공학석사 학위논문

Latent Semantic Analysis를 이용한 내용기반 정보검색 시스템

A Study on Content-Based Information Retrieval System using LSA

2000년 2월

서울대학교 대학원
협동과정 인지과학전공

신 동 호

Latent Semantic Analysis를 이용한 내용기반 정보검색 시스템

A study on Content-Based Information Retrieval System using LSA

지도교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

1999년 12월

서울대학교 대학원
협동과정 인지과학전공
신 동 호

신동호의 공학석사 학위논문을 인준함

1999년 12월

심사위원장	조명한印
심사위원	장병탁印
심사위원	김청택印

■ 목 차

요약

요 약

인간은 자연언어를 놀라우리만큼 쉽고 빠르게 습득한다. 특히 어린아이의 모국어 습득 능력은 가히 경이로울 정도이다. 심리언어학에서 이루어진 수많은 실험들은 인간의 언어습득 능력이 얼마나 훌륭한지를 증명하고 있다. 이와는 대조적으로 자연언어를 컴퓨터에 학습시키는 일은 대단히 어려운 일이다. 언어별로 약간씩 차이가 나기는 하지만 지난 수십 년간 자연언어처리 분야에서 행해진 수많은 노력들의 결과 이 일이 처음 예상했던 것보다 훨씬 어려우며 마법의 방망이처럼 일순간 해결할 수는 없는 문제라는 인식을 갖게 되었다.

자연언어가 가지는 복잡성과 예외성으로 말미암아 자연언어를 통계적 방법으로 학습시키는 것은 요원한 일이었다. 따라서 지금까지는 주로 인간의 직관에 의존하여 직접 규칙을 입력하거나 사전작업을 해 왔다. 그러나 규칙-기반적인 방법은 몇 개의 문장 수준에서는 완벽에 가까운 분석을 해낼 수 있다 하더라도 대규모 문서집합에서는 예외적인 경우가 많이 발생하기 때문에 일괄적으로 적용하기 어렵다. 이에 반해 통계적인 방법은 통사구조를 무시할 정도로 단순한 차원의 정보만을 이용하나 학습이 용이하고 대규모 텍스트 자료에 쉽게 적용될 수 있다. 통계적인 학습이란 공기정보(co-occurrence)와 같은 단순한 일차적 자료로부터 좀 더 일반적인 고차수준의 표상을 만들어 내는 과정으로 볼 수 있다.

이러한 통계적인 학습 방법 중 LSA(Latent Semantic Analysis)는 행렬연산인 SVD(Singular Value Decomposition)에 의해 고차

수준의 표상을 보다 직접적으로 계산해 낸다. 기존의 학습 방법들은 계산상의 어려움 때문에 대부분 미리 선택된 비교적 적은 수의 자질들을 바탕으로 학습이 진행되는 반면 LSA는 대규모 문서 코퍼스 전체를 통계적인 방법으로 처리하여 개념공간이라 부를 수 있는 고차원 벡터 공간을 생성한다. 순수한 통계적 처리 과정을 거쳐 만들어진 이 개념공간은 많은 문제 상황에 적용해 본 결과 사람의 장기기억 내지 심성 어휘집의 특성에 근사한 결과를 내었다. 또한 LSA가 제공해 주는 단어-문서 관련도, 단어-단어 관련도 및 문서-문서 관련도는 정보검색의 많은 문제들에 곧바로 적용 가능하다. 본 논문에서는 일단 작은 규모의 한국어 텍스트를 대상으로 LSA의 특성을 분석해 본 후, 서울대학교 중앙도서관에서 제공하는 학위논문 초록 20,000여 건을 모아 본격적인 정보검색 시스템을 구축하여 LSA의 성능을 검증하였다.

주요어 : LSA, SVD, 개념공간, 정보검색, 유사도, 자연언어, 심성어휘집

학 번 : 98132-504

1. 서 론

1.1 연구의 배경 및 필요성

자연언어의 의미를 처리하는 가장 대표적인 방법은 의미 정보들을 미리 사전에 잘 구축해 놓고 이를 사용하는 방법이다. 자연언어의 의미를 잘 파악하고 있는 사람이 직접 사전에 등록하기 때문에 정확하고 볼 수 있다. 보통 일반 사전에 기록이 되는 정의나 용례 이외에도 상위어 관계, 하위어 관계, 유의어 관계, 포함어 관계 등 다양한 단어들간의 관계성에 대한 정보가 들어간다. 이러한 사전을 시소러스(thesaurus)라 하는데 자연언어를 다루는 기본 관문이라 할 수 있다. 현재 가장 널리 사용되는 시소러스로는 프린스턴 대학의 Miller가 만든 WordNet이 있다[Miller, 1990]. WordNet은 의미망(semantic network) 형식으로 단어들을 색인하여 저장하여 두었기 때문에 일반 사전의 단편적인 알파벼식 사전보다 단어들간의 연관성을 알아내는 데에 훨씬 유용하다. WordNet의 범용성은 이미 자연언어처리 분야와 정보검색 분야에서 입증된바 있다. 그러나 불행히도 WordNet은 영어만을 대상으로 만들어졌기 때문에 한국어에는 적용할 수 없다. 한국어의 경우는 WordNet과 같이 범용적이고 정교한 시소러스가 아직 없는 실정이다.

사람에 의해 직접 만들어지는 시소러스는 비록 정확하기는 하지만 실제 사용에 있어서는 몇 가지 문제를 지니고 있다. 첫째, 시소러스의 구축 과정은 비용이 많이 든다. WordNet에 있는 8만여 단어를 사람이 일일이 의미에 맞춰 입력하기 위해서는 많은 사람들을

고용하여 사전 입력 작업을 하여야 하며 시간도 많이 걸린다. 시소러스를 만드는 작업 자체가 하나의 커다란 프로젝트이기도 하다. 둘째, 시소러스에 입력되는 정보는 아주 형식적인 체계에 맞춰 저장되기 마련이다. 사람의 직관 혹은 언어학적 이론을 바탕으로 체계적 분류를 하여 표준적인 어법에 대한 설명이 이루어진다. 그러나 우리가 다루어야 하는 문서 자료는 그 자체가 노이즈가 매우 심한 자료이다. 사람들은 표준 어법에 맞춰서 언어를 사용하는 경우는 오히려 예외적인 상황이고 상당한 부분의 경우 원래의 의미에서 벗어나거나 확장하는 등의 은유, 환유를 즐겨 사용한다. 시소러스에 의해서는 한 단어가 가지는 은유 및 환유를 잡아내기 어렵다. 셋째, 시소러스는 제작하고자 하는 의도에 맞추어 만들어진 것이기 때문에 변화하는 언어 환경에 적용하기 어렵다. 정보의 홍수 시대를 맞이하여 새로운 어휘들이 수없이 생성되고 지나간 단어들은 소멸해 버린다. 그러나 수동으로 구축된 시소러스는 급변하는 언어 환경을 제대로 반영하기 어렵다.

한편, WordNet처럼 사람이 직접 설계하고 입력하는 것이 아니라 컴퓨터에 의해 자동으로 구축하는 방법도 있다. 대규모 코퍼스에서 각 단어들의 공기(co-occurrence) 정보를 이용하여 각 단어들간에 어떠한 관련성이 있는지를 분석해 내는 것이다. 이 방법은 컴퓨터에 의해 완전 자동화된 방식으로 처리하기 때문에 시소러스 구축에 비용이 적게든다. 또한 영어든 한글이든 관계없이 동일한 방법으로 구축될 수 있고 새로운 언어 환경에 맞추어 계속 학습시킬 수도 있다. 다만, 구문론을 비롯한 언어학적 법칙을 전혀 고려하지 않기 때문에 종종 인간의 언어 직관에 어긋난 결과를 보이는 등의 문제점이 발생할 수 있다.

인간은 성장하면서 다양한 환경을 경험하며 이에 맞춰 언어를 습득하게 된다. 특정 상황하에서 그 상황에 맞는 언어표현은 함께 습득되기 때문에 인간의 언어생활은 삶의 양식과 떼어놓을 수가 없다. 반면, LSA에 의한 언어 학습은 인간 삶의 양식도 공유하지 않고 인간 언어의 중요 특징인 통사규칙도 고려하지 않은 채 오직 어떠한 단어와 어떠한 단어가 같은 문맥에서 함께 쓰이는가라는 공기(co-occurrence) 정보만을 이용한다. LSA는 이러한 단순 정보로부터 고차원의 개념공간을 생성해 낸다. 개념공간은 단순한 벡터공간에 불과한데 과연 인간의 복잡한 언어 습득을 흉내낼 수 있을까? 몇 가지 선행연구들은 LSA가 인간의 의미 습득에 필적하는 능력이 있음을 보여주었다.

1.2 관련 연구

개념학습 및 의미기억과 관련된 많은 계산론적 모델들은 임의로 설정한 특질들의 집합으로 개념을 명시적으로 나타내는 벡터 표현에 의존하고 있다. 이에 비해 LSA는 임의로 선정된 소규모 특질들의 집합이 아니라 대규모 코퍼스로부터 통계적인 분석을 거쳐 나온 고차원의 개념 벡터 공간으로 개념을 나타낸다. LSA를 이용해 만들어진 의미공간에서 개념들 간의 유사도는 두 벡터간의 코싸인 값으로 측정할 수 있다[Landauer, 1997].

단어들간의 연관성을 계산할 수 있으면 비슷한 단어들끼리 묶는 클러스터링(clustering)도 할 수 있다. 단어 클러스터링은 어떠한 단어들끼리 서로 뭉쳐서 저장되어 있는지를 보여주므로 인간의 심성어휘집(mental lexicon)의 구조를 탐구하는데 유용하다.

