論文 97-22-8-20

인터넷에서 잠재적 의미 분석을 이용한 지능적 정보 검색

正會員 임 재 현*, 김 영 찬*

Intelligent Information Retrieval using Latent Semantic Analysis on the Internet

Jae Hyun Lim*, Young Chan Kim* *Regular Members*

※이 연구는 95년도 한국과학재단 연구비지원에 의한 결과임(과제번호 : 95-0100-08-01-3).

요 약

인터넷에서 분산 정보를 검색하는 대부분의 시스템들은 사용자가 요구하는 검색 용어의 의미를 반영하지 못해 관련된 정보를 정확히 찾지 못하고 있다. 본 논문에서는 정보 검색 성능을 향상시키는 방안으로 검색 용어의 의미를 반영할 수 있는 용어 분포에 기반한 자동화 된 질의어 확장을 제안한다. 먼저, 사용자가 부여한 질의어와 전체 문서에서 용어의 중요도를 반영한 가중치(weight)를 계산하고, LSI의 SVD기법을 이용해 모든 문서에서 질의어와 유사하게 출현하는 용어의 분포를 측정하여, 이들 수치와 질의어 용어의 유사성을 측정하였다. 또한 자동적으로 추가할 용어를 줄이기 위한 방안을 연구하였으며 본 논문에서 제안한 방법을 사용해 검색 성능을 평가하였다.

ABSTRACT

Most systems that retrieve distributed information on the Internet have difficulties in retrieving relevant information for they are not able to reflect exact semantics of retrieval queries that users request. In this paper, we propose an automatic query expansion based on term distribution which reflects semantics of retrieval term to enhance the performance of information retrieval. We computed weight, indicating its overall importance in the collection documents and user's query and we use LSI's SVD technique to measure the term distribution which appears similar to query. And also, we measure the similarity to compared numerical value with query terms. Also we researched the method to reduce additional terms automatically and evaluated the performance of the proposed method.

*중앙대학교 컴퓨터공학과
論文番號:97162-0512
接受日字:1997年 5月 12日

Ⅰ. 서 론

현재 인터넷에서 정보 검색이 사용자에게 유용한 수단이 되기 위해서는 더욱 더 "지능화"되야 한다. 정보 검색에서 검색된 문서의 수는 적절한 탐색(search) 용어(term)와 관계가 있으며, 대부분의 정보 검색 시스템은 탐색 용어가 많을수록 검색된 문서는 많아진다. 그러나 검색의 목적은 많은 수의 문서를 검색하는 것이 아니고 검색된 문서의 높은 관련성에 있다. 하지만 이 전에 연구되었던 키워드 기반(key-based)의 연구는 질의어(query)의 용어가 포함되어 있는 문서만을 검색하여, 검색하고자 하는 문서의 용어를 정확히 알지 못하는 경우에는 관련성이 적은 문서가 더 많이 검색되고 있다. 또한 이 방법의 단점은 사용자가 부여한 탐색 용어와 시스템이 문서를 인덱스한 용어가 서로 일치하지 않아 동의어(synonymy)와 다의어(polysemy) 문제를 일으킨다. 이로 인해 사용자는 부적당한 정보를 검색하거나 원하는 정보를 찾지 못하는 용어 문제(vocabulary problem)가 발생한다. 기존의 용어 발생 확률을 이용하는 방법[5]은 용어가 출현하는 확률을 이용하는 것으로 자동화 된 검색을 지원하지만, 검색 효과가 높지 않았다. 또한 많이 이용되고 있는 관련성 피드백(relevance feedback) 기법[6]은 사용자의 관련 정보를 이용하는 것으로 사용자는 초기 질의어를 변화시킴으로써 질의어를 정제하거나 관련된 문서와 관련 없는 문서를 표시하여 다시 질의하는 방식이다. 그러나 이 방법은 사용자의 개입이 필요하며 관련된 정보가 있을 때는 효과가 있으나, 이용 가능한 정보가 없을 때는 사용할 수 없다.

