는 가중치 기법은 서로 다른 형태의 문서를 검색함을 설명한다. 또한 실험을 통하여 서로 다른 특성을 갖는 가중치 기법을 결합함으로써 보다 높은 검색 효과를 얻을 수 있음을 입증한다.

1. 서론

정보 검색 시스템의 중요한 역할 중의 하나는 문서가 질의를 만족하는 정도를 나타내는 문서값(document value)을 보다 정확히 계산함으로써, 높은 검색 효과(retrieval effectiveness) — 높은 재현율(recall)와 정확률(precision)를 제공하는 것이다. 최근 보다 높은 검색 효과를 얻기 위하여 다양한 검색 실험의 결과들을 결합하는 연구가 진행되어 왔으며, 이와 관련된 연구들은 “데이터 퓨전(data fusion)”이라는 용어에 의해 지칭되어 왔다.

McGill, Koll & Norreault [1]는 동일한 정보 요구에 대한 질의 표현 방법에 따라, 즉 서로 다른 사용자가 문서를 검색하거나 또는 통계어 사용 여부에 따라 상이한 문서들이 검색됨을 발견하였다. Katzer, et al. [2]는 질의 표현 방법 대신에 문서 표현 방법이 검색 효과에 미치는 영향을 조사하였으며, 서로 다른 문서 표현을 사용한 검색들이 매우 상이한 문서들을 검색함을 발견하였다. Saracevic & Kantor [3]는 여러 사용자에게 동일한 정보 검색 요구에 대하여 부울 질의를 생성하도록 요청하고, 생성된 다양한 질의를 검색에 이용하였다. 그결과 서로 다른 질의 표현이 서로 다른 문서들을 검색함을 재확인하였다. 그러나 또한 어떠한 문서가 보다 많은 질의에 의해 검색될수록 적합 문헌으로 판단될 가능성이 증가함을 추가로 발견하였다.

Turtle & Croft [4]는 서로 다른 문서 또는 질의 표현을 확률적 관점에서 결합할 수 있는 추론 네트워크 검색 모델(inference network-based retrieval model)을 개발하였다. 이 모델은 서로 다른 표현들을 문서가 질의를 만족할 수 있는 확률을 계산하기 위한 증거로서 사용한다. Turtle & Croft는 추론 네트워크 검색 모델을 기반으로 하는 INQUERY 시스템을 개발하여, 다중 증거의 사용이 검색 효과를 증가시킴을

입증하였다. Fox & Shaw [5]는 다양한 검색 실험들을 결합하는 방법에 대한 연구를 수행하였으며, 각각의 검색 실험에 비하여 결합된 검색 실험으로부터 보다 높은 검색 효과를 얻었다.

위에서 언급된 연구 결과들은 다양한 검색 수행을 결합함으로써 보다 높은 검색 효과를 얻을 수 있음을 보여준다. 그러나 기존의 연구들은 단지 서로 다른 표현 방법 또는 검색 기법만을 고려하였다. 본 논문에서는 단일 질의 표현, 단일 문서 표현, 그리고 단일 검색 기법의 사용만으로 서로 다른 집합의 문서들을 검색할 수 있는 방법을 제안한다. 문서가 다루는 주제들의 수와 문서에 포함된 색인어들의 수에 따라 문서의 형태를 분류하고, 코사인 정규화와 최대 정규화와 같은 가중치 기법의 특성들을 기술함으로써, 이러한 특성들이 검색되는 문서들의 형태에 어떠한 영향을 미치는가를 설명한다. 또한 서로 다른 특성을 갖는 가중치 기법이 서로 다른 집합의 문서들을 검색함을 설명한다. 실험 결과, 코사인 정규화를 사용한 검색 실험과 코사인 정규화를 사용하지 않는 검색 실험을 결합함으로써 검색 효과의 향상을 얻을 수 있었다.

2. SMART 시스템

SMART 시스템은 벡터 공간 모델을 기반으로 하며, 문서와 질의 모두 다음과 같은 벡터 형태로 표현된다 [6].

$$d_i = (w_{i1}, w_{i2}, \dots, w_{in})$$

여기에서 d_i 는 문서 또는 질의를 표현하고, w_{ik} 는 문서 d_i 에서 색인어 t_k 의 가중치이다. SMART 시스템에서 이러한 벡터들은 다음과 같은 텍스트 변환에 의해 생성된다.

