사전 의미 기반의 질의확장 검색에서 추가 용어 가중치 최적화

(Optimizing the Additional Term Weight Ratio in Query Expansion Search based on Dictionary Definition)

최 영 란*, 전 유 정**, 박 순 철***
(Young-Ran Choi, You-Jeong Jeon, Soon-Cheol Park)

요 약 본 연구가 갖는 중요성은 두 가지로 요약된다. 첫째는 질의 확장 검색 방법으로 사전에서 나타나는 용어를 질의의 추가용어로 채택하는 것이다. 이 방법은 기존의 피드백 확정 방법이 갖는 2차적 검색 과정을 줄인다. 둘째는 초기 질의어와 추가용어 사이에 가중치를 달리 적용하여 재현율과 정확률을 동시에 높일 수 있는 최적의 모델을 제시하였다. 이렇게 함으로써 정보 검색의 성능을 크게 향상시킬 수 있는 방법을 제시하고 있다.

핵심주제어 : 질의 확장, 정보 검색

Abstract The significances of this paper are of two points. One is that this research develops the query expansion search by adding the related terms based on the dictionary to the original query terms. This method shortens the process of the conventional model of query expansion utilizing the feedback data of the search. The other is that this research tries to find out the optimal point of precisions and recalls by differentiating the weight ratio between original query and additional terms. This method shows that the efficiency and precision of query expansion search increase.

Key Words : Query Expansion, Information Retrieval

1. 서 론

정보검색의 기본 목적은 이용자가 원하는 정보를 '정확하고 빠짐없이' 찾는 것이다. 최근 인터넷 사용이 보편화되고 일반 이용자들의 정보검색이 확대됨에 따라 보다 지능적이고 효과적인 정보검색에 대한 필요성이 더욱 커지고 있다. 전문가들과는 달리 일반 이용자들은 온라인 데이터베이스를 검색할 때 10개 이내의 질의어를 사용하거나 심지어는 2개 이하의 질의어를 사용하고 경우도 있다고 한다. 또한 이들은 검색결과를 보고 질의수정을 거의 하지 않는 것으로 분석되었다[2]. 이처럼 짧은 질의어를 가지고는 정확한 결과를 얻기 힘들다.

이 문제를 해결하기 위해서 이용자의 초기 질의와 관련된 용어를 새로운 질의에 추가하는 자동 질의확장에 대한 연구가 1970년대에 시작되었으며, 최근 들어 더욱 활발하게 연구되고 있다. 이용자들이 찾고자 하는 정보를 관련 문서 중에서 얼마나 찾았는가를 나타내는 재현율(Recall)과 검색된 문서 중에 관련 있는 문서가 얼마나 되는가를 나타내는 정확률(Precision)을 동시에 높이는 것이 정보검색의 성능을 향상시키는 방법이다. 그러나 재현율과 정확률은 역관계에 있으므로 재

* 전북대학교 정보통신학과
** 전북대학교 정보통신공학과
*** 전북대학교 전자정보공학부 부교수

현율이 오르면 정확률이 떨어지는 경향이 있다. 그러므로 재현율과 정확률을 동시에 올리기 위해서 질의를 확장(Query Expansion)하는 방법이 개발되어 왔다.

지금까지는 질의어를 확장하는 방법으로 본래의 질의로부터 도출된 결과를 피드백하여 초기질의어와 관련 있는 용어를 추가하여 재검색해서 재순위화하는 방법이 대표적인 것으로 알려져 있다.[1] 이러한 방법은 피드백결과를 놓고 재검색 처리하여야 하는 단점이 있다.

본 연구에서는 이 같은 단점을 보완하여 초기질의어에서부터 미리 사전적 의미를 이용하여 질의어확정을 유도하여 재처리 과정 없이 보다 정확한 결과를 얻을 수 있는 방법을 제시하고자 한다.

질의 확장에서는 크게 다음의 두 가지 문제가 제기되고 있다. 첫째는 주어진 질의와 가장 밀접한 확장 용어를 찾는 문제이고, 둘째는 추가된 용어의 가중치 할당에 대한 문제이다. 결국 추가된 용어를 원래 질의에 어떻게 참여시킬 것인가 하는 질의 수정 문제로 귀결된다.

본 연구에서는 사전을 이용하여 정보검색시 사용되는 질의어를 확장하여 용어를 찾고, 추가된 용어의 가중치와 원래 초기 질의어의 가중치를 변화시켜 재현율과 정확률을 동시에 높일 수 있는 가중치의 최적 상태를 찾아보았다.

