감성과학, Vol. 13, No. 1, pp.47-60, March 2010

선택적 자질 차원 축소를 이용한 최적의 지도적 LSA 방법*

Optimal supervised LSA method using selective feature dimension reduction

김정호** · 김명규**† · 차명훈** · 인주호**** · 채수환***

JungHo Kim** · MyungKyu Kim**† · MyungHoon Cha** · Joo-Ho In** · Soo-Hoan Chae***

한국항공대학교 컴퓨터공학과**

Department of Computer Engineering, Korea Aerospace University**

한국항공대학교 항공전자정보통신공학부***

School of Electronics, Telecommunication and Computer Engineering, Korea Aerospace University***

(주)이엠넷****

eMnet Inc.****

Abstract

Most of the researches about classification usually have used kNN(k-Nearest Neighbor), SVM(Support Vector Machine), which are known as learn-based model, and Bayesian classifier, NNA(Neural Network Algorithm), which are known as statistics-based methods. However, there are some limitations of space and time when classifying so many web pages in recent internet. Moreover, most studies of classification are using uni-gram feature representation which is not good to represent real meaning of words. In case of Korean web page classification, there are some problems because of korean words property that the words have multiple meanings(polysemy). For these reasons, LSA(Latent Semantic Analysis) is proposed to classify well in these environment(large data set and words' polysemy). LSA uses SVD(Singular Value Decomposition) which decomposes the original term-document matrix to three different matrices and reduces their dimension. From this SVD's work, it is possible to create new low-level semantic space for representing vectors, which can make classification efficient and analyze latent meaning of words or document(or web pages). Although LSA is good at classification, it has some drawbacks in classification. As SVD reduces dimensions of matrix and creates new semantic space, it doesn't consider which dimensions discriminate vectors well but it does consider which dimensions represent vectors well. It is a reason why LSA doesn't improve performance of classification as expectation. In this paper, we propose new LSA which selects optimal dimensions to discriminate and represent vectors well as minimizing drawbacks and improving performance. This method that we propose shows better and more stable performance than other LSAs' in low-dimension space. In addition, we derive more improvement in classification as creating and selecting features by reducing stopwords and weighting specific values to them statistically.

Keywords : Categorization, Text, Feature, LSA, SVD, Supervised learning

---

* 이 논문은 2009년도 고양시 산학관 공동 기술개발 사업으로 고양시 및 (주)RSN지원을 받아 연구되었음.

† 교신저자 : 김명규 (한국항공대학교 컴퓨터공학과)

E-mail : kimmk@kau.ac.kr

TEL : 02-300-0146

요 약

기존 웹 페이지 자동분류 연구는 일반적으로 학습 기반인 kNN(k-Nearest Neighbor), SVM(Support Vector Machine)과 통계 기반인 Bayesian classifier, NNA(Neural Network Algorithm)등 여러 종류의 분류작업에서 입증된 분류 기법을 사용하여 웹 페이지를 분류하였다. 하지만 인터넷 상의 방대한 양의 웹 페이지와 각 페이지로부터 나오는 많은 양의 자질들을 처리하기에는 공간적, 시간적 문제에 직면하게 된다. 그리고 분류 대상을 표현하기 위해 흔히 사용하는 단일(uni-gram) 자질 기반에서는 자질들 간의 관계 분석을 통해 자질에 정확한 의미를 부여하기 힘들다. 특히 본 논문의 분류 대상인 한글 웹 페이지의 자질인 한글 단어는 중의적인 의미를 가지는 경우가 많기 때문에 이러한 중의성이 분류 작업에 많은 영향을 미칠 수 있다. 잠재적 의미 분석 LSA(Latent Semantic Analysis) 분류기법은 선형 기법인 특이치 분해 SVD(Singular Value Decomposition)을 통해 행렬의 분해 및 차원 축소(dimension reduction)를 수행하여 대용량 데이터 집합의 분류를 효율적으로 수행하고, 또한 차원 축소를 통해 새로운 의미공간을 생성하여 자질들의 중의적 의미를 분석할 수 있으며 이 새로운 의미공간상에 분류 대상을 표현함으로써 분류 대상의 잠재적 의미를 분석할 수 있다. 하지만 LSA의 차원 축소는 전체 데이터의 표현 정도만을 고려할 뿐 분류하고자 하는 범주를 고려하지 않으며 또한 서로 다른 범주 간의 차별성을 고려하지 않기 때문에 축소된 차원 상에서 분류 시 서로 다른 범주 데이터간의 모호한 경계로 인해 안정된 분류 성능을 나타내지 못한다. 이에 본 논문은 새로운 의미공간(semantic space) 상에서 서로 다른 범주사이의 명확한 구분을 위한 특별한 차원 선택을 수행하여 최적의 차원 선택과 안정된 분류성능을 보이는 최적의 지도적 LSA을 소개한다. 제안한 지도적 LSA 방법은 기본 LSA 및 다른 지도적 LSA 방법들에 비해 저 차원 상에서 안정되고 더 높은 성능을 보였다. 또한 추가로 자질 생성 및 선택 시 불용어의 제거와 자질에 대한 가중치를 통계적인 학습을 통해 얻음으로써 더 높은 학습효과를 유도하였다.

주제어 : 범주화, 텍스트, 자질, LSA, SVD, 지도적 학습

1. 서론

인터넷이 보편화되어지면서 어느 누구나 공간상, 시간상 제약 없이 월드와이드웹을 통해 많은 정보를 손쉽게 얻을 수 있게 되었다. 특히 인터넷 사용자들은 상용화 되어 널리 사용되어 지는 검색 사이트를 통해 찾고자 하는 정보를 검색하고 검색 결과로부터 원하는 정보를 얻어낼 수 있다. 특히 요즈음 웹 2.0 시대에서는 사용자들이 블로그(blog)와 같은 웹 페이지를 생성하여 자신의 생각이나 지식을 인터넷 상에 올릴 수 있는 쌍방향 지식 체계를 이루고 있다. 이렇듯 인터넷 상의 무수히 많은 웹 페이지가 끊임없이 새로 생성되어지고 새롭게 갱신되어 짐에 따라 웹 페이지의 효율적인 관리가 필요하게 되었다. 예를 들어 검색엔진에서 크로울링(crawling)을 통해 얻은 다양한 웹 페이지들을 사전에 미리 정의한 범주(category)를 토대로 분류함으로써 사용자들이 찾고자하는 정보를 좀 더 정확하고 손쉽게 얻을 수 있도록 할 수 있다. 대표적인 검색 사이트인 Yahoo에서는 많은 사람들이 직접 일일이 웹 페이지를 분류하며 관리를 하고 있어서 많은 인적 자원이 소비된다(Qi & Davison, 2009). 하지만 나날이 늘어나는 웹 페이지를 인위적으로 관리하기에는 많은 비용과 시간이 들기 때문에 비효율적이다. 그렇기에 웹 페이지 자동 분류의 필요성이 대두되었고 이에 관한 많은 연구가 이루어 졌으며 현재 연구되어지고 있는 중이다.

기존의 웹 페이지 분류는 학습 기반인 kNN(k-Nearest Neighbor), SVM(Support Vector Machine)과 통계 기반인 Bayesian classifier, NNA(Neural Network Algorithm) 등의 분류 기법으로 수행되어 지고 있다(Selamat & Omatu, 2004)(조태호, 1998)(Chen & Hsieh, 2006). 하지만 이러한 분류 기법들은 자질의 수가 기하급수적으로 커지거나(data set이 커지거나) 분류해야 할 문서의 수가 많아질수록 계산량이 많아져 분류함에 있어서 시간적, 공간적 제약 문제가 존재하며 또한, 분류 대상을 표현하는 자질로 단일 자질 기반을 주로 사용하지만 이는 자질의 정확한 의미를 파악하지 못하기 때문에 자칫 분류 성능을 떨어트리는 원인이 된다(Kontostathis & Pottenger, 2006). 특히 한국어 웹 페이지의 경우 한글의 특성인 단어의 중의성으로 분류 시

단일 자질 기반을 사용하면 좋은 결과를 얻기 힘들다. 그래서 기존 몇몇 연구에선 자질 선택(feature selection) 시 n-gram 자질을 사용함으로써 자질의 정확한 의미분석을 가능하게 하였으나, 단일 자질에 비해 자질의 수가 급증하고 상대적으로 웹 페이지에서의 출현 빈도수(term frequency)가 낮아지기 때문에 드는 노력에 비해 성능 향상 정도가 효과적이지 않다(Li et al., 2007).

