정보 검색을 위한 용어 분포 임계치 모델

임 재 현† · 민 태 홍††

요 약

인터넷에서 전자 정보의 양이 증가함으로써 관련 정보만을 자동으로 검색하는 방법이 중요하다. 전통적인 정보 검색 시스템의 결점은 사용자가 부여한 탐색 용어가 시스템이 색인한 용어와 다르기 때문에, 부정확한 정보를 검색하거나 정확한 정보를 놓치게 된다. 본 논문에서는 검색 성능 향상을 위해 용어 분포에 기반한 질의어 확장을 사용하며, 용어 분포 임계치를 설정하여 효과적으로 검색 성능을 개선하는 방안을 연구한다.

Term Distribution Threshold Models for Information Retrieval

Jae-Hyun Lim† · Tae-Hong Min††

ABSTRACT

With the increasing availability of information in electronic form, it becomes more important and feasible to have automatic methods to retrieve relevant information in the Internet. A deficiency of traditional information retrieval systems is that search terms are often different from those indexed by the systems. Thus, users may either retrieve wrong information or miss what they really want. In this paper, we used an automatic query expansion based on term distribution to enhance the performance of information retrieval. Also this thesis proposed the method for setting the threshold according to area distribution in order to choose additional terms.

1. 서 론

인터넷에 대한 관심이 급격히 증가함에 따라 다양한 정보 서버 구축을 통한 고급의 정보 제공 서비스가 필요하게 되었고, 이들 서비스는 초고속 통신망의 구축으로 보다 가깝게 현실화되고 있다. 현재 이용 가능한 정보의 양은 5년마다 두배로 증가하며, 곧 4년마다 두배로 증가할 것이라고 한다[1]. 불행하게도 이용 가능한 정보의 양은 지수적 비율로 증가하는데, 정보를 탐색하고 유용한 팩터를 유도하는 능력은 감소하고 있다. 현재 인터넷에서 정보 검색의 문제는 정확도(precision)가 부족한 것으로 검색된 정보의 평균 50% 가량이 관련 없는 정보이다. 또 다른 문제는 재현도(recall)의 실패로 이용 가능한 관련된 정보의 20% 가량만을 검색하고 있다[6]. 이처럼 정확도 부족과 재현도 실패의 가장 큰 원인은 사용자가 부여한 탐색 용어와 시스템이 문서를 인덱스한 용어가 서로 일치하지 않아 동의어(synonymy)와 다의어(polysemy) 문제를 일으키는데 있다. 이로 인해 사용자는 부적당한 정보를 검색하거나 원하는 정보를 찾지 못하는 용어 문제

※ 이 논문은 1998년 학술진흥재단의 학술연구비에 의하여 지원되었음.
† 정 회 원 : 공주문화대학 컴퓨터정보과 전임강사
†† 종신회원 : 인하공업전문대학 컴퓨터정보과 교수
논문접수 : 1999년 12월 17일, 심사완료 : 2000년 4월 20일

(vocabulary problem)가 발생한다.

본 논문은 기존 연구의 문제점인 용어 문제, 검색 성능 향상 문제를 해결하기 위해 전체 문서에서 나타나는 용어 분포를 이용해 개념 기반(concept-based) 검색을 지원하는 질의어 확장 방법을 이용한다. 이들 용어의 분포를 파악하기 위해서 특이치 분해(SVD : Singular Value Decomposition) 기법[2]을 이용하고, 유사성(similarity) 측정을 위해서는 코사인 계수(cosine coefficient)를 사용한다. 특이치 분해를 이용하는 대부분의 연구[2, 7]가 질의어 벡터와 문서 벡터간의 유사성만을 측정하여 관련된 문서를 검색하였지만, 본 논문에서는 질의어 벡터와 용어 벡터간의 유사성을 먼저 측정하여 전체 문서에서 질의어와 사용 분포도가 유사한 용어를 선택하여 질의어에 추가하는 방법을 사용한다. 또한 기존의 연구에서는 질의어에 용어를 추가할 때 질의어 용어와 유사성이 높은 용어가 추가되는 경향이 있다[5]. 하지만 인덱스 한 용어의 수가 많을 때는 유사성 수치 값이 비슷한 것이 많아지고 이들 모두를 질의어에 추가하는 것은 비효율적일 수 있다. 오히려 자주 사용되는 용어가 관련 있는 문서와 관련 없는 문서를 식별하는데 어려움을 주기 때문에, 용어의 선택시 제거하는 것이 유용 할 수 있다[4]. 본 논문에서는 용어를 추가하는데 있어 임계치 방식을 통한 새로운 방식을 제안하여 사용자가 질의어를 부여하는 초기에 용어 문제를 해결하여 검색 성능을 개선하도록 한다.