1. 텍스트로부터 단어들을 인식한다.
2. 색인어로서 가치가 없는 불용어들을 제거한다.

3. 접미사들을 제거함으로써 어근들을 추출한다.
4. 각각의 어근들에 가중치를 부여한다.

문서 또는 질의에 대한 벡터들이 형성되면, 이후의 검색 과정은 벡터들의 연산에 의해 이루어진다. 문서 d 가 $(w_{d1}, w_{d2}, \dots, w_{dn})$ 로 표현되고, 질의 q 가 $(w_{q1}, w_{q2}, \dots, w_{qn})$ 로 표현되었을 때, 문서 d 와 질의 q 사이의 유사도를 의미한 문서 d 의 문서값은 다음과 같이 두 벡터들의 내적으로 계산된다.

$$Sim(d, q) = \sum_{i=1}^n (w_{di} \times w_{qi})$$

문서값은 색인어들의 가중치에 의해 결정되기 때문에, 가중치 부여 기법은 검색 효과에 영향을 미치는 중요한 요소이다.

정보 검색에 관한 많은 연구들은 색인어에 가중치를 부여하기 위하여 출현 빈도(term frequency), 장서 빈도(collection frequency), 정규화(normalization)의 세 가지 요소를 고려한다 [7]. 출현 빈도는 문서내에서 자주 출현하는 색인어에 보다 높은 가중치를 부여한다. 장서 빈도는 전체 문서들 중에서 작은 수의 문서에 출현하는 색인어에 보다 높은 가중치를 부여한다. 그리고 정규화는 작은 크기의 문서들이 문서값 계산에 있어서 불공평하게 취급되는 것을 피하도록 한다. (표 1)은 각각의 구성 요소에 대한 잘 알려진 공식들을 보여준다.

3. 가중치 기법들에 대한 분석

3.1 문서 형태 분류

출현빈도벡터길이(tf-vector length)는 출현 빈도의 합으로 정의된다. t_i 가 색인어의 출현 빈도이고, n 이 벡터를 구성하는 색인어들의 수일 때, 출현빈도벡터길이는 $\sum_{i=1}^n t_i$ 이다. 예를 들면, 문서 d_1 이 색인어와 가중치의 쌍으로 다음과 표현되었다고 가정하자.

$$d_1 = \{(t_1, 1), (t_2, 2), (t_3, 3), (t_4, 4), (t_5, 5)\}$$

이때 문서 d_1 의 출현빈도벡터길이는 $15 (= 1 + 2 + 3 + 4 + 5)$ 이다. 많은 현실적인 문서집합에서 문서들은 매우 다양한 출현빈도벡터길이를 갖는다. 문서들은 출현빈도벡터길이에 따라 다음과 같이 3가지 형태로 분류된다 [8].

- 짧은 출현빈도벡터길이 (short tf-vector length)
- 중간 출현빈도벡터길이 (median tf-vector length)
- 긴 출현빈도벡터길이 (long tf-vector length)

한편, 많은 경우에 긴 문서뿐만 아니라 짧은 문서들도 단일 주제가 아닌 다중 주제를 기술하는 것으로 알려져 왔다 [9]. 따라서 문서들을 문서가 다루는 주제들의 수에 다음과 같이 2가지 형태로 분류한다.

- 단일 주제 (single topic)
- 다중 주제 (multiple topic)

3.2 가중치 기법의 특성

출현빈도벡터길이를 정규화하지 않는 가중치 기법이 사용되었을 때, 긴 출현빈도벡터길이의 문서는 짧은 출현빈도벡터길이를 갖는 문서에 비하여 높은 검색 가능성을 갖는다. 예를 들면, 문서 d_2, d_3 가 다음과 같이 표현되어 있다고 가정하자.

$$d_2 = \{(t_1, 1), (t_2, 1), \dots, (t_n, 1)\}$$

$$d_3 = \{(t_1, 2), (t_2, 2), \dots, (t_n, 2)\}$$

출현빈도벡터길이 정규화 요소를 포함하지 않는 가중치 기법 l_{nn} ($\ln\{f + 1.0\}$)이 사용된다면, $d_{3.lnn}$ 의 색인어 가중치는 $d_{2.lnn}$ 보다 높다.