잠재적 의미 분석인 LSA 분류 기법은 벡터모델 분류 기법으로써, 선형 기법인 SVD을 이용해 행렬로부터 새로운 의미공간을 생성하고 그 중 큰 의미가 없는 즉, 분류 대상에 대한 표현력이 약한 차원을 제거한 최적의 차원만으로 분류 대상을 표현하여 분류함으로써 앞의 시간적, 공간적 제약 문제를 완화시킬 수 있다. 그리고 새로운 의미공간으로부터 문서 혹은 자질의 중의적인 의미를 고려하여 잠재된 의미분석을 가능하게 한다. 이는 이전 TOEFL(Test Of English as Foreign Language)의 동의어 테스트에 LSA를 적용하여 실험하였을 시 64%이상의 높은 정답률을 보임으로써 실험으로 입증되어 진 바가 있으며 그 외 동의어 구축 및 자질의 색인에서 좋은 결과를 보였다(이태헌과 김청택, 2004)(신동호, 1999).

하지만 LSA는 문서 또는 웹 페이지 분류에서 성능 향상의 한계를 보였다. 이러한 이유는 SVD를 통한 새로운 의미공간 생성 및 차원의 축소 과정에서 선택되는 저차원 공간은 오직 새로운 의미공간 상 분류 데이터를 표현하는 정도만을 고려할 뿐 분류하고자 하는 서로 다른 범주간의 구분을 고려하지 않기 때문이다(Sun et al., 2004)(Chakraborti et al., 2006). 그러므로 차원 축소로부터 생기는 서로 다른 범주의 분류 집단 간 구분의 모호성이 전체적인 분류의 성능을 저하시키게 된다. 기존 LSA 분류기법의 이러한 단점을 보완하기 위해 제안된 LSA 방법으로 범주 정보를 고려한 지도적 LSA와 차별적인 차원 축소를 이용한 지도적 LSA가 제안되었지만 여전히 차원 축소 시 범주 정보의 손실과 차원 선택 시 많은 시간이 드는 단점이 있었다.

본 논문은 위 단점들을 상호 보완하는 새로운 지도적 LSA로, 범주 정보를 고려하면서 SVD를 통한 차원 감소 시 서로 다른 범주의 데이터들 간 이질성을 최대화 시키는 차원을 선택하는 방법을 소개한다. 새로 생성된 의미공간을 이루는 각 차원이 서로 다른 범주에 미치는 영향력의 차이가 클수록 서로 다른 범주 간의 구분을 명확히 시킨다는 가정 하에 선택된 저차원공간에서 분류를 수행하여 기존 LSA보다 더욱 안정되고 높은 성능을 기대하였다. 또한 그보다 더 향상된 웹 페이지 분류를 위해 지도적 LSA 분류를 기반으로 하면서 통계적인 자질 선택 및 가중치 적용을 통한 분류 성능 향상을 제시한다.

이 논문의 구성은 2장에서 LSA 분류 기법을 소개하면서 그 특징과 약점에 대해 소개한다. 그리고 3장에서 이미 연구되어 진 지도적 LSA 기법과 본 논문이 제안하는 새로운 지도적 LSA 기법에 대해 소개하고 4장에서 이를 바탕으로 한 웹 페이지 자동 분류 시스템의 구성과 분류 방법을 제시한다. 5장에서는 분류 시스템을 이용하여 분류 실험을 수행한 결과와 그 결과에 대해 다른 분류 기법을 이용한 분류 성능과 비교 분석하고, 마지막으로 6장에서 결론 및 추후 연구에 대해 제시하겠다.

2. 관련연구

2.1. 잠재적 의미분석(LSA)

LSA(Latent Semantic Analysis)는 사전적 의미로는 잠재적 의미 분석이라 하며 다양한 분류에 적용 가능한 학습기반 및 벡터모델 분류기법이다. LSA는 선형대수인 행렬의 선형변환에 의거한 행렬 분해 기법인 SVD(Singular Value Decomposition)를 통해 한 행렬의 의미를 재해석한다. 여기서의 행렬은 기존 벡터모델과 마찬가지로 분류하고자 하는 대상을 특정 벡터 공간의 벡터로 표현한 벡터 집합을 의미한다. LSA를 사용한 기존 분류 연구들은 이 SVD기법으로부터 얻을 수 있는 이점들에 초점을 둔다.

2.1.1. 특이치 분해(SVD)

SVD(Singular Value Decomposition)은 행렬 분해 기법으로, 행렬의 대각화를 기반으로 하여 정방행렬(NxN) 뿐만 아니라 정방행렬이 아닌 모든 행렬을 3개의 행렬로 분해할 수 있다. 이러한 특징으로 SVD는 무제약 선형 최소 제곱 문제(unconstrained linear least squares problem), 해가 없는 방정식에 대한 근사해를 구하는 문제, 행렬의 rank 추정에 주로 사용되어 진다(신동호, 2000).

SVD은 아래의 식1 과 같이 mxn 행렬 A를 3개의 행

렬로 분해한다.

$$A = U\Sigma V^T \quad \text{(식 1)}$$

U : 행렬 A의 행공간의 정규 직교 기저 벡터
V : 행렬 A의 열공간의 정규 직교 기저 벡터
Σ : 행렬 A의 행공간(열공간)의 특이치

행렬 U는 mxn 행렬로, 행렬 A의 열벡터를 나타내는 열공간의 정규 직교 기저(standard orthonormal basis)[1]로 구성되어지며, 각 열벡터를 행렬 A의 left singular vector[2]라 한다. 행렬 V는 nxn 행렬로, 행렬 A의 행벡터를 나타내는 행공간의 정규 직교 기저로 구성되어지며, 각 열벡터를 행렬 A의 right singular vector[3]라 한다. 행렬 Σ는 nxn 대각 행렬로, 각 대각 값을 특이치(singular value)[4]라 하며 큰 값부터 내림차순으로 정렬되어져 있다. 그리고 행렬 U와 행렬 V의 각 열벡터인 정규 직교 기저 벡터는 행렬 Σ의 각 특이치에 대응되며 기저 벡터가 가지는 고유치에 맞게 내림차순으로 정렬된다. 특이치는 고유벡터와 고유값의 관계와 마찬가지로 대응되는 기저 벡터가 가지는 증감 정도 즉, 벡터에 미치는 영향 정도를 의미한다.

SVD는 행렬 축소라는 큰 특징을 가지고 있다. 행렬 축소는 행렬 분해를 통해 구한 정규 직교 기저 중 기저에 대응하는 특이치 값이 0이거나 0에 가까운 기저를 제거함으로써 기존 행렬이 가질 수 있는 잡음(noisy)을 제거한다는 의미를 가진다. 다른 말로, 특정 기저가 벡터공간의 벡터들에게 미치는 영향이 작을 때 이를 제거하여 벡터를 좀 더 잘 표현하는 저차원 공간으로의 투영을 의미한다. 예를 들면, 그림 1에서와 같이 2차원 공간상에 표현된 벡터 집합이 있다고 하자. 그리고 SVD를 통해 2차원 공간의 새로운 직교 기저인 v_1와 v_2를 구하였을 때, 벡터 집합들을 이 두 새로운 기저에 비춰 보면 상대적으로 기저벡터 v_1에 비해 기저벡터 v_2가 벡터 집합들의 분포에 영향력이 낮다. 그렇기 때문에 이와 같은 경우 기저벡터 v_2는 벡터 집합을 나타냄에 있어서 잡음으로 간주하여 제거하고 기저벡터 v_1로 투영시킨다.

