기계학습을 통한 디스크립터 자동부여에 관한 연구

A Study on automatic assignment of descriptors using machine learning

김 판 준(Pan-Jun Kim)*

목 차

* 연세대학교 문헌정보학과 시간강사(dpbluesea@hanmail.net)

초 록

학술지 논문에 디스크립터를 자동부여하기 위하여 기계학습 기반의 접근법을 적용하였다. 정보학 분야의 핵심 학술지를 선정하여 지난 11년간 수록된 논문들을 대상으로 문헌집단을 구성하였고, 자질 선정과 학습집합의 크기에 따른 성능을 살펴보았다. 자질 선정에서는 카이제곱 통계량(CHI)과 고빈도 선호 자질 선정 기준들(COS, GSS, JAC)을 사용하여 자질을 축소한 다음, 지지벡터기계(SVM)로 학습한 결과가 가장 좋은 성능을 보였다. 학습집합의 크기에서는 지지벡터기계(SVM)와 투표형 퍼셉트론(VPT)의 경우에는 상당한 영향을 받지만 나이브 베이즈(NB)의 경우에는 거의 영향을 받지 않는 것으로 나타났다.

ABSTRACT

This study utilizes various approaches of machine learning in the process of automatically assigning descriptors to journal articles. After selecting core journals in the field of information science and organizing test collection from the articles of the past 11 years, the effectiveness of feature selection and the size of training set was examined. In the regard of feature selection, after reducing the feature set by $\chi 2$ statistics(CHI) and criteria which prefer high-frequency features(COS, GSS, JAC), the trained Support Vector Machines(SVM) performs the best. With respective to the size of the training set, it significantly influences the performance of Support Vector Machines(SVM) and Voted Perceptron(VTP). but it scarcely affects that of Naive Bayes(NB).

키워드 : 디스크립터, 자동색인, 기계학습, 자질 선정, 분류기, 텍스트 범주화
descriptors, automatic indexing, machine learning, feature selection, classifier, text categorization

1 서 론

학술정보서비스를 목적으로 하는 온라인 데이터베이스의 경우 색인작업은 크게 두 가지 경로로 이루어진다고 할 수 있다. 먼저, 컴퓨터가 입력문헌의 텍스트를 분석하여 문헌의 내용을 대표하는 키워드(자연언어 색인어)를 일정한 기준에 의해 추출한다. 다음으로, 색인전문가는 해당 문헌의 내용을 분석하여 다루고 있는 주제를 판단한 다음, 통제어휘집에서 이를 표현할 수 있는 적절한 디스크립터(통제 색인어)를 부여한다. 예를 들면, 온라인 데이터베이스 서비스의 하나인 CSA Illumina에서는 데이터베이스 레코드를 구성하는 거의 모든 필드의 데이터로부터 색인어가 추출되며, 주제 색인어는 사람의 지적 작업이나 자동색인 프로그램에 의해 선정된다. 이처럼 대부분의 온라인 데이터베이스나 온라인 목록은 키워드 이외에 시소러스나 주제명표로부터 색인전문가가 선정한 디스크립터나 주제어를 추가적인 주제색인어로 수록하고 있다(정영미 2005).

색인 전문가는 전통적으로 문헌의 주제를 표현하기 위한 통제 색인어로서 디스크립터를 문헌에 부여하는 역할을 수행하여 왔으며, 인터넷을 통해 엄청난 양의 정보가 다양한 형태로 생산되어 유통되고 있는 현 상황에서도 이러한 역할은 여전히 중요한 것으로 인식되고 있다. 그러나 인터넷의 발달로 접근 가능한 정보의 양이 폭발적으로 늘어나면서 통제어휘 리스트로부터 주제를 부여하도록 훈련된 색인전문가에 의한 주제색인은 비용-효율적인 면에서 비현실적인 문제가 되고 있다(Chung, Pottenger, and Schatz 1998). 더구나 온라인 데이터베이스와 같이 대규모의 텍스트를 제한된 시간과 인력으로 처리해야 하는 환경에서는, 인간의 지적노력이 필수적으로 요구되는 이러한 색인작업을 종전과 같이 수작업에만 의존하여 수행하기는 거의 불가능할 것이다. 이러한 상황에서 지금까지 색인 전문가에 의해 전적으로 수행되어 온 이러한 역할을 일부 대체하거나 지원할 수 있는 효율적인 방법을 모색할 필요가 있다. 본 연구는 문헌정보학 분야 학술지 논문을 대상으로 통제 색인어인 디스크립터를 자동적으로 부여할 수 있는 가능성을 기계학습 접근법을 중심으로 검토하여 보는 것을 목적으로 한다.

본 연구의 제한점으로는 첫째, 사전에 색인전문가에 의해 부여되어 있는 디스크립터를 해당 문헌의 주제 범주로 가정하였기 때문에 성능 평가에서 색인작업에서의 일관성 문제를 다루지 않았다. 둘째, 색인 및 검색 단계에서 디스크립터를 거의 고려하지 않으면서 검색결과에 대한 범주화를 주된 목적으로 하는 웹 문서와는 달리, 학술지 논문 데이터베이스에서는 색인과 검색 단계에서 모두 디스크립터가 필수적인 요소로서 사용되고 있기 때문에 디스크립터 자동부여의 대상을 학술지 논문으로 제한하였다.