$$d_{2.lnn} = \{(t_1, 1), (t_2, 1), \dots, (t_n, 1)\}$$

$$d_{3.lnn} = \{(t_1, 1.69), \dots, (t_n, 1.69)\}$$

따라서 질의 $q_1 = \{(t_1, w_1), (t_2, w_2), \dots, (t_n, w_n)\}$ 에 대한 $d_{2.lnn}, d_{3.lnn}$ 의 유사도는 다음과 같은 관계를 갖는다.

$$Sim(d_{2.lnn}, q_1) \cdot 1.69 = Sim(d_{3.lnn}, q_1)$$

즉, 긴 출현빈도벡터길이를 갖는 문서 d_3 의 유사도가 짧은 출현빈도벡터길이를 갖는 문서 d_2 유사도의 1.69배이다. 그러나 출현빈도벡터길이를 고려할 때, 문서 d_2 와 d_3 는 거의 동일한 문서로 간주될 수 있다.

일반적으로 검색의 목적하에서 모든 문서는 동등하게 취급되어야 한다. 그러나 앞에서 설명된 바와 같이 출현빈도벡터길이에 대한 정규화가 가중치 기법에 고려되지 않는다면, 검색 시스템은 긴 출현빈도벡터길이를 갖는 문서를 선호한다. SMART 시스템의 가중치 기법에서 사용되는 코사인 정규화는 출현빈도벡터길이를 정규화하는 특성을 지니고 있으며, 많은 경우에 코사인 정규화를 하는 가중치 기법이 코사인 정규화를 하지 않는 가중치 기법보다 높은 검색 효과를 제공하는 것으로 알려져 있다.

코사인 정규화는 출현빈도벡터길이를 정규화하는 바람직한 특성을 지니고 있다할 지라도, 다중 주제를 다루는 적합한 문서의 검색을 어렵게 하는 특성을 지니고 있다. 이는 사용자 질의와 관련된 적합 색인어의 가중치가 비적합 색인어의 가중치에 의해 감소되기 때문이다. 예를 들면, 문서 d_4 가 단일 주제만을 다루고, 문서 d_5 는 문서 d_4 가 기술하는 주제를 포함하여 여러개의 주제를 다룬다고 가정하자. 즉, 문서 d_4, d_5 는 다음과 같이 표현될 수 있다.

$$d_4 = \{(t_1, 1), \dots, (t_m, 1), (t_{m+1}, 0), \dots, (t_n, 0)\}$$

$$d_5 = \{(t_1, 1), \dots, (t_m, 1), (t_{m+1}, 1), \dots, (t_n, 1)\}$$

가중치 기법 l_{nc} 가 적용된다면, 문서 d_4, d_5 는 다음과 같이 변환된다.

$$d_{4.lnc} = \{(t_1, \frac{1}{\sqrt{m}}), \dots, (t_m, \frac{1}{\sqrt{m}}), (t_{m+1}, 0), \dots, (t_n, 0)\}$$

$$d_{5.lnc} = \{(t_1, \frac{1}{\sqrt{n}}), \dots, (t_m, \frac{1}{\sqrt{n}}), (t_{m+1}, \frac{1}{\sqrt{n}}), \dots, (t_n, \frac{1}{\sqrt{n}})\}$$

따라서 질의 $q_2 = \{(t_1, w_1), \dots, (t_m, w_m), (t_{m+1}, 0), \dots, (t_n, 0)\}$ 에 대한 $d_{4.lnc}, d_{5.lnc}$ 의 유사도는 다음과 같다.

$$Sim(d_{4.lnc}, q_2) = \frac{1}{\sqrt{m}} \cdot \sum_{i=1}^m w_i$$

$$Sim(d_{5.lnc}, q_2) = \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^m w_i$$

n 이 항상 m 보다 크기 때문에, 단일 주제 문서 d_4 가 다중 주제 문서 d_5 보다 높은 순위를 부여받는다. 그러나 문서 d_4, d_5 는 질의 q_2 에 대하여 같은 양의 정보를 포함하고 있다. 이처럼 바람직하지 않은 결과는 문서 d_5 에 포함되어 있는 적합 색인어 t_1-t_m 의 가중치가 비적합 색인어 $t_{m+1}-t_n$ 의 가중치에 의해 감소되기 때문이다.

