다중 질의 결합을 통한 검색 효과의 개선

Improving Retrieval Effectiveness with Multiple Query Combination

이 기 호(Kyi-Ho Lee) *
이 준 호(Joon-Ho Lee) *
이 규 철(Kyu-Chul Lee) **

목 차

초 록

일반적으로 주어진 정보 요구에 대하여 서로 다른 사용자는 서로 다른 질의를 생성할 수 있으며, 또는 한명의 사용자가 통제어의 사용 여부에 따라 서로 다른 질의를 생성할 수 있다.
최근 정보 검색 분야의 연구들은 이러한 서로 다른 질의 표현은 서로 다른 문서 집합을 검색함을 보여준다. 본 논문에서는 하나의 사용자 질의에 대하여 다양한 적합성 피드백 방법을 적용함으로써 다중의 질의들을 자동으로 생성한 후, 생성된 다중 질의들을 다시 하나의 질의로 결합하는 방법을 제안한다. 또한 실험을 통하여 자동으로 생성된 다중의 질의들을 결합함으로써 보다 높은 검색 효과를 얻을 수 있음을 입증한다.

ABSTRACT

Different users or the same user using controlled versus free-text vocabularies could generate different queries for the same information need. It has been known in the information retrieval literature that different query representations may retrieve different sets of documents. In this paper, we first generate multiple query vectors from a given information problem by using different relevance feedback methods. Then, we combine the multiple query vectors into a single query vector. We also show through experiments that significant improvements can be achieved by the combination of the multiple query vectors.

· 연구개발정보센터 연구개발부
충남대학교 컴퓨터공학과 부교수
접수일자 1997년 8월 21일

1. 서 론

대용량의 데이터로부터 주어진 시간내에 원하는 정보를 발견하는 것은 매우 어려운 일이다. 이러한 문제점을 해결하기 위해 1960년도 초에 컴퓨터를 이용하여 지정된 정보를 검색하는 정보 검색이라는 연구 분야가 확립되었다 (Salton 1987). 문서 순위 결정 방법에 관한 연구는 정보 검색 분야의 중요한 연구 주제들 중의 하나이며 지금까지 이에 대한 많은 연구가 수행되어 왔다.

문서 순위 결정 방법은 각각의 문서에 대하여 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. 본 논문에서는 문서값 계산의 정확도를 개선함으로써 정보 검색 시스템의 검색 효과를 높일 수 있는 방법을 제안한다.

McGill, Koll & Norreault (1979)는 서로 다른 사용자가 문서를 검색할 때, 또는 한 명의 사용자가 문서를 검색할 경우에도 통제어의 사용여부에 따라, 동일한 정보 요구에 대해 검색되는 문서 집합들 사이에 중복이 적다는 것을 발견하였다. 이와 같은 현상은 동일한 정보 요구로부터 생성된 서로 다른 질의 표현들이 서로 다른 문서들을 검색함을 의미한다.

Saracevic & Kantor (1988)는 여러 전문가에게 동일한 정보 검색 요구에 대하여 부울 질의를 생성하도록 요청하였다. 그리고 생성된 다양한 질의에 대하여 검색을 수행하여 검색 결과의 중복 여부를 조사였다. 그 결과 이들은 서로 다른 질의 형태는 서로 다른 문서의 집합을 검색한다는 것을 재확인하였다.

위에서 언급된 연구 결과들은 동일한 정보 검색 요구에 대한 다양한 질의를 결합함으로써 보다 높은 검색 효과를 얻을 수 있음을 암시한다. 본 논문에서는 하나의 사용자 질의에 대하여 다양한 적합성 피드백 방법을 적용함으로써 다중의 질의를 자동으로 생성한 후, 생성된 다중 질의들을 다시 하나의 질의로 결합함으로써 보다 높은 검색 효과를 얻을 수 있는 방법을 제안한다.