2. LSI (Latent Semantic Indexing) 모델

2.1 LSI (Latent Semantic Indexing) Model

LSI (Latent Semantic Indexing)는 벡터 공간상에서 SVD (Singular Value Decomposition)를 이용한 개념 기반의 문서 검색 대수적 모델이다. 이를 이용하면 서술된 단어 자체뿐만 아니라 개념까지 비교가 가능하다. LSI에서는 유사 단어까지 고려하기 때문에, 사용자의 질의에 대한 결과 문서 도출에 있어서 개념이 비슷한 문서까지 찾을 수 있는 장점이 있다.

기존의 개념 기반 검색에서는, 사용자의 질의어가 부정확할 경우, 수많은 정보들에 존재하는 내포된 의미를 파악하여 사용자가 원하는 관련된 정보를 검색한다는 것이 쉽지 않다. 질의어 벡터와 문서 벡터간의 유사성에 초점을 두었기 때문이다. 그리고 관련 문서를 검색하기 위해 질의어 용어와 의미가 비슷하거나 소리가 비슷한 용어를 이용한 질의어 확장을 했다.

본 논문에서는 질의어 벡터와 문서 벡터간의 유사성을 계산하기 이전에, 질의어 벡터와 용어 벡터간의 유사성을 먼저 측정하고, 질의어와 유사도가 높은 단어들을 구한다. 그리고 질의어의 사전적 의미를 이용한 질의어 확장을 한 후, 재현율, 정확률 측면에서 기존의 개념 기반 검색과 비교하였다.

문서에서 색인어는 그 자체보다 개념을 분석하는 것이 중요하다. 문서에 따라 개념은 같지만 다른 형태의 색인어들이 많다. 특히 관용구의 경우 사전적 의미나 구문론적 분석만을 따진다면 그 의미를 이해하기가 어렵다. 문서의 내용은 서술된 색인어보다는 그 안에 내포된 개념에 더 관련되어 있으므로 색인어 대신에 개념에 기반을 두어야 한다. 이렇게 하면 문서들이 같은 색인어로 구성되어 있지 않더라도 연관성을 나타낼 수 있다. 어떤 문서가 다른 문서와 같은 개념을 공유한다면 유사한 문서라고 할 수 있다. 문서의 개념 기반에 대한 많은 시도들은 정보 검색 엔진에 있어서 많은 발전을 가져왔다. 이러한 개념 분석에 있어서 두드러진 시도들 중 하나는 LSI 기술이었다. 이것은 단순한 색인어를 사용하는 것이 아니라 문서에 있어서 주요 개념들을 분석하기 위하여 다차원 확장을 이용하는 기술이다.[5][6]

LSI는 문서들이 색인어들의 벡터로 표현된 벡터 공간 검색 모델에 기반을 둔다. 하지만 이것은 색인어와 문서로 구성된 행렬을 SVD를 통해서 축소한다는 점에서 벡터 공간 모델과 다르다.[5]

2.2 SVD (Singular Value Decomposition)

m개의 용어, n개의 문서로 구성된 전체 집합을 A라고 할 때, Simple Query Matching 방법은 (1)의 식을 이용한다.

$$Query = q^t * A \qquad (1)$$

그런데, 일반적인 A 집합은 약 30만개의 용어와 3억개의 문서로 이루어진다. 이렇게 큰 행렬을 이용하면 계산이 어려울 뿐만 아니라, 질의어 벡터의 정보에 비해 관련 없는 용어들과 문서들의 정보가 너무 많아진다는 문제점이 발생한다.

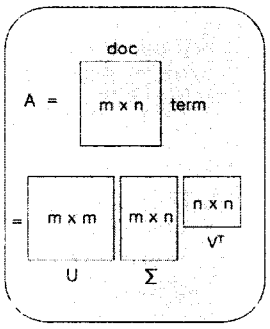
이를 해결하기 위해 수학적 Matrix Decomposition

중 SVD (Singular Value Decomposition)를 이용한다.

$$A = U\Sigma V^T \quad (2)$$

(2)식에서 U 행렬은 색인어간의 상관행렬, V 행렬은 문서간의 상관행렬, 그리고 Σ 행렬은 단일값을 갖는 대각행렬이다. [5]

SVD를 통해 분해 된 행렬을 이용해 LSI 모델을 적용한다. LSI 모델의 요점은 [그림1]에 보이는 것처럼 각 문서와 질의 벡터를 저차원 공간인 개념으로 사상시키는데 있다. 색인어 사이의 관계를 나타내는 행렬과 문서 사이의 관계를 나타내는 행렬을 같은 공간으로 사상시킨다.[5][6][7]