2. 기반 기술

2.1 용어 문제

정보 검색을 위한 대부분의 접근 방법은 사용자의 질의어와 문서 용어간의 정확한 키워드 매칭에 기반한다. 전형적으로 질의어 용어를 포함한 문서가 사용자에게 반환된다. 이 같은 방법은 관련된 문서가 사용자의 질의어에 포함된 용어를 갖지 못한 경우 검색에 실패한다. 실패하는 이유는 불리안, 표준 벡터, 확률 등을 이용하는 표준 검색 모델들이 용어들을 독립적으로 처리하기 때문이다. 예를 들어, "automobile," "car," "driver," "elephant"라는 용어를 생각해 보자. "automobile"과 "car"는 동의어이며 "driver"는 개념적으로 관련이 있고 "elephant"는 전혀 관련이 없다. 대부분의 검색 시스템에서 질의어 "automobile"은 "car"에 대한 문서를 검색하지 못한다. 그러나 "automobile"에 대한 질의어가 "car"를 포함한 문서와 내용 면에서 약간 관련이 있는 "driver"를 포함한 문서를 검색할 수 있는 것이 바람직하다.

결론적으로 기존의 키워드-기반 검색 시스템의 가장 큰 문제점은 사용자의 탐색 용어와 시스템이 문서를 인덱스 한 용어가 서로 일치하지 않는다는 것이다. 이로 인해 사용자는 부적당한 정보를 검색하거나 원하는 정보를 찾지 못하기도 한다. 이것을 용어 문제라고 한다[7]. 일반적으로 용어 문제에는 2가지 종류가 있는데 하나는 동의어 문제이고, 다른 하나는 다의어 문제이다. 동의어 문제는 다른 용어를 사용해 같은 정보를 표현하는 것이고, 다의어 문제는 다른 의미에 대해 동일한 용어를 사용하는 것이다. 동의어의 영향은 검색 시스템 성능의 재현도를 감소시키고, 다의어의 영향은 정확도를 감소시킨다.

2.2 개념 기반 검색

용어 문제를 해결하기 위해 본 논문에서는 문서 안에 출현하는 용어의 유사성을 측정하고, 유사한 문서에 분포하는 용어를 의미 공간(semantic space)에 표현하여 벡터 표현이 서로간에 근접하도록 하였다[5]. 이들 의미 공간을 이용해 동의어와 다의어 문제를 접근해 감으로써 검색 성능의 정확도와 재현도를 개선시킨다.

〈표 1〉 용어와 문서간의 관계

용어 \ 문서	문서1	문서2	문서3	문서4	문서5
automobile	1		1		
car		1			
information			1		1
oil	1	1		1	
HyunDai	1	1			
handle	1	1		1	1
air			1	1	

<표 1>을 이용해 용어의 출현 패턴에 내포되어 있는 의미 구조를 파악하기로 한다. 여기서 수치 값은 각각의 문서에 분포하는 용어의 빈도 수를 나타낸 것이다. 키워드 기반일 경우 "automobile"라는 용어를 사용해 검색한다고 할 때 문서1과 문서3이 검색된다. 그러나 문서1과 문서2에 출현하는 용어의 분포로 보아

거의 같은 내용을 갖는 문서들이다. 여기서 용어 분포를 이용한다면, 문서1과 문서2는 같이 검색될 것이고 자연스럽게 동의어("automobile"과 "car") 처리가 가능해 진다. 또 다른 예로써 "air"를 사용한다고 할 때, 이것의 다의어는 의미가 전혀 다르다. 키워드 기반인 경우 문서3과 문서4가 검색될 것이다. 그러나 이들의 용어 출현 분포를 보면 전혀 다른 내용을 갖는 문서들이다. 용어 분포에 따른 의미 구조를 이용해 검색을 한다면, 문서3과 문서4는 다른 내용으로 파악되어 완전하지는 않지만 다의어 처리도 가능해 진다.

2.3 특이치 분해

특이치 분해는 직사각형의 행렬을 매우 특별한 형태의 3개의 행렬로 분해하는 것으로 "특이치 벡터(singular vectors)"와 "특이치 값(singular values)"을 갖는다. 이들 행렬은 본래의 행렬이 갖고 있는 관련성을 독립적인 구성 요소 또는 팩터로 분해한다. 이들 구성 요소는 매우 작으며, 작은 차원(dimension)을 갖는 수많은 근사치 모델을 형성하게 한다. 이와 같이 축소된 모델에서 모든 용어-용어, 문서-문서, 용어-문서의 유사성이 차원을 구성하는 값에 의해 근사치 된다[2]. 두 객체간의 유사성은 이들을 표현하는 벡터간의 도트 적(dot product) 또는 코사인 값으로 알 수 있으며 기하학적으로 표현할 수 있다. 축소된 차원이 갖는 장점은 용어 사용에 약간의 차이가 있는 문서들도 같은 벡터를 가질 수 있다는 것이다. 바로 이 같은 특성이 신뢰하기 어려운 데이터를 효과적으로 개선할 수 있게 한다. 유도된 차원 또는 팩터를 사용해 원래의 용어-문서 행렬에 근사한 값을 얻을 수 있는데, 서로 다른 용어와 문서의 공통적 의미를 추출해 표현하기 때문에 이들 팩터는 지능적인 개념으로 생각할 수 있다. 여기서 용어와 문서 벡터는 이들 기반이 되는 각각의 개념과의 관련성 정도를 나타낸다. 다시 말해 용어, 질의어, 문서의 "의미(meaning)"는 팩터에 의해 정의된 k-공간에 벡터 위치로서 k 팩터 값을 표현한다. 의미 표현은 N개의 인덱스 용어를 $k < N$ 로 대체하여, 근사치로 접근하기 때문에 상당히 경제적이다.