본 연구는 기존 연구의 문제점인 용어 문제, 검색 성능 향상 문제와 자동화 문제를 해결하기 위해 전체 문서에서 나타나는 용어 분포 및 용어 가중치(weight)를 이용해 개념적(concept-based) 검색을 지원하는 질의어 확장 방법을 제안한다. 이 방법은 검색 성능을 향상시키기 위해서 먼저 사용자가 부여한 질의어와 전체 문서에서 용어의 중요도를 반영한 가중치를 계산한다. 용어-문서행렬을 구성할 때 단순히 문서내의 용어 발생 빈도수만을 기초로 하지 않고 문서내에서 용어의 중요도를 반영하여 용어-문서행렬을 구성 한다. 그런 다음에 질의어에 추가하려는 용어를 선택함에 있어 모든 문서에서 질의어와 유사하게 출현하는 용어 분포를 측정해 질의어의 잠재적 의미를 자연스럽게 반영토록하여 용어 문제를 해결한다. 이들 용어의 분포를 쉽게 파악하기 위해서 LSI(Latent Semantic Indexing)의 SVD(Singular Value Decomposition) 기법[3]을 이용하고, 유사성 측정을 위해서는 코사인 계수(cosine coefficient)를 사용한다. 그러나 용어의 수가 많을 때는 유사성 수치 값이 비슷한 것이 많아지고 이들 모두를 질의어에 추가하는 것은 비효율적이기에, 적은 용어를 추가하여 효과적으로 검색 성능을 개선하는 방안을 연구한다.

본 논문의 구성으로 2장에서는 본 연구의 기반이 되는 정보 검색을 위한 벡터-공간 모델을 기술하고, 3장에서는 본 연구에서 이용하고 있는 성능 향상 방법을 실험 DB에 적용하여 성능을 측정한다. 4장에서는 종합적인 평가를 하고, 5장에서 결론을 맺는다.

Ⅱ. 정보 검색을 위한 벡터-공간 모델

본 연구는 벡터-공간 모델을 기반으로 한다. 전통적인 키워드-기반의 정보 검색 기술은 전자(electronic) 정보의 양이 증가함에 따라 유용성이 떨어지고 있다. 사용자가 부여한 질의어에 대한 응답으로 반환되는 정보의 양이 너무 많기 때문에 이질적이고, 거대한 정보의 집합체를 탐색하기가 더욱 어려워지고 있다. 정보 검색을 위한 벡터-공간 모델(Vector-Space Model)은 키워드-기반이기보다는 개념적인 탐색에 이용하는 것으로 질의어에 대한 상대적인 유사성에 따라 탐색 결과를 반환한다. 특히 본 연구에서 용어의 분포를 측정하기 위해 사용한 LSI는 벡터-공간 접근방법의 하나로써 용어-문서 공간상에 용어와 문서를 표현한다[3].

2.1 LSI(Latent Semantic Indexing) 개념

LSI는 용어간의 의존성을 참조하는 벡터 검색의 일종이다. 기존의 대부분의 검색 모델이 용어들을 독립적인 것으로 처리하는데 반해 LSI의 핵심적인 주제는 용어간의 상호 관련성이 자동적으로 유도되고, 검색 효율을 개선시키는데 사용할 수 있다는 것이다. LSI는 연관된 관련성을 모델화하기 위해 통계적 기법인 SVD를 사용하며, 이를통해 거대한 용어-문서

행렬을 k값의 집합체로 분해하는데 보통 100-300 사이의 값을 갖는다. 각각의 용어와 문서는 k-차원의 LSI공간 안에 벡터로서 표현되며 유사한 내용의 문서에 사용된 용어들이 이들 공간 안에 유사한 값을 갖는다.

2.2 SVD를 이용한 용어의 구조 분석

잠재된 용어의 구조 분석은 기하학적인 표현인 용어-문서 행렬을 가지고 시작한다. 이 행렬은 잠재된 구조 모델을 유도할 수 있는 SVD에 의해 분석된다. SVD는 수학적이고 통계학적인 방법이며, 본 논문에서는 SVD를 적용한 수치 값을 얻기 위해 SVDPACKC[8]을 사용하였다.