신경심리학적 연구에 의하면 인공물 범주의 사물들에 대한 명명 능력은 그대로 있는데 자연종과 같은 특정 범주의 사물들에 대한 명명 능력에 장애를 보이는 환자들이 있다고 한다. (Warrington, 1984) 그런데 LSA를 이용해 클러스터링을 해 본 결과 자연종 범주들은 인공물 범주에 비해 더 많이 집중되는 양상을 보인다고 한다. 기억에 대한 두뇌 구조가 LSA 구조에 대응된다고 가정해 본다면, 개념의 표상이 부분적 손상을 입었을 때 띄엄띄엄 있는 인공물 범주보다 밀집되어 있는 자연종 범주에서 서로를 구별해 내기가 어렵기 때문이라는 해석을 내릴 수 있다. 그림 1은 LSA의 유사도 측정법에 의해 계산된 결과를 보여주고 있다. 자연종 범주인 요인1,2,3은 큰 loading값을 보여준 반면에 chair, table, desk 같은 인공물 범주

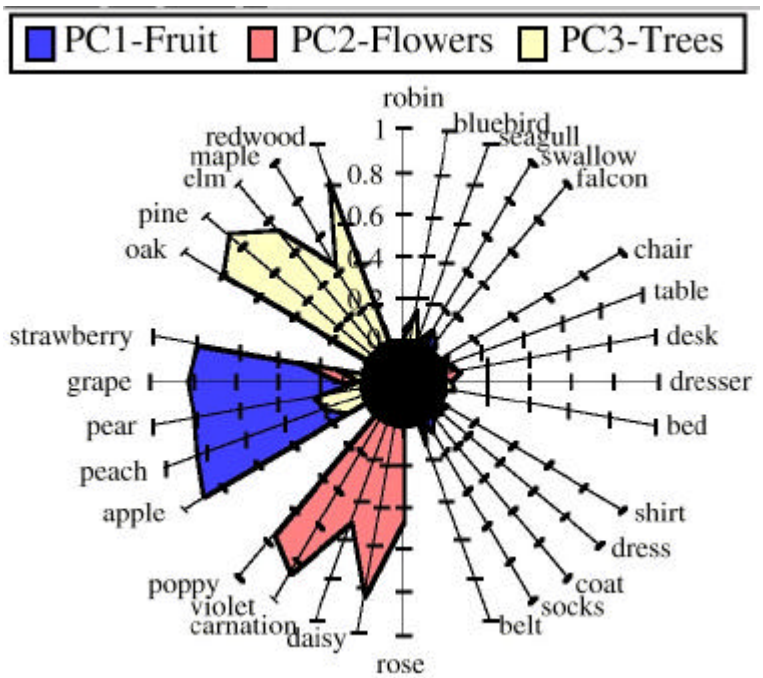

그림 1 Principal Components Factor Loadings

는 아주 작은 값을 가지고 있다[Laham, 1997].

LSA와 인간의 어휘지식간의 좀 더 직접적인 비교 실험도 있었다. Landauer 와 Dumais 는 고등학교 수준의 백과사전에서 5백만 단어를 뽑아 학습시킨 후 TOEFL(Test of English as a Foreign Language) 테스트를 해 보았다. 이 테스트는 단어 한 개나 구절을 제시하고 학생에게 이와 뜻이 가장 유사한 단어를 네 개중 고르도록 하는 테스트이다. LSA의 경우에는 제시된 단어나 구절을 벡터로 바꾸고 목표 단어와의 코싸인 유사도 계산법으로 답변하도록 했다. 그 결과 LSA는 80개 항목 중에서 64%의 정확률을 보였는데 이 수치는 비-영어권 외국인이 미국대학에 응시하기 위해 치르는 시험의 평균점수에 육박한다[Landauer and Dumais, 1997].

단순한 어휘테스트를 넘어 본격적인 인간 지식의 양상을 테스트해 본 실험도 있었다. 이 실험에서는 학생들에게 해부학과 심장의 기능, 심리학 개론에서의 현상, 파나마 운하의 역사 등과 같은 주제로 짧은 에세이를 쓰도록 했다. 그리고 LSA는 같은 주제 영역에서 뽑은 지시서들이나 관련 에세이들로 학습 시켜 100-350 차원의 개념 공간을 만들도록 했다. 이 공간에서 한편의 에세이는 단어들의 벡터를 단순히 합한 하나의 점으로 나타내어진다. 학생들의 에세이는 90-200 정도의 미리 채점된 에세이 중 가장 유사한 것 10개를 골라 이들을 평균한 점수로 매겼다. 실험 결과 에세이를 평가하는 사람들 사이의 상관계수는 각 주제별로 .83, .65, .82였었는데 LSA와 사람의 상관계수는 .80, .64, .84이었다. 비록 LSA의 평가는 단어의 순서를 무시한 상태에서 어떠한 어휘들이 주로 사용되었는가만 가지고 채점하는 것에 불과하지만 사람의 평가와 비슷한 결과를 내었다. 에세이는 인간의 수준 높은 지적 판단이 필요한 작업임을 감안

해 볼 때 통계적인 방법인 LSA가 보여준 결과는 놀랄만하다 [Landauer and Laham, 1997].

1.3 연구의 범위 및 논문의 구성

본 논문은 단어들의 국소 문맥(local context) 정보를 이용해 단어들간의 전역적(global) 의미관계를 찾아 주는 LSA 모델에 대해 다루고, 이를 한국어 텍스트 정보검색 시스템에 응용한 방법 및 결과를 분석한다. 1장의 서론에 이어 2장에서는 LSA의 수학적 기초가 되는 SVD 과정에 대해 설명한다. SVD는 메트릭스 연산의 일종인데 실제로 간단한 예를 중심으로 설명하여 나간다. 사용한 데이터는 인지과학회 논문 제목 50개와 이로부터 뽑은 25개의 색인어이다.

그 다음 본격적으로 LSA를 이용한 정보검색 시스템의 구축 방법 및 성능을 제시한다. 서울대학교 중앙도서관에서 제공하는 학위논문 초록 20,000여개를 기본 코퍼스로 삼아 정보검색 시스템을 구축한 후 이 검색 시스템을 바탕으로 LSA 실험을 진행한다. 우선 검색 시스템의 기능 중에 LSA의 문서-문서 유사도를 이용한 유사문서 찾기 기능을 추가하고 검색 시스템과는 별도로 LSA의 단어-단어 유사도 계산을 이용해 입력된 단어와 가장 유사한 단어들의 목록 출력 기능을 추가한다. 실험 결과는 여러 가지 경우를 표로서 제시하는데, 특히 단어-단어 유사도 계산의 결과를 통해 LSA가 인간의 어휘 습득에 얼마나 근사한 결과를 내는지 예증한다. 그밖에 본 시스템의 객관적 성능을 평가하기 위해 LSA를 초기부터 연구하여 온 Bellcore의 데모 사이트와의 비교를 시도한다.

결론 및 토론에서는 지금까지의 실험 결과를 요약하고 앞으로 발

전 및 적용시킬 수 있는 부분들에 대해 언급하였다.

2. Singular Value Decomposition (SVD)

2.1 개요

LSA는 다변량 통계분석 방법으로 고차원의 데이터 공간에 대해 축을 변경하여 그 데이터를 가장 잘 설명하는 새로운 축을 찾아내고 설명력이 낮은 축들은 걸러 내어냄으로서 데이터에 내재해 있는 구조를 밝히는 기법이다. 새로운 축을 찾아내기 위해서는 Singular Value Decomposition (SVD)이라는 통계적 기법을 사용하는데,

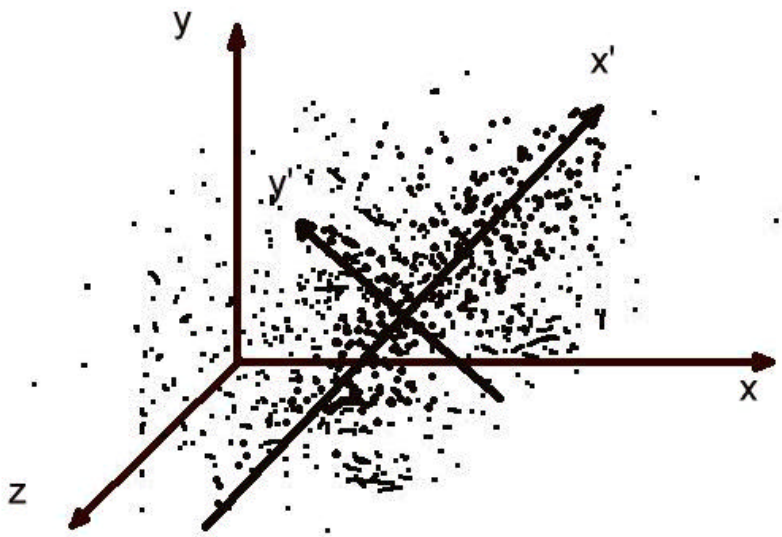

그림 2 LSA로 데이터의 축과 차원을 변경한 예

SVD를 거치면 설명력이 높은 순서로 원하는 수 만큼의 고유벡터를 얻을 수 있다. 이 때 설명력이 작은 축은 잡음으로 간주하여 제거함으로써 데이터의 차원을 줄이는 효과도 얻게 된다. 이 고유벡터들은 원래의 데이터 행렬에는 드러나 있지 않았던 의미구조를 표상하게 되므로 보다 심층적인 수준에서의 의미 분석을 가능하게 한다.

그림 2에서 보듯이 데이터들은 서로간의 연관성에 따라 일종의 구조를 지니게 되는데 LSA는 이 구조를 찾아낸다. 복잡하고 무작위적으로 보이던 x,y,z의 3차원 공간에 존재하던 데이터들이 LSA에 의해 차원이 축소된 x',y'로 투사(projection)된다. 즉, 본래는 3차원상에 존재하던 데이터들이었으나 실은 2차원의 구조 안에서 설명이 가능하게 된 것이다. 차원을 줄인다는 것은 단지 메모리의 효율을 올리기 위한 것만은 아니다. 일반적으로 데이터들이 가지게 되기 마련인 노이즈를 제거하여 좀 더 일반화되고 단순한 모델을 얻게 되는 이점이 생긴다.

2.2 SVD의 원리

Singular Value Decomposition(SVD)은 무제약 선형 최소 제곱 문제(unconstrained linear least squares problem), 행렬의 rank 추정, 정규 상관계수분석(canonical correlation analysis)에 주로 사용된다. 이를 응용한 정보검색, 실시간 신호처리의 문제에서는 최대한 빠른 시간 안에 계산을 해 내는 것이 중요하다. 따라서 현재에는 병렬 머신의 발달과 더불어 SVD를 병렬 컴퓨터에 구현하는 것이 최대의 관심사로 떠오르고 있다. 정보검색에서 다루는 거대한 단어-문

서 행렬은 매우 sparse한 행렬이다. 따라서 정보검색 분야에서는 sparse 행렬에 대한 SVD를 빠르게 계산하는 문제가 크게 부각되고 있다. 현재 제안된 대표적인 방법에는 Lanczos, block-Lanczos, 부분공간 반복법(subspace iteration), 추적 최소법(trace minimization method)등이 있다.

A를 m-by-n (m>>n)이고, rank(A)=r인 행렬이라 한다면, 행렬 A에 대한 SVD는 다음과 같이 정의된다.

$$A = U\Sigma V^T \qquad (1)$$

U와 V는 각각 orthonormal하기 때문에 $U^TU = V^TV = I_n$ 가 성립한다. $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$는 singular value들의 대각 행렬인데 $1 \le i \le r$ 사이에서는 양수값($\sigma_i > 0$)을 갖고 $i \ge r+1$ 에서는 영($\sigma_i = 0$)이다. orthogomal 행렬인 U와 V의 처음 r 개의 열들은 각각 AA^T와 A^TA의 r개의 eigenvalue들과 연관된 orthonormal eigenvector를 정의한다. A의 singular value는 AA^T의 n개의 eigenvalue들을 제곱한 양수들인 Σ의 대각 요소들로 정의된다. A의 i번째 singular triplet(vectors)은 집합 $\{u_i, \sigma_i, v_i\}$이다. 큰 값(작은 값)의 singular 값과 연관된 singular 벡터를 큰(작은) singular 벡터라고 한다. U의 처음 r 개의 열은 left singular vector라 불리고, V의 처음 r 개의 열은 right singular vector라 불린다.