첫째, 주어진 질의에 대한 질의 벡터를 생성하여 초기 검색을 수행한다. 둘째, 검색된 문서들 중에서 상위 30개의 문서를 적합 문서라고 가정하고, 다양한 적합성 피드백 방법들을 적용함으로써 다중 질의 벡터를 생성한다. 이러한 방법에 대한 기본적인 생각은 서로 다른 적합성 피드백 방법은 서로 다른 특성을 지니며, 따라서 서로 다른 피드백 질의 벡터를 생성한다는 것이다. 마지막으로, 생성된 다중 질의 벡터들을 결합하여, 하나의 질의 벡터를 생성한 후 검색을 수행한다.

본 논문의 구성은 다음과 같다. 2절은 실험에서 사용된 SMART 시스템에 대하여 기술한다. 3절은 다중 질의 생성에 사용된 다양한 적합성 피드백 방법에 대해

설명한다. 4절에서는 다양한 피드백 방법을 이용하여 다중 질의 벡터를 자동으로 생성하는 방법과 이와 관련된 특성에 대하여 서술한다. 5절에서는 다중 질의 벡터를 하나의 질의 벡터로 결합하고, 질의 벡터들에 대한 검색 실행 결과를 비교 분석한다. 마지막으로 6절에서는 결론을 맺는다.

2. SMART 시스템

SMART 시스템 (Salton 1983)은 지난 37년 동안 하버드와 코넬 대학에서 개발되어 왔다. 이 시스템은 질의와 문서의 색인을 완전히 자동으로 수행하므로, 데이터베이스 구축이나 질의 작성을 위한 전문가를 요구하지 않는다. 따라서 검색 결과가 데이터베이스에 속한 문서의 특성에 비교적 독립적이며, 다양한 데이터베이스에 폭 넓게 사용할 수 있다. SMART는 벡터 공간 모델 (Salton 1989)을 기반으로 하며, 질의와 문서를 모두 다음과 같은 벡터 형태로 표현한다.

$$d_i = (wi_1, wi_2, ..., wi_k)$$

여기서 d_i는 질의 또는 문서를 나타내고, 총 n개의 색인어가 질의 또는 문서의 표현을 위해 사용된다. wi_k는 문서 d_i내의 색인어 t_k의 가중치로서, 문서 d_i에서 색인어 t_k의 중요도를 나타낸다. 가중치 0은 해당 색인어가 문서 또는 질의에 할당되지 않았음을 의미하고, 양의 가중치는 해당 색인어가 문서 또는 질의의 표현을 위해 사용되었음을 나타낸다.

SMART에서 질의 또는 문서 벡터는 다음과 같은 텍스트 변환에 의해 생성된다.

1. 각각의 텍스트로부터 단어들을 인식한다.
2. 색인어로써 가치가 없는 불용어를 제거한다.
3. 접미사를 제거함으로써 어근을 추출한다.
4. 각각의 어근에 가중치를 부여한다.

SMART 시스템에서는 구문 분석 또는 단어 출현에 대한 통계적 분석에 의하여 생성된 구들을 질의 또는 문서 벡터의 표현을 위한 색인어로 사용할 수 있으나, 본 논문에서는 질의나 문서 벡터를 생성하는 데 있어서 단어 어근만을 사용하였다.

문서 또는 질의에 대한 벡터가 형성되면 이후의 검색 과정은 벡터 연산에 의하여 이루어진다. 문서 d가 벡터 (wd_1, wd_2, ..., wdn)로 표현되고 질의 q가 벡터 (wq_1, wq_2, ..., wqn)로 표현되었을 때, 문서 d와 질의 q 사이의 유사도를 의미하는 문서 d의 문서값은 다음과 같이 두 벡터들의 내적으로 계산된다.