차원 축소를 식2로 표현하면, 식1에서 행렬 Σ의 대각 값 S_k가 $1 \le k \le r$일 때 0이 아닌 값을 가지고,

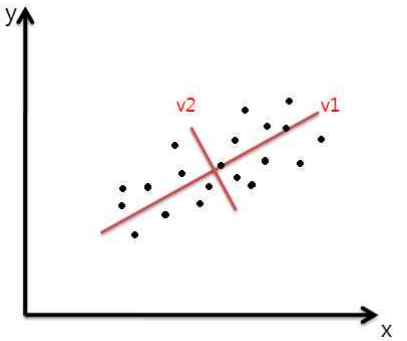

그림 1. 새로운 직교 기저에 대한 데이터 분포

$(r+1) \le k \le n$일 때 0값을 가질 경우, 0값을 가지는 (n-r)개의 특이치 혹은 0에 가까운 값을 가지는 특이치에 대응하는 행렬 U와 V의 기저 벡터(singular vectors)를 제거함으로써 행렬 축소를 수행한다. 행렬 축소를 적용하여 구한 식2로 기존 행렬 A에서 내부적으로 각 벡터에 큰 영향을 미치지 못하는 차원을 제거함으로써 행렬이 가지는 잡음 제거한 근사 행렬 A'를 구할 수 있다.

$$A' = \sum_{i=1}^{r} u_i s_i v_i^T \approx A \ \ (r \le n) \quad \text{(식 2)}$$

r : 선택된 차원의 수
u_i : 행렬 U의 i 번째 열벡터
s_i : 행렬 Σ의 i 번째 대각값
v_i : 행렬 V의 i 번째 열벡터

2.2. LSA를 이용한 분류

일반적으로, 웹 페이지 혹은 문서 분류에 벡터 모델 기법들을 사용할 때 가장 먼저 단어-문서 행렬을 구성한다. 단어-문서 행렬의 행벡터는 문서 좌표계 상에 표현된 단어 벡터를 의미하며, 열벡터는 단어 좌표계 상에 표현된 문서 벡터를 의미한다. 기존 벡터 모델 기법에서는 단어-문서 행렬을 어떠한 정제 작업 없이 그대로 사용하였으나, 이는 단어 혹은 문서의 수가 커질수록 행렬의 크기가 nm(n: 단어, m: 문서)만큼 커지기 때문에 대량의 문서를 분류할 때 계산량이 많아지는 단점이 있으며 또한 문서를 표현하기 위해 추출한 단어의 정확한 의미나 문서의 문맥을 고려하지 못한다는 단점을 가지고 있다. 하지만 LSA는 SVD기법을 적용하여 행렬 분해 및 차원 감소를 통해 고차원 데이터를 저차원 상으로 표현하기 때문에 분류 시 데이

1) 서로 직교하는 단위 기저 벡터
2) 정방행렬 및 대칭행렬인 AA^T의 고유벡터
3) 정방행렬 및 대칭행렬인 A^TA의 고유벡터
4) AA^T와 A^TA의 고유벡터에 대응하는 고유값의 제곱근

터의 크기에 상관없이 일정한 분류작업 시간을 가지며, 데이터를 표현함에 있어서 불필요한 차원을 잡음으로 간주하고 제거하여 문서의 내재적 의미를 파악하고자 한다.

다음 단어-문서 행렬 A에 대하여 SVD를 적용할 시 그림 2와 같이 표현되어진다. 행렬 A는 2.1절에서 소개한 SVD기법을 통해 행렬 A를 세 개의 행렬로 분해한다.

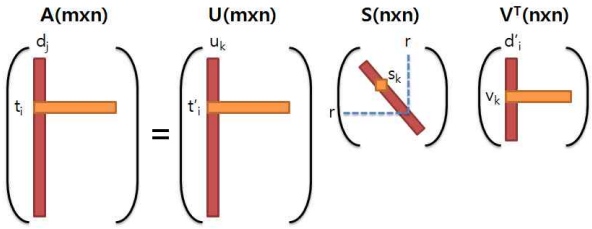

그림 2. SVD를 적용한 단어-문서 행렬

그림 2에서 단어-문서 행렬 A는 단어를 표현하는 행벡터, 이하 단어벡터, t_i와 문서를 표현하는 열벡터, 이하 문서벡터, d_j로 구성된다. 단어 벡터는 문서 좌표계(행공간)상에 표현된 점이며 문서 벡터는 단어 좌표계(열공간)상에 표현된 점이다. 그러므로 행렬 U는 행렬 A의 행공간을 이루는 정규 직교 기저 벡터 u_k로 구성되며, 행렬 V는 행렬 A의 열공간을 이루는 정규 직교 기저 벡터 v_k로 구성된다. 행렬 S는 U또는 V를 구성하는 정규 직교 기저에 대응하는 각 특이치 값 s_k를 대각 값으로 하여 구성된다. 즉, 행렬 S의 각 대각 값 s_k는 행렬 U와 V의 두 직교 기저 벡터의 각 공간상의 벡터에 대한 영향력을 의미하게 된다.

행렬 U와 행렬 V로부터 다음과 같이 새로운 의미를 추출해 낼 수 있다. 그림 2에서 행렬 U의 각 행 벡터와 S의 대각값의 곱인 $s_k t'_i$는 열공간의 기저 벡터인 v_k의 좌표벡터로 기존 단어 벡터 t_i를 기저 벡터 v_k가 구성하는 새로운 공간상의 단어 벡터로 표현한 것을 의미한다. 반대로 행렬 V의 각 행 벡터와 S의 대각값의 곱인 $s_k d'_i$ 역시 기저 벡터인 v_k가 구성하는 새로운 공간상의 단어 벡터로 표현한 것을 의미한다. 이를 각각 식으로 나타내면 다음 식3과 식4와 같다.

$$t_i = \sum_{k=1}^{r} t'_{ik} s_k v_k, \quad \mathrm{i} = 1, \ \ldots \ , \mathrm{m} \qquad \text{(식 3)}$$

$$d_j = \sum_{k=1}^{r} d'_{jk} s_k u_k, \quad \mathrm{j} = 1, \ \ldots \ , \mathrm{n} \qquad \text{(식 4)}$$

t_i : i 번째 단어 벡터
d_j : j 번째 문서 벡터
t'_{ik} : i 번째 단어 벡터의 새로운 공간 상 k 축 값
d'_{jk} : j 번째 문서 벡터의 새로운 공간 상 k 축 값

결론적으로 LSA는 SVD를 통해 행렬의 행공간과 열공간, 단어-문서 행렬에서는 단어벡터 공간과 문서벡터 공간의 각각의 정규 직교 기저를 구하여 두 기저 벡터로 구성 된 새로운 공간상에 단어 혹은 문서 벡터를 표현함으로써 기존 행렬로부터 얻을 수 없는, 표면적인 의미가 아닌 잠재적 의미를 분석할 수 있다. 또한 이 중 의미 있는 차원만을 선택하여 문서 벡터를 선택한 저차원 상에 표현함으로써 기존 행렬이 가지는 잡음을 제거하고 단어-문서 행렬의 새로운 의미를 파악할 수 있다. 그러므로 문서 분류는 기저 u_k가 구성하는 축소된 차원 상에 표현된 문서벡터 $s_k v_k^T$를 통해 이루어진다.

2.3. 분류에서의 LSA의 약점

LSA 또는 LSI(Latent Semantic Indexing)은 선형 기법인 SVD가 가지는 장점들로 많은 분류 및 인식 분야에서 사용되어지고 있다. 예를 들어 문서 혹은 웹 페이지 분류부터 시작해서 의견 마이닝, 이미지 인식 등 넓은 분야에서 유용하게 사용되어지고 있다(Zhang et al., 2008)(Wan et al., 2008)(Praks et al., 2003). 이 모든 것은 SVD의 행렬 분해로부터 얻을 수 있는 차원의 축소와 그에 따른 잡음 제거를 통한 의미 분석에 초점을 맞추고 있다.

하지만 사전에 정의한 범주로의 분류작업에서, LSA는 도리어 안 좋은 성능을 보일 수 있다. SVD를 통한 차원 축소 시 차원 선택의 기준은 식1에서 행렬 Σ의 대각값인 특이값에 따르며 각 특이값은 새로 생성한 의미공간을 이루는 각 차원 축이 분류 대상이 되는 벡터들을 표현하는 정도를 의미한다. 기존 LSA는 특이값이 높은, 표현력이 높은 차원의 축만을 고려하여 차원 축소를 수행하지만, 이는 분류해야 할 서로 다른 범주들 간의 구별을 고려하지 않는다. 예를 들면, 그림 3(a)에서 보면 2차원 공간상 서로 다른 두 범주 C_1과 C_2의 벡터들이 있고 SVD를 통해 얻은 새로운 의미공간을 구성하는 v_1과 v_2가 있다. 기존 LSA의 경우 새로운 의미공간을 구성하는 차원 축 v_1이 v_2에

비해 공간 상 벡터들을 잘 표현함으로 기저 v_2를 제거하고 기저 v_1으로 모든 벡터를 투영하게 된다. 이때 범주에 대한 정보를 고려하지 않았기 때문에 결과적으로 그림 3(b)와 같이 축소된 1차원 상 서로 다른 범주간의 구분이 모호한 상황을 초래한다. 이러한 경우 분류의 기준이 애매해 지므로 분류 성능에 악영향을 미칠 수 있다. 근본적으로 벡터모델 분류기는 학습 데이터의 종류와 이 학습 데이터를 표현하는 벡터 공간에 따라 분류 성능이 좌지우지되기 때문이다.