여기에서 LSA의 문제는 m-by-n 의 성긴(sparse) 행렬이 주어졌을 때 식 (1)에 정의된 대로 p개의 큰 singular 벡터들을 구하는 문제가 된다.

다음 두 정리는 SVD**가 어떻게 행렬 형태의 데이터 구조를 밝혀 낼 수 있는지를 보여준다.**

theorem 1. Let the SVD of A be given by (1) and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0$$

and let $R(A)$ and $N(A)$ denote the *range* and *null space of* A, respectively, then

1. *Rank property*:

$rank(A) = r, \quad N(A) \equiv span\{v_{r+1}, \cdots, v_n\}$,

$R(A) \equiv span\{u_1, \cdots, u_r\}$ where

$U = [u_1 u_2 \cdots u_m]$ and $V = [v_1 v_2 \cdots v_n]$.

2. *Dyadic decomposition:*

$$A = \sum_{i=1}^{r} u_i \cdot \sigma_i \cdot v_i^T$$

3. *Norms*:

$||A||_F^2 = \sigma_1^2 + \cdots + \sigma_r^2$, and $||A||_2^2 = \sigma_1$

*Rank property***는** A**의** singular **값들을 정성적**(qualitative)**인** rank**의 개념을 정량적**(quantitative)**인 측정치로 사용할 수 있도록 한다**. *Dyadic decomposition***은 데이터 압축의 근거를 제공해 주며**, **행렬을** singular **값들에 따라 내림차순으로 정렬된 벡터들의 합으로 표시할 수 있음을 보여준다**.

theorem 2. Let the SVD of A be given by (1) with

$r = rank(A) \leq p = \min(m, n)$ and define:

$$A_k = \sum_{i=1}^{k} u_i \cdot \sigma_i \cdot v_i^T \quad with \quad k < r,$$

then

$$\min_{r(B)=k} ||A - B||_F^2 = ||A - A_k||_F^2 = \sigma_{k+1}^2 + \cdots + \sigma_p^2 .$$

이 정리는 최소제곱법(least squares)**적 의미에서** A_k**가** A**에 대한 최상의** rank-k **근사치임을 보여준다**.

2.3 SVD의 계산방법

SVD**를 계산하는 과정은 다음과 같다**.

1. **행렬** A**는** bidiagonal **형태로** reduce **한다**. $A = U_1 B V_1^T$ U_1**과** V_1**은** othogonal **하며**, B**는** $m \geq n$ **일 경우 상위**-bidiagonal **이며**, $m < n$ **일 경우는 하위**-bidiagonal**이다**.

2. bidiagonal **행렬인** B**의** SVD**는** $B = U_2 \Sigma V_2^T$ **로 계산되는데**, U_2**와** V_2**는** othogonal **하며**, Σ**는** diagonal **하다**.

3. **그 다음**, **행렬** A**의** singular vector **들은** $U = U_1 U_2$, $V = V_1 V_2$**로 계산한다**.[1)]

1) http://nurapt.kaist.ac.kr/lapack/lug/node55.html

한편, SVD는 선형 요인화(factorization) 기법으로 전달함수가 선형인 다층 퍼셉트론과 동등하다[Landauer, 1997].

3. Latent Semantic Analysis (LSA)

3.1. 개요

앞서 살펴본 SVD에 의해 텍스트 코퍼스를 분석하여 의미 구조를 찾아내는 LSA에 대해 살펴보자. 코퍼스 안의 임의의 문서는 통사구조를 가지고 있는 문장들의 집합으로 되어 있다. 그러나 LSA에서는 통사구조를 무시하고 단어들만을 고려하기 때문에 문서 d_i 에는 어떤 단어들이 있는가의 정보만 사용하여 벡터로 표현할 수 있다.

$$d_i = <t_1, \ldots, t_n> \qquad (2)$$

t_j 는 문서에 그 단어가 포함되어 있으면 1 그렇지 않으면 0으로 나타낼 수도 있고 문서 안에서 몇 번의 빈도로 나타났는가에 따라 정수로 줄 수도 있는데, 현재 정보검색 분야에서는 대부분 $tf \cdot idf$ 로 계산된 가중치 값을 사용한다[Salton and McGill, 1983].

$$w_{ij} = tf_{ij} \cdot \log\left(\frac{N}{df_j}\right) \qquad (3)$$

w_{ij} 는 i 번째 문서의 j번째 단어의 가중치 곧 중요도를 나타내며, tf_{ij} 는 i번째 문서에서 j번째 단어가 나타난 횟수를, N운 전체 코퍼스에 존재하는 문서의 개수, df_j는 j번째 단어가 나타난 문서의 수를 나타낸다. 식 (3)을 잘 살펴보면 tf 항목은 문서에서 빈도수가 높은 단어에 가중치를 많이 주고 idf 항목은 반대로 여러 문서에 나타나는 단어의 가중치를 감소시킨다. 예를 들어 어떤 문서에 '경제' 라는 단어가 수십 번 나왔다면 그 문서는 경제와 관련된 문서일 가능성이 크나, 영어에서 정관서 'the'와 같은 단어는 그 단어가 아무리 많이 나와도 문서의 특성을 결정하지 못한다. 이와 같이 가중치를 주는 방법은 정보이론에 따라 정보량이 많은 단어에 많은 가중치를 주는 한편 간단한 계산으로 가능하다는 장점이 있다. 정보이론에 좀 더 충실한 방법은 엔트로피의 개념을 써서 다음과 같이 계산한다.

$$w_{ij} = \frac{\log(freq_{ij})}{-\sum\left(\left(\frac{freq_{ij}}{\sum freq_{ij}}\right) * \log\left(\frac{freq_{ij}}{\sum freq_{ij}}\right)\right)} \qquad (3) - 1$$

본 논문에서는 계산상의 경제성 때문에 (3)의 방법을 따랐다.

각 문서들은 단어들의 집합인 벡터이므로 문서들의 집합인 전체 코퍼스는 문서-단어의 행렬로 나타낼 수 있다. 보통 색인어(indexing

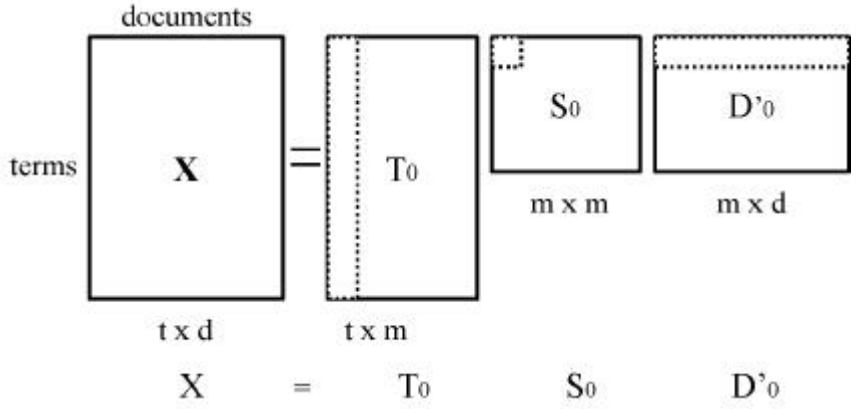

그림 3 단어-문서 행렬에 대한 SVD

terms) 수는 수만에서 수십만에 이르고 문서 수도 수만에서 수백만에 이르기 때문에 이 행렬은 대규모 sparse 행렬이 된다. 이제 이 행렬에 SVD를 적용만 하면 된다.

수식은 위와 동일하다. 예를 들어 단어의 수가 t 이고 문서의 수가 d 인 t x d 행렬 X는 식 (1)-1 과 같이 분해가 된다.

$$X = T_0 S_0 D_0 \qquad (1)\text{-}1$$

여기에서 T_0, D_0 는 orthonormal 열로 이루어지며 각각 left singular vector 와 right singular vector 들의 행렬이고, S_0 는 singular value의 diagonal 행렬인데 보통 singular value 값이 큰 순서대로 정렬되어 있다. 그림 3에 단어-문서 행렬에 대한 SVD의 도식이 나타나 있다.

T_0 : orthogonal, unit-length columns

D_0 : orthogonal, unit-length columns

S_0 : **고유치들의** diagonal **행렬**

t : **행렬** X**의 행 수(단어 수)**

d : **행렬** X**의 열 수(문서 수)**

m : **행렬** X**의** rank ($\leq$ min(t,d))

행렬 X**와 완전히 동일한 분해** $T_0 S_0 D_0$**를 얻기 위해서는 각 행렬들이 모두** full rank **이어야 한다.** full rank**를 모두 사용하지 않고 그보다 적은** k (k<m)**개 만큼의 벡터만을 사용하면 그것은 원래의 행렬** X**에 대한 근사치가 된다. 각 벡터에 대응되는 고유치들은 그 벡터의 설명력을 나타낸다. 고유치 값이 클수록 그 벡터는 주어진 데이터 공간을 잘 설명하는 것이다. 그러므로 여기서 우리가 관심을 갖는 것은 오히려 근사치 값이다. 일정한 수만큼의 벡터를 취하면 원래 데이터 행렬에 나타나 있던 노이즈를 제거하게 되고 좀 더 일반화된 정보를 취하게 되는 것이다.**

$$X \approx \hat{X} = TSD' \qquad (1)\text{-}2$$

T : t x k

S : k x k

D‘ : k x d

식(1)-2에서 X의 추정치 $\hat{X}$은 k개의 차원으로 축소(reduce) 되었다. k 의 값에 따라 얼마나 인간의 어휘지식과 유사한 표상을 하는지가 결정된다. SVD의 결과를 통해 우리는 3가지 비교를 할 수 있다: 단어 i와 단어 j는 얼마나 유사한가, 문서 i와 문서 j는 얼마나 유사한가, 단어 i와 문서 j는 얼마나 연관성이 있는가 ? 단어와 단어간의 유사도는 시소러스적 용도로 사용될 수 있고 문서와 문서간의 유사도는 문서 분류에, 단어와 문서간의 유사도는 문서 검색에 각각 사용될 수 있다.