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

문서 d의 문서값은 두 벡터에서 일치되는 색인어의 가중치에 의존하기 때문에,

가중치 부여 기법은 SMART 시스템의 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 고려하고 있다 (Salton 1988). 출현 빈도는 문서내에서 자주 출현하는 색인어에 보다 높은 가중치를 부여한다. 장서 빈도는 전체 문서들 중에서 적은 수의 문서에 출현하는 색인어에 보다 높은 가중치를 부여한다. 그리고 정규화는 데이터베이스 내의 모든 문서 벡터들의 길이를 일치시키는 요소로서, 작은 크기의 문서들이 문서값 계산에 있어서 불공정하게 취급되는 것을 피하도록 한다.

SMART 시스템에서 가중치 부여 기법은 앞에서 언급된 세가지 요소의 조합으로 구성된다. 본 논문에서는 대용량의 데이타 컬렉션에서 높은 검색 효과를 제공하는 것으로 알려진 가중치 기법 *lnc · ltc*를 사용하였다 (Lee 1995). *lnc · ltc*는 문서 벡터와 질의 벡터의 색인어들에 대하여 각각 *lnc*와 *ltc* 기법을 적용함을 의미한다. *ltc* 기법은 다음과 같이 색인어 출현 빈도와 역 문헌 빈도를 곱한 값을 코사인 정규화한다.

$$W_{ik} = \frac{(\log(tf_{ik})+1.0) \times \log(N/n_k]}{\sqrt{\sum_{j=1}^{n}[(\log(tf_{ij})+1.0) \times \log(N/n_j]^2}}$$

여기에서 tf_{ik}는 질의 q_i내 색인어 t_k의 출현빈도, N은 컬렉션내 문서의 총 수, 그리고 n_k는 색인어 t_k가 할당된 문서의 수를 나타낸다. *lnc* 기법은 역 문헌 빈도 요소가 사용되지 않는 점을 제외하고 *ltc* 기법과 동일하다.

3. 적합성 피드백 방법

적합성 피드백은 질의를 자동으로 재구성하는 기법으로 사용자의 편리를 도모한다. 사용자들은 대개 광범위한 정보 요구를 갖고 시스템에 접근하여 그 요구들을 질의로서 표현한다. 이러한 초기의 질의는 매우 임의적이며, 때때로 사용자들은 그들이 지니고 있는 문제점들 조차도 정확하게 표현할 줄 모른다. 따라서 사용자 정보 요구의 비정상적인 상태와 불확실성을 완화시키기 위해, 사용자와 정보 검색 시스템의 연속적인 상호작용을 가능하게 하는 전략이 필요하며, 이를 통해 검색 과정에서 불완전했던 초기 질의를 보완할 수 있다. 적합성 피드백은 이러한 상호 작용을 정교한 방법으로 지원할 수 있다.

지금까지 많은 사람들에 의해 다양한 적합성 피드백 방법들이 검증되어 왔으며, 각 방법의 장단점이 논의되어 왔다. 적합성 피드백에 관한 대부분의 연구는 질의를 가중치가 부여된 용어들의 벡터로서 표현한다. 적합성 피드백에 관한 연구들은 주로 새로운 질의의 재구성 과정에서 적합 문서들에 출현한 색인어들의 가중치를

높이고, 부적합 문서들에 출현한 색인어들의 가중치를 낮춘다는 원칙을 기초로 한다. 초기 질의 q가 $q=(wq_1, wq_2, ..., wqn)$로 표현되었을 때, 적합성 피드백의 적용은 다음과 같은 새로운 질의 벡터를 생성한다.

$$q' = (W_{q1}', W_{q2}', ..., W_{qn}')$$

여기에서 wqi'은 색인어 ti의 변화된 색인어 가중치를 나타낸다. 새로운 색인어들은 초기 가중치가 0인 색인어들에게 양의 가중치를 부여함으로써 질의 벡터에 할당되고, 또한 가중치를 0으로 줄임으로써 색인어들을 초기 질의 벡터로부터 삭제할 수 있다.