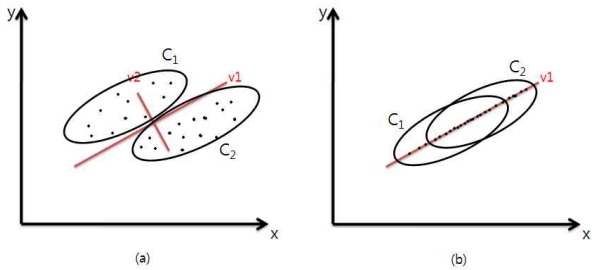

그림 3. (a) 새로운 의미공간 상 벡터 집합
(b) 차원 축소 후 벡터 집합

이러한 문제를 지도적으로 접근하여 해결하고자 하는 연구로 지도적 LSA인 Sprinkling이 있다. 이는 학습시킬 단어-문서 행렬을 생성 할 시 문서가 어떤 범주에 속하는지에 대해 인위적인 값을 추가 할당하는 방법으로, 기하학적으로 보았을 경우 같은 범주의 문서 집합끼리는 서로 강하게 묶이고, 서로 다른 범주에 속하는 문서 집합은 벡터 공간상에서 좀 더 명확히 구분시킬 수 있다.

다음 장에서 기존의 지도적 LSA 중 Sprinkling에 대해 소개하고 약점 및 본 논문이 제안한 새로운 지도적 LSA에 대해 소개한다.

3. 지도적 LSA

지도적 학습을 이용한 LSA 기법은 학습 대상인 문서 집합에 대한 단어-문서 행렬을 생성할 때 분류에 적합한 환경을 만들기 위한 것으로, 지도적으로 추가적인 정보를 삽입함으로써 서로 같은 범주에 속하는 문서간의 동질성을 높이고 다른 범주의 문서와의 이질성을 높여 더 명확히 분류될 수 있도록 한 대표적인 지도적 LSA 기법으로 Chakraborti 등(2006)이 제안한 Sprinkling 기법이 있다. Sprinkling 기법은 분류하고자 하는 범주의 수만큼 단어-문서 행렬에 새로운 단어를 추가하고 추가한 각 단어는 각 범주를 의미하도록 한다. 그리고 문서는 문서가 속하는 범주에 해당하는 추가된 단어에 대해서만 1의 값을 가지도록 하여 서로 다른 범주의 문서 간 차이를 둔다. Sprinkling을 응용한 예로, 이지혜(2009)는 한글 의견 문서의 자동분류를 Sprinkling을 사용하여 수행하였다(이지혜와 정영미, 2009).

그림 5는 그림 4에서 1차원 공간상의 서로 다른 두 범주 벡터들이 Sprinkling 후 추가 된 두 단어들로 인해 3차원으로 확장 된 공간상에서 두 범주의 차이를 확실히 보여준다.

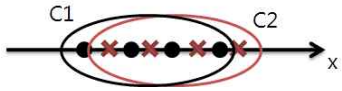

그림 4. 1차원 상 구분이 모호한 두 범주

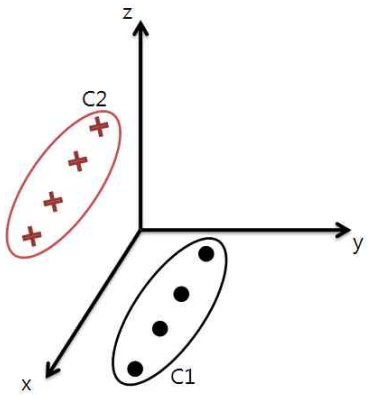

그림 5. 확장된 차원에서 명확히 구분된 두 범주

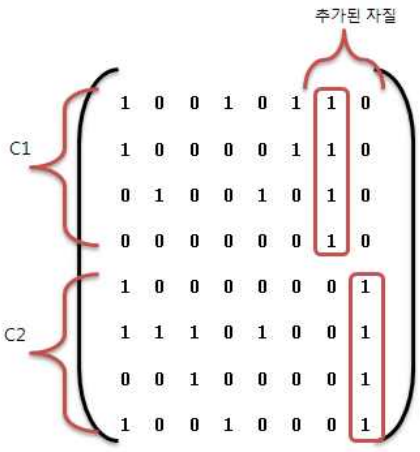

그림 6. Sprinkling한 행렬의 예

또한 그림 6에서 실제 단어-문서 행렬에 대해 Sprinkling 한 이후 추가된 범주에 대한 정보 자질을 확인할 수 있다.

기존 Sprinkling은 단어-문서 생성 시 단어에 대한 문서 표현에 4.2절에서 소개할 문서 벡터 표현법 중에 이진 표현법(binary representation)이 사용되어 진다. 이는 단어의 출현 여부만으로 문서를 표현하는 방법으로 문서 안에 한번이라도 출현한 단어는 1로 표현하는 것이다. 그렇기에 추가되는 단어의 값 역시 0또는 1로 표현할 시 명확하게 서로 다른 범주를 구별할 수 있게 할 수 있다. 그러나 이진 표현법은 각 단어의 특성을 반영하지 못하기 때문에 잘 사용되지 않는다. 그렇기 때문에 요즘은 4.2절에서 소개하게 될 각 단어의 의미를 반영하는 다양한 표현 방법을 사용하고 있다. 이때 Sprinkling 시 추가하는 범주에 해당하는 단어의 값을 1또는 0으로 표현하였을 시에는 추가한 범주에 대한 정보가 SVD의 자질 차원 축소 시 손실될 수 있다.

이에 본 논문에서는 기존 Sprinkling의 제한적이던 지도적 LSA방법에 비해 좀 더 범용적이고 유연하게 사용되어 질 수 있는, Sprinkling의 특징을 포함시킨 최적의 자질 차원 선택을 위한 지도적 LSA(OS-LSA, Optimal Supervised LSA)를 소개한다.

3.1 제안한 OS-LSA

기존 LSA와 지도적 LSA인 Sprinkling은 선형 기법인 SVD를 통해 얻은 새로운 의미공간을 분류할 문서 집합을 가장 잘 나타내는 저 자질 차원의 의미공간으로 투영하여 축소된 자질 차원상의 문서 벡터들로 분류를 수행한다. 이때 기존 LSA에서 자질 차원 축소의 기준이 되는 것은 의미공간을 구성하는 각 기저 벡터의 벡터에 대한 표현력인 특이치(singular value)이다. 하지만 각 기저 벡터의 표현력만을 고려할 뿐 서로 다른 범주의 문서집합들 간의 명확한 구분을 고려하지 않는다. 특히 학습 및 분류의 기준이 될 샘플 문서들은 의미공간 상에서 명확하게 구분되어져야 하나 그 구분이 모호해 질 수 있다. Sprinkling에서도 마찬가지로 자질 차원 축소 시 서로 다른 범주간의 구분을 고려하지 않아 추가한 범주 정보가 손실 될 수 있다.

이에 SVD의 자질 차원 축소에서 기존의 특이치에 의한 자질 차원의 선택이 아닌 서로 다른 범주의 문서집합을 잘 구분해 주는 자질 차원을 선택함으로써 서로 다른 범주의 문서집합의 이질성을 최대화 시키는 지도적 LSA(S-LSA, Supervised LSA) 기법이 제안되었다(Sun et al., 2004). 하지만 제안 된 방법은 t개의 새로운 의미 공간을 이루면서 서로 다른 범주를 더 확실히 구분시켜 주는 기저를 구하기 위해 매번 전체 단어-문서 행렬에 대해 t번 SVD를 수행해야하는 부담이 있다. 계산량이 분류해야 할 행렬의 크기에 따라 기하급수적으로 증가 할 수 있다.

본 논문은 계산량을 최소화하고 지도적 LSA가 추구하는 새로운 의미공간 상 서로 다른 범주끼리의 구분을 위해 기존 LSA의 표현력에 따른 자질 차원 축소뿐만 아니라 차별적 자질 차원 선택 방법을 이용하여 최적의 자질 차원을 선택하는 최적의 지도적 LSA 기법을 제안한다. 또한 제안하는 방법은 자질 차원 선택 전에 Sprinkling의 방법인 범주의 정보를 단어-문서 행렬에 추가함으로써 Sprinkling의 특징을 살려 더 정확한 차별적인 자질 차원 선택을 수행한다.