본 논문에서는 팩터 분석에 관심을 갖지 않으며 용어, 문서, 질의어를 의미 공간으로 표현하여 용어 사용의 문제점을 제거하는데 목적을 갖는다. 따라서 본 논문에서는 특이치 분해를 적용한 계산 결과를 얻기 위해 SVDPACKC[8]을 사용하였다.

용어-문서(m * n)행렬 A에 특이치 분해를 적용하면 식 (1)과 같이 3가지 행렬의 곱으로 분해된다[2]. 여기서 U와 V는 왼쪽과 오른쪽 특이치(singular) 벡터인 직교(orthogonal) 행렬이고 $\sum$는 특이치 값(singular value)으로 대각(diagonal) 행렬이다[3].

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

특이치 분해의 장점은 작은 행렬을 사용하여 최적의 근사치를 구하도록 제공한다[2]. 근사치 행렬을 생성할 때 중요한 것은 k 차원의 선택이며, k 차원이 선택되면 크기 순으로 정렬되어 있는 특이치 값 $\sum$에서, 처음 k개의 가장 큰 수만을 유지하고 나머지는 0으로 설정한다. 이것을 식으로 표현하면 식 (2)와 같다.

$$A_k = U_k\sum_k V_k^T \tag{2}$$

질의어의 처리는 축소된 용어-문서 공간 안에 질의어의 위치를 표현하기 위해 질의어를 다른 하나의 가상문서로 취급하여 k-차원 공간상의 벡터로 표현한다. 질의어 벡터는 식 (3)과 같이 가중치가 부여된 용어들의 벡터로써 정의한다[2].

$$q = q^T U_k \sum_k^{-1} \tag{3}$$

질의어 확장에 필요한 질의어 벡터와 용어 벡터간의 유사성 측정을 위해 각 용어 벡터와 질의어 벡터와의 관계는 식 (4)를 이용한다. t_k는 문서 안의 용어 k번째 값이고, q_k는 질의어안의 용어 값이다.

$$\mathrm{Sim}(d, q) = \frac{\sum_k t_k * q_k}{\sqrt{\sum_k t_k^2 * \sum_k q_k^2}} \tag{4}$$

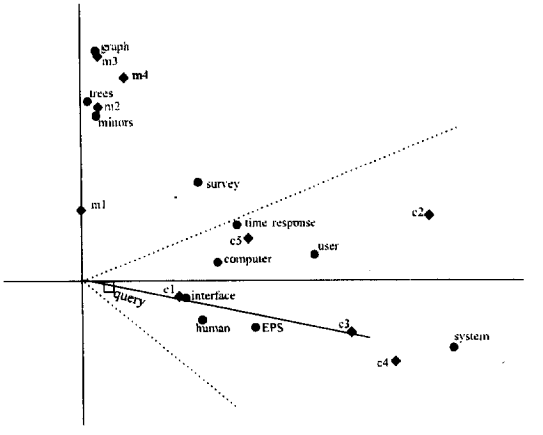
기존의 특이치 분해 기법을 이용한 정보 검색 방법을 예로 들어 설명한다. <표 2>와 같이 실험 데이터로는 9개의 문서 제목과 이 문서를 구성하고 있는 용어를 사용한다. 문서 집합체에 출현하는 29개의 용어 중 두 개 이상의 문서에 출현한 12개 용어만을 이용하고, 질의어로는 "human computer"를 사용한다. 이 문서 집합체는 두 부류로 구성되어 있는데 c1-c5 문서가 같은 내용의 문서이고, m1-m4가 같은 종류의 문서이다. 실제 질의어를 통해 검색하고자 하는 문서는 c1-c5이다.

〈표 2〉 문서의 용어와 질의어

문서	제 목
c1	*Human* machine *interface* for Lab ABC *computer* application
c2	A *Survey* of *user* opinion of *computer system response time*
c3	The *EPS user interface* management *system*
c4	*System* and *human system* engineering testing of *EPS*
c5	Relation of *user-perceived response time* to error
m1	The generation of random, binary, unordered *trees*
m2	The intersection *graph* of paths in *trees*
m3	*Graph minors* IV : Widths of *trees* and well-quasi-ordering
m4	*Graph minors* : A *survey*
질의어	*human computer*