본 논문에서는 주어진 정보 요구에 대한 다중 질의 벡터를 생성하기 위해 적합성 피드백 방법들을 사용한다. 다양한 적합성 피드백 방법들이 정보 검색 문헌에서 제시되어 왔지만, 본 연구에서는 평가 목적으로 SMART 버전 11.10에 구현된 5개의 적합성 피드백 방법들을 이용하였다. 5개의 방법중 2개는 벡터 수정에 의한 방법이고, 3개는 확률 피드백 방법이며, 다음과 같이 요약될 수 있다.

Rocchio 피드백 질의 벡터 Qnew는 초기 질의 벡터와 적합 및 부적합 문서 벡터들의 벡터합에 의해 생성된다 (Rocchio 1971).

$$Q_{new} = \alpha \cdot Q_{old} + \beta \cdot \sum_{r=1}^{n_{rel}} \frac{D_r}{n_{rel}} - \sum_{n=1}^{n_{nonrel}} \frac{D_n}{n_{nonrel}}$$

$\alpha \cdot \beta \cdot \gamma$ 상수
D_r 적합 문서 d_r에 대한 벡터
D_n 비적합 문서 d_n에 대한 벡터
n_{nei} 적합 문서수
n_{nonrei} 비적합 문서수

Ide Ide 공식은 Rocchio 공식을 수정함으로써 생성되었다. 즉, 적합 및 부적합 문서들의 수에 의한 정규화를 수행하지 않고, 최상위 순위 부적합 문서만을 피드백에 사용하였다 (Ide 1971).

$$Q_{new} = \alpha \cdot Q_{old} + \beta \cdot \sum_{r=1}^{n_{rel}} R_i - \gamma \cdot T_{nonrel}$$

T_{nonrei} 상위 순위 비적합 문서에 대한 벡터

Pr_cl Pr_cl 공식은 전형적인 확률 피드백 공식으로 확률 검색 모델에 근거하여 개발되었다(Croft & Harper 1979).

$$W_{qi}' = \log \frac{P_i(1-q_i)}{q_i(1-P_i)}$$

$$P^i = \frac{\gamma_i + 0.5}{R+1}$$

$$q^i = \frac{n_i - \gamma_i + 0.5}{N-R+1}$$

γ_i 색인어 ti를 갖는 적합 문서수
n_i 컬렉션내 색인어 t_i를 갖는 문서수
R 적합 문헌 총수
N 컬렉션내 문서수

〈표 1〉 초기 질의와 피드백 질의에 대한 11-포인트 평균 정확률
(TREC D1& D2; 50개 질의에 대한 평균)

Initial	Ide	Rocchio	Pr__cl	Pr__adj	S__rpi
0.2837	0.3523	0.3482	0.3361	0.3378	0.3301
	(+21.8%)	(+20.4%)	(+16.2%)	(+16.8%)	(+14.1%)

〈표 2〉 초기 질의 벡터와 피드백 질의 벡터 사이의 유사도
(TREC D1 & D2; 50개 질의에 대한 평균)

	Ide	Rocchio	Pr__cl	Pr__adj	S__rpi
Initial	0.5803	0.9566	0.2742	0.2522	0.2837

〈표 3〉 확장된 질의 벡터들 사이의 유사도
(TREC D1 & D2; 50개 질의에 대한 평균)

	Ide	Rocchio	Pr__cl	Pr__adj
Rocchio	0.7900(4)			
Pr__cl	0.6746(5)	0.4456(8)		
Pr__adj	0.6595(6)	0.4239(9)	0.9856(1)	
S__rpi	0.6470(7)	0.4091(10)	0.9725(2)	0.9403(3)

Pr__adj Pr__adi 공식은 Pr__cl 공식을 수정함으로써 개발되었다. 즉, 상수 0.5를 ni/N로 대치하였다 (Robertson 1986).

$$W\ qi' = \log \frac{P_i(1-q_i)}{q_i(1-P_i)}$$

$$P^i = \frac{r_i + n_i/N}{R+1}$$

$$q^i = \frac{n_i - r_i + n_i/N}{N-R+1}$$

S__rpi 공식은 Furh의 RPI 공식을 비선형 함수로 단순화시킨 식이다 (Fuhr & Buckley 1991).