α 로 표기되는 보강된 정규화 출현 빈도는 출현 빈도 tf 를 최대 출현 빈도 $\max tf$ 로 정규화한다. 이러한 최대 정규화는 특정 경우에 출현빈도벡터길이를 정규화할 수 있다. 예를 들

표 1: 가중치 부여 기법의 구성 요소

| | |
|--------------|--|
| 출현 빈도 | |
| b | 1.0 색인어 출현 빈도를 무시하고 벡터를 구성하는 색인어에 1의 가중치 부여 |
| n | tf 문서나 절의내에서 색인어의 출현 빈도 |
| a | $0.5 + 0.5 \frac{tf}{\max tf}$ 보강된 정규화 출현 빈도 (tf 를 $\max tf$ 로 나누고, 그 결과가 0.5에서 1.0 사이의 값을 갖도록 정규화) |
| l | $\ln tf + 1.0$ 색인어 출현 빈도에 로그함수 적용 |
| 장서 빈도 | |
| n | 1.0 색인어 출현 빈도(b, n, a, l)만으로 가중치 생성 |
| t | $\ln \frac{N}{n}$ 색인어 출현 빈도와 역문헌 빈도를 규한다 (N 은 전체 문서들의 수이며, n 은 그 색인어를 포함하고 있는 문서들의 수이다.) |
| 정규화 | |
| n | 1.0 색인어 출현 빈도와 장서 빈도만으로 유도된 가중치를 사용 |
| c | $\frac{1}{\sqrt{\sum_{vector} w_i^2}}$ 유클리디안 벡터 길이를 이용한 코사인 정규화 |

면, 문서 d_6, d_7 이 다음과 같이 표현되어 있다고 가정하자.

$$d_6 = \{(t_1, 1), (t_2, 1), \dots, (t_n, 1)\}$$

$$d_7 = \{(t_1, 2), (t_2, 2), \dots, (t_n, 2)\}$$

가중치 기법 $ann(0.5 + 0.5 \frac{tf}{\max tf})$ 이 적용된다면, 문서 d_6, d_7 은 다음과 같이 동일한 벡터 표현을 갖는다.

$$d_{6.ann} = \{(t_1, 1), (t_2, 1), \dots, (t_n, 1)\}$$

$$d_{7.ann} = \{(t_1, 1), (t_2, 1), \dots, (t_n, 1)\}$$

그러나 이러한 최대 정규화는 많은 경우에 출현빈도벡터길이를 정규화할 수 없다. 예를 들면, 문서 d_8, d_9 가 다음과 같이 표현되어 있다고 가정하자.

$$d_8 = \{(t_1, 1), (t_2, 1), \dots, (t_{100}, 1)\}$$

$$d_9 = \{(t_1, 2), (t_2, 1), \dots, (t_{100}, 1)\}$$

가중치 기법 ann 이 적용된다면, 다음과 같이 $d_{8.ann}$ 의 색인어 가중치가 $d_{9.ann}$ 의 색인어 가중치보다 크다.

$$d_{8.ann} = \{(t_1, 1), (t_2, 1), \dots, (t_{100}, 1)\}$$

$$d_{9.ann} = \{(t_1, 1), (t_2, 0.75), \dots, (t_{100}, 0.75)\}$$

따라서 임의의 절의에 대하여 문서 d_8 이 문서 d_9 보다 높은 순위를 부여받을 가능성이 크다. 그러나 색인어 t_1 의 출현 빈도만이 서로 다르기 때문에, 문서 d_8, d_9 은 거의 동일한 문서로 간주될 수 있다.

최대 정규화가 출현빈도벡터길이의 정규화에 있어 문제점을 가지고 있다할 지라도, 다중 주제 문서의 검색이라는 측면에서 코사인 정규화에 비하여 장점을 지니고 있다. 예를 들면, 문서 d_{10} 이 단일 주제를 다루고, 문서 d_{11} 이 문서 d_{10} 이 다루는 단일 주제를 포함한 여러 개의 주제를 기술한다고 가정하자. 즉, 문서 d_{10}, d_{11} 은 다음과 같이 표현될 수 있다.

$$d_{10} = \{(t_1, 1), \dots, (t_m, 1), (t_{m+1}, 0), \dots, (t_n, 0)\}$$

$$d_{11} = \{(t_1, 1), \dots, (t_m, 1), (t_{m+1}, 1), \dots, (t_n, 1)\}$$

가중치 기법 ann 이 적용된다면, 문서 d_{10}, d_{11} 은 다음과 같이 변환된다.