용어와 문서간의 관계를 행렬로 표현하고, 축소된 의미 공간을 구축하기 위해 식 (2)를 적용하며, k값을 2로 설정하여 근사치 행렬을 구한다. 특이치 분해를 통해 얻은 용어와 문서 벡터 값을 이용하여 2차원의 그래프를 그려보면 (그림 1)과 같다. 질의어 벡터 값은 식 (3)에 따라 계산하며, (그림 1)의 □는 질의어를 표현한다. 원점에서 질의어 좌표를 연결하여 기준 축으로 삼을 때, 점선으로 표현한 영역은 질의어를 중심으로 코사인 계수 값이 0.9내에 있는 문서들을 나타낸다. 이 영역 안에 포함된 문서 c1-c5가 질의어와 유사성이 높은 문서들이며, 문서 m1-m4는 질의어와 관련 없는 문서들이다. 본 논문에서는 유사성 측정으로 코사인 측정 방법을 사용했기 때문에 질의어를 중심으로 거리가 가까운 것이 유사성이 높은 게 아니다. 오히려 기준 축을 중심으로 근접한 각도 상에 위치한 문서가 유사성이 높다. 그래프 상에서 가장 근접한 각도에 위치한 문서는 c1과 c3이다. 여기서 중요한 사항은 c3와 c5 문서가 질의어와 관계된 용어를 하나도 갖고 있지 않다는 것이다.

그래프를 통해 용어간의 관계도 파악할 수 있다. 질의어 기준 축을 중심으로 점선으로 표현한 영역 안에 "system," "interface," "EPS," "human," "user," "computer"가 있다. 결국 이들 용어는 문서 집합체에서 사용 분포도가 유사하다는 것이다. 다른 지역에 있는 용어들은 서로간에 유사한 사용 분포도를 형성하지만 점선 내부에 있는 용어들과는 전혀 다르게 사용된 것이다.



(그림 1) 질의어 벡터와 용어, 문서 벡터간의 관계

3. 용어 분포 임계치를 이용한 검색 성능 개선

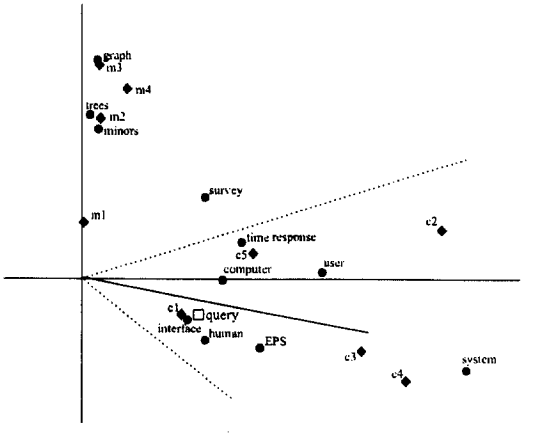
본 논문에서는 질의어와 유사하게 출현하는 용어의 분포를 파악하기 위해 특이치 분해를 사용하고 유사성을 측정한다. 유사성을 판단하는 대표적인 접근 방법은 거리와 각도를 측정하는 것이다. 거리 측정은 단지 문서의 그룹에 의존하는 것으로, 문서 공간의 한 점(point)에 대해 모든 방향은 동일하게 간주한다. 즉, 유사성은 한 점으로부터 문서가 얼마나 멀리 떨어져 있느냐가 기준이 된다. 그러나 각도 측정은 고정된 한 점으로부터 문서 공간의 뷰(view)를 표현하는 것으로, 원점으로부터 각 문서간의 거리보다는 방향을 중요하게 생각한다. 따라서 두 문서가 문서 공간에서 멀리 떨어져 있어도 원점으로부터 동일한 벡터를 가질 수 있다. 본 논문에서는 이와 같이 각도 측정에 기반 하는 코사인 계수의 특성을 질의어 용어 선택에 이용하고, 또한 확장된 질의어와 문서간의 유사성 측정에도 사용한다. 식 (4)를 이용해 질의어 벡터와 용어 벡터간의 유사성을 먼저 측정하고, 전체 문서에서 질의어와 사용 분포도가 유사한 용어를 선택하여 질의어에 추가한다[5]. 이렇게 함으로써 사용자가 질의어를 부여하는 초기에 용어 문제를 해결하여 검색 성능을 개선하도록 한다.

(그림 1)에서 질의어를 확장한다면 점선 내부에 있는 용어를 선택하는 것이 합리적일 것이다. 질의어 확장을 위해서 질의어 용어와 유사한 의미를 갖는 용어간의 유사성은 질의어 벡터로부터 용어 벡터가 얼만큼의 각도로 멀리 떨어져 있느냐가 기준이 된다. (그림

1)를 보면 질의어 기준 축에서 가장 근접한 각도에 위치한 용어는 "system"과 "interface"이다. 이것을 <표 3>에서 확인해 보면 유사성 측정값이 가장 높음을 알 수 있다. 따라서 질의어 확장을 위해 추가할 용어는 유사성 측정값이 0.9이상인 것을 대상으로 선택한다. 이들 용어는 <표 3>에서 밑줄이 있는 글자로 표현하였다. 선택된 용어를 갖고 질의어를 재구성해 확장된 질의어 벡터를 다시 계산한다. 확장된 질의어 벡터 값으로 이를 다시 그래프 상에 표현하면 (그림 2)와 같으며, 코사인 계수 값 0.9내에 있는 문서들의 범위가 약간 조정된다. (그림 1)과 비교해 보면 질의어 벡터와 가장 근접한 각도에 위치하는 문서가 c3에서 c1으로 변경되었다. 결국 질의어 벡터와 문서 벡터간의 유사성 값에 약간씩 차이가 생겼으며 유사성 순위에도 변화가 생긴다.