2 통제어휘 자동색인과 기계학습

문헌에 대한 통제 색인어의 자동부여는 “특정한 처리절차에 따라 시소러스 또는 주제명표와 같은 통제어휘집에서 색인어를 개개의 문헌에 자동적으로 부여하기 위한 것”(윤구호 1999)으로, 기존에 색인전문가에 의해 수작업으로 수행되어 온 통제어휘를 사용한 색인작업을 기계적인 방법으로 대체 또는 지원하는 것에 초점을 맞추고 있다. 또한, 최근 활발히 연구가 이루어지고 있는 텍스트 범주화(또는 문헌의 자동분류)는 사전 정의된 범주들로 문헌을 분류하는 것을 목적으로 하고 있는데(Joachims 1998), 여기서 사전 정의된 범주들을 문헌에 부여된 디스크립터(통제 색인어)로 본다면 거의 동일한 접근법을 공유할 수 있다. 따라서 시소러스, 주제명표의 용어 또는 분류체계의 엔트리로서 주제를 표현하는 통제 색인어 또는 분류기호의 부여는 텍스트 범주화로도 불린다(Moen 2002).

통제어휘 자동색인은 자동 메타데이터 생성(automated metadata generation)과도 긴밀히 관련되어있다. 전자도서관에서는 일반적으로 문서를 다양한 측면(예를 들면, 생성일자, 문서형태 또는 포맷, 이용가능성 등)으로 기술하는 메타데이터를 사용하는데, 주제를 표현하는 일부 메타데이터는 분류기호나 색인어를 사용하여 문서의 의미를 기술한다. 그러므로 이러한 종류의 메타데이터 생성은 통제어휘를 사용하여 문서 색인을 하는 것과 동일한 문제로 볼 수 있다(Sebastiani 2002).

통제어휘를 사용한 자동색인에 관한 연구는 1980년대의 지식공학적 접근법(Fuhr and Knorz 1984)에서 1990년대 이후 기계학습 기반 접근법(Lewis 1996; Yang 1997; Joachim 1998; Chung, Pottenger and Schatz 1998; Chang 2000; Ruiz and Srinivasan 2002)으로 중심이 바뀌어 다양한 응용과 시도가 이루어지고 있다. 기계학습을 통한 통제어휘 자동색인(디스크립터 자동부여)에서 기본적인 전제는 수작업으로 문헌에 부여된 통제 색인어(디스크립터) 정보를 프로그램을 통해 학습하여, 새로운 문헌에 적절한 통제 색인어를 부여할 수 있다는 것이다. 또한, 기계학습을 통한 디스크립터 자동부여를 위해서는 텍스트 범주화에서와 마찬가지로 세 가지 구성요소를 필요로 한다. 첫째, 색인전문가에 의해 디스크립터가 부여되어 학습과 검증에 사용되는 문헌집단, 둘째, 디스크립터 자동부여를 위한 단서를 제공하는 단어나 단어구(또는 개념)로 이루어진 자질집합, 셋째, 실제적으로 학습을 통해 디스크립터를 부여하는 기계학습 알고리즘(분류기)이 그것이다.

Tzeras and Hartmann(1993)은 사전 정의된 어휘들을 사용하여 문헌에 통제 색인어(디스크립터)를 부여하는데 베이지언 추론망(Baysian Inference Network) 모형을 적용하였다. 이 연구에서는 디스크립터 자동부여를 위한 구성요소로서 문헌에 수작업으로 할당

된 디스크립터, 문헌집합에 출현한 용어 리스트, 주제 분야를 디스크립터로 매핑하는 규칙이 포함된 색인어 사전 등을 사용한 결과를 보고하였다. 신경망에 기반한 디스크립터 자동부여를 제안한 연구로는 Chung, Pottenger, and Schatz(1998)가 있다. 개별 문헌에 대한 통제 색인어의 부여를 위해 문헌들로부터 추출된 개념들을 홉필드 망(Hopfield network)으로 표현하여 개념 공간(Concept Space: 일종의 시소러스)을 생성하였고, 개념부여기(concept assigner)를 통한 병렬 확장활성화(parallel spreading activation)를 수행하여 입력문헌과 가장 관련성이 높은 개념들을 출력하였다. 이들은 이러한 개념들이 문헌의 자동색인을 위한 후보 디스크립터로 사용될 수 있어, 색인전문가의 상호작용적인(또는 반자동) 색인작업에 사용하기에 적합하다고 주장하였다.

Plaunt and Barbara(1998)는 문헌표현에 포함된 어휘 정보(lexical clues)를 주제명표의 통제어휘로 매핑하는 어휘적 연어(lexical collocation) 기법에 기초한 2단계 알고리즘을 제안하였다. 이들은 INSPEC 문헌을 대상으로 우도비(likelihood ratio) 통계량에 기초하여 제목, 저자, 초록 필드의 어휘 요소(lexical items)와 이들의 관계를 표현하는 사전을 생성하였고, 이러한 사전을 입력문헌을 가장 잘 표현하는 주제명을 부여하기 위한 목적으로 사용하였다.

의학 분야 문헌(MEDLINE)에 MeSH의 주제명(디스크립터)을 자동부여하기 위한 목적으로 Ruiz and Srinivasan(1999)은 로치오 알고리즘과 역전파 신경망을 사용하였다. MeSH 디스크립터 119개를 선정하고 세 가지 방법(로치오 알고리즘, 신경망, 계층적 신경망)을 통하여 문헌에 부여한 결과, 계층적 신경망의 성능이 가장 좋은 결과를 보였다.