[그림1] SVD 분해

질의어는, 축소된 용어/문서 공간 안에서 식 (3)과 같이 질의어 위치로 표현된다.

$$q = q^T U_k \sum_k^{-1} \quad (3)$$

이 때, 각 용어 벡터와 질의어 벡터간의 유사성을 측정하기 위한 방법은 여러 가지가 있으나, 본 논문에서는 계산과 이해가 간편한 (4)의 코사인 유사도식을 이용하여 측정하였다.

$$Sim(d,q) = \frac{\sum_k t_k \times q_k}{\sqrt{\sum_k t_k^2 \times \sum q_k^2}} \quad (4)$$

3. 질의 확장 (Query Expansion)

3.1 질의 확장

질의 수정 과정은 대부분 처음 질의에 새로운 용어를 추가하거나 수정함으로써 이루어지며, 질의 수정 대상에 따라 검색 문헌 순위 재조정과 질의 확장이 있다. 검색 문헌 순위 재조정은 순위부여를 위한 확률 모델에 기반하고 있으며, 관련 문헌에 나타난 질의 용어의 가중치는 높이고, 비관련 문헌에 나타난 질의 용어의 가중치는 내림으로써, 문헌 순위를 재조정하는 작업이다. 즉, 이 방법은 이미 검색된 문헌집합의 순위만 재조정함으로써 관련문헌을 상위로 올리는 방법이다. 반면에 질의확장은 초기 질의에 새로운 용어를 추가함으로써 검색되지 않은 관련 문헌을 새로 찾아올 수 있으며, 질의 가중치를 조절하여 순위 조정도 할 수 있다.

새로 작성되는 질의에 추가될 용어를 어떻게 획득하는가에 따라 질의 확장 기법은 두 가지로 구분된다. 즉, 초기 질의에서 검색된 문헌들을 이용하는 지역적 (local) 또는 질의 기반 (query specific) 질의확장과 전체 문헌집단을 이용하는 전역적 (global)또는 말뭉치 기반 (corpus specific) 질의확장이 있다.

기존 연구에서는 주로 원 질의와의 유사도를 추가 질의어의 가중치로 사용하였다. 추가되는 질의어에 별도의 가중치를 주지 않고 초기 질의어와 추가 질의어를 대등하게 취급할 경우에는 오히려 성능이 저하되는 것으로 보고되었다[3]. 따라서 가중치를 서로 다르게 주고 각각 어떤 성능을 보이는지 살펴볼 필요가 있다.

3.2 사전적 의미를 이용한 질의어 확장

많은 용어와 문서로 이루어진 전체 집합에서, 보다 정확한 검색을 하기 위해서는 질의어 벡터가 보다 더 정확하고 더 많은 데이터를 가지고 있어야 한다.

이를 위해 질의어 확장을 수행한다. 본 논문에서는 질의어와 관련된 용어들, 그리고 질의어의 사전적인 의미와 관계된 용어들을 이용해서 질의어 확장을 한다. 따라서 보다 정확한 질의어 벡터 정보를 갖게 할 수 있게 된다.

$$Q = q_{original} + q_{dic\_func.} \quad (5)$$

식 (5)에서, $q_{dic_func.}$는 $q_{original}$ 의 사전적 정의에 포함된 단어들의 Parsing 및 Indexing 작업을 통해

얻은 벡터이다.

초기 질의어와 추가 질의어간의 가중치에 따른 성능 평가를 위해서 (6)의 식을 이용한다.

$$Q = \beta * q_{original} + (1 - \beta) * q_{dic\_func}(q_{original}) \quad (6)$$

여기에서, β는 0부터 1까지의 가중치를 의미한다.

4. 실험 환경 및 평가 척도

4.1 실험 환경 및 실험집단의 특성

실험을 위해 사용한 문서집합 KTSET93은 한국 통신에서 개발한 것으로 정보 검색을 위한 테스트 데이터 모음이며 주제는 정보과학이다.

KTSET93 문서들 중 문서번호 kt0001부터 kt0150 번까지의 150개 문서를 실험 데이터로 이용하였다. 이 연구에서 사용한 대상 실험집단의 특성은 <표 1>과 같다.

<표 1> 실험 집단의 특성

	언어	분야	성격	문헌수	용어수
KTSET 93	한국어 영어	전산학 정보학	초록	150	4667

그리고 앞서서 언급한 Parsing 및 Indexing 작업을 위하여 국민대학교 컴퓨터 학부의 자동색인기 HAM을 이용하였으며, HAM을 거친 150개의 문서에서는 4667개의 색인어가 추출되었다.