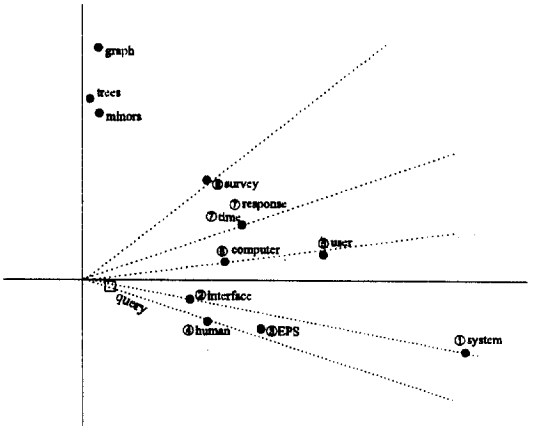
〈표 3〉 용어 벡터와 질의어 벡터의 유사성 값

용어	용어 벡터		유사성 값 (유사성 순위)
human	0.73	-0.28	0.9880(4)
interface	0.67	-0.18	0.9987(2)
computer	0.80	0.10	0.9443(6)
user	1.34	0.15	0.9484(5)
system	2.14	-0.43	0.9999(1)
response	0.90	0.28	0.8714(7)
time	0.90	0.28	0.8714(7)
EPS	1.00	-0.36	0.9910(3)
survey	0.70	0.69	0.5493(9)
trees	0.03	1.24	-0.1858(12)
graph	0.13	1.57	-0.1281(11)
minors	0.10	1.14	-0.1233(10)
질의어	0.14	-0.03	1



(그림 2) 확장된 질의어 벡터와 용어, 문서 벡터간의 관계

질의어 벡터와 용어 벡터간의 유사성 수치 값에 따라 (그림 1)을 다시 그려보면 (그림 3)과 같다. 질의어 기준 축을 중심으로 가장 가까운 각도에 위치한 "system," "interface," "EPS," "human," "user," "computer," "time," "response" 순으로 유사성 수치 값이 높게 나타난다. 이것은 코사인 측정 방법의 특성을 나타내는 것으로 질의어에 추가할 용어를 선택할 때 하나의 원리로 삼을 수 있다. 이 그래프에서는 용어의 수가 적기 때문에 가시적으로 질의어와 용어 벡터의 구별이 용이하다. 하지만 문서의 수가 많아지면 용어의 수가 수만 개에 이르기 때문에 그래프 상에 용어의 위치를 나타낼 경우에 서로 겹치거나 인접하는 일이 발생한다. 이것은 일정한 각도 안에 포함되는 용어들은 거의 같은 문서에서, 같은 분포를 갖는다는 것을 말한다. 예를 들어 "system," "interface"와 같은 용어는 그래프 상에서 위치는 다르지만, 유사성 벡터 값이 거의 같기 때문에 둘 중 하나만을 질의어에 추가하여도 검색 결과는 유사하다. 이들 같은 지역에 분포하는 용어들을 전부 질의어에 추가하는 경우에 검색 성능에 변화를 가져오지 않는다.

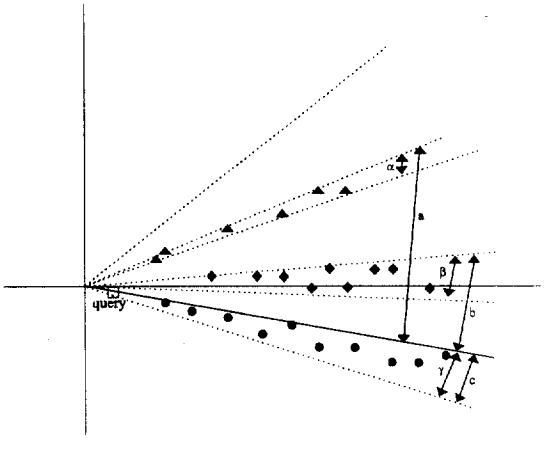


(그림 3) 질의어와 용어의 유사성 관계

보통의 질의어 확장 방법은 질의어 용어 가운데 하나와 밀접한 관련이 있는 용어가 추가되는 경향이 있다. 다시 말해 질의어 확장이 처리되는 동안 유사성 값이 어떤 임계치 값보다 적은 것이 전혀 고려되지 않는다. 이것이 검색 유효성을 개선하지 못하는 이유 중의 하나이다. 오히려 자주 사용되는 용어가 관련 있는

문서와 관련 없는 문서를 잘 식별하지 못하기 때문에, 용어의 선택시 제거하는 것이 유용하다[4].

