論文
3-3-14

용어분포 임계치를 이용한 정보검색 성능개선에 관한 연구 (A Study on Performance Improvement of Information Retrieval using Threshold of Term Distribution)

민 태 홍*
(Tae-Hong Min)

요 약

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

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, user 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. 서론

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. 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. 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(vocabulary problem)가 발생한다.

본 논문은 기존 연구의 문제점인 용어 문제, 검색

* 정회원 : 인하공업전문대학 컴퓨터정보과 교수

논문접수 : 2002. 2. 20.
심사완료 : 2002. 3. 14.

이 논문은 인하공업전문대학 교내연구비 지원에 의하여 연구되었음.

성능 향상 문제를 해결하기 위해 전체 문서에서 나타나는 용어 분포를 이용해 개념 기반(concept-based) 검색을 지원하는 질의어 확장 방법을 대상으로 한다. 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. 그러나 용어의 수가 많을 때는 유사성 수치 값이 비슷한 것이 많아지고 이들 모두를 질의어에 추가하는 것은 비효율적이기에, 본 논문에서는 임계치를 설정하여 효과적으로 검색 성능을 개선하는 방안을 연구한다.

2. 기반 기술

2.1 개념 기반 검색

<표 1>을 이용해 용어의 출현 패턴에 내포되어 있는 의미 구조를 파악하기로 한다. 여기서 수치 값은 각각의 문서에 분포하는 용어의 빈도수를 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.

2.2 특이치 분해 기반의 용어 분포도

질의어와 유사하게 출현하는 용어의 분포를 파악하기 위해 특이치 분해를 사용한다. 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].

$$A = U\sum V^T \qquad \text{<식 1>}$$

<표 1> 용어와 문서간의 관계

<Table 1> Relation of term and document

문서 \ 용어	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	

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.

$$A_k = U_k\sum_k V_k^T \qquad \text{<식 2>}$$

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].

$$q = q^T U_k \sum_k^{-1} \qquad \text{<식 3>}$$

질의어 확장에 필요한 질의어 벡터와 용어 벡터 간의 유사성 측정을 위해 각 용어 벡터와 질의어 벡터와의 관계는 <식 4>를 이용한다. tk는 문서안의 용어 k번째 값이고, qk는 질의어안의 용어 값이다.

$$Sim(d,q)=\frac{\sum_k t_k * q_k}{\sqrt{\sum_k t_k^2 * \sum_k q_k^2}} \qquad <식\ 4>$$

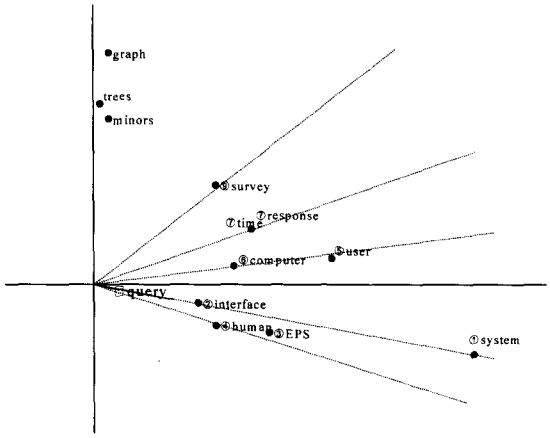
지금까지 기술한 내용을 예로들어 설명한다. <표 2>와 같이 실험 데이터로는 9개의 문서 제목과 이 문서를 구성하고 있는 용어를 사용한다. 문서 집합체에 출현하는 29개의 용어 중 두 개 이상의 문서에 출현한 12개 용어(밑줄)만을 이용하고, 질의어로는 "human computer"를 사용한다. 이 문서 집합체는 두 부류로 구성되어 있는데 c1-c5 문서가 같은 내용의 문서이고, m1-m4가 같은 종류의 문서이다. 실제 질의어를 통해 검색하고자 하는 문서는 c1-c5이다.

<표 2> 문서에서 용어와 질의어

<Table 2> Term and query in documents

문서	제 목
D1	Human machine interface for Lab ABC computer application
D2	A survey of user option of computer system response time
D3	The EPS user interface management system
D4	System and human system engineering testing of EPS
D5	Relation of user-perceived response time to error measurement
D6	The generation of rendom, binary, unordered trees
D7	The intersection graph of paths in trees
D8	Graph minors IV : Widths of trees and well-quasi-ordering
D9	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 벡터값을 이용하여 2차원의 그래프를 그려보면 [그림 1]과 같다. 질의어 벡터값은 <식 3>에 따라 계산하며, [그림 1]의 □는 질의어를 표현한다.