150개의 문서들은 크게 3개의 집단으로 분류되었다. 각각의 집단마다 질의어 확장 실험이 실시되었으며, 각 집단당 평균 관련 문서 수는 50개, 집단당 평균 색인어 수는 1556개, 그리고 질의어의 사전적인 의미에 관련되어 새로이 추가되는 용어들의 평균수는 31개로 나타났다. (<표 2> 참조)

<표 2> 실험 데이터 집단당 통계

	관련 문서	추가 용어
최대	56	67
최소	45	7
평균	50	30.833

(단위 : 개)

각 집단마다 각기 다른 질의어를 입력하고 그에 따른 사전적 의미를 이용한 질의어 확장 실험을 실시하였다. 그리고 집단별 성능 평가를 측정하였으며, 검색되어 나온 문서들을 전체 검색문서의 상위 10%부터 100%까지 10단계로 나누어 가장 높은 효율을 보이는 순위와 가중치를 알아보았다.

4.2 성능 평가 척도

검색효율은 이용자의 정보요구에 적합한 문헌을 검색해내는 검색 시스템의 능력을 의미하는 것으로 검색된 적합 정보와 부적합 정보, 검색되지 않은 적합 정보와 부적합 정보간의 비율로서 측정된다.

검색평가 척도로는 재현율과 정확률, E-척도, 조화평균, CACM, CISI, 만족도, 실망도, Cystic Fibrosis, TIPSTER/TREC, 등이 있다. 정보 검색과 텍스트 범주화의 평가에서 흔히 사용되는 정확률과 재현율은 각 문헌에 대한 적합 질의나 적합 범주가 미리 판정이 되어 있는 상태이므로 객관적이고 절대적인 평가가 가능하다.[8]

정확률은 검색된 문헌 가운데 적합문헌이 검색된 비율 즉, 시스템이 부적합 문헌을 검색해내지 않는 능력이다. 정확률은 식 (7)을 이용하여 계산된다.

$$\text{정확률}(P) = \frac{\text{검색된적합문헌수}}{\text{검색된문헌총수}} \quad (7)$$

재현율은 시스템내의 적합 문헌 가운데 검색된 적합문헌의 비율 즉, 시스템이 적합문헌을 검색해내는 능력이다. 예를 들어 재현율 80%라는 것은 시스템이 소장하고 있는 전체 적합문헌 10책 가운데 8책이 검색되었다는 것을 의미하는 것으로 식 (8)로 계산한다.

$$\text{재현율}(R) = \frac{\text{검색된적합문헌수}}{\text{적합문헌총수}} \quad (8)$$

재현율은 검색의 완전성을 측정하며 검색 효율 척도 중 가장 널리 사용되고 있다. 결국 재현율은 검색의 완전성을, 정확률은 검색의 정확성을 측정하는 것이라고 볼 수 있다.

4.3 가중치를 다르게 적용한 질의 확장 검색 실험

질의확장 정보검색 실험을 위해 질의어를 입력한다. 그리고 입력된 초기 질의어의 사전적인 의미를 HAM 프로그램을 거쳐 색인어 추출을 한 후, 추출된 용어들을 새로이 질의어에 추가하여 질의어 확장 실험을 실시한다.

이 때, 초기 질의어의 가중치와 새로 추가되는 용어들의 가중치를 다르게 하여 질의확장 검색 실험을 실시하였다.

가중치 β를 0부터 시작해서 1까지 0.2씩 증가시키면서 설정한 실험 결과를 가중치를 모두 1로 하였을 때의 실험 결과와 비교해보았다. 실험 결과를 살펴볼 때는 검색되어 나온 문서들의 순위를 상위 10%부터 100%까지 10%씩 증가시키면서 그룹별로 평균 정확률과 평균 재현율을 측정하고 비교하였다.

4.3.1 정확률

그림에서는 공간 문제로 상위 20%, 40%, 60%, 80%, 100%일 때의 평균 정확률을 그룹과 가중치별로 구분한 경우만 나타내었다.

가중치 β가 0인 경우는 (9)식과 같이 표현된다.

$$Q = q_{dic\_func}(q_{original}) \qquad (9)$$

즉, 사전적인 의미에 관련되어 새로 추가되는 용어들로만 이루어진 경우를 말하며, 가중치 β가 1인 경우는 (10)식으로 표현되어,

$$Q = q_{original} \qquad (10)$$

인 경우이다. 즉, 원래 질의어만 고려한 경우를 뜻한다.