$$d_{10.ann} = \{(t_1, 1), \dots, (t_m, 1), (t_{m+1}, 0), \dots, (t_n, 0)\}$$

$$d_{11.ann} = \{(t_1, 1), \dots, (t_m, 1), (t_{m+1}, 1), \dots, (t_n, 1)\}$$

따라서 절의 $q_3 = \{(t_1, w_1), \dots, (t_m, w_m), (t_{m+1}, 0), \dots, (t_n, 0)\}$ 에 대하여, 문서 d_{10}, d_{11} 은 다음과 같이 동일한 문서값을 부여받는다.

$$Sim(d_{10.ann}, q_3) = Sim(d_{11.ann}, q_3) = \sum_{i=1}^m w_i$$

비록 일반적이지 않을 지라도, 위의 예는 코사인 정규화의 문제점이 최대 정규화에 의해 완화될 수 있음을 보여준다.

지금까지 검색되는 문서 형태에 영향을 주는 코사인 정규화, 최대 정규화와 같은 가중치 기법의 중요한 특성을 설명하였다. 이러한 특성에 근거하여 가중치 기법은 다음과 같이 분류된다.

- 부류 C: 코사인 정규화를 수행하는 가중치 기법들
- 부류 M: 최대 정규화를 수행하고 코사인 정규화를 수행하지 않는 가중치 기법들
- 부류 N: 코사인 정규화와 최대 정규화 모두를 수행하지 않는 가중치 기법들

앞에서 설명된 바와 같이, 서로 다른 부류에 속하는 가중치 기법들은 서로 다른 형태의 문서들을 검색할 수 있다. 이를 확인하기 위하여 서로 다른 가중치 기법을 사용하는 두개의 검색 실행에 의해 공통적으로 검색되는 문서들의 수를 조사하였다. (표 2)는 같은 부류에 속하는 가중치 기법들에 의해 보다 많은 수의 문서들이 공통적으로 검색됨을 보여준다. 또한 코사인 정규화의 사용 여부가 서로 다른 집합의 문서들을 검색하는데 중요한 역할을 함을 알 수 있다.

4. 성능 평가

본 장에서는 WSJD2 자료 집합 [9]으로 서로 다른 가중치 기법을 사용하는 다양한 검색 실행을 수행하여, 그 결과를 결합한다. 각각의 검색 실행과 결합된 검색 실행은 상위 순위 200개 문서들을 검색하며, 11-포인트 평균정확률이 검색 효과를 측정하기 위해 사용되었다. 사용된 검색 실행의 결합 방법은 다음과 같다. 첫째, 각각의 검색 실행이 생성한 문서값을 그 검색 실행이 생성한 최대 문서값으로 나눈다. 둘째, 결합된 검색 실행에서 각각의 문서에 대한 문서값은 결합에 참가하는 검색 실행들이 그 문서에 대해 생성한 문서값들의 합이다.

서로 다른 가중치 기법을 사용하는 6개의 검색 실행을 수행하고, 그들의 쌍들을 결합하였다. (표 3)는 그 결과를 보여준다. % 변화는 각 쌍에서 높은 검색 효과를 보이는 검색 실행에 대하여 계산되었다. (표 2)와 (표 3)의 비교로부터, 결합되는 검색 실행이 유사한 검색 효과를 제공하고 상이한 문서들을 검색할수록, 그들을 결합으로부터 얻을 수 있는 검색 효과의 향상이 큼을 알 수 있다. 또한 (표 3)은 부류 C에 속

표 2: 서로 다른 가중치 기법을 사용하는 두개의 검색실행에 의해 공통적으로 검색되는 문서들의 수 (WSJ.D2: 상위 순위 200개 문서가 100개의 질의에 의해 검색됨)

| | inc · ltc (C) | anc · ltc (C) | lnc · ntc (N) | ltn · ntc (N) | ann · ntc (M) |
|---------------|---------------|---------------|---------------|---------------|---------------|
| anc · ltc (C) | 15183 | . | . | . | . |
| lnc · ntc (N) | 9911 | 8200 | . | . | . |
| ltn · ntc (N) | 10573 | 9032 | 15937 | . | . |
| ann · ntc (M) | 10106 | 10395 | 13069 | 13310 | . |
| atn · ntc (M) | 10301 | 10627 | 11745 | 13733 | 16069 |

