잠재적 의미 색인을 이용한 유사 질의어 확장

임태훈, 안동언, 정성종
전북대학교 전자 정보 공학부
methodic19@naver.com, duan@chonbuk.ac.kr, sjchung@chonbuk.ac.kr

Query expansion by Similar words Using LSI

Tae Hun Lim, Dong un An, Seong Jong Chung
Division of Electronic & Information, Chonbuk Univ.

요약

오늘날 인터넷 검색은 하루가 다르게 발전되고 있다. 주로 키워드 매칭에 의존을 둔 지금의 검색 서비스들은 사용자 중심의 아이템들을 개발해 정보검색의 경과시간 및 결과의 분류면에서 우수함을 보여주고 있다. 질의어의 의미에 유사한 검색은 아직은 발전하는 단계로, 내용에 기반을 둔 검색 환경에 초점이 맞춰지고 있다. 이와 관련하여 행렬의 특이치 분해(SVD)를 이용한 잠재적 의미 색인 기법(LSI)을 본 연구에서 다루고자 한다. 구축한 시스템의 성능 평가는 재현도 계산으로 비교되었는데 작은 크기의 특이값(singular value)들 생략에 의한 SVD의 성능과 그것을 재이용, 질의어에 대한 의미 구조상 근접한 용어들을 찾아 질의어를 확장한 후 적합한 문서들의 검색을 사용한 특이값 개수, 유사단어 확장 개수를 달리하여 실험하였다. 실험 결과, 특이값 2개를 사용한 잠재적 의미 색인이 특이값 3개를 사용한 잠재적 의미 색인보다 보다 나은 성능을 보였다. 그리고 조건을 달리한 모든 잠재적 의미 색인의 경우 단어 매칭에 의한 적합문서 검색보다 별 뚜렷한 나은 결과는 보이지 않았다. 하지만 의미적으로 관계가 깊은 유사어들을 찾아냈고, 의미적으로 가장 관계 깊은 문서를 대부분의 경우에서 순위 1위로 찾아내는 부분적 우수함을 보였다.

1. 서론

정보 지식 사회의 발전으로 수많은 데이터들을 형식기반으로 분류함도 중요하지만, 구체적인 내용면으로 접근할 때는 양상이 또 달라진다. 오래된 인터넷 또는 어플리케이션 엔진에 의한 검색은 수많은 정보를 양적으로 분별없이 모으기만 할 뿐이었다.

또한, 다른 용어들이라도 같은 의미를 가질 수 있는 동의어(synonymy) 문제 처리나 같은 용어라도 다른 의미를 포함하는 다의성(polysemy) 문제 처리 등은 기존의 단어 매칭 검색으로는 한계가 있는 문제들이다.

위 문제점들을 해결하기 위한 방법으로 외부적으로 나타나는 단어들 사이의 의존성을 기반으로 한 벡터검색의 한 방법인 잠재적 의미 색인이 제안 된다.[1] 즉, 특이치 분해를 이용한 작은 크기의 특이값들 생략에 의한 근사로 각 문서, 용어, 질의 벡터를 더 저 차원 의미공간에 사상시킨 의미적 유사에 초점을 둔 잠재적 색인 기법을 설명하여 실험하고 의미구조상 근접한 용어들로 질의어를 확장하여 다른 실험 환경 조건들 속에서 비교 평가해보았다.

2. 질의 확장에 관한 관련연구

질의 확장이란 용어 사전 등을 사용하여 동의어를 추가하거나, 동일한 범주에 속하는 보다 광의의 용어를 추가하거나, 다른 적절한 용어로 대치하여 질의어를 확장하는 것이다.[2]

사용자로부터의 피드백에 의한 질의 확장은 관련 피드백(Relevance Feedback)으로 널리 알려진 방법이다. 사용자가 처음 질의를 주고 검색된 문서 중에서 관련된다고 생각하는 문서를 선택하면 이 문서의 단어들이 원래 질의에 더해지고 처음 질의의 용어들의 가중치를 변화 시키는 방법이다. 관련 피드백은 정확률과 재현율을 높일 수 있는 효과적이 방법으로 알려져 있지만 사용자와 대화식으로 작용해야 하기 때문에 사용자에게 부담을 줄 수 있다는 단점이 있다.[3]

위 단점을 극복할 수 있는 그 밖의 질의 확장으로 초기에 검색된 최상위 문서를 분석해서 질의를 자동으로 확장하는 지역분석 방법과 유사 시소러스를 구축하거나 통계적 시소러스에 바탕을 두고 질의를 확장하는 문서집합 전체에 대한 전역분석 방법이 있다.[3]