SVD는 문서간의 의미 구조를 파악하기 위해, 용어-문서(m*n)행렬에 SVD를 적용하여 k개의 벡터를 생성한다. k차로 분해된 벡터는 동일한 의미 공간 안에 문서와 용어를 표현하는데 사용한다. n개의 문서와 m개의 용어를 m*n ($m \geq n$) 행렬(A_k)로 나타내고, k는 인자(factor)의 수, r은 A의 범위(rank)이다. A_k는 3가지 행렬의 곱으로 표현한다.

$$A_k = U \sum V^T \tag{1}$$

여기서 $U^TU = V^TV = I_n$이고, $\sum = diag(\sigma_1, \ldots, \sigma_n)$, $(\sigma_i > 0,\ 1 \leq i \leq r)$, $(\sigma_j = 0,\ j \geq r+1)$이다. 직교(orthogonal) 행렬 U와 V^T의 첫 번째 r열은 AA^T와 A^TA의 r고유치와 관련된 왼쪽과 오른쪽 특이(singular)벡터이고, $\sum$는 대각(diagonal) 행렬로써 특이치 값(singular value)이다.

2.3 측정 요소

검색 시스템의 성능을 평가하는 방법에서 중요하게 고려해야 할 점은 검색 과정의 효율성(efficiency)과 유효성(effectiveness)이다. 사용자 입장에서 고려해 본다면, 일반적으로 가능한 한 관련된 문서들을 많이 검색하고, 관련되지 않은 문서는 가능한 한 적게 검색하는 것이 최적일 것이다. 대표적으로 이 개념을 반영하여 수치적으로 평가하는 방법이 정밀도(precision)와 재현도(recall)이다. 이 두 가지 방법은 서로간의 상반(trade-off)되는 의미를 지닌다. 효과적인 검색을 하기 위해서는 정밀도 값은 감소하지 않으면서 또는 증가하면서, 재현도 값이 증가하도록 하는 것이다. 이는 사용자의 질의에 대해 검색된 문서가 관련되지 않은 문서보다 관련된 문서가 많을 때 가능하며 검색 성능면에서 매우 바람직하다[9].

$$\text{정밀도(Precision)} = \frac{\text{검색된 문서중 관련된 문서의 합}}{\text{검색된 문서의 합}} \tag{2}$$

$$\text{재현도(Recall)} = \frac{\text{검색된 문서중 관련된 문서의 합}}{\text{전체 문서중 관련된 문서의 합}} \tag{3}$$

정밀도와 재현도 값만을 이용하여 검색의 유효성을 평가하기에는 미비한 점이 있다. 예를 들어, 단지 재현도 값만 고려했을 경우, 이 값이 증가한다고 검색 성능이 좋다고 할 수 없다. 정보 검색의 전체적인 성능을 평가하기 위해서 정밀도와 재현도 값을 종합하여 측정한 수 있게 유효성을 측정(E, effectiveness measure)[4]하였다. 이를 정의하면, 식(4)와 같다.

$$E = 1 - \frac{(1+\beta^2)PR}{\beta^2 P + R} \tag{4}$$

여기서, P와 R은 정밀도와 재현도 값이고, β는 사용자가 정의하는 것으로 정밀도에 대한 재현도의 상대적인 중요도를 나타낸다. 예를 들어 $\beta = 2.0$은 정밀도보다 재현도가 두배 더 중요도가 높다는 것이다. 유효성 값은 $(0 \leq E \leq 1)$ 사이의 값을 나타내며, 낮은 수치가 효율적인 검색이다[2][4].

Ⅲ. 성능 향상 방법

기존 연구의 문제점인 용어 문제, 검색 성능 향상 문제와 자동화 문제를 해결하기 위해 전체 문서에서 나타나는 용어 분포 및 용어 가중치를 이용해 개념적 검색을 지원하는 질의어 확장을 수행한다. 본 논문에서는 용어를 추출하는 과정으로 불 용어(stop word) 제거 과정과 스테머(stemmer) 과정[7]을 거쳐 용어의 수를 줄이고 문서내에서 같은 용어이면서 다르게 인덱스 되는 경우를 방지한다. 불 용어로서는 682개를 사용하였고, 스테머 과정을 위해 포터(porter) 알고리즘을 사용하였다.

