잠재적 의미와 k-means 군집화를 이용한 개념추출 검색[1]

장유진[0], 임호섭, 박기림, 김민구
아주대학교 정보통신전문대학원
(xaritas,mcdonald,hoop,minkoo)@madang.ajou.ac.kr

Extraction of Concept by Latent Semantic Indexing and k-means Clustering

You-Jin Chang[0], Ho-Seop Lim, Ki-Rim Park, Min-Koo Kim
Graduate School of Information and Communication, Ajou University

요 약

정보검색 시스템에서 사용자의 질의어가 불완전함에 따라 생기는 검색 효율의 저하를 줄이기 위하여 용어의 상호관련성을 반영함과 동시에 벡터의 공간을 축소하는 LSI 모델을 사용하여 문서 집합으로부터 잠재적 의미 공간을 구축하였다. 또한 의미 공간상에 있는 문서의 분포에 따라 '개념'을 추출하기 하기 위해 k-means algorithm을 사용하여 군집화 시켰다. 이로부터 불완전한 초기 사용자 질의어를 의미 공간에 구축된 클러스터링 정보로 수정하여 새로운 질의어를 생성함으로 검색의 효율을 높이고자 하였다. 검색 효율을 측정하기 위해 TREC 데이터를 이용하여 분석하였으며 결과는 질의어의 성격에 따라 달라졌으나 대체적으로 우수한 성능을 보였다.

1. 서 론

정보검색 시스템의 사용자는 자신의 원하는 정보를 얻기 위해서 정보요구(information need)를 하게 된다. 정보검색 시스템은 수집된 정보 또는 정보자료의 내용을 분석한 뒤, 가공하여 축적해 놓은 정보로부터 사용자가 원하는 정보를 되돌려준다. 인터넷이 발달된 요즘 정보의 홍수시대에 살고 있는 사용자들이 필요로 하는 것은 많은 양의 정보가 아니라 정확한 양질의 정보이다. 문제는 비록 사용자가 자신의 원하는 요구를 검색 시스템에게 넘겨 주었다 하더라도 그것이 제대로 자신의 정보요구를 충분히 반영해주지 못했을 때에 일어난다. 정보화 시대가 되면서 사용자들은 단순한 검색보다는 개념적 정보요구를 필요로 하게 되었다. 그러나 사용자가 제시하는 한두개의 개별적인 단어만으로는 개념적인 표현이나 문서의 의미들에 대해서 신뢰할 수가 없다[1]. 이는 정보 검색의 단계에서 사용되는 어휘 통제 및 개념표현에 대한 지원이 시급하다는 것을 뜻한다.

이러한 단점을 보완하기 위해서 많은 연구가 있었다. 대표적인 예로 시소러스(thesaurus)는 용어와 용어간의 계층적 구조를 나타내며[8] 동의어(synonym), 상위어(broader terms), 하위어 (narrower terms) 및 관계어 등으로 질의어 및 검색된 용어를 대체할 수 있는 용어사전이다. 시소러스는 이렇게 미리 구축된 사전 지식(preknowledge)를 사용하여 사용자의 정보요구를 확장하는 역할을 한다. 그러나 시소러스의 구축에는 많은 비용과 시간이 드는 관계로 필요한 모든 분야에 적절한 시소러스를 구축하지 못하고 있다. 또한 어렵게 구축한 시소러스도 시간이 흐름에 따른 해당 분야의 변.화를 신속하게 반영하여 개정하지 못하고 있는 실정이다. 이러한 이유로 시소러스는 specific한 범위에 사용될 때에는 효과를 보여주지만 general한 범위 즉, 웹과 같이 매우 크고 광범위한 자료 그리고 새로운 문서가 유입되는 공간에서 사용될 때에는 오히려 검색효율을 저하시킨다[6].

또 다른 연구로 Latent Semantic Indexing(LSI)[1]은 잠재적 의미 분석이라고 불리며 용어간의 의존성을 참조하는 벡터검색의 일종으로 용어와 문서간에서 "의미적 공간"을 구성하는 역할을 한다. 기존의 대부분의 검색 모델이 용어들을 독립적으로 처리하는데 반해 LSI는 용어간의 상호관련성을 유도하여 검색효율을 개선시키는 장점을 가진다[7].

본 연구에서 LSI를 사용하여 잠재적 의미로 표현한 새로운 의미의 벡터 공간에 클러스터링 기법을 적용하여 개념 공간을 구성하고 이를 이용한 개념 기반 검색을 시도하였다.

본 논문의 구성은 다음과 같다. 2장에서는 잠재적 의미 분석인 LSI에 대한 연구와 클러스터링 기법 중 k-means algorithm을 설명하였으며 3장에서는 앞의 두 단계를 거쳐 구성된 공간에서 개념을 추출하는 과정을 설명한다. 4장에서는 이에 따른 실험 및 결과를 분석한다.