$$W\ qi' = \log \frac{P_i(1-q_i)}{q_i(1-P_i)}$$

w_n 적합 문서 d_r에서 색인어 t_i의 가중치
w_{ni} 비적합 문서 d_n에서 색인어 t_i의 가중치
n_{rei} 적합 문서의 수
$nnonrei$ 비적합 문서의 수

$$P_i = \sum_{r=1}^{n\_rel} \frac{W\_ri}{n\_rel}$$

$$P_i = \sum_{n=1}^{n\_nonrel} \frac{W\_ni}{n\_nonrel}$$

4. 다중 질의 생성

본 절에서는 하나의 사용자 질의에 대해 앞 절에서 언급한 5개의 피드백 방법을 적용하여 다중의 질의 표현을 자동으로 생성한다. 이에 대한 기본적인 생각은 서로 다른 피드백 방법들은 서로 다른 특성을 지니기 때문에, 서로 다른 피드백 질의 벡터들을 생성하며, 이러한 질의 벡터들은 서로 다른 문서들을 검색한다는 것이다. 이러한 생각을 검증하기 위하여 다음과 같은 실험을 수행하였다.

1. 주어진 질의에 대한 초기 질의 벡터를 생성한다.
2. 초기 검색을 수행하고 상위 30개 문서를 적합 문서로 가정한다.
3. 다양한 적합성 피드백 방법들을 사용하여 새로운 피드백 질의 벡터들을 생성한다.
4. 새로운 피드백 질의 벡터들에 의한 피드백 검색을 수행한다.

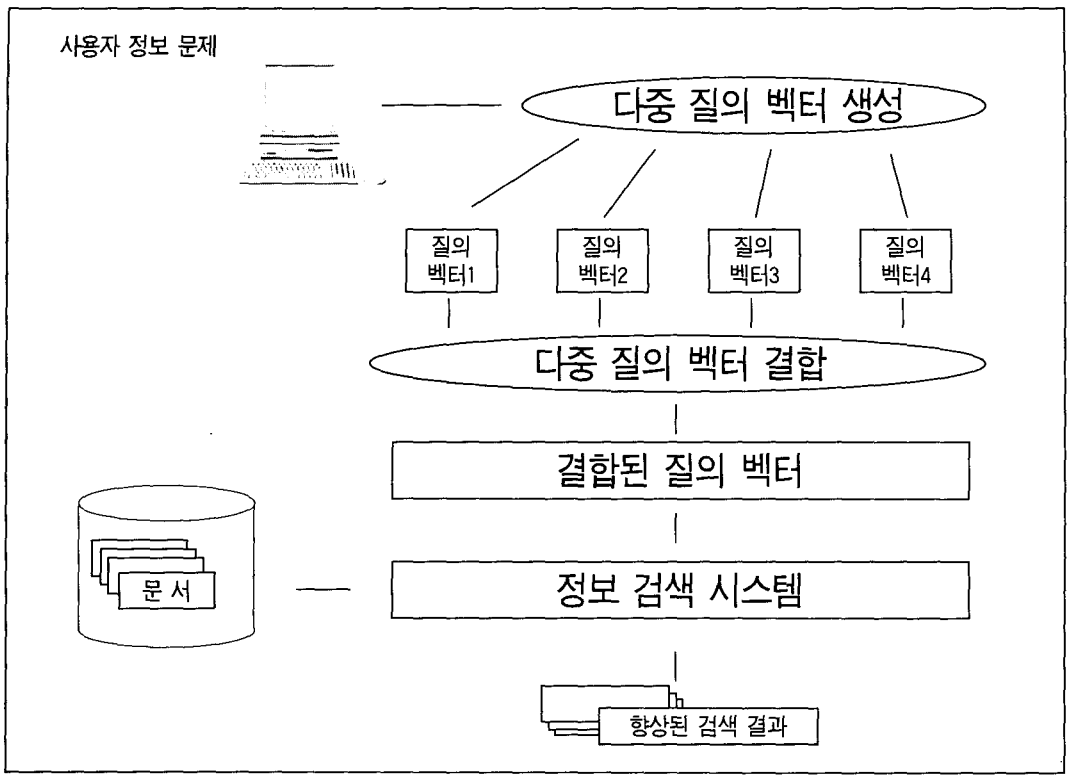
실험 환경은 문서 집합으로 TREC D1 & D2를, 질의 집합으로 TREC Q151-Q200 50개 질의를 사용하였고, 각각의 질의에 대해 상위 1000개 문서들을 검색하였으며, 11 포인트 평균 정확률을 사용하여 검색 성능을 평가하였다. <표 1>에 제시된 실험 결과는, 5개 피드백 검색 실행 모두가 초기의 검색 실행보다 15% - 20% 정도의 높은 검색 효과를 제공하는 것을 보여준다.