Humprey(1999)의 연구는 의학 분야 학술지(MEDLINE)에 부여된 디스크립터들을 학술지 논문에 부여하는 것으로, 텍스트의 단어와 학술지 색인어(디스크립터) 간의 통계학적 관계에 기반한 문헌의 자동색인 방법을 제안하였다. 학술지에 부여된 디스크립터는 연속간행물 전거 데이터베이스의 레코드로 유지되고 해당 주제분야의 하위 영역들을 대표하고 있어, 학술지 논문에 대한 일차적인 자동색인에 유용할 것이다. 또한, 생명정보학(Bioinformatics) 분야의 논문에 MeSH 디스크립터를 부여하기 위한 연구를 수행한 Chang(2000)은 나이브 베이즈(Naive Bayes) 알고리즘과 최근접 이웃(kNN: k Nearest Neighbor) 알고리즘[1]을 계층적, 비계층적 두 가지 방법으로 적용하였다. 그 결과, kNN보다는 나이브 베이즈가 더 좋은 성능을 보였고, 계층적 방법보다는 비계층적인 방법이 더 나은 성능을 보여주었다.

학술지가 아닌 잡지(『*Knack*』, 『*Weekend Knack*』, 『*Trends*』, 『*Cash!*』) 기사에 대한 디스크립터 자동부여를 다룬 (Moens 2000)의 연구는 학습 알고리즘으로 로치오, 베이지언 독립 분류기, 카이제곱(χ^2)의 세 가지 알고리즘을 사용하였다. 이 연구에서

---

1) 논문의 나머지 부분에서 나이브 베이즈와 최근접이웃을 NB와 kNN으로 약칭함.

는 특히 가장 효과적인 자질 선정 기준의 하나로 알려진 카이제곱(χ^2) 통계량을 디스크립터 자동부여를 위한 가중치 벡터의 생성에 적용하여 만족스러운 결과를 얻었다고 보고하였다.

Lauser and Hotho(2003)는 지지벡터기계(SVM: Support Vector Machine)[2]를 사용하여 복수의 디스크립터를 여러 언어(영어, 불어, 스페인어)로 작성된 문헌에 자동부여하는 접근법을 제안하였다. 시소러스 또는 온톨로지에 표현된 다중언어 배경 지식을 색인 목적의 텍스트 문헌 표현으로 통합하기 위한 가능성을 알아보기 위한 이 연구에서, 이들은 SVM이 여러 언어로 구성된 문헌에 대하여 복수의 디스크립터를 자동부여하기 위한 목적에 가장 부합하는 방법이라고 주장하였다. 최근에는 단순한 용어 리스트 형태의 사전을 이용하여 화학 분야 웹문서(web pages)를 분류하는 연구가 Liang 등(2006)에 의해 수행되었는데, 이들은 기존의 화학 분야 사전들을 통합하여 생성한 기계가독형 사전을 기반으로 kNN 알고리즘을 적용하였다. 여기서 잠재의미색인과 투표(voting) 방법을 적용한 접근법이 일반적인 kNN 알고리즘을 사용한 것보다 나은 성능을 보인 것으로 나타났다.

본 연구는 지금까지 살펴본 이전의 연구 결과를 바탕으로 보다 실용적인 측면에서 기계학습을 통한 디스크립터 자동부여의 가능성을 살펴보고자 하였다. 즉, 많은 온라인 데이터베이스에서 이미 통제어휘집을 사용하고 있거나 데이터베이스의 필수 요소로서 디스크립터 필드를 유지하고 있는 상황에서, 이를 기반으로 하는 자동색인은 보다 실제적인 대안이 될 수 있을 것이다. 따라서 기존의 통제어휘집이나 디스크립터 필드에 있는 통제 색인어 정보를 학습하여 새로운 문헌에 적절한 디스크립터를 부여하기 위한 목적으로 기계학습 접근법을 적용하였다. 특히, 통제어휘를 사용한 디스크립터 자동부여 측면에서 좋은 자질 선정 기준을 식별하고, 이전 몇 년간의 데이터로 학습하여 디스크립터를 부여하는 것이 가장 효과적일 수 있는가를 살펴보았다. 또한, 실험결과를 바탕으로 여러 분류기의 특성을 검토하여, 디스크립터 자동부여에 실제적으로 적용하기 위한 고려사항들을 도출하였다.

3 실험 설계

3.1 문헌집단

실험에 사용된 문헌집단은 Journal of Citation Reports(2001년~2004년)의 데이터에 기초하여 문헌정보학 분야의 핵심 학술지들을 선정하여 구성하였다. 이를 위해

---

2) 논문의 나머지 부분에서 지지벡터기계를 SVM으로 약칭함.

ISI Web of Knowledge 사이트(http://portal.isiknowledge.com)에서 Journal of Citation Reports(JCR Social Science edition)을 선택한 다음, 주제범주로 "Information Science & Library Science"를 선택한 결과로 출력된 상위의 학술지들을 우선적인 고려대상으로 삼았다. 그리고 학술지의 영향력을 Impact Factor(IF), Immedicacy Index(II), Cited Half-Life(CH)의 세 가지 요소를 중심으로 식별하기 위하여 각 요소를 조건으로 상위 50위까지 학술지 리스트들을 출력하였다. 이에 따라 2001년부터 2004년까지 4년간의 3개 요소별 순위에 기초하여 생성한 리스트에서 최상위 5위까지 2회 이상 출현한 학술지를 선정한 결과는 <표 1>과 같다.

〈표 1〉 JCR 데이터에 기초한 문헌정보학 분야 학술지 종합 순위('01년~'04년)

학술지명	2001	2002	2003	2004	평 균	순위
JASIST	1	2	2	2	1.75	1
ARIST	0	1	3	3	1.75	1
JD	4	3	4	4	3.75	2
IPM	5	4	1	5	3.75	2

이 중에서 ARIST를 그 성격이 비평지란 측면에서 제외하고 IPM(Information Processing & Management), JASIST(Journal of the American Society for Information Science & Technology), JD(Journal of Documentation)의 3개 학술지를 정보학 분야의 핵심 학술지로 선정하였다.