2. 기존연구에 관한 고찰

본 장에서는 본 연구에서 사용한 두 방법론인 잠재적 의미 분석과 k-means 군집화에 대한 방법을 살펴보기로 하겠다.

2.1 잠재적 의미 분석(LSI)에 대한 연구

[1] 본 연구는 정보통신연구진흥원 과제 번호 AA-00-2060-00의 재정적 지원을 받아 수행되었음

LSI는 용어와 문서간의 상호 관련성을 참조하는 벡터 검색의 일종이며 모델화를 위해서 통계적 기법인 Singular Vector Decomposition(SVD)를 사용한다. SVD는 거대한 용어-문서의 행렬을 k-차원의 행렬로 분해하며 이때 보통 k는 100-300사이의 값을 갖는다.[7] 각각의 용어와 문서는 SVD를 통해서 k-차원의 LSI공간 안에 벡터로서 표현되며 유사한 내용의 문서에 사용된 용어들이 이들 공간 안에 유사한 값으로 표현된다. SVD는 데이터 공간에서 major associative pattern들을 반영하도록 용어와 문서를 벡터공간에 배치시킨다. 이 과정을 통해서 smaller, less important influence들을 무시하게 된다. LSI에서는 용어-문서간의 행렬이 주어지고 이를 수학적이고 통계학적인 방법인 SVD로 적용시킨 후 결과를 식(2)처럼 벡터 유사도 계산법으로 얻을 수 있다. SVD를 수행하기 위해서는 SVDPACKC[3]를 사용하였다. SVD는 행렬분해를 위해서 초기 용어-문서 행렬 X를 t × d로 표현하며 X는 다음 그림 1과 같이 T_0, S_0, D_0' 3개의 서로 다른 행렬로 분해된다. T_0와 D_0는 직교하는 열의 행렬이며 S_0는 m차원으로 singular value가 대각선방향으로 차있는 대각선 행렬이다.

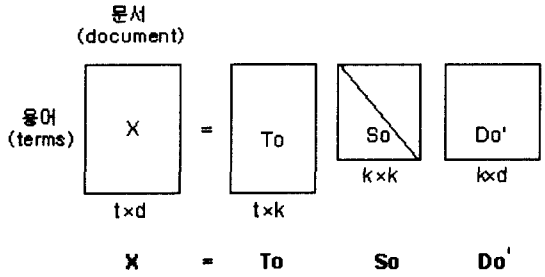

그림 1 Reduced singular value decomposition

여기서 t는 행렬 X의 행 개수이고 d는 열 개수이다. 행렬 X의 rank로 m값(≤min(t,d))이 주어지는데 k값은 m보다 작은 reduced model에서의 차원(dimension)을 결정하는 수이다. $T'T = D'D = I$ 이며 $S=diag(\sigma_1, \ldots, \sigma_k)$, $(\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0)$는 특이치 값(singular value)의 대각선 행렬을 뜻한다. SVD가 수행되고 나면 용어와 문서가 k-차원의 공간 안에 벡터형태로 표현되어 존재하게 된다. 이렇게 용어와 문서간에 잠재되어 있는 의미를 파악할 수 있는 분포 수치를 구하고 나면, 질의어 벡터값을 구해야 한다. LSI에서는 질의어를 또 하나의 가상 문서(pseudo document)로 취급하여 마찬가지로 k-차원상의 벡터형태로 표현한다. 이때 가상문서는 용어 벡터들의 가중치들의 합이며 다음의 식(1)과 같이 표현된다. D_q는 가상문서의 벡터이며 X'_q는 색인되어 있는 용어가 질의어에 존재할 때 값을 1로, 없을 때는 0으로 갖는 전치행렬(transpose matrix)이다. T는 앞에서 구한 용어 벡터이고 S^{-1}는 특이치 벡터로 S행렬의 역행렬이다.

$$D_q = X'_q T S^{-1} \quad (1)$$

질의어와 용어들간의 유사성(similarity)을 구하기 위해서 두 벡터의 코사인 계수식을 사용한다.

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}} \quad (2)$$

LSI는 순수한 벡터모델이 문서를 표현하는 차원을 용어의 개수만큼 고차원 벡터로 처리했던 어려움을 SVD의 rank수만큼 줄이므로 계산의 효율을 높였으며 벡터모델이 용어사이의 관계를 독립적으로 처리한 것과 달리 LSI는 용어간의 상호관련성을 고려했다는 점에서 개념적인 공간을 구축하는 데에 의미가 있다고 할 수 있다.