상위 100%는 검색되어 나온 모든 문서들을 뜻한다. 평균 정확률의 결과를 살펴보면 상위 40%와 50%를 제외한 나머지 경우에서 가중치 β가 1보다 작은 값을 가질 때 즉, 가중치 β가 0.2, 0.4, 0.6, 0.8 일 때의 평균 정확률이 가중치 β가 1인 경우보다 높았다.

그리고, 상위 10%부터 20%까지의 문서들의 평균 정확률을 살펴봤을 때, 가중치 β가 0인 경우 즉, 질의어의 사전적인 의미로만 이루어진 질의어로 검색된 결과가 현저히 높은 평균 정확률을 보였다.

전반적으로 살펴보면 가중치 β가 0.2와 0.4 일 때의 평균 정확률이 다른 경우보다 우세하게 나타났다.

그룹 A의 정확률이 다른 그룹들에 비해서 우세하지 못한 것은 그룹별로 새로이 추가되는 용어들의 개수가 원인인 것으로 짐작된다. 그룹 A에 새로 추가되는 용어의 개수는 평균 45개이고 그룹 B와 C에 새로 추가되는 용어의 개수는 각각 평균 33개 16개이다.

추가되는 질의어 수가 많을수록 정확률이 감소하는 이유는 질의어가 많을 경우에는 추가되는 용어를 더 신뢰할 수 있음에도 불구하고 오히려 용어들 간에 더 낮은 평균 유사도를 가지게 되기 때문으로 추정된다.[8]

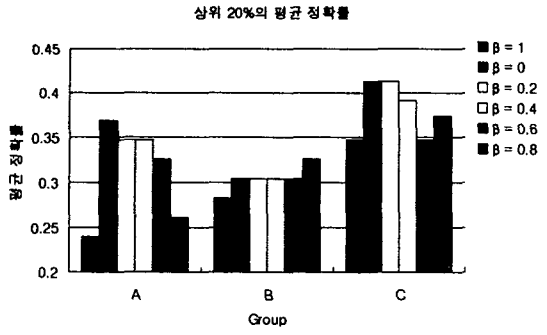
4.3.2 재현율

가중치 β를 0부터 시작해서 1까지 0.2씩 증가시키면서 설정한 실험 결과를 가중치를 모두 1로 하였을 때의 실험 결과와 비교해보았다. 그림에서는 역시 공간 문제로 상위 20%, 40%, 60%, 80%, 100%일 때의 평균 재현율만 나타내었다.

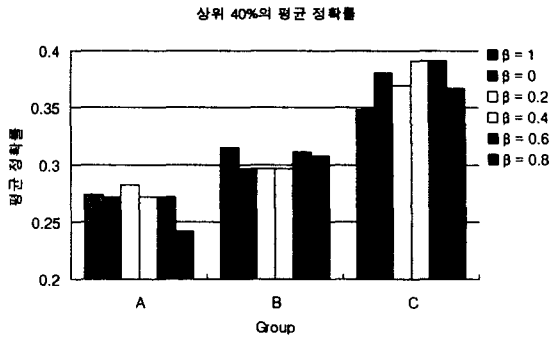
평균 재현율의 결과를 살펴보면 상위 30%와 40%, 50%를 제외한 나머지 경우에서 가중치 β가 1보다 작은 값을 가질 때 즉, 가중치 β가 0.2, 0.4, .0.6, 0.8 일 때의 평균 재현율이 가중치 β가 1인 경우보다 높았다. 그리고, 상위 10%와 20% 그리고 30% 일부 문서들의 평균 재현율을 살펴봤을 때, 가중치 β가 0인 경우 즉, 질의어의 사전적인 의미로만 이루어진 질의어로 검색된 결과가 현저히 높은 평균 재현율을 보였다. 평균 정확률과 마찬가지로 전반적으로 살펴보면 가중치 β가 0.2와 0.4 일 때의 평균 재현율이 다른 경우보다 우세하게 나타났다.

그룹 C의 재현율이 다른 그룹들에 비해서 우세한 것은 그룹별로 새로이 추가되는 용어들의 개수 때문인 것으로 짐작된다. 추가되는 질의어 수가 적을수록 재현율이 증가하는데, 그 이유는 용어들 간에 더 높은 평균 유사도를 가짐으로써 관련 문서들을 많이 찾아주기 때문이라고 추정된다. 가장 좋은 성능을 보이는 경우를 알아보기 위해서는, 그룹별로 추가되는 용어들의 개수와 가중치와의 관계, 그리고 추가되는 용어들의 개수와 정확률, 재현율과의 관계 등을 알아볼 필요가 있다.