[그림 1] 질의어와 용어의 유사성 관계

[Fig. 1] Similarity relation of query and term

질의어 확장을 하기 전에 앞에서 구축한 의미 공간에서 문서-질의어, 용어-질의어 벡터간의 유사성을 측정하면 <표 3>, <표 4>와 같다. <표 3>에서 키워드 기반 검색을 수행하면 질의어 용어를 포함하고 있는 c1, c2, c4만이 검색될 것이다. 하지만 의미 공간을 구축하여 검색하면 c1-c5의 문서를 검색한다.

그래프를 통해 용어간의 관계를 파악할 수 있다. 질의어 기준축을 중심으로 점선으로 표현한 영역 안에 있는 용어는 문서 집합체에서 사용 분포도가 유사하다. 따라서 질의어를 확장한다면 점선 내부에 있는 용어를 선택하는 것이 합리적일 것이다. 이들 용어는 <표 4>에 밑줄이 있는 글자로 표현하였다. 선택된 용어를 갖고 질의어를 재구성해 확장된 질의어 벡터를 다시 계산한다.

<표 3> VSM과 의미 공간을 이용한 검색 유사성

<Table 3> Retrieval similarity using SVM and semantic space

문서와 질의어	VSM(키워드 기반 검색)		의미 공간(개념 기반 검색)	
	벡터	유사성 값 (유사성 순위)	벡터	유사성 값 (유사성 순위)
c1	1 1 1 0 0 0 0 0 0 0 0 0	0.8165(1)	0.67 -0.15	0.9999(2)
c2	0 0 1 1 1 1 1 0 1 0 0 0	0.2887(2)	2.03 0.43	0.9132(4)
c3	0 1 0 1 1 0 0 1 0 0 0 0	0	1.54 -0.33	1(1)
c4	1 0 0 0 2 0 0 1 0 0 0 0	0.2887(2)	1.80 -0.58	0.9949(3)
c5	0 0 0 1 0 1 1 0 0 0 0 0	0	0.94 0.28	0.8773(5)
m1	0 0 0 0 0 0 0 0 0 1 0 0	0	0.00 0.48	-0.2095(9)
m2	0 0 0 0 0 0 0 0 0 1 1 0	0	0.07 1.18	-0.1513(7)
m3	0 0 0 0 0 0 0 0 0 1 1 1	0	0.07 1.57	-0.1685(8)
m4	0 0 0 0 0 0 0 0 1 0 1 1	0	0.27 1.35	-0.0137(6)
질의어	1 0 1 0 0 0 0 0 0 0 0 0	1	0.14 -0.03	1

<표 4> 용어 벡터와 질의어 벡터의 유사성 값

<Table 4> Similarity value of term vertor and query vector

용어	용어 벡터	유사성 값(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

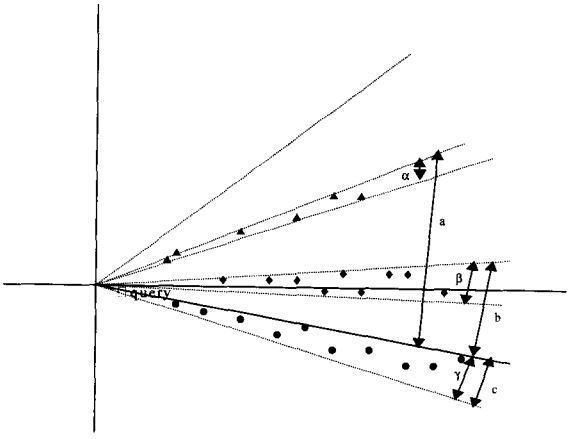
2.3 임계치를 이용한 용어의 선택

[그림 1]에서는 용어의 수가 적기 때문에 가시적으로 질의어와 용어 벡터의 구별이 용이하다. 하지만 문서의 수가 많아지면 용어의 수가 수만 개에 이르기 때문에 그래프상에 용어의 위치를 나타낼 경우에 서로 겹치거나 인접하는 일이 발생한다.