3.2. LSA를 적용한 텍스트 분석 예

SVD을 이용하여 LSA를 구하는 간단한 예제를 살펴보자. 다음 자료는 인지과학 관련 논문 제목 50개를 뽑아 놓은 것이다. 각 논문 제목들은 사용된 단어들에 대해 국소적 문맥을 제공하여 준다.

c01 = 글 이해시의 상황모형과 대용어 참조해결
c02 = 글의 첫 정보가 이해와 표상에 미치는 효과
c03 = 글 이해에서 떨어져 있는 인과정보의 통합
c04 = 글자의 이중상에 의한 심리적 혼란감과 지각 통합 과정
c05 = 대뇌반구간 한글 단어처리와 색채처리 특성
c06 = 대명사의 참조관계처리시의 맥락의 역할
c07 = 대형구조가 글 이해에 미치는 영향
c08 = 마음, 지능, 인공지능의 정의에 관한 철학적 반성
c09 = 사전지식과 정교화가 기억에 미치는 영향
c10 = 상호작용 활성화 모형과 초보 지각자·기억자 모형의 비교 검증
c11 = 상호정보를 이용한 구조적 모순성 해소와 결과에 대한 확신도 측정
c12 = 신경망 모형에 의한 범주화 연구
c13 = 신경망을 이용한 인쇄체 한글 인식 시스템
c14 = 심성표상의 심리학적 고찰: 지식표상체계들의 특성
c15 = 연결주의 표상의 계산신경과학적 배경
c16 = 연결주의와 원자주의
c17 = 연결주의의 철학적 기초
c18 = 원정시합: 연결주의자와 구성적 구조
c19 = 의미 네트워크를 이용한 단어 의미의 모호성 해결 방법
c20 = 의미와 표상: 논항 구조의 문제
c21 = 이해능력에 따른 대용어 처리시 억압기제의 효율성 차이
c22 = 인과적 계산이론과 언어이해
c23 = 인지과학의 방법론- 계산주의와 연결주의를 중심으로
c24 = 인지과학의 지각심리학적 기초
c25 = 인지체계 속의 언어기능
c26 = 자극의 감정가가 정보처리에 미치는 효과
c27 = 자연언어 이해의 전산적 가능성
c28 = 자연주의적 인식론의 한계
c29 = 점진적 개념 집단화에 의한 문헌정보의 분류
c30 = 정보 가변추리의 특징과 논리적 구조
c31 = 정보검색 모델에서 개념적 거리를 이용한 추론
c32 = 정서가 정보처리과정에 미치는 영향에 대한 의미망 모형 검증
c33 = 주의 수준이 암묵 기억과 외현 기억에 미치는 영향
c34 = 지식 갱신의 기저구조로서의 상황모형

c35 = 지식 베이스에 의한 합성 명사의 의미 관계 분석 시스템
c36 = 지식의 표현의 모델
c37 = 지식처리를 위한 퍼지 논리와 신경망의 결합모델 고찰
c38 = 직선적 삼단논법의 형상효과: 언어 이해설
c39 = 처리깊이에 따른 학습단어의 반복제시가 단어완성검사와 단서회상검사에 미치는 효과
c40 = 추론모형에 기초한 믿음과 지식의 증명
c41 = 통사구조와 인지구조와의 관계
c42 = 퍼지 추론을 이용한 한글 문자 인식: 최대 길이 투영에 의한 한글 문자 유형분류
c43 = 학습과 기억의 생물학적 기초: 신경심리학적 개관
c44 = 한글과 한자 단어 재인 비교 연구
c45 = 한글의 이해와 산출의 심리적 과정
c46 = 한글 글자의 내부 구조와 글자 인지과정
c47 = 한글 글자의 통합 처리에 대한 반복효과 연구
c48 = 한글 낱말의 처리단위
c49 = 한글 다의 단어의 의미적 활성화와 맥락효과
c50 = 한글 단어 재인에 있어서 글자처리와 낱자의 지각

t01=**개념**, t02=**관계**, t03=**구조**, t04=**기억**, t05=**단어**,
t06=**마음**, t07=**모델**, t08=**모형**, t09=**신경**, t10=**심리**,
t11=**언어**, t12=**연결**, t13=**의미**, t14=**이해**, t15=**인식**,
t16=**인지**, t17=**정보**, t18=**주의**, t19=**지각**, t20=**지능**,
t21=**지식**, t22=**처리**, t23=**표상**, t24=**학습**, t25=**한글**

표 1 색인어 목록

	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	...
t01	0	0	0	0	0	0	0	0	0	0	...
t02	0	0	0	0	0	1	0	0	0	0	...
t03	0	0	0	0	0	0	1	0	0	0	...
t04	0	0	0	0	0	0	0	0	1	1	...
t05	0	0	0	0	1	0	0	0	0	0	...
t06	0	0	0	0	0	0	0	1	0	0	...
t07	0	0	0	0	0	0	0	0	0	0	...
t08	1	0	0	0	0	0	0	0	0	2	...
t09	0	0	0	0	0	0	0	0	0	0	...
t10	0	0	0	1	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...

표 2 단어-문맥 행렬

	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	...
t01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
t02	0.00	0.00	0.00	0.00	0.00	2.77	0.00	0.00	0.00	0.00	...
t03	0.00	0.00	0.00	0.00	0.00	0.00	1.79	0.00	0.00	0.00	...
t04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.48	2.48	...
t05	0.00	0.00	0.00	0.00	2.08	0.00	0.00	0.00	0.00	0.00	...
t06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.91	0.00	0.00	...
t07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
t08	2.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.16	...
t09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
t10	0.00	0.00	0.00	2.30	0.00	0.00	0.00	0.00	0.00	0.00	...
...	...	...	...	...	...	...	...	...	...	...	...

표 3 tf*idf로 가중치를 준 후의 단어-문맥 행렬

표 1은 위 문맥에서 색인어로 뽑아 놓은 25개의 단어(색인어)들이다. 각 문맥에 대하여 각 단어의 나타난 횟수에 따라 표 2와 같은 단어-문맥 행렬이 만들어진다. 가로축은 문맥을 나타내고 세로축은 단어를 나타낸다.

각 문맥에 등장한 단어들의 가중치는 앞서 언급한 식 (3)의 $tf*idf$로 계산하여 준다.

표 3은 식 (1)-1의 SVD에 의해 다음 3개의 행렬로 분해가 된다.

	v01	v02	v03	v04	v05	v06	v07	v08	v09	v10	...
v01	9.41	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
v02	0.00	8.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
v03	0.00	0.00	8.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	...
v04	0.00	0.00	0.00	7.68	0.00	0.00	0.00	0.00	0.00	0.00	...
v05	0.00	0.00	0.00	0.00	7.43	0.00	0.00	0.00	0.00	0.00	...
v06	0.00	0.00	0.00	0.00	0.00	6.90	0.00	0.00	0.00	0.00	...
v07	0.00	0.00	0.00	0.00	0.00	0.00	6.67	0.00	0.00	0.00	...
v08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.53	0.00	0.00	...
v09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.08	0.00	...
v10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.97	...
...	...	...	...	...	...	...	...	...	...	...	...

표 4 diagonal eigen value S

	t01	t02	t03	t04	t05	t06	t07	t08	t09	t10	...
v01	0.03	- 0.00	0.02	- 0.06	- 0.09	0.00	0.02	0.41	- 0.14	- 0.17	...
v02	0.11	0.00	0.05	- 0.01	- 0.23	- 0.24	- 0.01	- 0.08	0.23	- 0.21	...
v03	0.21	- 0.00	- 0.04	- 0.03	- 0.41	- 0.41	- 0.19	- 0.07	0.18	- 0.2	...
v04	0.26	- 0.00	- 0.08	- 0.37	0.49	- 0.15	0.13	- 0.09	- 0.04	- 0.21	...
v05	0.28	- 0.00	0.43	0.34	0.13	0.06	0.16	- 0.13	- 0.18	- 0.17	...
v06	- 0.00	- 0.45	0.00	0.00	0.00	- 0.00	- 0.00	0.00	- 0.00	- 0.00	...
v07	0.08	- 0.00	0.05	- 0.12	- 0.07	0.09	0.04	0.48	0.12	- 0.21	...
v08	0.21	- 0.00	0.09	- 0.41	0.15	- 0.37	0.14	0.10	- 0.01	0.47	...
v09	0.22	0.00	- 0.02	- 0.12	0.13	0.23	- 0.06	0.18	0.24	- 0.01	...
v10	0.19	0.00	0.03	- 0.24	0.04	0.25	- 0.30	- 0.27	- 0.05	- 0.1	...
...	...	...	...	...	...	...	...	...	...	..	...

표 5 othonormal eigen vector T

	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	...
v01	0.06	- 0.00	0.03	- 0.13	0.02	- 0.11	- 0.01	0.05	- 0.13	0.22	...
v02	0.11	- 0.00	0.00	- 0.13	- 0.21	0.19	- 0.06	0.05	- 0.23	0.09	...
v03	0.04	0.00	0.02	- 0.04	- 0.08	- 0.02	- 0.05	0.13	- 0.21	0.04	...
v04	0.09	0.00	0.04	- 0.12	0.07	0.04	- 0.18	- 0.16	- 0.05	- 0.04	...
v05	0.19	- 0.00	0.28	0.23	0.09	0.04	- 0.07	0.10	- 0.03	- 0.03	...
v06	0.07	0.00	0.08	0.04	- 0.07	- 0.09	- 0.02	0.03	0.09	- 0.12	...
v07	0.05	- 0.00	- 0.00	- 0.03	- 0.13	- 0.11	- 0.11	- 0.00	- 0.07	- 0.00	...
v08	- 0.00	- 1.00	- 0.00	0.00	0.00	0.00	- 0.00	0.00	- 0.00	- 0.00	...
v09	0.11	- 0.00	- 0.01	- 0.20	0.13	- 0.03	0.08	0.00	0.08	- 0.10	...
v10	0.20	- 0.00	0.05	- 0.40	0.31	- 0.32	0.06	- 0.04	- 0.05	0.26	...
...	...	...	...	...	...	...	...	...	...	...	...

표 6 othonormal eigen vector D

행렬 T,S,D는 원래의 데이터를 정확하게 복원해 낼 수 있다. 그러나 LSA에서 관심을 가지는 사항은 전체 텍스트 집합이 몇 개의 개념 공간으로 분석될 수 있는가이므로 이중 상위 k개의 벡터만을 사용한다.

표 7는 k=3으로 차원을 줄였을 때의 단어-문맥 행렬을 보여준다. 차원을 줄이기 전의 행렬인 표 3와 차원을 줄인 후인 표 7를 비교해 보면 즉각 0 아닌 요소들이 많다는 것을 알 수 있다. 이는 국소적 문맥만을 고려하여 계산된 원래의 행렬에서는 알 수 없었던 단어와 문서와의 심층적 관련성이 전체 문맥을 고려한 LSA에 의해 새롭게 출현한 것이다.

	c01	c02	c03	c04	c05	c06	c07	c08	c09	c10	...
t01	0.02	0.03	0.01	0.03	0.11	0.04	0.01	0.00	0.03	0.07	...
t02	0.07	0.11	0.04	0.10	0.30	0.11	0.05	-0.00	0.10	0.22	...
t03	0.11	0.21	0.06	0.16	0.27	0.12	0.11	0.00	0.22	0.39	...
t04	0.13	0.26	0.07	0.19	0.26	0.13	0.13	0.00	0.27	0.46	...
t05	0.26	0.28	0.17	0.36	1.54	0.49	0.13	0.00	0.24	0.71	...
t06	0.00	0.00	-0.00	-0.00	0.00	-0.00	0.00	3.91	0.00	0.00	...
t07	0.06	0.08	0.04	0.08	0.26	0.09	0.04	0.00	0.08	0.18	...
t08	0.14	0.21	0.09	0.20	0.59	0.21	0.11	0.00	0.20	0.44	...
t09	0.12	0.21	0.07	0.17	0.33	0.14	0.11	-0.00	0.22	0.41	...
t10	0.11	0.19	0.07	0.17	0.40	0.15	0.10	-0.00	0.19	0.37	...
...	...	...	...	...	...	...	...	...	...	...	...