표 1은 본 논문의 실험을 위해 사용한 데이터 집합

체와 질의어의 특성을 요약한 것이다. CISI는 과학과 관련된 내용으로 문서 요약만을 갖고 있으며, TIME은 타임 잡지의 기사를 발췌하여 만든 실험 모델이다. 성능 향상을 위한 모든 실험은 SUN Enterprise3000 (솔라리스 2.5.1)에서 구현하였으며, 프로그래밍 언어로는 C언어를 사용하였다.

표 1. 데이터 집합체의 특성

종 류	CISI	TIME
- 문서의 수	1460	425
- 용어의 수	7063	14007
- 질의어의 수	112	83
- 질의어와 관련있는 평균 문서 수	50	4
- 문서당 용어의 평균 수	45	190
- 용어당 문서의 평균 수	13	8
- 질의어당 용어의 평균 수	8	8

3.1 인자 수의 선택

거대한 용어-문서 행렬을 k값의 집합체로 분해하는데 있어 차원(dimension) 또는 인자 수의 선택은 매우 중요한 일이다. 차원을 축소하는 것이 발생할 수 있는 많은 잡음(noise)을 제거하지만, 너무 작은 차원을 사용하면 중요한 정보를 잃어버릴 수 있다[1]. 본 연구에서는 성능을 최대화시킬 수 있는 효과적인 k 값을 설정하기 위해 인자수의 다양한 범위를 사용하여 검색 성능을 평가하였고, 이를 통해 상대적으로 작은 인자수를 구하였다. 그림 1은 CISI 실험 DB에 적용해 본 결과로써 재현도의 전 구간(0부터 1까지)을 0.1 간격으로 하여 평균 정밀도를 계산한 것이다. 이와

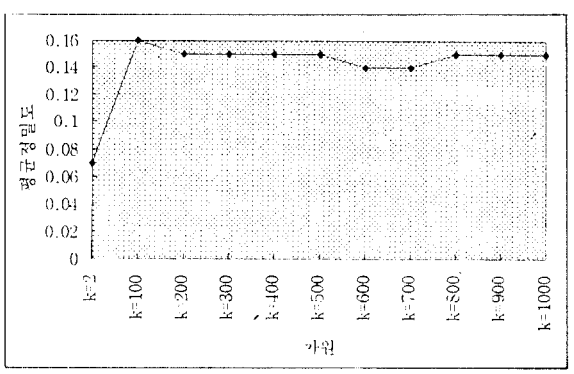

그림 1. CISI에서의 차원 선택에 따른 정밀도 변화

같은 성능의 변화는 TIME 실험DB에서도 관찰되었으며 본 논문에서의 실험은 k값을 100으로 설정하여 평가한다.

3.2 질의어 확장

용어와 문서의 관계를 표현하기 위해 용어-문서 행렬 m(용어의 개수)*n(문서의 개수)을 만들고, 특이치 분해기에 의해 SVD를 적용하면 용어와 문서간에 잠재되어 있는 의미를 파악할 수 있는 분포 수치를 구할 수 있다. 식(1) $A = U \sum V^T$에 의해 세 개의 벡터로 분해되고, k-차원의 공간을 위해 각 벡터에서 k개의 요소들만을 이용하여 용어와 문서의 내용을 반영하고 있는 근사치 행렬을 형성한다.

용어와 문서 벡터를 구한 다음, 질의어 벡터 값을 구하기 위해 먼저, 질의어를 다른 하나의 문서로 취급하여 k-차원의 공간상의 벡터로 표현한다. 이런 가상의 문서는 용어 벡터들의 가중치들의 합이다. 질의어 벡터는 식(5)와 같이 가중치가 부여된 용어들의 벡터로써 정의한다[3].

$$q = q_v^T U_k \sum_k^{-1} \tag{5}$$

여기서, q^T는 인덱스 되어 있는 용어에 질의어가 존재할 때 1값을, 없을 때 0값을 갖는 전치행렬이다. U_k는 앞에서 구한 용어 벡터의 k개 열이고, $\sum_k^{-1}$는 특이치 벡터의 k개 고유벡터의 역행렬이다.