이것은 일정한 각도 안에 포함되는 용어들은 거의 같은 문서에서, 같은 분포를 갖는다는 것을 말한다. 이들 같은 지역에 분포하는 용어들을 전부 질의어에 추가하는 경우에 검색 성능에 변화를 가져오지 않는다.

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 2]를 통해 설명하기로 한다. 그림에서 •, ◆, ▲는 유사성이 비슷한 용어들을 나타낸다. 용어 벡터들이 기준축에서 a각도 떨어진 α영역, b각도 떨어진 β영역, c각도 떨어진 γ영역에 분포되어 있다. 기준축을 중심으로 각 방향의 각도 크기에 따라 질의어와 유사성이 결정되기 때문에 기준축에서 c각도 떨어진 γ영역 안에 포함된 용어들(•)이 가장 유사성이 높다. 이들 영역에 포함된 용어를 전부 질의어에 추가하기보다는 이 중에 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 2] 유사성 값에 따른 용어 선택
[Fig. 2] Term selection by similarity value

2.4 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 서로간의 상반(trade-off)되는 의미를 지닌다.

$$\text{정확도} = \frac{\text{검색된 문서 중 관련된 문서}}{\text{검색된 문서}} \quad \text{<식 5>}$$

$$\text{재현도} = \frac{\text{검색된 문서 중 관련된 문서}}{\text{관련된 문서}} \quad \text{<식 6>}$$

3. 검색 성능 평가

3.1 실험 문서 집합체

본 논문에서 제안하고 있는 용어 분포도에 기반한 개념적 정보 검색의 성능을 확인하기 위해 TIME과 CACM 실험 문서 집합체를 사용하였다. 이들 문서 집합체는 널리 사용되고 있으며 사전에 각각의 질의어가 검색해야 할 문서들이 결정되어 있다. 성능 평가을 위한 모든 실험은 SUN Enterprise3000(솔라리스 2.5.1)에서 구현하였으며, 프로그래밍 언어로는 C언어를 사용하였다.

3.2 실험 결과

문서 집합체 TIME과 CACM에서 k값을 100으로 설정하여 질의어 확장을 수행하였다. 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. 실험 결과는 질의어 확장을 위해 추가되는 용어 개수가 50-200개 사이일 때 최적의 성능 개선을 나타내었다. <표 5>는 두가지 실험 모델에서 질의어에 추가되는 용어의 개수가 50, 100, 150, 200일 때 평균 정확도 개선율을 나타낸 것이다.

<표 5> 평균 정확도 개선율

<Table 5> Improvement rate of average precision

문서 집합체	질의어 확장 모델	임계치 모델
TIME	15%	17%
CACM	16%	19%
평균 정확도 개선율	15.5%	18%

정확도 개선율은 키워드-기반인 벡터 공간 모델과 비교하여 측정된 수치이다. 용어 출현 빈도수 분포도를 이용하여 단순히 질의어 확장을 한 경우 TIME에서는 평균 15%, CACM에서는 16% 개선되었고 임계치를 설정하여 질의어 확장을 한 경우 TIME에서는 17%, CACM에서는 19%의 성능이 개선되었다. 따라서 임계치를 이용하는 방법은 실험 문서 집합체의 크기가 클수록 검색 성능의 개선에 효과가 클 것이다.

4. 결론

본 논문은 기존 연구의 문제점인 용어 문제, 검색 성능 향상 문제를 해결하기 위해 전체 문서에서 나타나는 용어 분포를 이용해 개념 기반 검색을 지원하는 질의어 확장 방법을 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 질의어와 용어간의 유사성을 측정한 후 이것을 정보 검색에 활용하는 방법을 제안하였다. 또한 질의어 확장시, 질의어에 추가할 용어를 선택할 때 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 평가하였다. 그 결과 단순히 질의어를 확장하는 모델보다는 3% 정도의 검색 성능 개선을 가져왔다.

※ 참고 문헌

[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, pp391-407, 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), pp378-383, 1991.

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

[6] 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.

[7] 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.

[8] 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.

민 태 홍

1981년 중앙대학교 전자계산학과 졸업(이학사)

1983년 중앙대학교 대학원 전자계산학과(이학석사)

1992년 중앙대학교 대학원 컴퓨터공학과(공학박사)

1984년-현재 인하공업전문대학 컴퓨터정보과 교수

관심분야 : 분산운영체제, 정보검색, 무선인터넷