표 7 차원을 줄였을 때(k=3)의 단어-문맥 행렬

LSA의 결과 이전에는 서로 관련이 없던 단어들끼리도 어떠한 관련을 맺게 되었다.[2] 표 8은 LSA의 최대 미스테리라 할 수 있는

선택된 차원의 수에 따라 단어-단어 유사도가 바뀌는 것을 보여준다. 본 논문에서의 유사도 계산은 몇 경우를 제외하고는 일반적으로 가장 많이 사용하는 코싸인 방법을 사용하였다.[3)]

$$\text{유사도}(a, b) = \cos\theta = \frac{a \cdot b}{||a|| ||b||} \qquad (4)$$

식 (4)에 의해 사물 a와 사물 b의 유사도는 그것들이 표현되어 있는 벡터 공간상에서 이루는 각도의 코싸인 값으로 계산된다. 그러므로 동일할 때에는 1이 되고 무관 또는 독립일 경우에는 0이 된다.

코싸인 값을 취하면 어떤 경우에서든 자기 자신과의 유사도는 항상 1이 되기 때문에 표 8에서는 다른 단어들 중에 가장 유사한 단어 두 개씩을 골라 나타내었다. '마음'에 대해서는 '지능'만이 나왔고 '지능'에 대해서는 '마음'만 나온 이유는 이 두 단어가 사용된 문맥은 오직 'c08 = 마음, 지능, 인공지능의 정의에 관한 철학적 반성'뿐인데 이 문맥에서는 다른 문맥과 연결되는 단어들을 가지고 있지 않기 때문이다. 즉, c08 문맥은 완전히 고립되어 있다.

한편, '주의'라는 단어는 '연결', '인지'와 높은 유사도를 나타내는데, 각각 신경망의 '연결주의(ism)'와 심리학의 '주의(attention)'가 반영된 결과이다. LSA에서는 이처럼 동음이의어(同音異議語)의 문제가 반영된다. 이로부터 파생된 문제로 '연결'이란 단어의 유사 목록

---

2) 여기서 단어들간의 관련이 있고 없고는 사전적 의미나 인간의 언어지식에 의한 것이 아니라 그 두 단어가 같은 문맥에서 사용된 적이 있는가 없는가만을 말한다.

3) 정보검색 분야에서는 질의어와 문서와의 유사도를 계산하는 무수한 방법들이 제안되었는데 코싸인 방법은 간편하게 사용할 수 있으면서도 통계적 의미가 있는 값을 산출한다. 때로는 더 간단하게 두 벡터의 내적에 의해 계산하기도 한다.

	LSA로 계산된 유사도가 높은 단어들		
	k=25	k=5	k=3
개념	정보(0.50) 모델(0.41)	정보(0.97) 이해(0.96)	학습(1.00) 정보(1.00)
관계	구조(0.35) 인지(0.26)	구조(0.95) 언어(0.95)	이해(1.00) 모형(1.00)
구조	인지(0.40) 관계(0.35)	언어(0.98) 관계(0.95)	언어(1.00) 표상(1.00)
기억	학습(0.27) 모형(0.25)	지각(0.87) 신경(0.87)	표상(1.00) 구조(1.00)
단어	학습(0.47) 처리(0.46)	한글(0.99) 처리(0.98)	처리(1.00) 의미(0.99)
마음	지능(1.00)	지능(1.00)	지능(1.00)
모델	지식(0.44) 개념(0.41)	지식(0.98) 이해(0.95)	지각(1.00) 정보(1.00)
모형	지각(0.33) 기억(0.25)	심리(0.96) 지각(0.90)	관계(1.00) 이해(1.00)
신경	학습(0.32) 인식(0.26)	심리(0.88) 지각(0.88)	언어(1.00) 구조(0.99)
심리	지각(0.45) 표상(0.34)	모형(0.96) 지식(0.89)	인식(1.00) 지식(0.99)
언어	이해(0.50) 인지(0.22)	구조(0.98) 관계(0.95)	구조(1.00) 신경(1.00)
연결	주의(0.87) 인지(0.20)	주의(0.99) 인지(0.82)	주의(1.00) 인지(0.93)
의미	단어(0.35) 관계(0.20)	관계(0.86) 정보(0.81)	한글(1.00) 처리(1.00)
이해	언어(0.50) 정보(0.24)	정보(0.98) 개념(0.96)	관계(1.00) 모형(1.00)
인식	한글(0.48) 신경(0.26)	한글(0.81) 단어(0.74)	심리(1.00) 지식(0.98)
인지	구조(0.40) 관계(0.26)	구조(0.83) 연결(0.82)	기억(0.97) 표상(0.96)
정보	개념(0.50) 이해(0.24)	이해(0.98) 개념(0.97)	모델(1.00) 지각(1.00)
주의	연결(0.87) 인지(0.25)	연결(0.99) 인지(0.75)	연결(1.00) 인지(0.94)
지각	심리(0.45) 모형(0.33)	모형(0.90) 신경(0.88)	정보(1.00) 모델(1.00)
지능	마음(1.00)	마음(1.00)	마음(1.00)
지식	모델(0.44) 표상(0.29)	모델(0.98) 이해(0.91)	관계(0.99) 이해(0.99)
처리	단어(0.46) 한글(0.38)	한글(0.98) 단어(0.98)	의미(1.00) 단어(1.00)
표상	심리(0.34) 지식(0.29)	언어(0.94) 이해(0.93)	구조(1.00) 기억(1.00)
학습	단어(0.47) 심리(0.32)	지각(0.87) 처리(0.77)	개념(1.00) 모델(0.99)
한글	인식(0.48) 처리(0.38)	단어(0.99) 처리(0.98)	의미(1.00) 처리(1.00)

표 8 선택된 singular vector의 수에 따른 단어-단어 유사도

을 살펴보면 사실상 별 관련성이 없는 '인지'와 높은 유사도를 내게 된다. '연결'은 '주의(ism)'와 높은 유사도를 갖고, '주의(attention)'는 '인지'와 높은 유사도를 갖기 때문에 전이적(transitive)으로 '연결'이 '인지'와 높은 유사도를 갖게 되었다. 그런데 '인지'의 경우를 보면 '연결'과는 별다른 유사성을 갖지 않음을 볼 수 있다. 이러한 비대칭성(asymmetric)은 '연결'의 경우 오로지 '주의(ism)'와만 문맥을 공유하는 반면, '인지'는 '주의(attention)'이외의 다른 단어들과도 많은

eigen vector 수	그룹1	그룹2	그룹3	그룹4	그룹5	그룹6
k=25	구조 연결 의미 이해 정보 주의 지식 표상	개념 관계 마음 모델 모형 언어 인식 인지 지능	기억 신경 심리	단어 학습	처리 한글	지각
k=5	연결 주의	개념 관계 구조 모델 모형 신경 심리 언어 이해 인지 정보 지식 표상	기억	단어 의미 지각 처리 학습 한글	마음 지능	인식
k=3	구조 기억 신경 언어 연결 인식 인지 주의 표상	개념 관계 단어 모델 모형 의미 이해 정보 지각 지식 처리 학습 한글	심리		마음 지능	

표 9 LSA의 결과를 Self Organizing Map으로 클러스터링한 결과

문맥을 공유하고 있기 때문에 중의적인 문제가 어느 정도 해소된 것으로 보인다[4].

표 9에서는 표 8에서 나열한 단어들간의 전체적인 유사도를 살펴보기 위해 Self Organizing Map으로 클러스터링을 해 보았다. 사용된 singular value의 수가 적을수록, 즉 고려하는 개념 공간의 차원이 작을수록 단어들은 집중화되었고 클수록 넓은 분포를 보였다. 표 9를 볼 때 유의해야 할 점은 이 결과가 대규모의 코퍼스에서 도출

4) **인간의 유사도 판단에서 보이는 비대칭성은 카네만과 트버츠키에 의해 지적된 바 있다.**

5) **본 실험은** http://krcogsci.snu.ac.kr/~krishna **에서 직접 수행하여 볼 수 있다.**

된 것이 아니라 단 50개의 문맥에서 나왔다는 것이다. 그러므로 표 11을 해석할 때에는 일반적인 의미보다는 '인지과학 관련 논문 제목 50개'라는 상황하에서 해석하는 것이 좋다. 대체적으로 보았을 때 k=5인 경우 가장 잘 분류를 해 낸 듯 하다. 각 클러스터를 '인지과학 관련 논문 제목 50개'라는 상황에 맞춰 해석해 보면, 각각 신경망 분야, 인지심리학, 학습, 한글처리분야, 인공지능 및 철학, 인지과학 일반으로 분류했다는 것을 알 수 있다.

4. 정보검색 시스템에의 응용[5)]

4.1 개요

정보검색 시스템이란 대량의 자료들을 색인 및 저장한 후 질의어(query)로 표현되는 사용자의 정보요구(information need)에 부합되는 자료를 찾아주는 시스템이다. 현재 이미지 검색, 음성 자료 검색을 포함하는 멀티미디어 정보검색에 대한 연구가 활발하지만 아직까지는 텍스트를 중심으로 하는 정보검색이 주도적이다. 텍스트 정보검색은 한마디로 도서관의 사서와 같은 역할을 한다. 도서관에서는 대량의 텍스트 자료들을 주제에 따라 듀이 10진 분류법(DDC)으로 보관하여 놓아 사용자가 관련 서적을 쉽게 찾을 수 있도록 해 놓았다. 그러나 정보검색 시스템에서는 텍스트의 내용을 사전에 분석하여 주제별로 나누어 놓기 어렵기 때문에 오로지 색인어들에 기반한(term-based) 포스팅화일(posting file)이라는 역화일(inverted file) 구조로 저장한다[6)]. 이 과정을 색인(indexing)이라 하는데 전통적으로는 글의 저자나 사서들이 직접 내용에 알맞는 색인어들을 추출하여 색인 하여 왔으나 현재에는 자동화된 색인 기법을 많이 사용한다.

정보검색의 또 다른 과정은 검색(retrieval)이다. 검색과정은 사용자의 정보요구를 질의어로 만드는 질의어 생성과정, 질의어의 조건에 맞는 문서를 찾는 유사도 계산 및 우선 순위 계산 과정, 결과를

---

5) 본 실험은 http://krcogsci.snu.ac.kr/~krishna 에서 직접 수행하여 볼 수 있다.