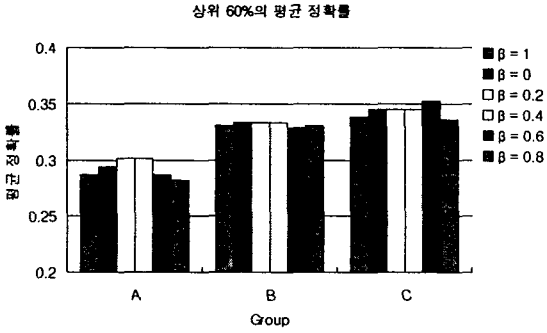
그래서 그룹별로 새로 추가되는 용어들의 개수와 가중치에 따른 정확률, 재현율과의 관계를 알아보았다. 정확률과 재현율 모두에서 우세한 성능을 나타낸, 가중치 β값이 0.2인 경우와 0.4인 경우에 대해 추가된 용어들의 개수와의 관계를 구해보면, [그림15, 16, 17, 18]과 같이 추가 용어 개수가 약 30개 정도일 때 가중치 β가 0.2인 경우와 0.4인 경우 모두에서 정확률, 재현율이 모두 증가되는 것으로 나타났다.



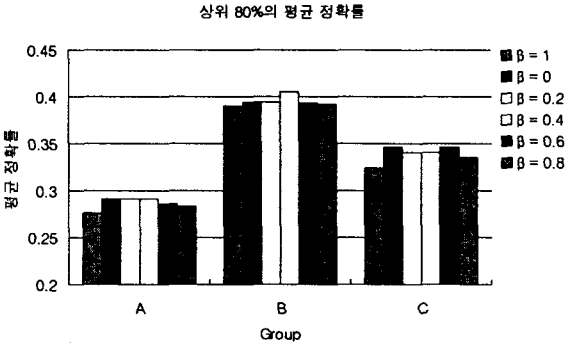
[그림 2] 가중치 β를 다르게 했을 때
-각 그룹별 상위 20%의 평균 정확률



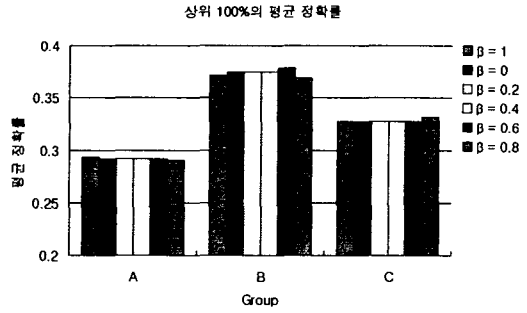
[그림 3] 가중치 β를 다르게 했을 때
-각 그룹별 상위 40%의 평균 정확률



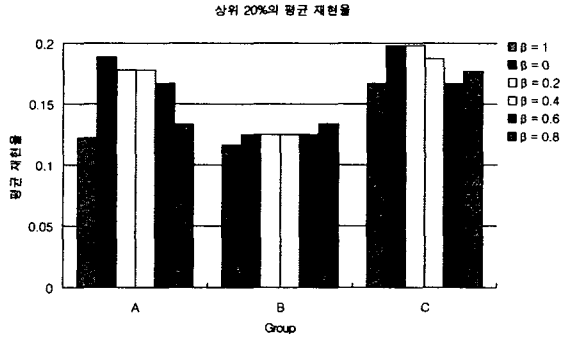
[그림 4] 가중치 β를 다르게 했을 때
-각 그룹별 상위 60%의 평균 정확률



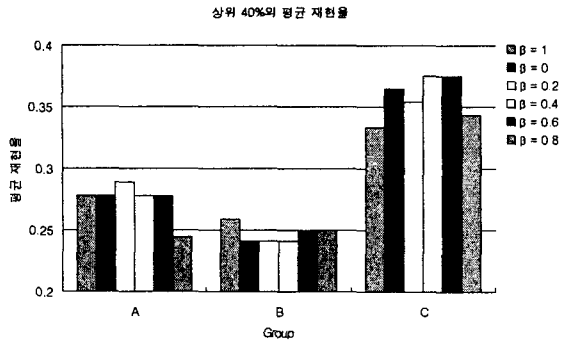
[그림 5] 가중치 β를 다르게 했을 때
-각 그룹별 상위 80%의 평균 정확률