질의어 확장에 필요한 질의어와 용어들간의 유사성 측정을 위해 각 용어와 질의어 벡터와의 관계를 코사인 계수 식(6)을 이용하였다[2].

$$\mathrm{SimTQ}(t_i, q) = \frac{\sum t_i * q_i}{\sqrt{\sum t_i^2 * \sum q_i^2}} \tag{6}$$

여기서, $\mathrm{SimTQ}(t_i, q)$는 i번째 용어와 질의어 벡터의 유사성 값, t_i는 i번째 용어 벡터를 의미하고, q는 질의어 벡터를 말한다.

표 2는 두 개의 실험 DB CISI와 TIME에서 k값을 100으로 설정하여 질의어 확장을 수행한 결과이다. 재현도의 구간을 전구간과 부분구간으로 나누어 실험을 하였으며, 전구간의 간격은 0.1로하였고, 부분구간은 재현도의 값이 0.2, 0.5, 0.7일 때 정밀도 평균값을

측정한 것이다. 괄호 안의 수치는 키워드-기반인 VSM과 비교하여 향상된 정밀도 개선율을 나타내며, 확장되는 용어의 수는 유사성 수치가 높은 순서대로 상위 100개부터 400개를 선택하여 실험하였다. 확장되는 용어 수에 대한 표기는 "qe숫자"를 사용하였으며, 표 2에서 알 수 있듯이 추가되는 용어의 수가 200(qe200)일 때 평균 21%의 높은 검색 개선율을 보이고 있다.

표 2. LSI를 이용한 질의어 확장 결과

항목	VSM	qe100	qe200	qe300	qe400
CISI (전구간)	0.142	0.179 (26.06%)	0.179 (26.06%)	0.173 (21.83%)	0.166 (16.9%)
CISI (부분구간)	0.147	0.183 (25%)	0.187 (27.27%)	0.187 (27.27%)	0.187 (27.27%)
TIME (전구간)	0.628	0.718 (14.33%)	0.725 (15.45%)	0.713 (13.54%)	0.71 (13.06%)
TIME (부분구간)	0.69	0.777 (12.56%)	0.793 (14.98%)	0.793 (14.98%)	0.79 (14.49%)
정밀도 개선율		19.49%	20.94%	19.41%	17.93%

표 3. LSI와 유사성 수치 간격을 이용한 질의어 확장결과

항목	VSM	iqe30	iqe50	iqe70	iqe100
CISI (전구간)	0.142	0.172 (21.13%)	0.173 (21.83%)	0.178 (25.35%)	0.175 (23.24%)
CISI (부분구간)	0.147	0.163 (11.36%)	0.173 (18.18%)	0.18 (22.73%)	0.183 (25%)
TIME (전구간)	0.628	0.727 (15.76%)	0.733 (16.72%)	0.736 (17.2%)	0.729 (16.08%)
TIME (부분구간)	0.69	0.783 (13.53%)	0.797 (15.46%)	0.793 (14.98%)	0.79 (14.49%)
정밀도 개선율		15.45%	18.05%	20.07%	19.68%

문서의 수가 많아지면 용어의 수가 수십만에 이르기 때문에 유사성 수치 값이 서로 겹치거나 인접하는 일이 발생한다. 이것은 같은 지역에 속하는 용어들은 거의 같은 문서에서, 같은 분포(빈도수)를 갖는다는 의미이기 때문에 같은 지역에 분포하는 용어들을 전부 질의어에 추가하는 경우 검색 성능에 큰 변화를 가져오지 않는다. 따라서 LSI를 적용하는 경우에는 같은 지역에 분포하고 있는 일정량의 용어들은 제거하고, 그 중의 대표가 되는 용어만을 적용하는 것이 효과적이다. 실험을 통해 알아본 결과, 코사인 계수 값이 0.9정도면 질의어에 추가할 용어의 대부분을 선택할 수 있다. 대표가 되는 용어를 선택하기 위해 용어-문서 행렬에 SVD를 적용하여 나온 용어 벡터 값에서 일정 값(0.001)이내에 속하는 것들을 제거한다. 그러면 그 중 가장 큰 값이 선택된다. 이 같은 방법으로 용어를 선택함으로써 본 연구에서 제안한 방법이 다른 방법들보다 훨씬 적은 용어를 질의어에 추가하더라도 검색의 성능은 떨어지지 않는다는 장점을 갖는다.