6) 단어를 기반으로 한 검색은 중의성등의 문제로 인해 한계를 지니고 있다. 본 논문에서 구현한 문서-문서 유사도에 의한 검색은 이런 한계를 벗어나 사람의 검색 방식과 비슷한 주제 중심, 내용 중심 검색이라 할 수 있다.

사용자에게 보여주는 제시(presentation) 과정으로 나누어 볼 수 있다. 질의어에 의한 검색 부분을 따로 떼어놓고 보면 - 특히 불리언(Boolean) 검색일 경우, 정보검색의 문제는 기존 데이터베이스 분야에서 발달되어온 SQL(Structured Query Language)에 의한 검색과 다를 바가 없어 보이나 정보검색에서 전제로 하고 있는 몇 가지 사항들에 의해 데이터베이스와 구별해 볼 수 있다. 첫째, 정보검색의 사용자들은 자신의 정보요구가 무엇인지 정확히 모른다. 대충 어떤 주제의 문서를 찾으려 하는 것이지 정확하게 '1998년에 출판된 김철수의 책'만을 검색하는 것은 아니다. 다음은 데이터베이스 분야에서 사용하는 SQL의 예제이다.

```
SELECT BOOK FROM LIBRARY WHERE
YEAR=1998 AND AUTHOR=김철수
```

둘째, 정보검색의 사용자들은 자신의 정보요구를 정확한 질의어로 만들 능력이 없다. 이 문제는 검색 시스템이 전적으로 사용자의 질의어에만 의존해서는 안 된다는 것을 말해준다. SQL의 경우는 WHERE 이하의 조건 식에 맞는 사례들은 모두 정확한 것들이므로 모두 검색되어야 하고 조건에 맞지 않는 사례들은 검색되어서는 안 된다. 그러나 정보검색의 문제는 이와 같은 가정을 하지 않기에 검색의 목표를 질의어에 맞추는 것이 아니라 사용자의 정보요구에 맞추고 있다. 이는 곧 정보검색 시스템이 어느 정도는 인간의 언어사용 및 개념형성에 대해 학습하고 있어야만 된다는 것을 뜻한다.

4.2 LSA와 정보검색 시스템

정보검색 분야에서는 정형화된 불리언 검색의 한계를 극복하고자 많은 노력들이 시도되었다. 확장 불리언 검색, 벡터 공간 모델, 확률 모델, 퍼지모델, 나아가 단어가 아닌 사실, 키팩트(key fact)를 중심으로한 내용기반 정보검색까지 연구되고 있다.

앞서 보인 LSA와 관련된 실험 결과들을 볼 때, LSA에 의한 어휘 학습은 인간의 어휘습득과 매우 비슷한 결과를 내었다. 인간은 어휘를 학습할 때 그 단어가 사용되는 맥락을 파악하고 그 단어의 개념을 습득한다는 점을 고려해 볼 때 LSA는 내용기반 정보검색에 매우 적합한 방법이라 할 수 있다.

4.3 실험 환경 및 시스템 설명

앞의 실험 결과들은 단지 LSA가 어떤 식으로 동작하는지의 원리를 파악하기 위해 소규모의 데이터를 대상으로 하였다. 그러나 인간이 가지고 있는 풍부한 어휘 및 배경지식을 고려해 볼 때, 좀 더 인간과 유사한 결과를 내기 위해서는 그에 상응하는 대규모 텍스트의 학습이 필수적이다. 본 실험을 위해 서울대학교 중앙 도서관에서 제공하는 1988년에서 1998년 사이의 석,박사 학위 논문 초록 21,556개(100MB)를 모았다.[7] 이 문서를 대상으로 색인기를 돌려 뽑은 단어들은 총 60만개 정도가 되는데 이중 문서출현빈도(document frequency)가 4이하인 단어와 1,500번 이상인 단어들을 제외한

7) 실제 도서관에서 제공되는 논문 수는 이보다 약간 많으나 입력상의 오류가 있어 제외했다.

51,738개의 색인어를 골랐다. 이를 앞서 언급한 행렬로 나타내면 51,738- by- 21,556의 거대한 행렬이 된다.

정보검색 시스템은 그림 4와 같은 구조로 되어 있다. 사용자의 질의어를 받아들이는 사용자 인터페이스는 보통 웹인터페이스이며 이로부터 시스템에 적합한 형태로 질의어를 생성시켜주는 Query Compiler가 있는데 특히 불리언 질의어의 경우나 유의어, 교차어 확장을 할 경우 필수적이다. 정

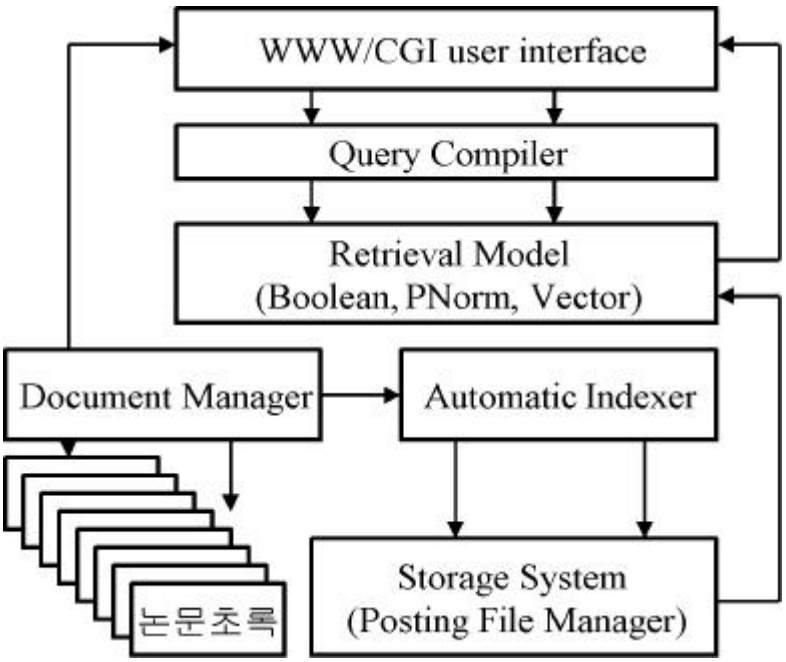

그림 4 정보검색 시스템 구조

보검색 시스템은 많은 시간을 들여 미리 문서집합으로부터 색인화일을 생성시키는데 보통 빠른 속도를 보장하기 위해 역화일 구조로 저장한다.[8] 본 논문에서는 B 트리 자료구조를 이용하여 구현하였기 때문에 빠른 검색이 가능하다. 검색모델에서는 사용자의 질의어와 문서간의 유사도를 계산한다.

8) 복잡한 질의어의 경우라도 1초를 넘지 말아야 한다.

유사도를 계산하는 방식에 따라 검색의 특색이 달라진다.

정보검색 초기에서부터 사용된 불리언 검색은 데이터베이스 검색과 유사하며 숙련된 사용자에게는 강력한 방법이지만 불리언 표현(boolean expression)에 익숙치 못한 사람들에게는 적합하지 않다. 한마디로 불리언 검색은 정보검색을 데이터베이스 검색으로 보고, 검색 시스템의 역할을 최소한으로 한 채 사용자에게 모든 권한과 책임을 넘기는 방법이라 할 수 있다. 이론적으로 불리언 표현은 임의의 문서집합을 나타낼 수 있을 만큼 표현력이 크나 실제 그러한 불리언 표현을 만들기는 매우 어렵다. 그 이유는 불리언 표현에서 사용되는 단어들이 자연언어 특유의 애매성 및 중의성을 가지게 되는데 불리언 표현은 각 단어들을 너무 엄격하게(strictly) 해석하기 때문이다. 예를 들면, '백신 **AND** 프로그램'이라는 질의어를 주었을 경우 '프로그램' 대신 '소프트웨어'가 사용된 문서는 검색되지 않는다. 이 문제를 해결하기 위해서는 '백신 **AND** (프로그램 **OR** 소프트웨어)'라는 질의어를 주어야 하나 이 경우에도 '백신' 대신 '바이러스 검사'로 되어 있을 수도 있다. 이렇듯 애매성을 지닌 자연언어를 불리언 표현으로 나타내기에는 한계가 있다. 퍼지(fuzzy) 검색이나 PNorm 검색 등은 불리언 검색의 문제를 해결하기 위해 나온 방법들로 불리언 연산자들을 좀 더 부드럽게 해석한다.

불리언 연산자들을 아예 무시하고 질의어를 문서와 똑같이 단어들로 이루어진 벡터로 보고 벡터 공간상의 거리를 이용해 유사도를 구하는 방법을 벡터 스페이스 모델(vector space model)이라 한다. 벡터 공간 거리 계산은 각 단어의 특색을 살린 다기보다는 질의어에 포함된 전체 단어 목록의 총체적 비교를 한다. 그러나 벡터 공간 모델의 경우에서도 질의어의 어느 단어에 더 많은 가중치를 주느냐에 따라 검색 결과에 큰 영향을 미친다[Shin and Zhang, 1998]. 본 논문에서 구현된 LSA는 벡터 공간 모델을 발전시킨 방법으로 볼 수 있다.

기타 본 시스템을 이루는 문서 관리기(Document Manager)는 색인시와

원문 제공 시에 필요하고, 자동 색인기(Automatic Indexer)는 한국어 텍스트의 형태소 분석을 통해 색인어들을 자동 추출해 낸다. 한국어의 경우 복합명사를 적절히 분리해 낼 수 있어야 하는데, 색인이 잘 안될 경우는 노이즈가 많아져 정보검색 시스템이 엉뚱한 결과를 내놓게 된다[9]. 포스트화일 관리자(Posting File Manager)는 그림 5와 같이 B 트리 구조로 되어 있어 빠른 검색 및 새로운 문서의 추가가 용이하도록 설계되었다.

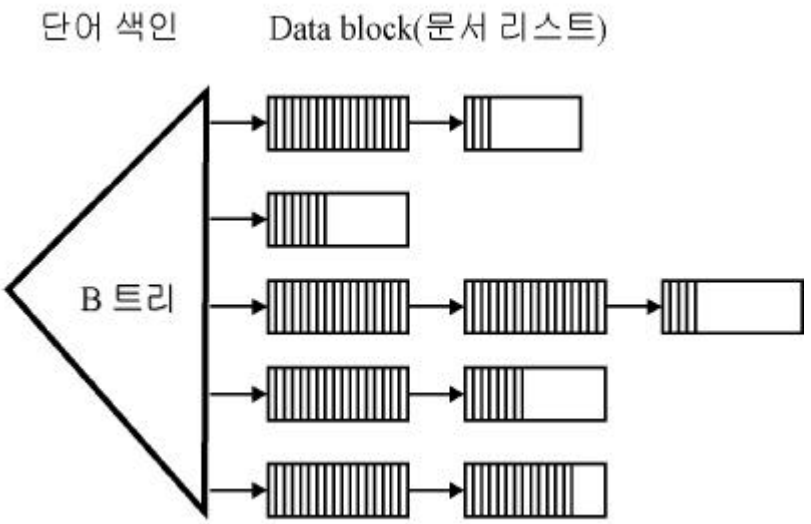

그림 5 Posting File Manager의 구조

9) 본 시스템에서는 (주)코난테크놀로지에서 개발한 색인기를 사용하였다.