제안하는 방법은 다음의 순서로 분류를 수행하기 전 단어-문서 행렬로부터 효율적으로 서로 다른 범주간 정확한 구별을 고려한 최적의 자질 차원을 선택한다.

먼저 학습 및 분류의 기준이 될 단어-문서 행렬에 각 문서가 포함되는 범주에 대한 정보를 추가한다. 범주의 정보가 추가된 행렬로부터 SVD를 이용하여 문서 벡터를 표현하기 위한 새로운 의미공간을 생성한다. 생성된 의미공간으로부터 기존 SVD의 자질 차원 축소를 바탕으로 모든 벡터 집합에 대해 일정 수의 표현력이 큰 자질 차원들만을 선택하여 자질의 저차원 공간에 벡터 집합을 투영한다. 여기서 축소 될 일정 수는 경험에 의해 얻은 값으로 설정하였다. 이후 축소된 자질 차원으로부터 서로 다른 범주를 가장 잘 구분시켜 주는 자질 차원을 선택하여 최적의 자질 차원을 선택함으로써 문서를 가장 잘 표현하면서 서로 다른 범주를 가장 잘 구별 해 주는 자질의 저차원의 의미공간에 문서를 표현한다.

각 자질 차원의 축이 서로 다른 범주의 문서집합을 얼마만큼 구분시켜 주는지를 나타내는 정도는 다음 식6로 구할 수 있다.

$$e_i^c = \frac{1}{N_c} \sum_{v \in D_c} b_j^T v \qquad \text{(식 5)}$$

e_i^c : 기저 벡터 b_j가 범주 c인 단어 벡터들에 미치는 영향

N_c : 범주 c에 속하는 문서의 총 수

v : 범주가 c인 문서 집합 D_c에 속하는 문서 벡터

b_j : j 번째 기저 벡터

$$dw_i^{(c_1,c_2)} = | e_i^{c_1} - e_j^{c_2} | \quad \text{(식 6)}$$

$dw_i^{(c_1,c_2)}$: 기저 벡터 b_i가 두 범주 c_1과 c_2 간 차이를 나타내는 정도

한 기저 벡터가 서로 다른 두 범주에 속하는 문서 집합에 미치는 영향 정도를 식5를 통해 각각 구한 뒤 해당 기저 벡터가 서로 다른 범주간의 차이를 표현하는 정도를 두 값의 차이값을 통해 얻는다. 식 6을 통해 모든 기저벡터에 대해 서로 다른 범주의 문서집합 간의 차이를 표현하는 정도를 구하여 알고리즘 1과 같이 그 값이 가장 큰 기저 벡터들을 내림차순으로 정렬하여 구하고자 하는 찾고자 하는 t개의 최적의 자질 차원을 구성하는 기저 벡터만을 선택한다.

For each sorted dw_i ,
Select t vectors of dw_i

(알고리즘 1)

선택된 기저 벡터들로 구성된 새로운 자질의 저차원 공간에서 다시 모든 범주의 문서집합을 표현함으로써 각 범주별 문서집합 간의 종속성을 높이고 다른 범주의 문서집합 간의 이질성을 최대화 시켜 분류에 있어서 명확한 분류가 가능하게 한다.

4. OS-LSA를 이용한 웹 페이지 분류 시스템

본 논문에서 제안한 OS-LSA를 기반 분류기법으로 한 웹페이지의 분류를 위하여 시스템을 구성하였다. 시스 템은 그림 7과 같이 전 처리기로 크로울러 및 웹 페이지 파서와 크게 3개의 주요 처리기로 구성하였다. 주요 처리기는 각각 자질선택(Feature Selection), 자질 색인(Feature Indexing), 분류기(classifier)로 대부분의 분류 시스템과 거의 동일한 구조를 가진다. 자질선택은 분류 대상을 표현할 자질들을 선택하는 작업을 수행하며, 자질 색인은 선택한 자질을 이용하여 분류 대상을 표현하는 것을 수행한다. 그리고 마지막으로 분류기는 자질 색인의 결과인 자질들로 표현된 분류 대상을 사전에 정의한 범주로 분류하는 작업을 수행한다.

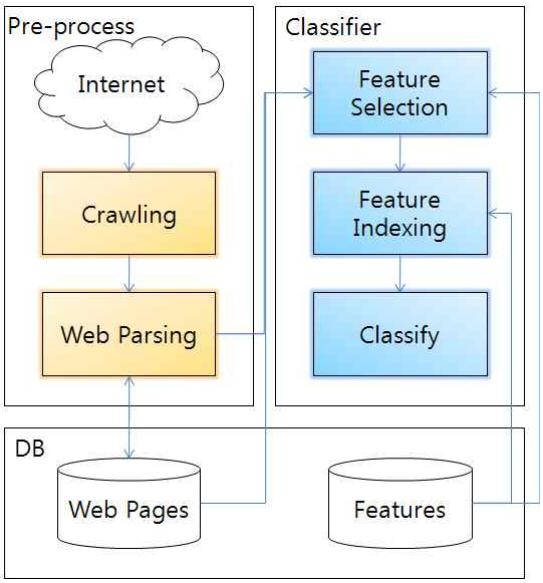

그림 7. 웹 페이지 분류기 시스템 구조도

4.1. 자질 선택

분류작업을 수행하기 전에 전처리 작업으로 문서 혹은 웹 페이지를 표현할 자질의 선택이 이루어져야 한다. 자질이란 분류의 대상이 되는 객체를 나타내기 위한 객체를 대표할 수 있는 특성(characteristic)으로, 이 논문에서는 표현하고자 하는 대상인 문서를 나타내는 자질로 단어를 사용하며 웹 페이지의 경우 단어뿐 만 아니라 HTML 태그, 그림 등을 자질로 사용한다. 본 논문의 분류 대상이 문서, 웹 페이지이므로 앞으로 자질과 단어를 동일한 의미로 혼용하여 사용하겠다. 분류작업 전 분류하고자 하는 대상의 단어의 종류에 따라 대상을 잘 표현함이 결정되기 때문에 단어 선택이 분류의 성능에 영향을 미치는 가장 큰 요인 중 하나이다.

대표적인 단어 선택 기법으로 문서 빈도(document frequency)와 카이제곱 통계량 및 정보 획득량(information gain)이 있다(고영중과 서정연, 2002). 예전엔 문서 빈도 기법이 가장 쉽고 간단한 기법으로 자주 사용되었지만 특정 단어가 분류범주를 나타내는데 큰 영향을 미치더라도 문서 빈도가 낮은 경우 단어 선택에서 제외되는 문제가 있다. 이러한 문제를 다

소 해결하기 위해 통계적 접근방법으로 학습 문서 집합으로부터 추출된 단어들의 출현 빈도의 통계값을 이용한 카이제곱 통계량과 정보 획득량이 자주 사용되고 있다.

본 논문에서는 단어 선택 시 단어의 품사를 명사로 한정지어 선택하였다. 명사가 사물 또는 사람을 가리키는 명칭을 나타내는 품사이므로 특정 범주에서 사용되어지는 명사가 그 범주를 가리키는 대표 단어이기 때문이다. 그리고 웹 페이지의 효과적인 단어 선택을 위해 문서 빈도와 통계적인 기법을 접목하여 두 선택 기법이 가지는 장점을 최대화하였다. 문서 빈도 방법은 기존의 방법과 마찬가지로 알고리즘 2와 같이 자질이 문서에서의 출현 빈도가 일정 경계값(threshold) 이하일 경우 분류할 범주를 나타내는 정도가 낮다고 가정하여 자질에서 제외시키는 방법으로, 본 논문에서는 문서 빈도에 의한 자질 선택뿐만 아니라 사전에 정의한 특정 범주에서만 출현하는 자질에 대해서는 그 범주를 잘 나타내는 자질이라 간주하여 문서 빈도가 낮아도 자질로 선택하였다.

if $df_i \leq threshold$ (i : 단어 번호)
    selected
else
    not selected

(알고리즘 2)

단어 선택에서 제외된 단어들은 대부분 품사가 고유명사인 단어들이다. 이는 고유명사가 고유한 사물 또는 사람을 나타내는 것이므로 일반명사에 비해 더욱 특정 범주에 한정되며 그에 따라 문서에서의 출현 정도가 매우 미비하기 때문이다. 본 연구에서는 문서 빈도가 매우 낮은 고유명사 중 앞서 제시한 문서 빈도가 낮더라도 특정 범주에서만 출현하는 고유명사의 경우 그 범주를 잘 나타내므로 자질로 선택하고 그렇지 않은 고유명사는 불용어로 처리하여 불용어 사전에 추가하였다. 그리고 불용어로 판정 된 고유명사는 매 학습 시 출현 빈도수를 갱신시켜 자질 조건을 만족 할 시 다시 자질로 선택 될 수 있게 하였다.