표 3은 유사성 수치 값의 간격을 0.001로 하여 추가되는 용어를 선택하고 질의어 확장을 수행한 결과이다. 여기서 "iqe숫자"는 유사성 수치 간격을 이용하여 질의어 확장이 되는 용어의 수를 나타내며, 결과에서 보듯이 70개만을 추가하여도 표 2의 결과에 못지 않은 20%의 검색 성능 향상을 나타냈다.

3.3 용어의 가중치(weight)

정보검색에서 검색 성능을 향상시키기 위해서 용어-문서행렬을 구성할 때 단순히 문서내의 용어 발생 빈도수만을 기초로 하지 않고 문서내에서 용어의 중요도를 반영하여 용어-문서행렬을 구성할 수 있는데 이것을 용어 가중치라고 한다[2]. 용어 가중치에는 모든 문서에서의 용어의 중요도를 반영하는 전역 가중치(global weight)와 한 문서내에서의 용어의 중요도를 반영하는 지역 가중치(local weight)가 있다. 보통 용어-문서행렬을 구할 때는 용어의 전역 가중치와 지역 가중치를 곱하여 사용한다. 본 연구에서는 [1]의 연구 결과에 따라 가장 우수한 것으로 알려져 있는 지역 가중치 $\log(TF+1)$와 전역 가중치 $1-entropy$를 사용 하였다.

표 4는 가중치를 부여하고 LSI(k = 100)를 적용하여 실험한 결과이다. 여기서 "wqe숫자"는 가중치를 이용해 질의어 확장이 되는 용어의 수를 나타내며, 결과에서 알 수 있듯이 모든 경우에서 단순히 빈도수만을 기초로하여 질의어 확장을 한 것 보다 우수한 결과를 보인다. 표 3과 동일하게 추가되는 용어의 수가 200개일 때 34%의 검색 성능 개선율을 갖으며, 단순히 질의어 확장을 한 경우보다 13%의 성능 향상을

보인다.

표 4. LSI와 가중치를 이용한 질의어 확장 결과

항목	VSM	wqe100	wqe200	wqe300	wqe400
CISI (전구간)	0.142	0.187 (31.69%)	0.196 (38.03%)	0.198 (39.44%)	0.2 (40.85%)
CISI (부분구간)	0.147	0.197 (34.09%)	0.21 (43.18)	0.207 (40.91%)	0.217 (47.73%)
TIME (전구간)	0.628	0.813 (29.46%)	0.792 (26.11%)	0.795 (26.59%)	0.768 (22.29%)
TIME (부분구간)	0.69	0.897 (29.95%)	0.88 (27.54%)	0.873 (26.57%)	0.83 (20.29%)
정밀도 개선율		31.3%	33.73%	33.38%	32.79%

표 5. LSI, 가중치 및 유사성 수치간격을 이용한 질의어 확장 결과

항목	VSM	wiqe30	wiqe50	wiqe70	wiqe100
CISI (전구간)	0.142	0.181 (27.46%)	0.192 (35.21%)	0.194 (36.62%)	0.189 (33.1%)
CISI (부분구간)	0.147	0.19 (29.55%)	0.207 (40.91%)	0.22 (50%)	0.203 (38.64%)
TIME (전구간)	0.628	0.795 (26.59%)	0.806 (28.34%)	0.806 (28.34%)	0.801 (27.55%)
TIME (부분구간)	0.69	0.877 (27.05%)	0.88 (27.54%)	0.877 (27.05%)	0.87 (26.09%)
정밀도 개선율		27.66%	33%	35.5%	31.35%

표 5는 표 3과 동일하게 유사성 수치 간격을 0.001로 하여 용어를 선택하고 질의어 확장을 수행한 결과이다. 여기서 "wiqe숫자"는 가중치와 유사성 수치 간격을 이용하여 질의어 확장이 되는 용어의 수를 나타내며, 추가되는 용어의 수가 70개 정도일 때 평균 검색 성능 향상 36%라는 가장 우수한 결과를 보인다. 결국 용어의 수가 많을 때는 유사성 수치 값이 비슷한 것을 제거하여 질의어에 추가하는 것이 효율적이며, 적은 용어를 추가하여도 효과적으로 검색 성능을 개선함을 알 수 있다.