실제 데이터 수집을 위해서는 Cambridge Scientific Abstracts(CSA)사에서 제공하는 데이터베이스인 CSA Illumina(http://www.consortia.co.kr/csa)를 이용하였다. 이 서비스는 LISA, ERIC 등 사회과학 분야의 주요 데이터베이스들을 제공하고 있는데, 이 중에서 LISA는 앞에서 선정한 3개 학술지의 논문을 모두 검색할 수 있다. 따라서 LISA 데이터베이스를 대상으로 1994년부터 2004년까지 3개 학술지에 실린 논문들을 검색한 결과를 다운로드하여 문헌집단을 구성하였다. 여기서 3개 학술지의 전체 논문이 아닌 최근 11년간의 논문들을 대상으로 문헌집단을 구성한 이유는 다음과 같다.

첫째, 3개 학술지의 1990년대 이전 수록 논문은 실험을 위한 필수적인 데이터인 논문의 초록이 검색결과에 포함되어 있지 않은 경우가 많았다.

둘째, 최근의 문헌에 대한 자동 디스크립터 부여를 목적으로 문헌집단을 구성하는 것이므로, 너무 오래전의 논문들을 포함시켜 학습하는 것은 실효성에 의문이 있을 수 있다.

셋째, 본 연구가 진행되고 있는 2005년의 논문들은 문헌집단을 구성하는 시점까지 출판이 진행되고 있어 대상에서 제외하였다.

1994년부터 2004년까지 출판된 3개 학술지의 논문 중에서 필수 데이터 요소(제목, 초록, 디스크립터)가 없는 논문을 제외한 최종 문헌집단은 1,858개 문헌으로 구성되었는데, 이 중에서 JASIST에 실린 논문은 1,055개(56.8%), IPM에 수록된 논문은 507개(27.3%), JD에 포함된 논문은 296개(15.9%)였다. 3개 학술지의 구성 비율에서 차이가 나는 이유는 각 학술지의 연간 발행 건수와 수록 논문 수에서 차이가 있기 때문이다. 즉, JASIST는 1년에 14번 출판되며 각 호당 약 8편, IPM은 1년에 6번 출판되고 각 호당 약 9편, JD는 2000년 이전에는 연간 5번, 이후에는 연간 6번 출판되고 각 호당 5~6편의 논문을 수록하고 있다. 실험 문헌집단을 구성하고 있는 논문들의 주제 배경이 대부분 문헌정보학 전반이 아닌 정보학 분야가 주종을 이루고 있는 것도 여기에서 원인을 찾을 수 있다.

3.2 사전처리 및 소프트웨어

기계학습을 기반으로 하는 디스크립터 자동부여를 위한 세 가지 기본 요소는 (1) 문헌집단(collection), (2) 자질집합(feature set), (3) 분류기(classifier)이다. 여기서 디스크립터 자동부여를 위한 문헌집단은 크게 학습집합과 검증집합으로 구분된다. 본 연구에서 학습집합은 1994년부터 2003년까지 이전 10년 동안 발행된 학술지에 수록된 논문들로, 검증집합은 가장 최근인 2004년 한 해 동안 발행된 학술지에 수록된 논문들로 구성하였다.

LISA 데이터베이스의 모든 논문들에는 기본적인 서지사항과 키워드 필드 이외에 별도로 디스크립터 필드가 제공되고 있어 이용자들로 하여금 주제를 통한 검색을 가능하도록 하고 있다. 디스크립터 필드에 출현한 디스크립터들 가운데 학습집합에 속한 논문들에 한 번 이상 부여된 디스크립터는 총 1,565개이며, 6개 이상의 논문에 부여된 디스크립터는 218개이다. Ruiz & Srinivasan(1998)은 신경망의 적절한 학습을 위한 최소 긍정사례의 수가 20인 것으로 선행연구에서 보고하였는데, 본 연구에서는 학습집합 내 문헌빈도가 최소 15이상인 디스크립터 중에서 임의 선정한 33개의 디스크립터를 사용하였다. 또한, 문헌에 디스크립터를 부여하기 위한 단서가 되는 자질집합은 각 문헌의 제목과 초록을 대상으로 Porter Stemmer를 사용한 어간/어근 분리(stemming)와 전치사, 조사, 숫자 등의 불용어 제거 절차를 거친 다음, 여러 자질 선정 기준에 따라 구성하였다.

실험 문헌집단의 구체적 내용은 <표 2>와 같다.

〈표 2〉 실험 문헌집단

항 목	내 역
전체 문헌 수(학습/검증집합의 문헌 수)	1,858(1,666/192)
디스크립터 종수	33
디스크립터 당 학습문헌 수(최대/최소/평균)	409/15/68.3
학습문헌 당 디스크립터 수(최대/최소/평균)	12/1/4
학습집합의 색인어 종수	6,344
학습집합의 용어 종수(최대/최소/평균)	119/11/46.1
학습집합의 용어 수(최대/최소/평균)	184/18/69.1

실험집단에 대한 사전처리와 자질선정을 위한 프로그램은 Python 및 Visual FoxPro로 구현된 프로그램을 사용하였고, 디스크립터 자동부여 실험을 위한 프로그램(분류기)은 공개된 기계학습 실험 패키지인 WEKA Version 3.4(Witten & Frank 2005, http://www.cs.waikato.ac.nz/~ml/weka/)를 사용하였다. 패키지에 포함된 여러 분류기들 중에서 실제 실험에 사용된 분류기는 WEKA에서 제공하는 여러 분류기들 중에서 나이브 베이즈(NB: Naive Bayes; John and Langley 1995), 지지벡터기계(SVM: Support Vector Machine; Platt 1998), 투표형 퍼셉트론(VPT: Voted Perceptron; Freund and Schapire 1998)의 3개 분류기를 사용하였다. 이 중에서 NB와 SVM은 기계학습을 통한 텍스트 범주화 실험에서 가장 널리 사용되고 있는 분류기들로, 디스크립터 자동부여를 위한 목적으로의 적용 가능성을 알아보기 위해 선정하였다. 또한, VPT는 기본적인 신경망 모형 중 하나인 퍼셉트론의 결과를 다수결 투표(majority voting) 방식으로 출력하는 분류기로서, SVM과 거의 대등하거나 약간 떨어지는 성능을 보이면서도 계산상의 복잡도가 낮고 처리속도가 상대적으로 빠르다는 장점을 가지고 있어 실험을 위한 분류기로 선정하였다.