이들의 관계를 분명히 하기 위해 (그림 4)를 통해 설명하기로 한다. 그림에서 ●, ◆, ▲는 유사성이 비슷한 용어들을 나타낸다. 용어 벡터들이 기준 축에서 a각도 떨어진 α영역, b각도 떨어진 β영역, c각도 떨어진 γ영역에 분포되어 있다. 기준 축을 중심으로 각 방향의 각도 크기에 따라 질의어와 유사성이 결정되기 때문에 기준 축에서 c각도 떨어진 γ영역 안에 포함된 용어들(●)이 가장 유사성이 높다. 이들 영역에 포함된 용어를 전부 질의어에 추가하기보다는 이중에 대표가 되는 용어를 선택하여 사용하는 것이 합리적이다. 따라서 임계치를 두어 같은 지역에 분포하는 일정량의 용어들은 제거하고, 그 중의 대표가 되는 용어만을 적용하는 것이 효과적일 것이다.



(그림 4) 유사성 값에 따른 용어 선택

4. 검색 성능 평가

본 논문에서 제안한 용어 분포 임계치의 검색 성능을 평가하기 위해 3가지 실험을 수행한다. 실험 대상은 세 가지의 문서 집합체를 사용하며, 결과의 비교는 ①벡터 공간 모델(즉, 키워드 기반)의 검색 결과 ②특이치 분해를 이용하여 질의어와 문서간의 유사성만을 측정하여 검색을 수행한 결과 ③특이치 분해를 이용하여 질의어와 용어간의 유사성을 먼저 측정한 후 용어 분포 임계치를 이용하여 질의어 확장을 수행한 결과를 비교한다.

4.1 측정 요소

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 정확도와 재현도이다. 이 두 가지 방법은 서로 간의 상반(trade-off)되는 의미를 지닌다.

A를 관련된 문서의 개수, B를 검색된 문서의 개수, A^-를 관련되지 않은 문서의 개수, B^-를 검색되지 않은 문서의 개수 그리고 N를 문서의 총 개수라 할 때, 다음과 같이 <표 4>로 나타낼 수 있다.

〈표 4〉 정확도-재현도 관계표

	관련된 문서	비관련된 문서	
검색된 문서	$A \cap B$	$A^- \cap B$	B
검색되지 않은 문서	$A \cap B^-$	$A^- \cap B^-$	B^-
	A	A^-	N

검색된 문서 중 관련된 문서 = $A \cap B$

정확도 = $\frac{A \cap B}{B}$, 재현도 = $\frac{A \cap B}{A}$

4.2 실험 문서 집합체

본 논문에서 제안하고 있는 용어 분포도에 기반한 개념적 정보 검색의 성능을 확인하기 위해 CISI, TIME과 CACM 실험 문서 집합체를 사용하였다. <표 5>는 본 논문의 실험을 위해 사용한 문서 집합체와 질의어의 특성을 요약한 것이다. CISI는 과학과 관련된 내용으로 문서 요약만을 갖고 있으며, TIME은 타임 잡지의 기사를 발췌하여 만든 것이며, CACM은 "Communication of ACM" 논문지의 문서 요약을 갖고 있는 문서 집합체이다. 이들 문서 집합체는 널리 사용되고 있으며 사전에 각각의 질의어가 검색해야 할 문서들이 결정되어 있다. 성능 평가를 위한 모든 실험은 SUN Enterprise3000(솔라리스 2.5.1)에서 구현하였으며, 프로그래밍 언어로는 C언어를 사용하였다.

〈표 5〉 문서 집합체의 특성

종 류	CISI	TIME	CACM
• 문서의 수	1,460	425	3,204
• 용어의 수	7,063	14,007	7,733
• 사용한 질의어의 수	50	40	32
• 질의어와 관련 있는 평균 문서 수	50	4	12
• 문서당 용어의 평균 수	45	190	24
• 질의어당 용어의 평균 수	8	8	11

4.3 실험 결과

거대한 용어-문서 행렬을 k값의 집합체로 분해하는데 있어 차원(dimension) 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[9]. 실험에 의하면 50-200 사이에 있을 때 최고의 성능을 나타내다가 그 이후에 감소하기 시작한다[5]. 본 논문에서의 모든 실험은 k값을 100으로 설정하여 평가하였으며, 검색 결과는 정확도와 재현도의 성능을 쉽게 파악할 수 있도록 재현도의 전구간을 0.1 단위로 세분하여 표현하는 11-포인트 평균 정확도(11-point average precision)를 사용하였다.

본 논문에서는 임계치가 검색 유효성에 어떤 영향을 미치는지 알기 위해 몇 가지 실험을 하였다. 실험을 통해 알아본 결과, 코사인 계수 값이 0.9정도면 질의어에 추가할 용어의 대부분을 선택할 수 있다. 사용자가 부여한 질의어에 추가할 용어를 선택하기 위해 세 가지 실험 집합체에서 질의어 벡터와 용어 벡터간의 유사성 값이 0.9이상인 것을 대상으로, 유사성 수치의 임계치 간격을 0.01, 0.05, 0.001, 0.005, 0.0001로 하여 평가하였다. 그 결과 모든 실험 집합체에서 임계치 간격을 0.001로 설정했을 때 검색 성능 개선이 가장 우수하였다. 이와 같은 결과가 의미하는 것은 임계치 간격이 너무 크면 질의어와 유사성 값이 비슷한 상위 순위에 해당하는 용어가 많이 삭제되어 검색 유효성이 감소한 것이며, 반대로 임계치 간격이 너무 작으면 같은 지역에 분포하는 용어가 많이 추가되어 검색 유효성이 감소했기 때문이다.