4.4 LSA에 의한 단어-단어 유사도 실험

식 (1)-2에 의해 LSA 결과로 나온 새로운 행렬을 $\hat{X}$ 이라 하면 단어들간의 유사도는 서로간의 코싸인 각도도 계산된다. D는 othonormal 이기 때문에 $DD' = I$ 이다.

$$\begin{aligned} \hat{X}\hat{X}' &= (TSD')(TSD')' \\ &= TSD'DST' \\ &= TS^2T' \end{aligned} \qquad (5)$$

식 (5)는 사실상 내적을 계산하는 셈인데 이 값을 식 (4)와 같이 벡터의 크기로 나누어주면 코싸인 값을 얻는다. 이 코싸인 값에 따라 내림차순으로 정렬한 후 10개씩 뽑아 보았다.

표 10은 단어-단어 유사도를 LSA로 계산한 결과를 보여준다. 구체적인 명사뿐만 아니라 추상적인 명사에 이르기까지 LSA의 결과는 우리가 두뇌 속에 개념화하고 있는 어휘들의 관계를 잘 반영하고 있다. 더욱 놀라운 것은 비단 한국어 단어 사이에서만 연관성을 반영하는 것이 아니라 한자 및 영어에 이르기까지 교차-언어적(cross-linguistic)인 연관성도 찾아내었다. 이는 최근 인터넷의 성장과 함께 정보검색 분야에서 관심을 받고 있는 교차언어 검색[10]에

10) 교차언어검색(cross linguistic retrieval): 모국어로 질의어를 입력해도 외국어로 된 문서를 검색해 준다. 예를 들면 한국어로 질의어를 입력했을 때, 한국어, 중국어, 일본어, 영어로 된 인터넷 문서들을 검색해 준다. 이는 외국어에 익숙하지 않은 사용자들에게 큰 도움이 될 뿐만 아니라, 신문 기사 검색

신경망	신경망 인식자 다층인식자 신경회로망 backpropagation 인식률 은닉층 패턴인식 인공신경회로망 신경회로
가구	가구 평수 가구주 가구원 공공주택 households 입주 생애 생애주기 확대기
철학	철학 사상 학문 걸음 지음 성찰 배타 퇴계 일탈 학자
언어학	언어학 음성학 등위 접속사 접두 접두사 의미기능 문법 지시 일상언어
심리	심리 정서 개인 emotion 소통 배경 일상 긴장 언어행위 관적
문화	문화 수용양상 전국시대 고유성 문화가 대성 정체 청동기 무문 문화권
이데아	이데아 Forms 레아 재론 해탈 입론 논증 보편자 현상세계 실재론
상대성	상대성 시공간 우주론 상대론 이율 이율배반 선천 存在의 하학 칸트
논리	논리 방법론 실현 론적 보편 實在 재론 명제 유기체론 양자간

표 10 LSA로 계산된 단어-단어 유사도 실험 결과

까지 LSA가 훌륭하게 적용될 수 있다는 가능성을 보여준다.

본 시스템은 또한 LSA를 최초로 개발한 Bellcore의 시스템과 성능을 비교하여 보았다. 본 시스템은 한국어를 중심으로 학습한 것이기 때문에 영어를 대상으로 학습한 Bellcore의 시스템과는 비교하기가 어려우나 영어의 경우에 있어서도 본 시스템이 Bellocore의 시스템보다 월등한 결과를 보였다. 표 11에서 그 결과를 보이고 있다.[11]

의 경우는 동일한 사건에 대한 세계 여러 언론들의 보도를 동시에 알아 볼 수 있다.

11) 결과는 직접 테스트해 볼 수 있다.
논문에서 구현된 시스템: http://krcogsci.snu.ac.kr/~krishna/termsim.html
Bellcore에서 구현한 시스템: http://superbook.bellcore.com/lsi-bin/lsiQuery

question	**본 시스템**	question, essence, notion, communicative, conception, sciences, usages, arrive, understanding, vision
	Bellcore	question, ma, adresse, mr, my, the, le, directed, of, que
law	**본 시스템**	law, legal, Court, prosecutor, criminal, legality, reserved, judicial, Crime, provisions
	Bellcore	law, grief, real, employees, employee, grievance, problemes, dise, precise, based
faith	**본 시스템**	faith, faithful, debated, principle, originally, ruler, bringing, backdrop, grounded, Confucian
	Bellcore	faith, confiance, document, coding, respond, vast, classification, histoire, biggest, analyser
economy	**본 시스템**	economy, domestic, Economies, economic, underdeveloped, protectionism, reinforce, initiative, Privatization, internationalization
	Bellcore	economy, economie, derniers, compared, heavily, cereales, subvention, americans, subsidized, speaking
politics	**본 시스템**	politics, political, republic, faction, politically, ruling, democracy, corrupt, regime, regionalism
	Bellcore	politics, assainissement, risque, progressiste, heart, donnent, defend, ecouter, laissee, themselves

표 11 Bellcore **시스템과 비교해 본 단어-단어 유사도**

4.5 LSA에 의한 문서-문서 유사도 실험

본 실험을 위해서는 일반 인터넷 정보검색 시스템과 같은 형태의 정보검색 시스템을 구축했다. 논문의 각 섹션별로 검색이 가능한데 빈칸에 질의어를 입력하고 검색 버튼을 누르면 해당 논문 리스트의 제목들이 나온다. 검색 섹션을 '저자'로 선택하고 빈칸에 '배문정'이라고 쓴 후 검색 버튼을 누르면 아래의 결과가 뜬다. 그후 '심리학과'란 링크를 누르면 초록이 보인다. 그림 3의 화면에서 오른쪽 [현

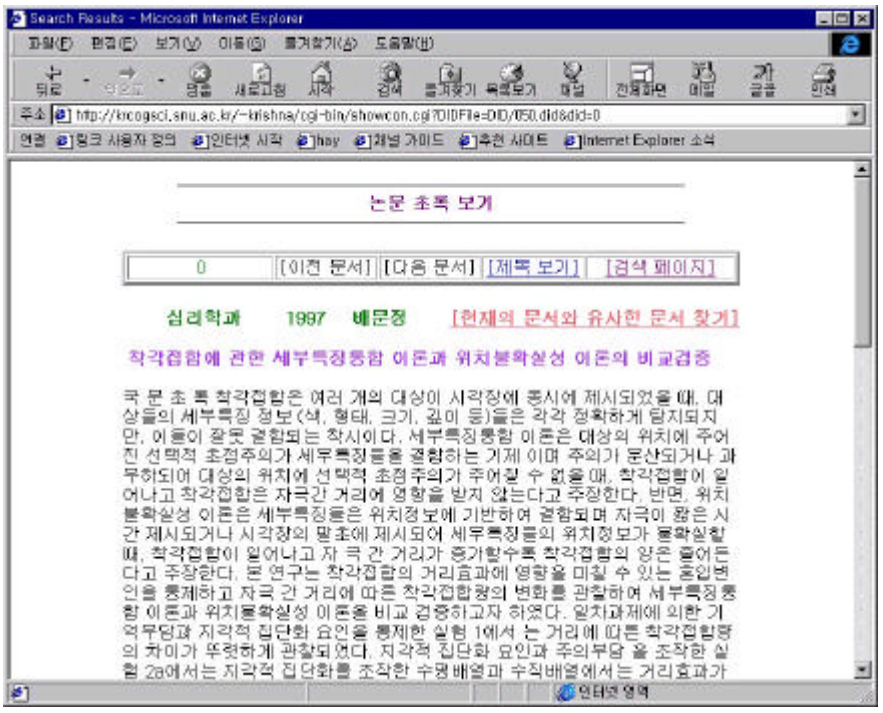

그림 6 검색 결과의 논문 초록 보기 화면

재의 문서와 유사한 문서 찾기]를 누르면 LSA로 계산된 유사한 문서 100개를 보여준다.

그림 7에서 보는 바와 같이 결과는 매우 훌륭하다. 특히 상위 3개의 문서들은 제목만 보아도 원래의 논문과 유사한 논문이라는 것을 알 수 있다. 재미있는 사실은 문서 4번의 경우처럼 전혀 다른 학과에서도 비슷한 문서를 찾아냈다는 것이다.

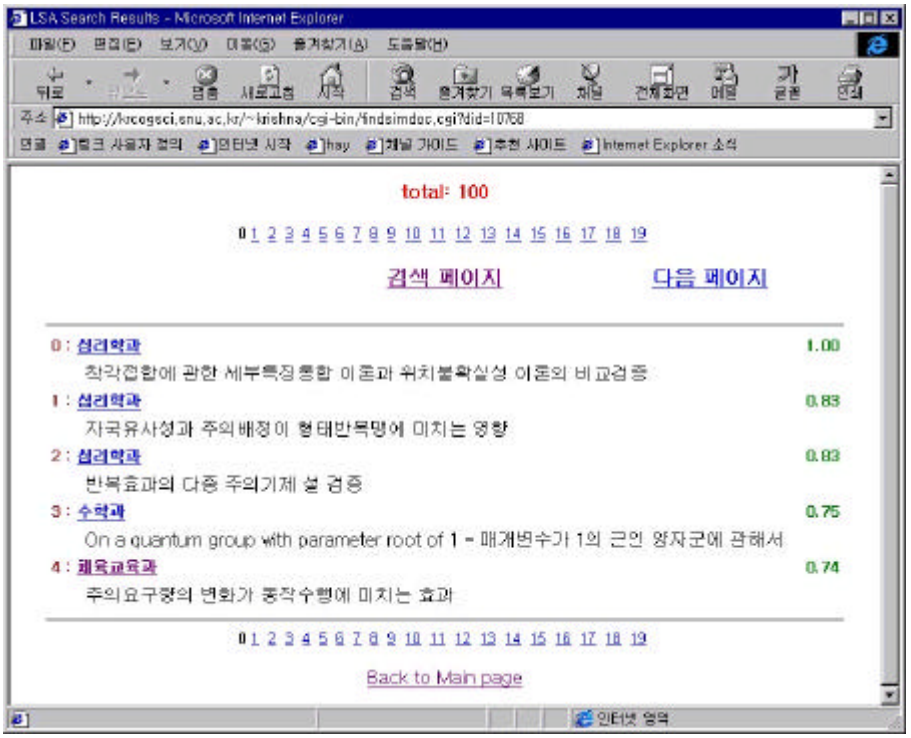

그림 7 LSA로 찾은 100개의 유사한 문서 목록

5. 결론 및 토의

자연언어는 그 구조가 매우 복잡하고 예외도 많기 때문에 통계적인 방법으로 자연언어를 분석하는 작업은 매우 어려운 일이다. 그러나 본 논문에서 살펴본 바와 같이 LSA는 인간의 직관에 상당하는 수준의 결과를 보여 주었다. LSA가 인간의 장기기억 학습과 유사한 면을 보인다는 것은 이미 선행 연구들에서 지적되었다. 앞으로도 LSA 및 그와 유사한 통계적 방법들이 언어분석의 유용한 기법으로 발전해 나갈 것으로 전망된다.