3.3 실험 구성

디스크립터 자동부여를 위한 실험은 각 디스크립터별로 전체 검증집합에 대한 해당 디스크립터의 부여 여부를 결정하는 이원 판정 방법으로 설계하였다. 따라서 분류기의 학습 및 검증을 위해 문헌집단의 문헌들은 기본적으로 용어들로 구성된 자질벡터로 표현되고, 특정 디스크립터의 부여 여부에 대한 정보가 추가되었다. 여기서 각 자질에 대응하는 값은 해당 용어의 출현빈도(tf)이다. 자질 벡터의 구성에서 가중치 부여공식(log tf, tf·idf 등)을 사용하지 않은 것은 최근 연구에서 단순하게 용어의 출현빈도를 사용하는 것이 다른 복잡한 가중치를 사용하는 것과 유사하거나 더 나은 결과를 산출하는 것으로 보고되었기 때문이다(Lan 2005). 따라서 문헌집단을 구성하는 각 문헌은 용어 가중치(tf)로 구성되는 문헌벡터와 이원 판정의 쌍으로 표현된다.

$$\vec{d} = (t_1\, t_2\, t_3 \cdots\, t_m),\ binary\ decision = (Yes \text{ or } No)$$
$$document:\ \vec{d},\ binary\ decision$$

문헌 자동분류에서와 마찬가지로 디스크립터 자동부여에서도 자질 선정과 학습집합의 크기(긍정/부정 사례의 수)는 분류기의 성능에 큰 영향을 주는 요소라고 할 수 있다. 따라서 디스크립터 자동부여에서 분류기의 학습 이전의 요소로서 자질 선정기준과 학습 상황에서의 요소로서 학습집합의 크기에 따른 여러 분류기의 성능을 비교하였다.

3.3.1 자질 선정 기준에 따른 성능 비교 실험

먼저, 디스크립터 자동부여를 위한 기본적인 구성요소 중 하나인 자질 선정을 위한 실험을 수행하였다. 자질선정은 크게 전체 자질집합 중에서 유용한 하부집합을 선정(feature subset selection)하거나 새로운 자질을 구축(feature construction)하는 방법이 있다(Joachims 2001). 본 연구에서는 자질 선정 기준에 따른 성능의 비교를 위하여 전자의 방법을 적용하였다.

선행 연구에서는 가장 효과적인 자질 선정 기준으로 카이제곱(χ^2) 통계량, 정보획득량(Information Gain)이 가장 좋고, 처리속도 또는 계산량 측면에서 문헌빈도를 이용한 방법도 추천할 만한 것으로 보고되었다(Yang and Pedersen 1997). 또한, Ruiz and Srinivasan(2002)의 연구에서는 상관계수(Correlation Coefficient), 상호정보량(Mutual Information), 승산비(Odds Ratio) 등 세 가지 자질 선정 기준이 비슷한 성능을 보인다고 하였다. 그리고 앞에서 나온 여러 자질 선정 기준들을 조합한 결과를 평가한 최근의 연구에서는 카이제곱(χ^2) 통계량을 포함하는 조합들이 대체로 좋은 성능을 보이는 것으로 나타났다(Rogati and Yang 2003). 따라서 본 연구에서는 선행 연구에서 좋은 성능을 보인 자질 선정 기준으로 지금까지 많이 사용되어 온 카이제곱(χ^2) 통계량, 문헌빈도(DF: Document Frequency), 로그승산비(LOR: Log Odds Ratio) 이외에 고빈도어 선호 기준이라고 할 수 있는 코사인 계수, 자카드 계수, GSS 계수를 추가하여 자질 선정을 해보았다(이재윤 2005).

<표 3>은 자질 선정 기준에 적용하기 위하여 특정 문헌에 대한 디스크립터 부여와 용어 출현 정보에 따라 작성한 2×2 분할표와 각 셀의 값이다.

〈표 3〉 자질 선정 기준별 수식 계산을 위한 2×2 분할표

	용어 출현(Yes)	용어 출현(No)
디스크립터	a	b

부여(Yes)		
디스크립터 부여(No)	c	d

- a: 동시출현빈도(디스크립터 & 키워드)
- d: N − a − b −c
- a+b: 키워드 출현 빈도
- a+c: 디스크립터 출현(부여) 빈도
- N: 전체 문헌 수

자질 선정 기준의 수식과 위의 분할표에 따라 각 용어 자질의 값을 계산하기 위한 분할표 표현은 <표 4>와 같다. 여기서 주의할 것은 로그승산비(LOR)의 경우 수식의 특성상 각 셀의 값이 0이 되면 계산이 불가능해지므로 이를 전체 문헌 수인 N으로 나눈 값(1/N)으로 대체하여 계산하였다.