그러나 서론에서 언급한 것처럼 실제적으로 LSI에서 의미적 공간을 구축한다 하더라도 사용자의 질의어 자체에 있는 불완전함을 보완하기 위한 방법은 고려되지 않았다. 따라서 새로운 의미적 공간에서 문서집합을 군집화시켜 질의어를 보완하는 방법을 다음에서 연구하였다.

2.2 k-means 군집화(clustering)에 대한 연구

클러스터링 기법 중 k-means algorithm은 Euclidian distance(거리)를 이용하여 가깝게 위치한 점들을 찾아 군집으로 묶어주는 기법으로 차원의 제약이 전혀 없고 간단한 알고리즘으로 가장 널리 알려져 있다.

LSI를 통해 의미적(semantic)으로 분해된 공간에서 서로 가까운 거리에 있는 문서끼리 군집화를 시켜 개념을 구축할 수 있다. 용어-문서간의 상호관련성을 반영한 의미적 공간에서 군집화 된 클러스터들은 문서집합이 가진 속성들을 표현하는 역할을 한다. 사용된 k-means 알고리즘은 다음과 같다.

Input : list of m-dimension vector
Output : k lists of m-dimension vector

Step 1: Select K seeds from the data set
Step 2: Allocate each record to one of K seeds, which is most similar to the record
Step 3: Compute the centroid of K groups of records
Step 4: Repeat the steps 2 and 3 until the centroids do not change

3. 잠재적 의미와 k-means 군집화를 통한 개념 추출

LSI를 거쳐 분해된 k-차원의 벡터는 단순한 분해의 의미뿐만 아니라 LSI의 특징인 유사한 내용의 문서들끼리의 상호연관성이 고려된 상태의 공간이다. 따라서 이 벡터들끼리의 군집화를 거쳐서 형성된 각 클러스터의 중심값(centroid)을 문서집합이 표현하는 '개념(concept)'이라고 생각할 수 있다. 따라서 초기 제시했던 사용자의 불완전한 질의를 클러스터의 중심값으로 치환 또는 보정하여 검색의 효율을 높이는 방법을 제시하고자 한다. 먼저 k-차원의 벡터를 앞에서 제시한 k-means algorithm을 사용하여 거리상 가까운 벡터끼리 k개의 군집을 만든다. 각 군집의 중심값을 대표값으로 하여 질의어 벡터와 유사도 구해 가장 가까운 중심값을 질의어 보정에 사용한다. 보정식은 다음과 같다. D_q는 식(1)에 의해 표현된 질의어 벡터이며 C_i는 D_q와 유사도 계산을 해서 가장 근사값으로 선택된 클러스터의 대표값(centroid)이다.

$$D'_q = \alpha D_q + \beta C_i \quad (3)$$

새로 보정된 질의어 D'_q는 α와 β값에 따라 다른 결과를 갖는다. α가 1이고 β가 0인 경우는 클러스터링의 결과에 영향을 받지 않고 사용자가 던진 초기 질의를 LSI의 결과로만 재조정한 형태이며, 반대로 α가 0이고 β가

1인 경우는 사용자의 질의를 클러스터의 중심값으로 치환하여 검색하는 의미가 된다.

4. 실험 및 결과분석

테스트한 문서집합은 TREC에서 제공한 DOE 문서의 일부였으며 문서의 개수는 10,000개이고 포함하는 용어의 수는 42,849개였다. TREC의 topic 중 96번,134번, 135번을 가지고 실험했으며 각각의 topic은 문서 10,000개에 대해 relevant document를 29개,25개,40개로 가지고 있다. 본 실험에서는 식(3)에서 주어진 질의어를 바탕으로 α 와 β 의 값을 1.0~0.0으로 조정하면서 recall, precision을 계산하여 아래와 같은 결과를 얻었다. LSI 수행단계에서는 SVDPACKC[3] package을 사용하였다. SVDPACKC의 특징은 large sparse matrices에 대하여 singular value decomposition을 수행한다는 것이며 패키지에서는 총 4개의 numerical (iterative) 방법을 제공하는데 그 중에서 Single-Vector Lanczos Methos를 사용하였다.

군집화에서는 여러 실험을 통해 k=50 에 대해 적절한 결과를 얻었으며 원소 개수 5개 이하의 군집에 대해서는 무시하였다. 실험한 시스템은 Linux server(kernel 2.2.16-3kr2버전)와 Ultra spac 2였다.