3. 잠재적 의미 색인(LSI)

3.1 Singular Value Decomposition (SVD)

MXN 행렬 C의 rank를 r이라 하면 행렬 C의 singular-value decomposition이 다음과 같은 형식으로 존재하게 된다. (여기서, U=용어벡터, ∑=특성벡터, V=문서벡터)

$$C = U \sum V^T$$

여기서, CC^T의 eigenvalue들인 $\lambda_1, \lambda_2, \cdots, \lambda_r$ 은 C^TC의 eigenvalue들과 같다. 또, $1 \le i \le r$에 대해서, $\sigma_i = \sqrt{\lambda_i}$ (단 $\lambda_i \ge \lambda_{i+1}$)이라면 MXN인 행렬 Σ는 $1 \le i \le r$이고 나머지들은 zero일 때, $\Sigma_{ii} = \sigma_i$로 구성된다. σ_i 값들을 행렬 C의 특이값들(singular values)이라 한다.

3.2 Low-rank approximation

앞에서 언급한 식을 다시 기술한다면, 다음과 같이 풀이된다.

$$C_k = U \sum\nolimits_k V^T = U \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & \sigma_k & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} V^T = \sum_{i=1}^{k} \sigma_i \vec{u_i} \vec{v_i^T}$$

여기서 $\vec{u_i}, \vec{v_i}$ 는 각각 U, V의 i번째 행 값 들이다.

이미 언급했듯이 특이값들(σ_i) 중 행렬에 덜 영향을 주는 크기가 작은 특이값들을 생략해서 그것으로 $\sum_2, C_2, (V')^T$ 값들을 다시 구해 새로운 SVD를 구할 수 있게 된다. 이렇게 각 문서와 질의 벡터를 더 저차원 공간인 개념으로 사상 시키는데 이는 색인어 벡터를 저차원 공간에 사상시키는 것이 가능하고, 축소 공간에서의 검색이 색인어 공간 검색 보다 우월하다고 주장되기 때문이다.[2] 즉 차원을 줄임으로써 검색효과에 방해되는 많은 잡음을 줄어들게 돼는 것이다.

$$C_k \quad = \quad U \quad \Sigma_k \quad V^T$$

(그림 1) Low rank approximation using the SVD

3.3 질의문

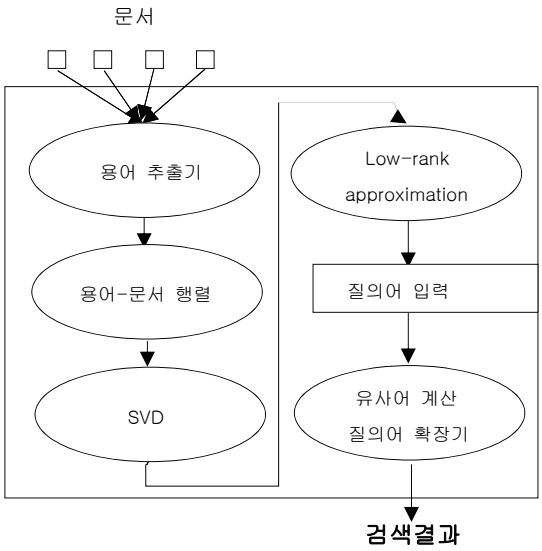
위에서 기술했던 low-rank representation에 아래식과 같은 형식으로 query를 던지면 질의문, 문서, 용어들을 벡터공간에 투사시켜 query-document 유사도(similarity)를 계산 할 수 있게 된다.

$$\vec{q}_k = \sum\nolimits_k^{-1} U_k^T \vec{q}$$

이러한 과정을 latent semantic indexing(LSI, 잠재적 의미 색인) 라고 한다.

4. 시스템의 설계

4.1 시스템의 구성



(그림 2) 검색 시스템

4.2 실험 집단 및 환경

본 실험에서는 한글 Data Collection으로 HANTEC-2.0을 사용하였다.[5] 문서 내용은 주로 정치, 외교에 관한 내용들이다.

재현율을 측정하고자 단어 12469개, 문서 61개의 문서 집합을 사용 하였고, 초기 질의어는 총 28개의 단일어를 사용하였다. (HANTEC-2.0에 이미 질의 SET이 있지만 메모리 문제로 임의의 단일어 질의어들을 사용하였다.)

실행된 시스템은 MATLAB 7.5.0 (R2007b) 환경이었다.