4.2. 자질 색인

자질 선택을 통해 사전에 정의한 범주로 문서를 분류하기에 가장 적절한 자질들을 선택한 뒤 선택된 자질들로 분류 대상인 문서를 알맞게 표현하는 색인 작업 역시 분류에 앞서 중요한 작업 중 하나이다. 벡터 모델 분류기를 사용하기 때문에 문서는 선택 된 단어들로 이루어진 벡터 공간상의 벡터로 표현된다. 그렇기 때문에 선택된 각 단어에 대해 벡터값 즉, 단어가 문서를 표현하는 정도를 의미하는 가중치를 주어야 한다. 보통은 다음 3가지 방식으로 단어에 대한 가중치를 주어 문서를 표현한다.

a) 이진 표현(Binary Representation)
b) 단어 빈도(Term Frequency)
c) 역문헌 빈도(Term Frequency inverse Document Frequency)

이진 표현은 단어가 문서에서의 출현여부만을 고려하여 문서 안에 단어가 존재하면 1, 그 반대의 경우 0의 값으로 나타낸다. 단어 빈도는 문서 안에서 단어가 총 출현한 횟수를 값으로 표현하는 방법이며 가장 흔히 사용되는 방법이다. 하지만 단어 빈도 방법의 경우 출현 빈도는 높지만 의미 없는 단어의 경우 높은 값을 가지게 되어 분류 성능에 악영향을 미칠 수 있다. 이러한 문제를 해결하기 위해 나온 색인(indexing) 방법인 역문헌 빈도수는 식7과 같이 특정 문서에서의 단어 출현 빈도수에 해당 단어를 포함하는 문서의 수를 나누어 줌으로써 의미 없이 많이 사용되는 단어를 작은 값으로 표현하여 문서 벡터에 미치는 영향을 줄이게 하고 반대로 출현빈도는 다소 낮지만 의미 있는 단어에는 높은 값을 주어 문서의 정확한 의미를 나타내도록 한다.

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad \text{(식 7)}$$

$tf_{i,j}$: 문서 j에서 단어i가 출현하는 빈도
df_i : 단어 i를 포함하는 문서의 수
N : 문서의 총 수

위 세 가지 색인 방법들은 항상 단어의 값으로 양의 정수를 가진다. 이러한 경우 유사한 자질들을 가지는 서로 다른 범주의 문서 벡터를 표현 할 시 양의 공간상에서 표현되므로 자칫 잘못하면 서로 다른 범주의 문서들이 섞이거나 분리하기 힘든 경우가 발생한다. 앞선 그림 3에서 보듯이 유사한 자질을 가지는 서

로 다른 범주 C_1과 C_2의 문서 벡터 집합들이 뒤엉켜 분류하기 애매한 상황이 발생할 수 있다.

이 문제를 다소 해결하기 위한 방법으로 앞서 언급한 지도적 LSA 기법인 Sprinkling이 있다. 추가적으로 본 논문에서는 통계적 접근방법으로 학습 문서로부터 각 자질의 범주별 출현 빈도수를 구한 뒤 이를 바탕으로 한 식8를 이용해 서로 다른 범주에 따라 같은 자질일지라도 다른 가중치 값을 주어 좀 더 명확히 범주를 나눌 수 있도록 하였다. 학습으로 얻은 각 자질의 범주별 빈도수의 비를 이용하여 어느 범주에서 해당 자질이 더욱 의미가 있는지를 나타낼 수 있도록 함으로써 동일한 자질을 가지면서 서로 다른 범주에 속하는 문서들을 구분지어 벡터 공간상에 표현할 수 있다.

$$w_{(i,j)} = \frac{f_i^{c_2}}{f_i^{c_1}} \times tf_{i,j} \qquad \text{(식 8)}$$

$f_i^{c_j}$: 학습 결과 범주 c_j에서 단어 t_i의 출현 빈도수
$tf_{i,j}$: 단어 t_i의 문서 d_j에서의 출현 빈도 수

4.3. 분류기

분류기는 본 논문에서 제안한 분류 기법인 OS-LSA를 사용하여 분류하기 전 처리 작업인 자질 선택과 자질 색인의 결과인 단어-문서 행렬을 입력값으로 받아 사전에 정의한 두 범주로 문서의 분류를 수행한다. 기본적으로 LSA는 MATLAB 2007의 svd 함수를 통해 수행되어지며 svd의 결과로 얻은 새로운 의미를 가지는 축소된 행렬로부터 각 문서 벡터의 유사정도를 계산한다.

유사도 계산 기법은 먼저 유사 정도의 척도로 거리계수와 유사계수로 나누어지며 각각 거리계수는 벡터 공간상에서 각 벡터간의 상이성 정도를 측정하며, 유사계수는 각 벡터의 일치 정도를 측정한다. 일반적으로 분류에서 거리계수로는 유클리드 거리(euclidean distance) 기법이 사용되며 유사계수로 코사인 유사도(cosine similarity) 기법을 사용한다. 각 기법들의 벡터간 유사 정도를 계산하는 식은 식9와 식10이다.

$$\cos(p,q) = \frac{p \bullet q}{\|p\| \bullet \|q\|} \qquad \text{(식 9)}$$

$$d(p,q) = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \qquad \text{(식 10)}$$

$p,\ q$: 서로 다른 두 벡터
$\cos(p,q)$: 벡터 p와 q의 코사인 값
$d(p,q)$: 벡터 p와 q의 유클리드 거리 값

본 논문에서는 두 유사도 계산 기법 중 자질의 저차원 공간에서도 문서 벡터의 분류를 안정적으로 분류할 수 있는 코사인 유사도 기법을 사용하여 분류를 수행하였다.

5. 실험 및 분석

5.1. 실험 데이터

본 논문은 분류 실험을 위해 인터넷 상의 조선일보 사이트로부터 20만개의 신문기사 웹 페이지를 크로울링 하여, 분류 실험을 위해 표 1에서 보이는 바와 같이 사전에 정의한 세 개의 범주인 경제, 정치, 스포츠에 속하는 각각 1000개의 신문기사 웹페이지를 임의적으로 선택하여 학습 데이터로 각 범주 당 500개의 페이지를 사용하였다. 그리고 학습 데이터를 제외한 각 범주에 속하는 나머지 500개씩의 신문기사 웹페이지를 가지고 분류 실험을 수행하였다. 분류 실험은 서로 다른 두 범주로의 분류에 초점을 맞추었다.

분류할 두 범주의 구성은 서로 상이한 자질을 가지는 경제와 스포츠, 스포츠와 정치로 구성하였으며 또한 서로 유사한 자질을 가지는 경제와 정치를 분류할 범주 구성으로 구성하여, 범주 간 자질의 유사성에 따른 분류에서도 어느 정도 효과적으로 분류를 수행하는지 실험을 통해 확인해 보았다.

표 1. 실험 데이터의 종류 및 개수

	학습 페이지	실험 페이지 1	실험 페이지 2
경제	500개	500개	500개
스포츠	500개	500개	500개
정치	500개	500개	500개

표 2. 학습 페이지로부터 추출한 자질 수와 분류를 위해 선택한 자질 수

	명사	명사, 문서 빈도
경제	19454개	6202개
스포츠	10154개	4613개
정치	16180개	5810개
전체	29059개	7269개

5.2. 자질 선택

먼저 학습을 위해 수집한 각 범주 당 500개의 신문기사 웹 페이지들로부터 품사가 명사인 단어들을 추출하였다. 명사 추출을 위해 지능형 형태소 분석기를 이용하였으며, 명사는 세부적으로 일반명사와 고유명사만을 추출하였다.