Ⅳ. 평 가

정보 검색의 전체적인 성능을 평가하기 위해서 정밀도와 재현도 값을 종합하여 측정할 수 있게 식(4)를 이용하여 유효성을 측정하였다.

대부분 검색 시스템의 단점은 재현도 값이 상승할 때, 정밀도 값이 떨어진다는 것이다. 본 논문에서는 정밀도와 재현도의 중요도를 동일하게 설정하고, 유효성 측정(E) 값을 얻기 위해 β를 1로하여 검색의 성능을 측정한다. X축은 재현도 값이 0.2, 0.5, 0.7일 때를 기준으로 하고, Y축을 E 변동률로 하였다.

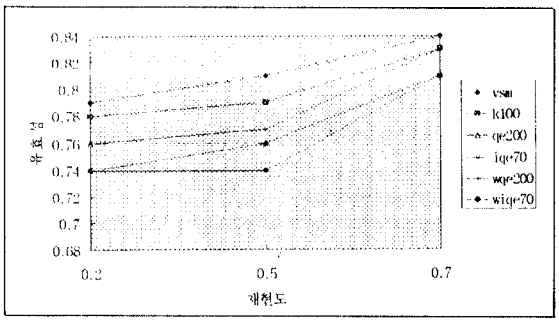

그림 2. CISI에서의 유효성

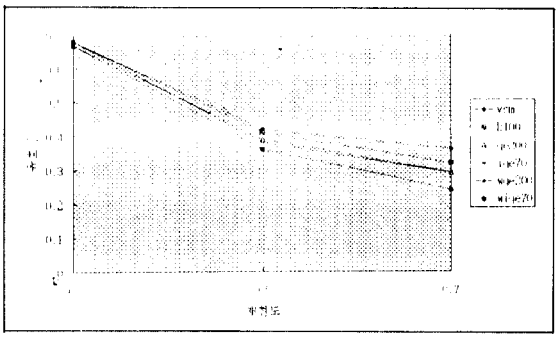

그림 3. TIME에서의 유효성

그림 2는 실험 DB CISI를 대상으로 측정한 결과이고 그림 3은 실험 DB TIME을 대상으로 측정한 것이다. 단순히 질의어 확장을 수행하면서 유사성 수치값에 따라 200개 용어를 질의어에 확장한 것(qe200), 유사성 수치 간격을 0.001로 하여 70개 확장한 것(iqe70), 가중치를 이용하면서 유사성 수치값에 따라 200개를 질의어에 확장한 것(wqe200)과 유사성 수치 간격을 0.001로 하여 70개를 확장한 것(wiqe70)을 측정하였

다. 결과의 해석은 유효성 값이 낮은 것이 검색 성능이 높은 것이다. 그림에서 알 수 있듯이 가중치를 부여하고 유사성 수치값을 0.001 간격으로 하여 용어를 선택하고 질의어 확장을 수행한 결과(itleg70)가 가장 우수하다.

지금까지의 실험을 종합하여 평가하면 그림 4와 같다. 정밀도 개선율은 키워드-기반인 VSM과 비교하여 측정된 수치이며, LSI를 이용해 용어 출현 분포도만을 기준으로 질의어 확장을 한 경우 평균 20%개선되었고 가중치를 부과하여 질의어 확장을 한 경우 평균 36% 개선 되었다.