4.3 실험과정

질의어와 용어, 문서들 간의 유사성 측정은 각 용어, 문서들과 질의어의 벡터관계를 이용한 코사인 계수를 사용하였다.

$$SiTQ(T_i \vec{q}) = \arccos\left(\frac{q_1 \times t_{i1} + q_2 \times t_{i2}}{\sqrt{q_1^2 + q_2^2} \times \sqrt{t_{i1}^2 + t_{i2}^2}}\right)$$

여기서, $SiTQ(T_i \vec{q})$ 는 i번째 용어와 질의어 벡터의 유사값, T는 i번째 용어 벡터를 의미하고, t_{i1}과 t_{i2}는 T_i의 x축, y축 좌표 값이며, $\vec{q}$는 질의어 벡터를 의미한다.

위 식을 활용해 초기 질의어와 의미 구조상 근접한 용어들을 순서대로 찾아 질의어를 확장하여 다시 시스템에 입력, 의미상 가까운 문서들을 찾는다.

(표 1) 문서들 내의 용어들과 질의어

문서	용어들
Doc1	태환정책 3주년 맞는 아르헨티나
Doc2	미국서 대파키스탄 F16판매 논란
Doc3	한보그룹, 대련서 목재관련 프로젝트
Doc4	한국-미국, 이달 말께 대북한 종합대책
Doc5	러시아 여객기추락, 전원사망 추정
Doc6	한국-일본, 군사교류 추진
.....	.....
질의어	'양국'

표1은 문서집합내의 문서번호와 각 문서의 중심내용을 일례로 열거해보았다. 이러한 문서집합에 일례로 이표에서와 같이 질의어('양국')를 투사한다.

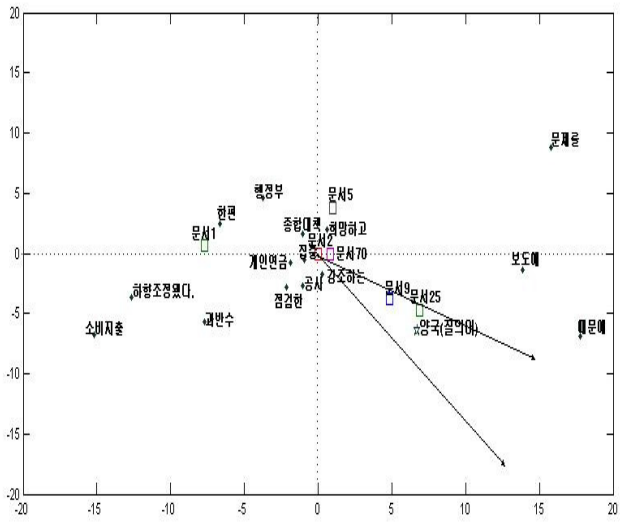
표2는 어떠한 용어가 각 문서에 몇 개씩 존재하는가를 일례로 나타낸 표이다. 아래 표를 기반으로 행렬을 구성하여 특이치 분해를 계산할 수 있게 된다.

(표 2) 용어 발생 빈도수와 문서간의 관계 행렬

용어	문서 번호											
	1	2	3	4	5	6	7	8	9	10	11	12
과반수	1	0	0	0	0	0	0	0	0	0	0	0
국민들은	1	0	0	0	0	0	0	0	0	0	0	0
인플레율이	2	1	0	0	0	0	0	0	0	1	0	0
또	0	1	0	0	0	0	0	0	0	0	0	0
한편	0	0	0	0	0	0	2	0	0	0	0	0
문제를	0	0	0	0	0	0	0	0	0	0	0	0
전해졌다.	0	1	0	0	0	0	0	0	0	0	1	0
한국인	0	0	0	0	0	0	1	0	0	0	0	0
보내는	0	0	0	0	0	0	0	0	2	0	2	4
.....	.....											

그림3은 특이치 분해후, Low-rank approximation 과정을 거쳐 새롭게 생성된 질의어, 문서 ,용어들간의 위상을 2차원으로 투사한 그래프이다.

위에서 언급했듯이. 질의어('양국')의 위치가 계산되면, 질의어 위치에 가까운 각도에 놓여있는 문서, 용어들이 의미 구조적으로 가까운 거리에 있는 문서, 용어들이 된다.



(그림 3) 질의어-용어-문서간의
2차원 관계 그래프

재현율은 검색의 완전성을 측정하며 검색 효율 척도 중 가장 널리 사용 되고 있다. [2]

$$\text{재현율(Recall)} = \frac{\text{검색된 문서중 적합문서수}}{\text{적합문서총수}}$$

5. 실험 및 평가

실험 결과는 아래와 같다.
특이값들은 크기가 큰 순서대로 2개, 3개가 사용되었다.