한편 새로이 생성된 Rocchio, Ide, Pr__cl, Pr__adj, S__rpi 5개 피드백 질의가 어느 정도 상이한가를 알아보기 위하여 피드백 질의 벡터들 사이의 유사도를 측정하였다. 피드백 질의 벡터들 사이의 유사도는 질의와 문서 사이의 유사도를 계산할 때와 동일하게 피드백 질의 벡터들 사이의 내적으로 계산하였으며, 피드백 질의 벡터의 정규화를 위해 코사인 정규화 방법을 사용하였다. 유사도는 0과 1 사이의 값을 지니며, 유사도 1은 두개의 피드백 질의 벡터들이 동일함을 의미한다.

<표 2>는 초기 질의 벡터와 확장된 질의 벡터들 사이의 유사도를 나타내고, <표 3>은 확장된 질의 벡터들 사이의 유사도를 보여준다. <표 1>, <표 2>, <표 3>으로부터 서로 다른 적합성 피드백 방법은 서로 다른 피드백 질의 벡터를 생성하며, 생성된 피드백 질의 벡터들은 동일 수준의 검색 효과를 제공함을 알 수 있다.

5. 다중 질의 결합

본 절에서는 다양한 적합성 피드백 방법들을 이용하여 생성된 다중의 질의 벡터들을 결합하고, 결합된 질의 벡터들에 대한 검색 실행 결과를 분석한다. <그림 1>은 다중 질의 벡터들의 자동 생성 및



〈그림 1〉 다중 질의 벡터들의 자동 생성 및 결합

결합 과정을 보여준다. 먼저 하나의 정보 요구에 대해, 다양한 적합성 피드백 방법들을 사용하여 다중 질의 벡터들을 생성한다. 그리고 생성된 다중 질의 벡터들을 다양한 결합 방법을 통하여 단일 벡터로 결합하고, 이에 대해 검색을 수행한다.

Fox & Shaw (1994)는 SMART 시스템을 사용하여 검색 실행들을 결합하는 다양한 함수들을 실험하였다. 이들은 6개의 결합 함수에 대한 실험을 수행하였으며, 문서값들을 더하는 합계 함수(summation function)가 거의 모든 TREC 문서 집합에서 좋은 결과를 제공함을 발견하였다. Belkin et al. (1993)도 TREC 주제들로부터 생성된 다중 부울 질의를 결합하기 위해 INQUERY 시스템에서 지원하는 합계 함수를 사용하였다. 본 논문에서도 피드백 질의 벡터들의 결합을 위해 각 피드백 질의 벡터의 가중치를 더하는 합계 함수를 사용하였다. 예를 들어, 3개의 피드백 질의 벡터에서 색인어 ti의 가중치가 각각 0.3, 0.7, 0.5 일 때, 합계 함수에 의해 결합된 질의 벡터에서 색인어 *ti*의 가중치는 1.5이다.