4.3.1 CISI 문서 집합체

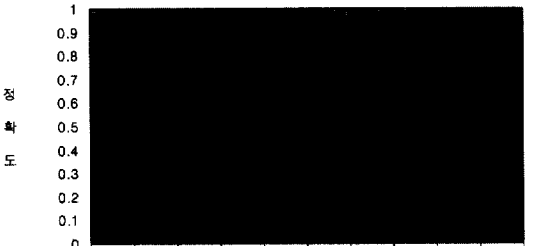
(그림 5)에서 X축은 재현도, Y축은 정확도를 나타내며 결과 그래프의 위치가 높을 수록 검색 성능이 좋은 것이다. 확장되는 용어는 질의어와 유사성 값이 높은 순서대로 40개에서 200까지 실험하였다. 그 결과 70개 이상이면 검색 성능이 우수하게 나타난다. 범례에서 "VSM (Vector Space Model)"은 키워드 기반 검색 결과이며, "K100"은 특이치 분해를 이용해 질의어와 문서간의 유사성만을 측정하여 검색을 수행한 결과이고, "t숫자"는 용어 분포 임계치를 이용하여 검색한 결과이다.

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
vsm	0.34	0.23	0.2	0.15	0.12	0.11	0.09	0.06	0.06	0.04
k100	0.4	0.24	0.22	0.18	0.13	0.11	0.1	0.08	0.06	0.04
t70	0.46	0.29	0.25	0.21	0.15	0.12	0.1	0.09	0.06	0.05

(그림 5) CISI에서 실험 결과

4.3.2 TIME 문서 집합체

확장되는 용어는 질의어와 유사성 값이 높은 순서대로 40개에서 200까지 실험하였다. 그 결과 40개 이상이면 검색 성능이 우수하게 나타난다. (그림 5)와 (그림 6)의 그래프 모습이 차이가 나는데 이것은 실험 문서 집합체 TIME이 재현도의 전 구간에 걸쳐 정확도가 높고 변동의 폭이 심하지 않기 때문이다.



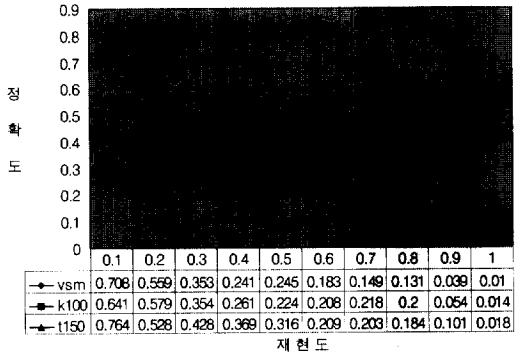
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
vsm	0.87	0.8	0.76	0.68	0.68	0.64	0.59	0.55	0.41	0.3
k100	0.87	0.84	0.82	0.81	0.72	0.69	0.67	0.57	0.43	0.35
t40	0.93	0.85	0.88	0.88	0.81	0.78	0.72	0.61	0.52	0.38

(그림 6) TIME에서 실험 결과

4.3.3 CACM 문서 집합체

확장되는 용어는 질의어와 유사성 값이 높은 순서대로 40개에서 200까지 실험하였다. 그 결과 150개 이상이면 검색 성능이 우수하게 나타난다. (그림 7)에 나타난 결과를 보면 CACM이 다른 문서 집합체에 비해 확장되는 질의어 개수에 따라 정확도 개선율 폭이 상당히 차이가 있다. 이것은 문서 집합체의 특성에 기인하는 것으로 CACM 자체가 문서 개수에 비해 용어의 개수가 작기 때문이며, 문서 집합체가 동일 영역의 내용을 담고 있기 때문에 용어 문제를 발생하는 동의어와 다의어가 많이 사용되지 않은 것으로 해석할 수 있

다. 이런 경우 키워드 기반의 검색과 개념 기반의 검색 결과 사이에는 큰 차이가 나타나지 않는다. 반대로 CISI나 TIME의 경우 문서 개수에 비해 용어 개수가 많이 나타나는데 이것은 문서 집합체가 좀더 다양한 영역의 내용을 많이 담고 있기 때문이다.



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
vsm	0.708	0.559	0.353	0.241	0.245	0.183	0.149	0.131	0.039	0.01
k100	0.641	0.579	0.354	0.261	0.224	0.208	0.218	0.2	0.054	0.014
t150	0.764	0.528	0.428	0.369	0.316	0.209	0.203	0.184	0.101	0.018

(그림 7) CACM에서 실험 결과

4.4 평 가

정확도 개선율은 키워드-기반인 벡터 공간 모델과 비교하여 측정된 수치이다. 기존의 특이치 분해를 이용하여 질의어와 문서간의 유사성만을 측정하여 검색을 수행한 결과 CISI에서 10%, TIME에서 8%, CACM에서는 16% 개선되었고, 용어 분포 임계치를 설정하여 질의어 확장을 한 경우 CISI에서 25%, TIME에서 17%, CACM에서는 19%의 성능이 개선되었다.