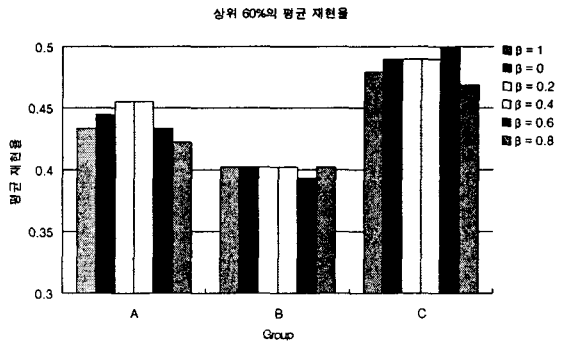
[그림 6] 가중치 β를 다르게 했을 때
-각 그룹별 상위 100%의 평균 정확률



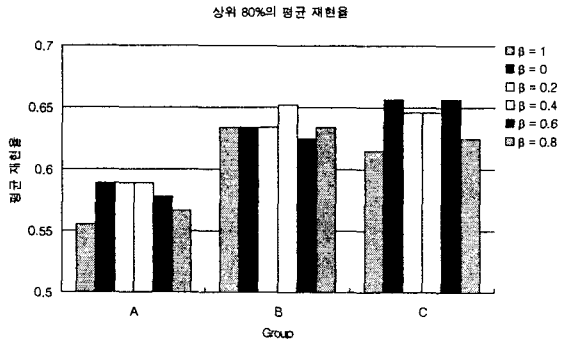
[그림 7] 가중치 β를 다르게 했을 때
-각 그룹별 상위 20%의 평균 재현율



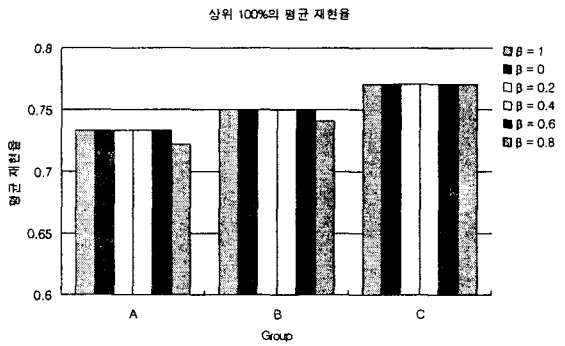
[그림 8] 가중치 β를 다르게 했을 때
-각 그룹별 상위 40%의 평균 재현율



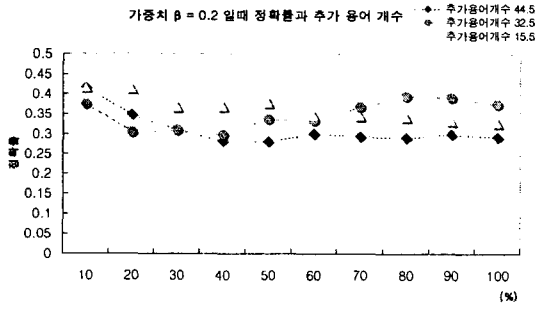
[그림 9] 가중치 β를 다르게 했을 때
-각 그룹별 상위 60%의 평균 재현율



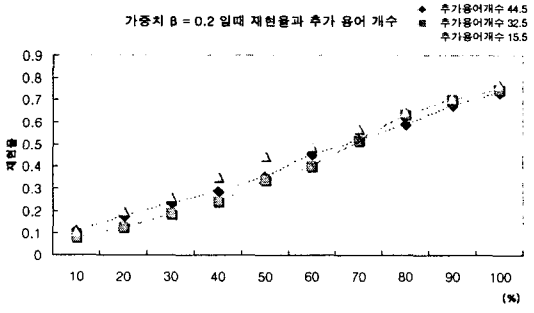
[그림 10] 가중치 β를 다르게 했을 때
-각 그룹별 상위 80%의 평균 재현율



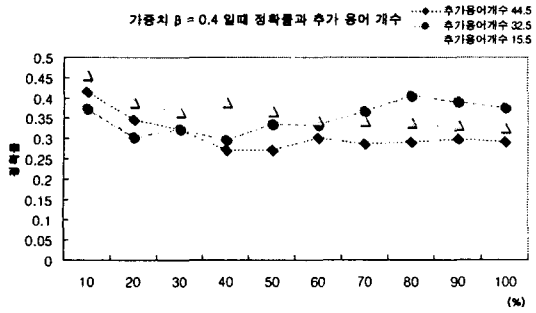
[그림 11] 가중치 β를 다르게 했을 때
-각 그룹별 상위 100%의 평균 재현율