위에서 설명된 합계 함수를 적용하여 3절에서 설명된 5개의 적합성 피드백 방법

에 의해 생성된 5개의 피드백 질의 벡터들의 쌍을 결합하였다. 그리고 결합된 질의 벡터들에 대해 검색을 수행하였다. 〈표 4〉는 결합된 질의 벡터들의 검색 실행 결과를 나타내고 있으며, 여기에서 % 변화는 초기 검색 결과의 11-포인트 평균정확률 0.2893에 대한 변화율이다.

〈표 4〉로부터 Rocchio와 Pr_cl의 결합이 11-포인트 평균 정확률 0.3649로 가장 높은 검색 효과를 보여주며, 이것은 Rocchio(0.3482) 보다 4.8%, Pr_cl (0.3361) 보다는 8.6%의 검색 효과 향상을 제공한다. 또한 Pr_cl과 S_rpi의 결합이 0.3343으로 가장 낮은 검색 효과를 보여주며, S_rpi(0.3301) 보다는 약간의 검색 효과 향상을, Pr_cl(0.3361) 보다는 약간의 검색 효과 감소 현상을 보이고 있다.

한가지 고려할 현상은 확률 피드백 방법들 사이의 결합이 나머지 결합들보다 낮은 검색 효과를 제공한다는 것이다. 즉, Pr_cl, Pr_adj, S_rpi 사이의 결합에 의한 검색 효과의 향상은 15% 정도인데, 나머지 결합에 의한 검색 효과의 향상은 23-26%의 범위를 나타내고 있다. 이러한 원인은 〈표 3〉에서 보여주는 바와 같이 Pr_cl, Pr_adj, S_rpi에 의해 생성된 피드백 질의들 사이의 유사도가 0.9 이상인데 비하여, Ide, Rocchio의 피드백 질의와 Pr_cl, Pr_adj, S_rpi의 피드백 질의 사이의 유사도가 0.4-0.6 정도로 훨씬 작다는 현상으로 유추할 수 있다.

〈표4〉 2개의 질의 벡터 결합시의 검색 효과
(TREC D1 & D2; 50개 질의에 대9한 평균; 11-포인트 평균 정확률 0.2893을 제공하는 초기 검색 결과를 기준으로 검색 효과 향상의 % 변화를 계산)

	Ide 0.3523 (+21.8%)	Rocchio 0.3482 (+20.4%)	Pr_cl 0.3361 (+16.2%)	Pr_adj 0.3378 (+16.8%)
Rocchio 0.3482 (+20.4%)	0.3544 (+22.5%)			
Pr_cl 0.3361 (+16.2%)	0.3568 (+23.3%)	0.3649 (+26.1%)		
Pr_adj 0.3378 (+16.8%)	0.3574 (+23.5%)	0.3647 (+26.1%)	0.3375 (+16.7%)	
S_rpi 0.3301 (+14.1%)	0.3560 (+23.1%)	0.3635 (+25.6%)	0.3335 (+15.3%)	0.3343 (+15.6%)

〈표 5〉 3개 이상의 질의 벡터 결합시의 검색 효과
(TREC D1 & D2; 50개의 질의에 대한 평균)

	Initial	1-way	2-way	3-way	4-way	5-way
average	0.2893	0.3409 (+17.8%)	0.3523 (+21.8%)	0.3565 (+23.2%)	0.3589 (+24.1%)	0.3590 (+24.1%)
best	0.2893	0.3523 (+21.8%)	0.3649 (+26.1%)	0.3637 (+25.7%)	0.3640 (+25.8%)	0.3590 (+24.1%)