(표 3) 적합문서 검색 결과 비교

검색엔진 / 질의환경		LSI (특이값 2개)	LSI (특이값 3개)	단어 매칭
재현율 (%)	질의어	24.30	21.93	22.55
	질의어 +유사어1	23.70	21.58	
	질의어 +유사어1 +유사어2	23.64	22.00	

예상과는 달리, 특이값 2개를 사용한 잠재적 의미 색인이 특이값 3개를 쓴 잠재적 의미 색인보다 보다 나은 성능을 보였다. 그리고 가장 좋은 성능을 보인 조건은 질의어를 확장하지 않은 특이값 2개를 사용한 잠재적 의미 색인이었다. 하지만 나머지 조건들에 비해 아주 근소하게 나았기 때문에 단정적으로 이 조건이 가장 좋은 성능을 보인다고 말할 수는 없다. 문서 집합의 수가 엄밀히 실험하기에는 수가 작고, 적합문서의 판단기준이 주관적인 면이 많으며, 질의어들과 문서간의 유사하다는 수치가 얼마큼이 돼야지 의미상 유사하다고 보는가도 수치를 얼마로 적용 하냐에 따라 달라지므로(본 실험에서는 잠재적 의미 색인의 경우 유사값 크기대로 10위 안에 드는 문서를 의미상 유사하다고 판단했다.) 근소한 차이가 나는 위의 결과는 어떤 조건이 확실히 검색 성능에 낫다고 할 수는 없다. 하지만, 비록 단어매칭 검색에 의한 적합문서 검색성능보다 나쁘게 나타나는 잠재적 의미 색인검색조건들도 있었지만, 모든 조건의 잠재적 의미 색인의 검색결과는 의미상 아주 관계가 깊은 유사어들을 찾아내는 우수한 결과를 보여줬다.

(표 3)의 결과에서 보듯이 유사어들 추가에 의한 질의어 확장이 특이값을 2개, 3개를 쓰느냐 보다는 결과에 훨씬 더 적은 영향을 미쳤다. 즉 유사어 한 개를 추가한 문서검색 결과와 유사어 두 개를 추가한 문서검색결과는 거의 같았다.

6. 결론

본 실험의 처음 의도대로, 질의어 확장 조건과 특이값 사용 개수 조건을 달리하여 잠재적 의미 색인의 의미적으로 적합한 문서를 찾아내는 검색결과의 우수함을 보여주려 했던 의도는 재현율 척도로는 뚜렷이 나타나지 않았다.

유사어들 추가에 의한 질의어 확장 결과도 다른 조건의 검색 성능보다 별 우수한 성능을 보여주지 않았다.

하지만, 모든 조건의 잠재적 의미 색인은 확연히 관계가 깊은 유사어들을 찾아냈고,

모든 조건의 잠재적 의미 색인은 의미상 가장 관계 깊은 유사값 순위 1의 적합문서를 대부분의 질의에서 1위로 찾아내는 등 부분적인 검색성능의 탁월함을 보였다.

본 실험결과, 형태소 분석기 등을 이용한 문서 집합에서의 보다 정제된 용어추출이 없었음이 실험 결과에 보이지 않는 상당한 영향을 미친 것으로 보고 있으며, 실험 문서의 개수를 시스템 환경의 메모리 문제 처리 미숙으로 보다 많이 늘리지 못한 것이 엄밀한 성능 결과 추출에 불리한 작용을 미쳤으리라 본다. 마지막으로 초기 질의어를 단일어로 한정해 어구나 문장을 질의어로 입력하지 못해 실험의 현실적인 면을 감소시켰다.

7. 참고문헌

[1] 음대호, “잠재적 의미색인을 이용한 지능형 정보 검색 시스템”, 한양 대학교 대학원 전파공학과, 석사논문, 3쪽, 40쪽, 1998년 12월

[2] 최영란, “잠재적 의미 색인기반의 정보검색에서 사전의미를 이용한 질의어 확장”, 전북 대학원 정보통신학과, 석사논문, 6쪽, 16쪽, 42쪽, 40쪽, 2003년 02월 2일

[3] 안성수, "SVD를 이용한 질의 확장", 한양 대학교 대학원 전자계산학과 , 석사논문, 18쪽, 19쪽, 20쪽, 1999년 12월

[4] Christopher D.Manning, Prabhakar Raghavan, Minrich Schutze, “Introduction Information Retrieval”, CAMBRIDGE, Chapter 18, First published 2008

[5] 한국과학기술정보연구원과 충남대학교 공동 개발, HANTEC(HANgul TEst Collection) Version2.0, www.kristalinfo.com/download/