표 3: 서로 다른 가중치 기법을 사용하는 검색 실행의 결합

| | inc · ltc (C) | anc · ltc (C) | lnc · ntc (N) | ltn · ntc (N) | ann · ntc (M) |
|-------------------------|--------------------|--------------------|-------------------|-------------------|-------------------|
| | 0.3284 | 0.3034 | 0.2661 | 0.3028 | 0.3002 |
| anc · ltc (C) 0.3034 | 0.3205 (-2.1%) | . | . | . | . |
| lnc · ntc (N) 0.2661 | 0.3378 (+2.9%) | 0.3433 (+13.1%) | . | . | . |
| ltn · ntc (N) 0.3028 | 0.3464 (+5.5%) | 0.3517 (+15.9%) | 0.2887 (-4.7%) | . | . |
| ann · ntc (M) 0.3002 | 0.3575 (+8.9%) | 0.3473 (+14.5%) | 0.2982 (-0.7%) | 0.3162 (+4.4%) | . |
| atn · ntc (M) 0.3198 | 0.3627 (+10.4%) | 0.3451 (+7.9%) | 0.3130 (-2.2%) | 0.3217 (+0.6%) | 0.3148 (-1.6%) |

하는 가중치 기법과 다른 부류에 속하는 가중치 기법 사이의 결합에 대해서만 의미있는 검색 효과의 향상을 얻을 수 있음을 보여준다. 즉, 서로 다른 2개의 검색 실행의 결합시, 다 음과 같은 조건하에서 의미있는 검색 효과의 향상을 얻을 수 있다.

- 결합되는 검색 실행은 유사한 수준의 검색 효과를 제공한 다.
- 하나의 검색 실행의 가중치 기법은 코사인 정규화를 수행 하고, 다른 하나의 검색 실행의 가중치 기법은 코사인 정 규화를 수행하지 않는다.

5. 결론

지금까지 다양한 질의 및 문서 표현법과 검색 기법들이 정보 검색 분야에서 개발되었다. 최근 이러한 방법들을 결합하기 위한 연구들이 수행되었으며, 서로 다른 검색 실행들의 결합 으로 보다 높은 검색 효과를 얻을 수 있음이 확인되었다. 본 논문에서는 서로 다른 가중치 기법을 사용하는 검색 실행을 결합함으로써 보다 높은 검색 효과를 얻을 수 있는 방법을 제안하였다. 출현빈도벡터질어와 문서가 기술하는 주제의 수 에 따라 문서의 형태를 분류하고, 코사인 정규화와 최대 정규 화와 같은 가중치 기법의 특성들을 설명하였다. 또한, 문서 형태와 가중치 기법의 특성을 기반으로 하여 서로 다른 형태 의 문서들이 서로 다른 특성의 가중치 기법에 의해 검색될 수 있음을 설명하였다. 제안하는 방법은 문서 및 질의의 단일 표 현과 단일 검색 기법을 사용하는 시스템에 쉽게 수용될 수 있 다.

참고 문헌

[1] M. McGill, M. Koll and T. Norreault, "An evaluation of factors affecting document ranking by information re-

trieval systems," Syracuse, Syracuse University School of Information Studies, 1979.

[2] J. Katzer, M.J. McGill, J.A. Tessier, W. Frakes and P. Dasgupta, "A study of the overlap among document representations," *Information Technology: Research and Development*, Vol. 1, No. 2, pp. 261-274, 1982.

[3] T. Saracevic and P. Kantor, "A study of information seeking and retrieving. III. Searchers, searches, overlap," *Journal of the American Society for Information Science*, Vol. 39, No. 3, pp. 197-216, 1988.

[4] H. Turtle and W.B. Croft, "Evaluation of an inference network-based retrieval model," *ACM Transactions on Information Systems*, Vol. 9, No. 3, pp. 187-222, 1991.

[5] E.A. Fox and J.A. Shaw, "Combination of multiple searches," *Proceedings of the 2nd Text REtrieval Conference (TREC-2)*, National Institute of Standards and Technology Special Publication 500-215, pp. 243-252, 1994.

[6] G. Salton, *Automatic Text Processing - the Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley Publishing Co., Reading, MA, 1989.

[7] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," *Information Processing and Management*, Vol. 24, No. 5, pp. 513-523, 1988.

[8] J.H. Lee, "Combining Multiple Evidence from Different Properties of Weighting Schemes," *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995 (to be published)

[9] J.P. Callan, "Passage-level evidence in document retrieval," *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 302-310, 1994.