〈표 4〉는 전반적으로 피드백 질의 벡터들의 결합에 의한 검색 실행이 결합 이전의 피드백 질의의 검색 실행보다 높은 검색 효과를 제공함을 보여준다. 결론적으로 실험 결과는 서로 다른 적합성 피드백 방법에 의해 생성된 피드백 질의 벡터를 합계 함수를 이용하여 결합하고, 결합된 질의를 실행함으로써 보다 높은 검색 효과를 얻을 수 있음을 입증한다. 마지막으로 본 논문에서는 3개 이상의 피드백 질의 벡터들을 결합했을 때 검색 효과의 향상에 대하여 살펴보았다. 즉, 5개의 피드백 질의 벡터들을 2단계에서는 2개씩 결합하여 10개, 3단계에서는 3개씩 결합하여 10개, 4단계에서는 4개씩 결합하여 5개, 5단계에서는 5개 모두를 결합하여 1개의 결합된 질의 벡터를 생성한 후, 생성된 질의 벡터들에 대한 검색 실행을 수행하고 각 단계에서 최대 검색 효과와 검색 효과들의 평균을 조사하였다.

〈표 5〉는 그 결과를 보여주며, 이 표로부터 검색 효과의 평균은 단계가 증가함에 따라 단순하게 증가함을 알 수 있다. 그러나 최대 검색 효과는 2단계가 나머지 단계보다 우수하다. 이러한 결과는 Belkin et al. (1993)이 다중 부울 질의의 결합에서 얻은 결과와 일치한다.

6. 결 론

본 논문에서는 적합성 피드백 방법을 이용하여 다중의 질의 벡터들을 자동으로 생성하고, 생성된 질의 벡터들을 결합함으로써 검색 효과를 개선할 수 있는 다음과 같은 방법을 제안하였다. 첫째, 주어진 질의에 대한 초기 질의 벡터를 생성한다. 둘째, 초기 검색을 수행하고 상위 30개 문서를 적합 문서로 가정한다. 셋째, 다양한 적합성 피드백 방법들을 사용하여 새로운 피드백 질의 벡터들을 생성한다. 마지막으로 새로운 피드백 질의 벡터들에 의한 피드백 검색을 수행한다.

실험 결과, 피드백 질의 벡터들의 검색 실행은 비슷한 수준의 검색 효과를 제공함을 알 수 있었고, 피드백 질의 벡터들 사이의 유사도 계산을 통하여 서로 다른 적합성 피드백 방법들은 서로 다른 피드백 질의 벡터들을 생성함을 확인하였다. 또한 피드백 질의 벡터들의 결합에 의해

생성된 질의 벡터들의 검색 실행은 결합 이전의 피드백 질의 벡터들의 검색 실행보다 높은 검색 효과를 제공하였다.

참고문헌

Belkin, N.J., Cool, C., Croft, W.B. and Callan, J.P. 1993. "The effect of multiple query representations on information retrieval performance", Proceedings of the 16th Annual International ACM Development in Information SIGIR Conference on Research and Retrieval 339-346.

Croft, W.B. and Harper, D.J. 1979. "Using probabilistic models of document retrieval without relevance", Journal of Documentation, Vol. 35, 285-295.

Ide, E. 1971. "New experiments in relevance feedback", The Smart system - experiments in automatic document processing, Englewood Cliffs, NJ: Prentice Hall Inc., 337-354.

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

Fuhr, N. and Buckley, C. 1991. "Aprobabilistic learning approach for document indexing", ACM Transactions on Information Systems, Vol. 9, No. 3, 223-248.

McGill, M., Koll, M. and Norreault, T. 1979. "An evaluation of factors affecting document ranking by information retrieval systems", Syracuse, Syracuse University School of Information Studies.

Lee, J. H. 1995. "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, 180-188.

Robertson, S.E. 1986. "On relevance weight estimation and query expansion", Journal of Documentation, Vol. 42, 182-188.

Rocchio, J.J.Jr. 1971. "Relevance

feedback in information retrieval", The Smart system - experiments in automatic document processing, Englewood Cliffs, NJ: Prentice Hall Inc., 313-323.

Salton, G. and McGill, M.J. 1983. Introduction to Modern Information Ptrieval, McGraw-Hill, Inc.

Salton, G. "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.

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

Salton, G. 1989. Automatic Text Processing The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley Publishing Co., Reading, MA.