한편, 본 논문에서 구현한 단어-단어 유사도와 문서-문서 유사도는 일반 정보검색 시스템에 즉각 응용 가능성을 보여 준다. 단어-단어 유사도는 사용자의 질의어 생성과 관련하여 시소러스의 역할을 함으로써 자동 질의어 확장 및 피드백을 가능하게 해 준다. 문서-문서 유사도의 경우는 사용자가 정보검색을 할 때 원하는 문서를 찾기 위해 계속 질의어를 변경해 나가는 번거로움 없이 즉각적으로 관련된 문서들을 찾아 줄 수 있다. 문서-문서 유사도를 이용하여 검색을 할 경우 또 한가지 유리한 점은 국소적인 유사도를 비교하는 질의어-문서 유사도에 비해 전역적인 유사도 계산을 하므로 문서 전체의 특성을 반영한 정보검색이 가능하다[신동호 장병탁, 1999].

참고문헌

[신동호 장병탁, 1999] 신동호, 장병탁, 계층적 클러스터링을 이용한 텍스트 정보 검색 모델, *JCEANF-99*, PP. 293-296. 1999.

[Berry et al, 1993] Berry, M.W., Do, T., O'Brien, G.W., Krishna, V., and Varadhan, S, SVDPACKC (Version 1.0) User's Guide. 1993.

[Berry et al, 1995a] Berry, M.W., Dumais, S.T., and Letsche., T.A., Computational Methods for Intelligent Information Access, *Proceedings of Supercomputing'95*, San Diego, CA. 1995.

[Berry et al, 1995b] Berry, M.W., Dumais, S.T., and Shippy, A.T., A Case Study of Latent Semantic Indexing. 1995.

[Berry et al, 1995c] Berry, M.W., Dumais, S.T., and O'Brien, G.W., Using Linear Algebra for Intelligent Information Retrieval. *SIAM Review* 37:4, pp. 573-595, 1995.

[Berry et al, 1996] Berry, M.W. and Ricardo D.F., Low-Rank Orthogonal Decompositions for Information Retrieval Applications., *Numerical Linear Algebra with Applications* 3:4, pp. 301-328. 1996.

[Jiang, 1997] Jiang, J., Using Latent Semantic Indexing for Data Mining., MS Thesis, Department of Computer Science, University of Tennessee. 1997.

[Landauer and Dumais, 1997] Landauer, T. K. and Dumais, S. T. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240. 1997.

[Landauer and Laham, 1997] Landauer, T. K. and Laham, D. Learning Human-like Knowledge by Singular Value Decomposition: A

Progress Report. *NIPS*, 10, 45-51. 1997.

[Laham, 1997] Laham, D., Latent Semantic Analysis Approaches to Categorization. *Proceedings of the Cognitive Science Society*. 1997.

[Letsche, 1996] Letsche, T.A., Toward Large-Scale Information Retrieval Using Latent Semantic Indexing. MS Thesis, Department of Computer Science, University of Tennessee. 1996.

[Letsche and Berry, 1997] Letsche, T.A. and Berry, M.W., Large-Scale Information Retrieval with Latent Semantic Indexing, *Information Sciences - Applications* 100, pp. 105-137. 1997.

[Miller, 1990] Miller G., Five Papers on WordNet, *Special Issue of International Journal of Lexicography* 3(4). 1990.

[O'Brien, 1994] O'Brien, G.W., Information Management Tools for Updating an SVD-Encoded Indexing Scheme. 1994.

[Salton and McGill, 1983] Salton, G. and McGill, M.J., Introduction to Modern Information Retrieval. McGraw-Hill. 1983.

[Shin and Zhang, 1988] Shin, D.H. and Zhang, B.T., Automatic Query Generation for the TREC-7 Ad-Hoc Task, *Proceedings of the 7th Text REtrieval Conference (TREC-7)*, NIST Special Publication, 1998.

[Warrington and Shallice, 1984] Warrington, E. K. and Shallice, T. Category-specific semantic impairments. *Brain*, 107, 829-853. 1984.

Abstract

Natural language processing is one of the most difficult research areas because of its complexity and diversity. So, syntactic rules based on human intuition have been mainly used to analyze human language until now. Whereas a rule-based approach is effective for the analysis of a small corpus, its performance degrades rapidly as the corpus size increases (over tens of thousands of documents) due to the many exceptional cases. Contrary to this, a statistical approach can attack large text corpus easily using simple information such as co-occurrence. A statistical approach can be viewed as processing lower level data to produce more higher and more general level representation. LSA (Latent Semantic Analysis) can compute higher level representation directly by statistical method of SVD(Singular Value Decomposition).

In the previous learning method, only some pre-selected features are considered in learning concept for its computational burden. But LSA produce high dimensional vector space called 'concept space' through the entire large text corpus statistically. These 'concept space', made by purely statistical method, showed similar results with human mental lexicon in many experiments. In this paper, The human-like property (semantic property) of LSA is applied to information retrieval system. Information

retrieval system using LSA can really be called 'content-based information retrieval system'. Experiments were done under the 20,000 paper abstracts provided by Seoul National University Library (http://solarsnet.snu.ac.kr). The results were excellent, even better than Bellcore LSI demo system.

Key words: LSA, SVD, IR, mental lexicon, concept space, similarity measure, learning

student number : 98132-504

감사의 글

지난 2년 동안의 생활을 총 정리한 것이 본 논문이라 생각하니 완성의 기쁨보다는 좀 더 잘할 수 있었을 텐데라는 반성과 부끄러움이 앞섭니다. 다만 그 동안 음으로 양으로 많은 도움을 주신 분들이 아니었다면 오늘의 성과물도 없었을 것이라는 생각에 이렇게 감사의 글을 올립니다.

먼저 지난 2년 동안 옆에서 세심한 지도를 해 주신 장병탁 교수님께 감사드립니다. 매사를 긍정적으로 보시는 교수님은 조그마한 실패에도 좌절하고 부정적인 사고에 빠지던 제게 큰 힘이 되었습니다. 또한 성가신 논문 심사를 흔쾌히 받아주신 조명한, 김청택 교수님께 감사드립니다. 조명한 선생님의 수업과 김청택 선생님과의 대화가 본 논문을 시작하는 밑거름이 되었습니다. 더불어 언제나 인지과학 협동과정의 발전을 위해 힘쓰시는 김영정 선생님께도 감사의 말씀을 드립니다. 젊은 사람 못지 않은 패기와 학문에 대한 열정은 제게 좋은 귀감이 되었습니다. 대학원보다는 오히려 학부 때 더 많은 가르침을 받은 김정오 교수님께도 감사의 말씀을 드립니다. 묘한 인연으로 계속 선생님의 과목을 수강하며 심리학에 대한 관심을 꾸준히 유지할 수 있었습니다. 다만 마지막 1년 동안은 해외로 나가 계시는 바람에 지도를 받을 수 없게 된 것이 아쉬울 따름입니다. 그밖에 언어학을 가르쳐 주신 이정민 교수님, 신경과학을 가르쳐 주신 이춘길 교수님, 인류학을 가르쳐 주신 왕한석 교수님 등 제가 인지과학자의 길로 들어서는데 도움을 주신 협동과정의 모든 교수님들께 감사드립니다.

각자 다른 학문적 배경에서 모여 함께 토론하던 협동과정 학생들과의 기억도 제겐 잊을 수 없는 추억으로 남을 것입니다. 근 1년 동안 '의식 세미나'를 함께 한 김병우 선배에게 감사를 드립니다. 그 기간은 저의 대학원 생활 중 순수한 학문적 열정이 가장 크게 타올랐던 시기였습니다. 또한 깁슨과 군이론 및 표현론에 대한 토론은 제게 새로운 세계를 보여주었습니다. 인지과학 티타임을 정기적으로 열어 각 분야의 사람들과 의견 공유를 할 수 있도록 한 배문정 선배에게도 큰 감사를 드립니다. 어려운 여건에도

꿋꿋하게 학문에 매진하시는 모습에 큰 감명을 받았습니다. 늘 명랑하고 성실한 손정우 선배와의 많은 토론 덕에 신경 모델링에 대해 많이 배웠고, 김윤현 선배와 장시간 벌인 유비사고에 대한 토론은 저를 유비의 문제에 관심을 갖도록 해 주었습니다. 그밖에 협동과정 학생들과 함께 여러 주제에 대해 벌인 토론들과 채터를 통해 나누었던 많은 이야기들도 잊지 못할 추억일 될 것입니다. 무엇보다 2년 생활에 동고동락했던 98 동기분들에게 감사를 드립니다.

인공지능 연구실 1기로서 함께 생활했던 서영우, 김석준, 정제균형들의 끝임 없이 연구하는 모습과 신수용, 조동연, 김정집군들의 성실한 모습은 항상 저를 부끄럽게 하였습니다. 김선형과 엄재홍군이 후배로 들어오면서 연구실 분위기가 화기애애하게 살아났지만, 정작 선배로서 별다른 도움을 주지 못해 미안할 뿐입니다. 함께 정보검색을 공부했던 김유환군과 신형주 양은 PCA 관련된 많은 자료를 제공해 주어서 꼭 감사의 글에 넣어 주겠다고 약속했음을 밝힙니다.

본 논문의 완성에는 ETRI 재직 시절부터 도움을 아낌없이 주신 김영섬 박사님과 양승현 박사님의 은덕이 있었음을 밝힙니다. 학교라는 울타리 안에서 안이하게만 지내오던 저에게 정보검색 분야의 여러 문제들을 피부로 느낄 수 있도록 도와 주셨으며, 또한 엔지니어로서의 길을 보여 주셨습니다.

고등학교 동창들인 강민철, 인재우 군은 저의 자랑스러운 지음(知音)의 친구들입니다. 정보 및 사이버를 주제로 한 1년여 동안의 토론은 끊임없는 지적 호기심을 불러일으키도록 도왔습니다. 서로 공부하고 있는 내용도 다르고 관심사도 다르지만, 가끔씩 만나 세상의 온갖 일들에 대해 밤새워 토론하던 일들은 아름다운 기억으로 간직될 것입니다.

끝으로 먼 대전에서 믿음직스럽게 저를 지켜 보아준 친형 신동명과 오늘도 당진에서 농사일로 고생하시는 부모님께 이 글을 바칩니다.