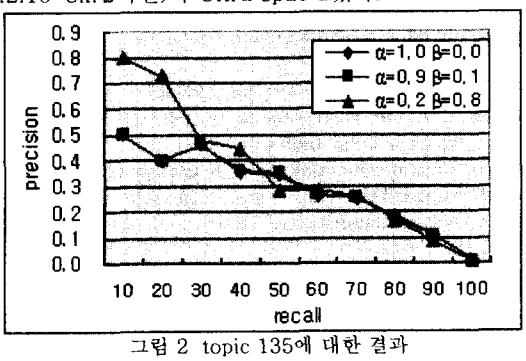

그림 2 topic 135에 대한 결과

위와 같이 topic 134와 135의 실험에서는 β 값이 높고 α 값이 작을수록 좋은 결과를 나타내었다. 즉 클러스터링의 정보에 영향을 받았을 때 검색결과가 높아졌음을 의미한다. 특히 topic 135의 경우, α =0.2이고 β =0.8일 경우, α =1.0으로 적용했을 때보다 정확도에 있어서 25%의 향상을 보여주었다. 이 경우는 사용자의 질의가 불완전하여 문서로부터 만들어진 클러스터의 영향을 받아 표현했던 것이 개념적으로 잘 매칭된 경우라고 분석할 수 있다. 그러나 반대로 topic 96에 대해서는 α 값이 높고 β 값이 작을수록 좋은 결과를 보였다. 이 경우는 사용자의 질의어가 정보요구에 맞게 적절히 표현된 경우이므로 β 값을 높여 클러스터의 영향을 받는 것이 오히려 정확도를 떨어뜨린 것으로 보인다. 그림 3은 α ,β 비율에 따른 topic 별 검색 결과의 변화를 보여준다. 따라서 클러스터링 정보의 적용은 질의어와 클러스터의 관계에 따라 다른 결과를 나타낸다고 볼 수 있다. 초기 질의어와 각 클러스터들의 유사도 및 질의어와 상위 relevant한 문서와의 유사도를 계산하여 그 상관관계를 조사한 뒤 새로운 질의어 생성에 대한 판단 기준으로 삼을 수 있다.

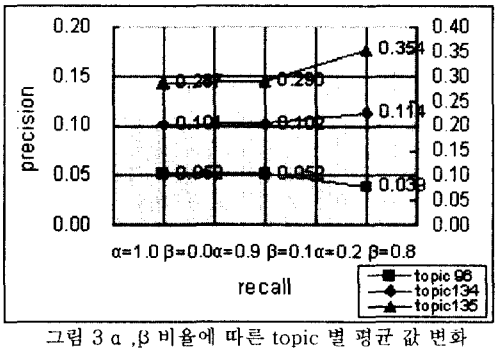

그림 3 α ,β 비율에 따른 topic 별 평균 값 변화

5. 결론 및 향후과제

용어-문서간의 상호관련성을 반영해주는 LSI 모델을 통하여 잠재적 의미를 분석한 공간을 구축함과 동시에 의미들을 군집화 함으로써 개념을 추출하는 연구를 수행하였다. 문서집합에 클러스터가 잘 구축되어 있고 사용자 질의가 정보요구에 정확히 맞지 않을 때, 클러스터링의 정보가 개념적 매칭(matching)을 도와줌으로 질의어 표현 및 검색에 효과가 있다는 것을 알았다.

향후과제로는 위 실험에서 사용자의 초기 질의어에 따라 다른 결과가 유도되므로 질의어가 문서 및 클러스터와 어떤 상관관계가 있는 지 연구할 필요가 있다. 또한 k-means 군집화에 있어서 클러스터 k값을 이론적으로 산출하는 방법을 연구하는 것과 군집화 기법에 있어서 문서 클러스터링에 많이 사용되는 계층적 클러스터링 방법 등 다양한 방법을 적용해보아 가장 효율적인 군집화 기법을 찾아내고자 한다. 마지막으로 위 실험을 객관적 데이터 set인 TREC 전체에 적용해보고자 한다.

참고문헌

[1] Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A., "Indexing by la tent semantic analysis." Journal of the Society for Information Science, 41(6). 391-407, 1990

[2] Michael R. Anderberg, Cluster Analysis for application, 156-162, 1973

[3] Michael B., Theresa D., SVDPACKC(Version 1.0), Computer Science Department, CS-93-194, 1993

[4] Warren Sarle.,The number of clusters from the SAS/STAT User's Guide (1990) and Sarle and Kuo (1993) ,comp.ai.neural-nets,1996

[5] Ian H.W., Data Mining practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 75. 210-211, 2000

[6] Ricardo B.Y, Modern Information Retrieval, Addison Wesley, 19-34, 1999

[7] 임재현, 김영찬, 인터넷에서 잠재적 의미분석을 이용한 지능적 정보검색, 한국통신학회논문지, Vol.22 No.8, 1782-1789, 1997.08

[8] 정영미, 정보검색론, 구미무역, 1993