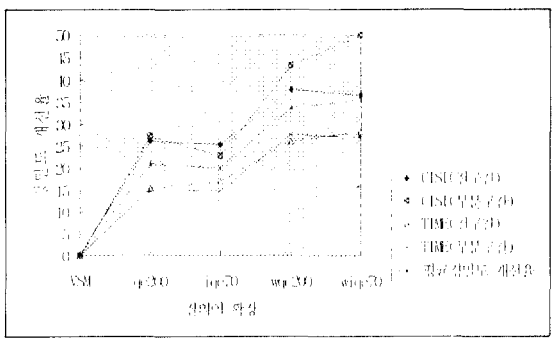

그림 4. 종합평가

V. 결 론

본 논문은 인터넷상에서 개념적 검색을 지원하기 위한 방안으로 수많은 정보들에 존재하는 잠재적 의미를 파악하여 사용자가 원하는 관련된 정보를 지능적으로 검색하는 자동화된 질의어 확장을 연구하였다. 이를 위해 용어-문서행렬을 구성할 때 단순히 문서내의 용어 발생 빈도수만을 기초로 하지 않고 문서내에서 용어의 중요도를 반영하는 가중치를 적용하였고, 사용자의 질의어 개념을 반영하는 용어의 분포를 측정하기 위해 LSI의 SVD 기법을 이용하였다. 단지 사용자가 초기에 검색하기 위해 부여한 질의어만이 아니라 질의어 개념을 내포하고 있는 유사한 용어를 찾아내어 초기 질의어에 추가함으로써 사용자가 원하는 관련된 문서를 더 많이 검색할 수 있도록 하였다. 또한 질의어 확장시 서로 유사한 유사성 값을 가지는 용어를 질의어에 모두 추가한다는 것은 상당한 오버헤드가 생기기 때문에, 본 논문에서는 이들 용어들 중에서 대표가 될 수 있는 용어를 임계치를 주어 선택하였다. 유효성 측정 결과에서 알 수 있듯이 적은 용어를 추가하여 확장시킨 본 연구의 결과가 키워드-기반인 VSM보다 평균 36% 향상되었음을 알 수 있다. 하지만 본 논문에서 제안한 방법은 검색 성능의 유효성만을 고려하여 초기 질의어보다 더 정확하게 관련된 문서를 검색할 수 있지만, 검색 효율성(efficiency)을 고려한다면 몇 가지 문제들이 있다. 향후 계획으로 정보 검색의 효율성을 위한 연구를 추가하여 사용자 입장에서 검색의 유효성과 효율성을 모두 만족시킬 수 있는 방안을 연구하고자 한다.

참 고 문 헌

1. M. W. Berry, S. T. Dumais, and T. A. Letsche, "Computational Methods for Intelligent Information Access", Proceedings of Supercomputing'95, San Diego, CA, December 1995.
2. Willian B. Frakes, Ricardo Baeza-Yates, "Information Retrieval: Data Structure and Algorithms", Prentice Hall, 1995.
3. S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshan, "Indexing by latent semantic analysis", Journal of the American Society for information Science, 41(1990).
4. Shih-Hao Li, "Internet Resource Discovery-topical clustering and visualization using latent semantic indexing", Southern California Univ., August 1996.
5. Qiu, Y. and Frie, H. p., "Concept based query expansion", In Proc. of the 16th International ACM SIGIR Conf. on R & D in Information Retrival, pp. 160-169, New York. 1993.
6. Gerard Salton, Chirs Buckery, "Improving Retrieval Performance by Relevance Feedback", Journal of the American Society for Information Science, pp288-297,41(1990).
7. David A. Hull, "A Detailed Analsis of English Stemming Algorithms", Rank Xerox Research Centre, January 1996.
8. M. W. Berry al./SVDPACKC: Version 1.0 User's Guide, Tech. Rep. CS-93-194, University of Ten-

nessee, Knoxville, TN, October 1993.
9. Michael Gordon, Manfred Kochen, "Recall-Precision Trade-off: A Derivation", Computer and information Systems, The Univ. of Michigan, 1989.

임 재 현(Jae Hyun Lim) 정회원
1986년 2월 : 중앙대학교 전자계산학과(이학사)
1988년 8월 : 중앙대학교 대학원 전자계산학과(이학석사)
1994년 9월~현재 : 중앙대학교 일반대학원 컴퓨터공학과 박사과정
※주관심분야 : 운영체제, 분산시스템, 인터넷

김 영 찬(Young Chan Kim) 정회원
1965년 2월 : 연세대학교 전기공학과(공학사)
1968년 8월 : 연세대학교 대학원 전자공학과(공학석사)
1983년 2월 : 연세대학교 대학원 전자공학과(공학박사)
1982년 12월~1983년 12월 : 프랑스 Grenoble 대학 연구교수
1996년 1월~1996년 12월 : 한국정보과학회 회장
1995년 9월~현재 : 중앙대학교 정보산업대학원 원장
1973년 3월~현재 : 중앙대학교 컴퓨터공학과 교수
※주관심분야 : 운영체제, 분산시스템, 망관리