〈표 6〉 평균 정확도 개선율

문서 집합체	기존 특이치 분해 모델	용어 분포 임계치 모델
CISI	10%	25%
TIME	8%	17%
CACM	5%	19%

5. 결 론

본 논문에서는 개념적 검색을 지원하기 위한 방안으로 수많은 정보들에 존재하는 내포된 의미를 파악하여 사용자가 원하는 관련된 정보를 검색하는 자동화된 질의어 확장을 대상으로 하였다. 이를 위해 벡터 공간 모델을 기반으로, 질의어와 문서 또는 용어간의 의미적 유사성을 파악하기 위해 특이치 분해를 이용하였다. 기존에 특이치 분해를 이용한 방법들은 질의어와 문서간의 유사성에 초점을 두어 연구하였으나, 본 논문에서는 특이치 분해 기법을 이용해 사용자의 질의어 개념을 반영하는 용어 분포를 먼저 측정하였다. 이들 방법에 따라 사용자가 부여한 초기 질의어에, 전체 문서에서 질의어 용어와 사용 분포도가 유사한 용어를 찾아내어 질의어에 추가함으로써 사용자가 원하는 관련된 문서를 정확하게 검색할 수 있도록 하였다. 이때 질의어 확장시, 질의어에 추가할 용어를 선택할 때 질의어와 용어들간에 유사성이 매우 밀접한 것들이 많이 발생하였다. 유사성 값이 비슷한 용어를 질의어에 모두 추가한다는 것은 검색 성능 개선에 큰 도움이 되지 않기 때문에, 본 논문에서는 이들 용어 중에서 임계치를 설정하여 검색 성능을 개선할 수 있는 방법을 연구하고 그 결과를 세 가지 문서 집합체에서 평가하였다. 그 결과 기존의 특이치 분해를 이용해 질의어와 문서간의 유사성만을 측정하여 검색을 수행한 결과는 10% 가량의 성능 개선을 가져왔지만, 용어 분포 임계치를 이용한 검색 결과는 20% 가량의 성능 개선을 가져왔다. 용어 분포 임계치를 이용하는 방법은 실험 문서 집합체의 크기가 클수록 검색 성능의 개선에 효과가 클 것이다.

참 고 문 헌

[1] Todd A. Letsche, "Toward Large-Scale Information Retrieval Using Latent Semantic Indexing," Master Thesis, University of Tennessee, Knoxville, Aug. 1996.

[2] 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, pp.391-407, 1990.

[3] Gavin W. O'Brien, "Information Management Tools for Updating an SVD-Encoded Indexing Scheme," Master's thesis, the Univ. of Tennessee, Knoxville, Dec. 1994.

[4] Peat, H. J., Willett, P., "The limitations of term co-occurrence data for query expansion in document retrieval system," J. of the ASIS, 42(5), pp.378-383, 1991.

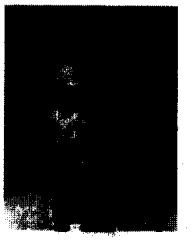
[5] Qiu, Y. and Frie, H. P., "Concept based query expansion," In Proc. of the 16th International ACM SIGIR Conf. on R & D in Information Retrieval, pp160-169, New York. 1993.

[6] Dumais, S. T., "Using LSI for Information Retrieval, Information Filtering, and Other Things," Talk at Cognitive Technology Workshop, April 4-5, 1997.

[7] Shih-Hao Li and Peter B. Danzig, "Vintage : A Visual Information Retrieval Interface Based on Latent Semantic Indexing," Technical Report USC-CS-96-632, Uni. of Southern Califonia, 1996.

[8] Michael W. Berry, Theresa Do, Gavin W. O'Brien, Vijay Krishna, and Sowmini Varadhan, SVDPACKC (version 1.0) user's guide, Technical Report CS-93-194, University of Tennessee, Knoxville, October 1993.

[9] 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.



임 재 현

e-mail : nolboo@munhwa.kongju-c.ac.kr

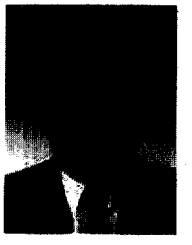
1986년 중앙대학교 전자계산학과 졸업(이학사)

1988년 중앙대학교 일반대학원 전자계산학과(이학석사)

1998년 중앙대학교 일반대학원 컴퓨터공학과(공학박사)

1998년~현재 공주문화대학 컴퓨터정보과 전임강사

관심분야 : 클라이언트/서버 시스템, 정보검색, 분산운영체제



민 태 홍

e-mail : thmin@true.inhatc.ac.kr

1981년 중앙대학교 전자계산학과 졸업(이학사)

1983년 중앙대학교 일반대학원 전자계산학과(이학석사)

1992년 중앙대학교 일반대학원 컴퓨터공학과(공학박사)

1984년~현재 인하공업전문대학 컴퓨터정보과 교수

관심분야 : 분산실시간 시스템, 멀티미디어, 정보검색시스템