〈표 4〉 자질선정 기준별 수식과 분할표 표현

명칭	약호	공식	분할표 표현
문헌빈도	DF	$P(t_k)$	$a+b$
코사인 계수	COS	$\frac{P(t_k,c_i)}{\sqrt{P(t_k)P(c_i)}}$	$\frac{a}{\sqrt{(a+b)(a+c)}}$
자카드 계수	JAC	$\frac{P(t_k,c_i)}{P(t_k)+P(c_i)-P(t_k,c_i)}$	$\frac{a}{a+b+c}$
GSS 계수	GSS	$P(t_k,c_i)P(\overline{t_k},\overline{c_i})-P(t_k,\overline{c_i})P(\overline{t_k},c_i)$	$\frac{ad-bc}{N}$
카이제곱 통계량	CHI	$\frac{N(P(t_k,c_i)P(\overline{t_k},\overline{c_i})-P(t_k,\overline{c_i})P(\overline{t_k},c_i))^2}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})}$	$\frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$
로그 승산비	LOR	$\log\frac{P(t_k,c_i)P(\overline{t_k},\overline{c_i})}{P(t_k,\overline{c_i})P(\overline{t_k},c_i)}$	$\log(\frac{ad}{bc})$

3.3.2 학습집합의 크기에 따른 성능 비교 실험

디스크립터 자동부여에서 학습 데이터의 증가에 따른 3개 분류기의 성능을 비교해 보았다. 이를 위해 학습집합의 논문에 15회 이상 부여된 디스크립터들 중에서 임의 선정한 33개 디스크립터를 사용하였고, 이전 실험에서 좋은 성능을 보여 주었던 자질 선정 기준 중의 하나로 자카드 계수(JAC)를 사용하여 자질집합을 구성하였다. 실험 문헌집단은 모두

1,858개 문헌으로 구성되어 있는데, 이 중 학습집합은 1993년부터 2003년까지 10년 간 발행된 3개 학술지의 논문(1,669)으로 구성되어 있고, 검증집합은 2004년 한 해 동안 발행된 3개 학술지의 논문(192)으로 구성되어 있다. 먼저, 학습집합의 크기에 따른 성능 비교 실험을 위해 <표 5>와 같이 학습집합을 년도 별로 10개의 하위집합으로 구분하였다. 다음으로 가장 최근의 2003년부터 1년씩 추가된 학습집합을 사용하여 각 분류기별로 학습한 다음, 검증집합의 논문들에 디스크립터를 부여한 결과를 비교하였다. 즉, 이를 통해 이전 몇 년간의 논문을 이용하는 것이 새로운 문헌에 대한 디스크립터 자동부여에 가장 좋은 가를 알아보고자 하는 것이다.

〈표 5〉 학습/검증집합의 년도 및 문헌 수

약호	학습집합(년) vs. 검증집합(년)	문헌 수
1t1	2003년(1년) vs. 2004년(1년)	162/192
2t1	2002년~2003년(2년) vs. 2004년(1년)	334/192
3t1	2001년~2003년(3년) vs. 2004년(1년)	512/192
4t1	2000년~2003년(4년) vs. 2004년(1년)	672/192
5t1	1999년~2003년(5년) vs. 2004년(1년)	870/192
6t1	1998년~2003년(6년) vs. 2004년(1년)	1,045/192
7t1	1997년~2003년(7년) vs. 2004년(1년)	1,224/192
8t1	1996년~2003년(8년) vs. 2004년(1년)	1,375/192
9t1	1995년~2003년(9년) vs. 2004년(1년)	1,519/192
10t1	1994년~2003년(10년) vs. 2004년(1년)	1,666/192

3.3.3 성능 평가 척도

문헌의 자동분류 또는 텍스트 범주화에서 주로 사용하는 성능 평가 척도로서 정보검색의 성능 평가에서 사용되는 재현율, 정확률이 있다. 분류기의 출력이 범주들의 순위 리스트 또는 범주 할당에 대한 이원 판정인가에 따라 다른 성능 평가 척도를 사용할 수 있지만, 본 연구에서는 재현율과 정확률을 기본적인 성능 척도로 사용하였다.

정보검색에서의 재현율과 정확률을 디스크립터 자동부여에 사용된 분류기의 성능 평가를 위해 적용하면 다음과 같이 표현할 수 있다(정영미 2005, 162).

$$\text{분류 재현율} = \frac{\text{부여된 적합범주 수}}{\text{전체 적합범주 수}}$$

$$\text{분류 정확율} = \frac{\text{부여된 적합범주 수}}{\text{부여된 총 범주 수}}$$

본 연구에서 사용하는 분류기들의 출력이 문헌에 대한 디스크립터 부여의 이원 판정이므로 분류 재현율과 분류 정확률은 <표 6>과 같이 각 디스크립터에 대한 2×2 분할표를 이용하여 계산하는 것이 편리하다.

〈표 6〉 분류기 성능 평가를 위한 2×2 분할표

	옳은 디스크립터	틀린 디스크립터
디스크립터 부여(Yes)	a	b
디스크립터 부여(No)	c	d

- a: 옳은 디스크립터를 제대로 부여한 문헌 수
- b: 틀린 디스크립터를 잘못 부여한 문헌 수
- c: 옳은 디스크립터를 부여하지 않은 문헌 수
- d: 틀린 디스크립터를 부여하지 않은 문헌 수

<표 6>의 분할표를 이용하여 분류 재현율과 분류 정확률을 구하는 공식은 다음과 같다.

$$Recall = \frac{a}{a+c}, \qquad \Pr ecision = \frac{a}{a+b}$$

이러한 분류 재현율과 분류 정확률은 서로 상충되는 부분이 있고, 양자 간의 trade-off를 찾기가 어려워 분류기의 성능을 한눈에 판단하기가 어렵다. 이에 따라 이원분류기의 성능 평가를 위한 여러 단일 척도가 제안되어 사용되고 있는데, 대표적으로 F_1, BEP(Break-Even Point), Accuracy[3]를 들 수 있다. 이 중 BEP는 재현율과 정확률이 같아지는 지점($r=p$)의 값을 말하는데 실제 값을 계산하기가 쉽지 않고, 보간법을 사용하는 경우 가장 근접한 재현율과 정확률 값이 너무 멀리 떨어져 있으면 분류기의 실제 성능을 반영할 수 없다는 문제가 있다(Yang 1997). 또한, F_1척도와의 관계에서 BEP 점수는 F_1값과 유사하거나 더 작은 값을 취한다(Ruiz and Srinivasan 1999).