학습 페이지로부터 추출한 명사 단어는 각 범주별 출현 빈도수를 가지고 있으며 이 빈도수를 바탕으로 전체 범주에 대한 페이지에서의 출현 빈도가 낮은 단어를 제외하였다. 또한 추가로 각 범주별 출현 빈도의 비율이 1:1에 비슷한 경우 두 범주를 분류함에 있어서 큰 영향을 미치지 않는다는 가정 하에 제외하였다. 그 외에 직접 정의한 불용어 사전을 통해 불용어를 제거하여 최종 분류에 사용할 자질 즉, 단어를 선택하였다.

표 2와 같이 각 범주별 명사 단어만을 추출하였을 시, 중복을 포함하여 경제에서 19454개, 스포츠 10154개, 정치 16180개의 단어가 추출되었으며, 이 중 이지혜와 정영미(2009)의 연구에서 문헌 빈도가 5 이하인 단어들로 분류를 수행하였을 때 가장 좋은 분류 결과를 얻은바와 같이 본 논문에서도 문헌 빈도가 5 이하인 단어를 제거하여 전체 범주에 대해 약 1 만개의 단어를 추출하였다. 이때 제거된 단어들 중 대부분이 고유명사이며, 본 연구에서는 제거된 고유명사 중, 분류에 있어서 의미가 있는 일부 고유명사를 추출하였다. 마지막으로 각 범주별 빈도수를 토대로 의미 없는 단어를 제거하여, 중복을 포함하여 범주별로 각각 경제 6202개, 스포츠 4613개, 정치 5810개로 전체적으로 7269개의 단어를 분류에 사용할 자질로 선택하였다.

표 3은 자질 선택 후 범주간의 중복되는 자질의 수를 보여주며, 중복되는 정도를 통해 각 범주간 유사 정도를 파악할 수 있다. 경제와 스포츠, 스포츠와 정치의 경우 사용되는 자질의 중복 정도가 50%대인 반면, 경제와 정치는 중복 정도가 76% 이상으로 두 범주의 유사 정도가 높다는 것을 확인 할 수 있다. 분류 실험에서는 분류 시 분류의 대상이 되는 두 범주에 속하는 단어들만으로 자질을 구성하여 단어-문서 행렬을 생성하여 이를 바탕으로 분류를 수행하였다.

표 3. 범주 간 중복되는 단어 수 및 비율

	단어 수	중복 단어 수	중복 비율
경제&스포츠	6870 개	3945 개	57.4%
스포츠&정치	6769 개	3654 개	54.0%
경제&정치	6800 개	5212 개	76.6%

5.3. 다른 분류기법들과의 성능 비교

본 논문에서의 분류 성능 측정의 척도로 이진 분류(binary classification) 테스트에서 사용하는 정확도(accuracy)를 사용하였다. 정확도는 범주를 알 수 없는 전체 데이터에 대해서 자동으로 분류한 결과 중 정확하게 분류한 정도를 의미한다. 식11은 서로 다른 두 범주에 대한 정확도 측정을 나타낸다.

$$Accuracy = \frac{Correct(C_1) + Correct(C_2)}{N(C_1) + N(C_2)} \quad \text{(식 11)}$$

$Correct(C)$: 정확히 분류된 범주가 C인 데이터의 수
$N(C)$: 범주가 C인 데이터의 수

5.3.1. 기존 LSA 방법들과의 비교

분류 실험을 하기 위해 사전에 구성해 놓은 테스트 집합을 기존 LSA, Sprinkling, 지도적 LSA, 그리고 본 논문에서 제안한 OS-LSA로 각각 분류 실험을 하였다.

그림 8~10과 그림 11~13는 각각 분류 실험을 위해 사전에 구성한 두 실험 데이터에 대해 4가지 종류의 LSA로 분류 실험한 결과를 보인다. 평균적으로 모든 실험에서 본 논문이 제안한 OS-LSA가 다른 LSA 분류 기법에 비해 높은 성능을 보였다. 특히 다른 LSA 기법 보다 적은 자질 차원의 수에서 높은 정확도를 보였다. 특히 의미 공간의 자질 차원 수 10~30에서 OS-LSA는 다른 LSA 기법들 보다 높은 분류 성능을 보이며, 반면 다른 LSA 기법들은 자질 차원의 수가 100 이상일 때 자체적으로 높은 분류 성능을 나타냈지만 OS-LSA의 분류 성능에 비해 비교적 낮은 결과를 보였다. 이는 Rosario(2000)가 LSA에 대해 조사한 결과 일반 LSA는 70~100 자질 차원에서 가장 높은 성

능을 보임을 여러 실험군에 대한 실험으로 관측된 바와 동일하다. 하지만 자체적으로 높은 성능이지만 제안한 OS-LSA가 10~30 자질 차원에서 보이는 성능에 비해 낮은 성능을 보인다.

다음 표 6은 각 실험 데이터를 OS-LSA로 분류한 결과 10~30 자질 차원에서의 평균 정확도를 보인다. 대체적으로 80% 내외의 평균 성능을 보이며, 스포츠와 정치와 같이 분류하는 범주간의 이질성이 큰 데이터에 대해서는 90%에 가까운 분류 성능을 보였다.

결론적으로, 본 논문이 제안한 OS-LSA는 낮은 자질 차원일지라도 분류할 데이터를 가장 잘 표현하는 최적의 자질 차원만으로 구성되기 때문에 다른 LSA 기법보다 분류에서 높은 성능을 보였다.

표 5. 자질 차원 10~30에서의 OS-LSA의 평균 정확도

	실험 데이터 1	실험 데이터 2
경제&스포츠	0.859	0.888
경제&정치	0.765	0.809
스포츠&정치	0.899	0.94

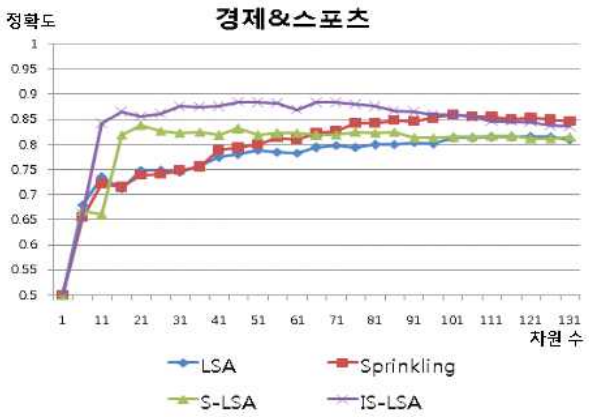

그림 8. 실험 데이터 1 (경제&스포츠)

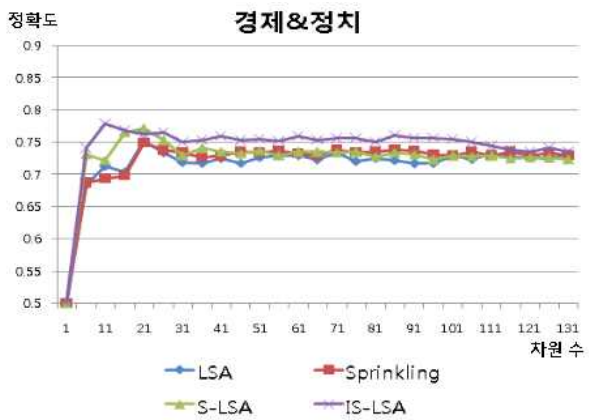

그림 9. 실험 데이터 1 (경제&정치)

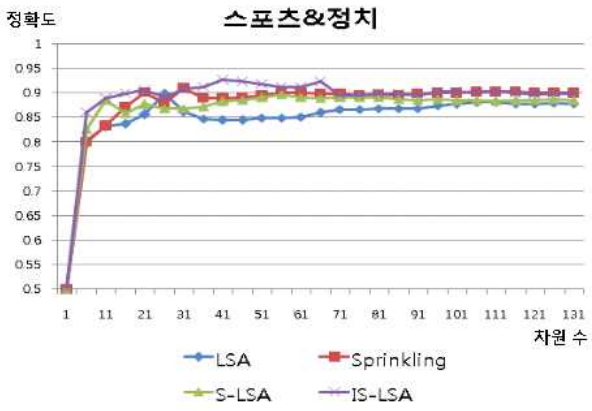

그림 10. 실험 데이터 1 (스포츠&정치)