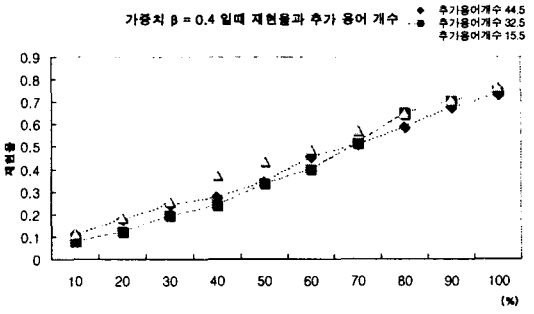
[그림 12] 정확률과 추가용어개수
(가중치 β= 0.2)



[그림 13] 재현율과 추가용어개수
(가중치 β= 0.2)



[그림 14] 정확률과 추가용어개수
(가중치 β= 0.4)



[그림 15] 재현율과 추가용어개수
(가중치 β= 0.4)

5. 결 론

질의어 확장을 하면서 초기 질의어와 새로 추가되는 용어들로 이루어진 질의어에 가중치를 부여하는 것은 정확률과 재현율 측면에서 성능의 변화를 가져온다.

이 때, 새로 추가되는 용어들로 이루어진 질의어들의 가중치를 초기 질의어와 동일하게 하는 것이 최선의 선택이 아님을 알 수가 있었다.

본 연구의 실험 결과를 살펴보면, 초기 질의어의 가중치가 0.2이고 새로 추가되는 질의어의 가중치가 0.8인 경우, 그리고 초기 질의어의 가중치가 0.4이고 새로 추가되는 질의어의 가중치가 0.6인 경우에서 정확률과 재현율이 우세한 성능을 나타냈다. 그리고 이 때 새로 추가되는 용어들의 개수가 오히려 너무 많거나 적게 되면 효율이 저하된다는 것을 알 수 있었다.

본 연구에서는, 전체 문헌 집합 내의 문서와 질의 벡터에 LSI 모델을 적용하여 문서간의 관계를 나타내어 실험을 진행하였다. 그러나 다차원 확장을 위한 LSI의 행렬 계산을 위해서 시스템 상의 많은 리소스가 요구되므로, 데이터의 양을 충분히 확장하는 것이 어려웠다.

그러므로, 대규모 데이터 문헌 집합에 적합한 질의어 확장을 위해서는 일반 IR 시스템에로의 접근이 필요하다. 향후 과제로서, 현재 개발되어 있는 Condor 검색 시스템에 본 연구의 질의어 확장을 적용한 결과를 도출해 낼 것이다.

참 고 문 헌

[1] 김명철, "공기기반 용어간 유사도를 이용한 정보검색 질의확장 비교 연구", 한국과학기술원, 박사학위 논문 1998.11

[2] Fenichel, C. H. 1981. "Online searching: measures that discriminate among users with different types of experiences." Journal of the American Society for Information Science, 32(1): 23-32.

[3] Jansen, B. J. Spink, A., & Saracevic, T. 2000. "Real life, real users, and real needs: a study and analysis of user queries on the Web." Information Processing & Management, 36(2): 207-227.

[4] Kim, M. C., & Choi, K. S. 1999. "A comparison of collocation-based similarity measures in query expansion." Information Processing & Management, 35(1): 19-30.

[5] Richardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[6] 고지현, 오형진, 박순철, "LSI를 이용한 가중치 변화에 따른 클러스터링 결과 분석", 정보처리학회지, 제 9권, 제 2호, pp. 1009-1012, 2002

[7] Gerald J. Kowalski, Mark T. Maybury, Information Storage And Retrieval Systems. Kluwer Academic Publishers, 2000.

[8] Michael W. Berry, Susan T. Dumais, Todd A. Letsche, "Computation Methods For Intelligent Information Access" ACM, 1995.

[9] 정영미, 이재윤, "질의확장 검색에서의 추가용어

가중치 최적화", 한국 정보관리학회지, pp. 241-246, 2001.

최영란 (Young-Ran Choi)
2001년 전북대학교 정보통신공학과 졸업
2003년 전북대학교 대학원 정보통신학과 졸업
현재 (주)ITRONICS 연구원 재직 중

관심분야 : 데이터베이스, 보안

전유정 (You-Jeong Jeon)
2002년 전북대학교 정보통신공학과 졸업
현재 전북대학교 대학원 정보통신 공학과 석사과정
관심분야 : 데이터베이스, 정보검색, 질의어처리, 문서요약

박순철 (Soon-Cheol Park)
1979년 인하대학교 공과대학 졸업
1991년 미국 루이지애나 전자계산학 박사
현재 전북대학교 전자정보공학부 교수

관심분야 : 데이터베이스, 정보검색, 문서요약, 문서클러스터링, 문서분류