Accuracy는 디스크립터처럼 대부분이 저빈도에 속하는 범주를 대상으로 성능을 평가할 때는 전혀 분류를 하지 않은(디스크립터를 부여하지 않은) 경우에 오히려 성능이 상당히 높은 것으로 나타나는 문제가 있다. 즉, 검증집합의 문헌에 특정 디스크립터를

3) Accuracy는 〈표 6〉의 분할표 상에서 다음과 같은 공식에 의해 계산할 수 있다.

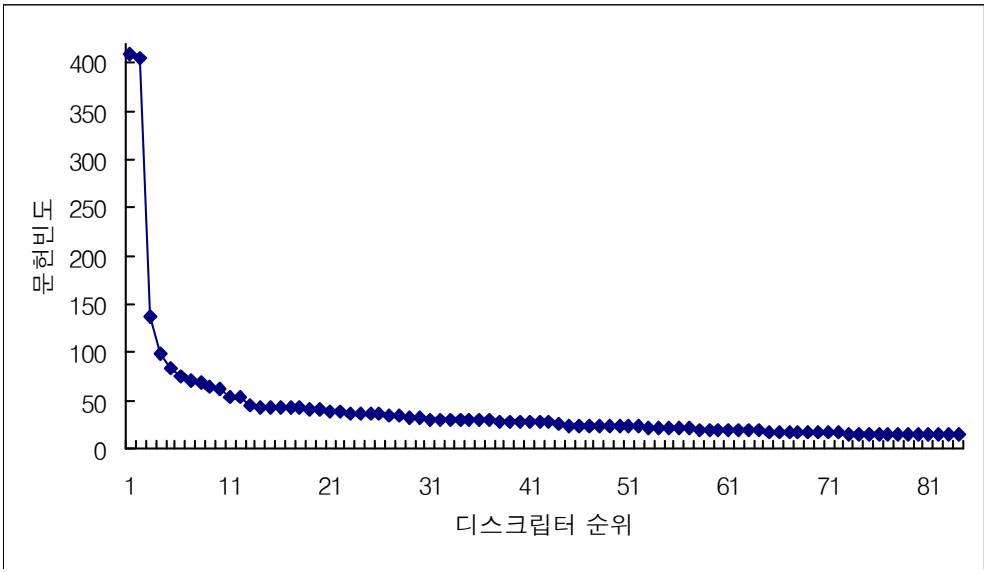
$$accuracy = \frac{a+d}{a+b+c+d}$$

전혀 부여하지 않는 분류기를 'trivial rejector'(Yang 1999)라고 하는데, Accuracy는 이러한 분류기의 성능을 아주 높게 산출한다. 따라서 본 논문에서는 분류기의 성능 평가를 위한 기본 척도로서 분류 재현율과 분류 정확률의 효과를 동등하게 반영할 수 있는 F_1 척도(Lewis 1996)를 사용하였다.

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} = \frac{(\beta^2+1)a}{(\beta^2+1)a + b + \beta^2 c}$$

$$F_1 = \frac{2PR}{P+R} = \frac{2a}{(2a+b+c)}$$

각 디스크립터에 대한 이원 분류의 성능을 계산한 후에는 분류기의 전체 성능을 알기 위해 평균을 계산하여야 한다. 평균을 계산하는 방법은 디스크립터를 중심으로 계산하는 매크로 평균과 문헌을 중심으로 계산하는 마이크로 평균이 있다. 이 중 매크로 평균은 저빈도 범주의 성능에 크게 영향을 받으므로(Yang and Liu 1999;), 저빈도인 경우가 대부분인 디스크립터 부여의 결과에 대한 분류기의 성능 평가에는 마이크로 평균을 사용하는 것이 적절할 것이다. 학습집합의 문헌빈도를 기준으로 상위의 디스크립터를 나타낸 <그림 1>에서 보는 바와 같이, 학술지 논문에 부여된 디스크립터는 상대적으로 고빈도인 극히 일부를 제외하고는 대부분이 저빈도에 속한다. 따라서 본 연구에서는 전체적인 분류기의 성능 평가를 위한 기본적인 성능 척도로 마이크로 평균 F_1을 사용하였다.