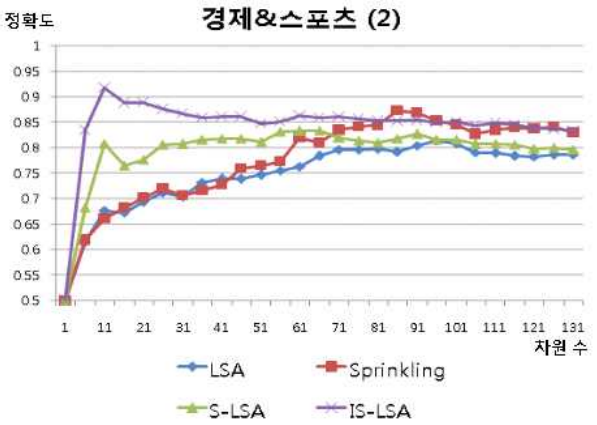

그림 11. 실험 데이터 2 (경제&스포츠)

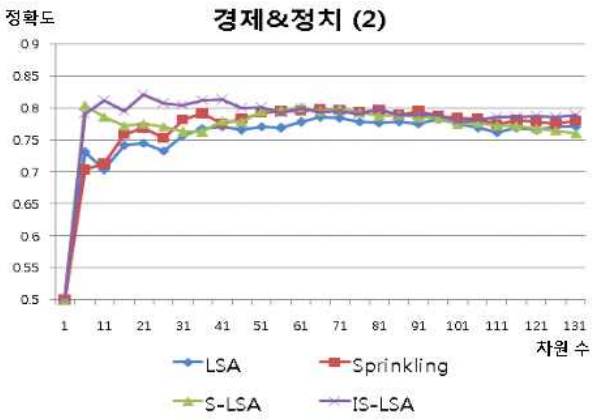

그림 12. 실험 데이터 2 (경제&정치)

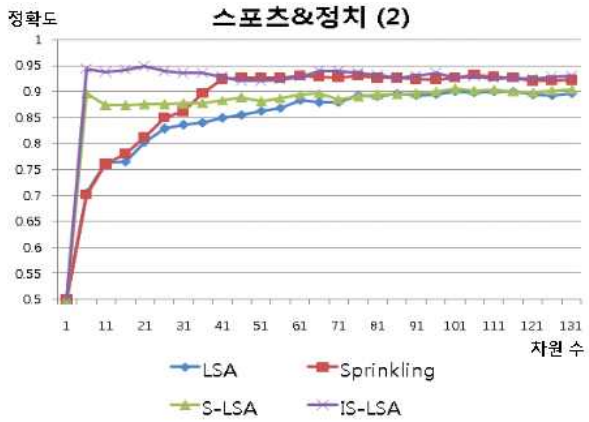

그림 13. 실험 데이터 2 (스포츠&정치)

6. 결론 및 향후 연구

LSA는 자질 차원의 축소 및 자질의 잠재적 의미 분석을 가능하게 해 주는 LSA의 특성으로 여러 패턴 인식 및 분류분야에서 사용되었고 현재 연구 중에 있다. 하지만 이 중 웹 페이지 또는 텍스트 문서 분류 분야에서 LSA를 사용하였을 경우 분류 성능에 한계가 있었다. LSA의 특성인 자질 차원 축소 시 분류할 서로 다른 범주의 구별을 고려하지 않아 분류의 기준이 될 학습 문서가 구성된 벡터공간에서 명확한 기준을 제공하지 못하는 경우가 발생하여 애매한 분류의 기준에 의해 분류의 정확도를 떨어트리기 때문이다. 이를 해결하는 방법으로 기존 논문은 지도적으로 서로 다른 범주 간 구분을 고려한 LSA에 대한 몇몇 연구가 진행되었다. 하지만 기존 지도적인 LSA 방법은 응용이 제한적이며 높은 복잡도로인해 드는 노력에 비해 성능이 비효율적이다.

본 논문에서는 기존 LSA의 자질 차원 축소 특성을 유지하면서 기존 지도적 LSA의 단점을 최소화하고 장점을 최대화 할 수 있는 지도적 LSA를 이용하여 보다 향상된 분류 성능을 얻을 수 있었다. 선택한 일정 자질 차원 수에서 기존 LSA에 비해 비교적 안정된 분류 성능을 보였으며 전반적인 분류 성능역시 입증된 다른 분류 기법에 상응하거나 높게 나왔다.

하지만 본 논문에서 사용한 지도적 LSA는 오직 서로 다른 범주를 잘 구별하기 위해 지도적 자질 차원 선택에만 초점을 맞췄을 뿐, 서로 다른 두 범주의 밀집 정도에 따른 분류의 편향성을 고려하지 않으므로, 추후 두 각 범주의 밀집 정도를 고려한 자질 차원 선택을 통해 좀 더 정확한 분류 성능을 기대할 수 있다. 그리고 본 논문의 경제, 정치, 스포츠 신문의 한정된 실험 범위에서 좀 더 일반화된 분류 성능의 입증을 위하여, 이것을 감성 용어간의 유사성의 응용분야로써 감성분류에 적용하는 것을 추후 연구로 확대할 수 있다.

참고문헌

고영중, 서정연 (2002). 문서관리를 위한 자동문서범주화에 대한 이론 및 기법. *정보관리연구, 33(2),* 19-32.

김도겸, 엄재홍, 장병탁 (2009). Latent Semantic Analysis를 이용한 Naive Bayes 텍스트 분류. *한국컴퓨터종합학술대회 논문집, 36(1),* 149-150

신동호 (1999). LSA를 이용한 정보검색용 시소러스 구축. *지식표현 및 추론 기말 보고서,* 서울: 서울대학교 인지과학 협동과정

신동호 (2000). LSA를 이용한 내용기반 정보검색 시스템. *서울대학교 공학석사 학위논문.*

이경찬, 강승식 (2002). 범주 대표어의 가중치 계산 방식에 의한 자동 문서 분류 시스템. *한국정보과학회 2002년도 봄 학술발표논문집, 29(1),* 475-477

이태헌, 김청택 (2004). LSA 모형에서 다의어 의미의 표상. *인지과학, 15(2),* 23-31.

이지혜, 정영미 (2009). 지도적 잠재의미색인(LSI)기법을 이용한 의견 문서 자동 분류에 관한 실험적 연구. *정보관리학회지, 26(3)*, 2009, 451-462.

조태호 (1998). 신경망 또는 k-NN에 의한 신문기사 분류와 그의 성능 비교. *한국정보과학학회 가을 학술발표논문집, 25(2),* 363-365.

Chakraborti, S., Lothian, R., Wiratunga, N. & Watt, S. (2006). Sprinkling: Supervised Latent Semantic Indexing. *Lecture Notes in Computer Science, 3936,* 510-514.

Chen R. C. & Hsieh C. H. (2006). Web page classification based on a support vector machine using a weighted vote schema. Expert *Systems with Applications 31,* 427-435.

Kontostathis A. & Pottenger W. M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing and Management, 42(1),* 56-73.

Landauer, T. K, Foltz, P. W., & Laham, D. (1998). An

Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

Li, S., Zong, C., & Wangm, X. (2007). Sentiment Classification through Combining Classifiers with Multiple Feature Sets. *Natural Language Processing and Knowledge Engineering International Conference,* 135-140.

Praks, P., Dvorsky, J., & Snasel, V. (2003). Latent Semantic Indexing for Image Retrieval Systems. In Proccessing SIAM Conference on Applied Linear Algebra.

Qi, X., & Davison, B. D. (2009). Web Page Classification : Features and Algorithms. *Association for Computing Machinery,* 41(2), 1-31.

Rosario B. (2000). Latent Semantic Indexing: An overview. INFOSYS, Spring Final Paper.

Selamat A. & igeru Omatu S. (2004). Web page feature selection and classification using neural networks. *Information sciences, 158,* 69-88.

Sun, J. J., Chen, Z., Zeng, H. J., Lu, Y. C., Shi, C. Y. & Ma, W. Y. (2004). Supervised Latent Semantic Indexing for Document Categorization. *Fourth IEEE International Conference on Data Mining,* 535-538.

Wan, Y. & Tong, H. (2008). Categorization and Monitoring of Internet Public Opinion Based on Latent Semantic Analysis. *International Seminar on Business and Information Management, 2,* 121-124.

Zhang, Y., Fan, B. & Xiao, L. B. (2008). Web Page Classification Based-on A Least Square Support Vector Machine with Latent Semantic Analysis. *Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2,* 528-532.

원고접수 : 09.11.01
수정접수 : 10.01.12
게재확정 : 10.01.22