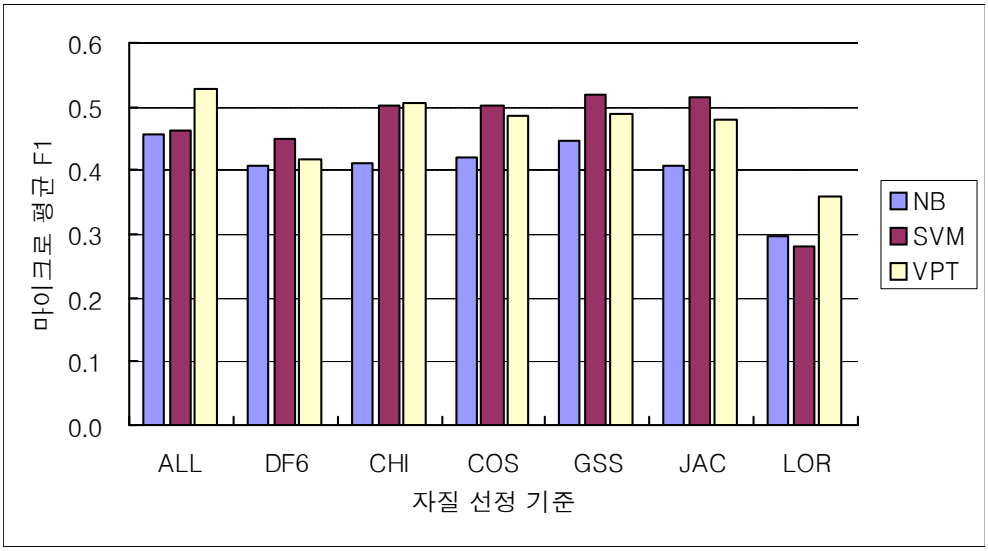
〈그림 1〉 상위 84개 디스크립터 빈도 분포(DF ≥15)

4 실험 결과 및 분석

4.1 자질 선정 기준에 따른 결과 및 분석

디스크립터 자동부여의 주요 구성요소 중 하나인 자질 선정 측면에서 3개 분류기(NB, SVM, VPT)의 성능을 알아보았다. 사용된 자질 선정 기준은 문헌번도(DF), 코사인계수(COS), 자카드계수(JAC), GSS계수(GSS), 카이제곱 통계량(CHI), 로그승산비(LOR)로서 로그승산비와 카이제곱 통계량을 제외하고는 모두 고빈도어를 선호하는 기준들이다.

<그림 2>는 1994년부터 2004년까지 LISA의 논문들에 부여된 디스크립터들을 대상으로 문헌빈도 순으로 순위화한 상위의 디스크립터(DF≥15) 중에서 임의로 선정한 11개 디스크립터에 대하여, 전체 자질집합을 사용한 경우의 성능 및 6개 자질 선정 기준에 따라 자질집합을 축소한 경우의 성능을 각각 마이크로 평균 F_1으로 산출한 결과이다. 모든 자질을 이용한 경우는 전체 6,344개 자질을 모두 사용하였고, 문헌빈도(DF≥6)를 비롯한 나머지 자질 선정 기준에서는 전체 자질의 약 25%에 해당하는 1,611개 자질을 사용하였다.



〈그림 2〉 자질 선정 기준에 따른 분류기별 성능

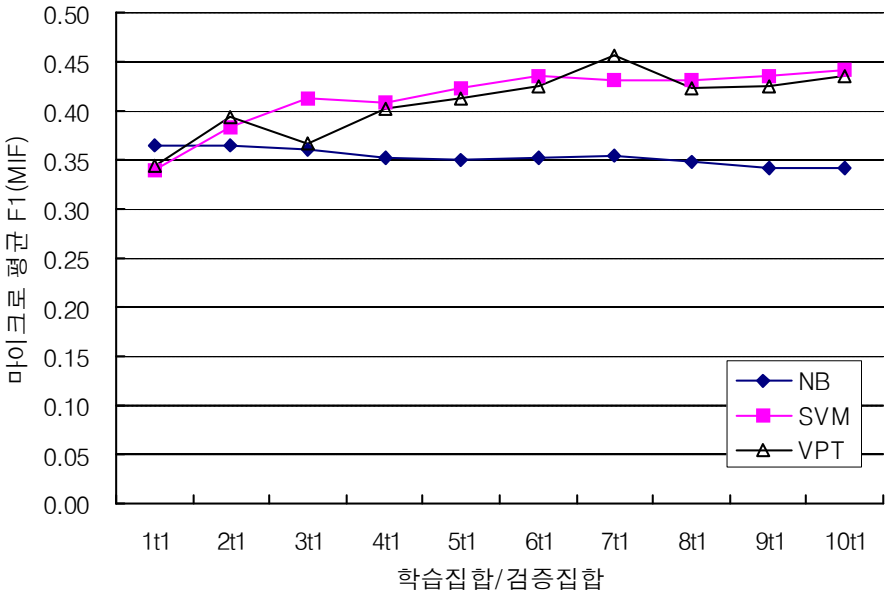
<그림 2>의 결과에서 발견된 사실로는 첫째, 선행 연구에서 좋은 결과를 보였던 기준들(CHI, COS, GSS, JAC)이 대체로 좋은 성능을 보여주었다(Yang and Pedersen. 1997; Joachims 1997; Sebastiani 2002; Rogati and Yang 2003; 이재윤 2005). 이 중에서 카이제곱 통계량(CHI)을 제외한 나머지 기준들은 모두 고빈도 선호 기준(COS, GSS, JAC)이다. 둘째, 문헌빈도(DF≥6)를 제외한 고빈도 선호 기준들(COS, GSS, JAC)은 전체 자질(ALL)을 사용하였을 때와 비교하여도 거의 유사한 성능을 나타내었으며, SVM으로 학습한 경우에는 오히려 더 높은 성능을 보여주었다. 셋째, 모든 자질 선정 기준 가운데 저빈도 선호 기준이라고 할 수 있는 로그승산비(LOR)가 가장 낮은 성능을 보였다.

4.2 학습집합의 크기에 따른 결과 및 분석

학습집합을 2003년(1년)부터 1994년(10년)까지 1년씩 년도 별로 추가한 10개의 하위집합으로 구분하고 각각 분류기를 학습시킨 다음, 검증집합의 논문들에 대하여 디스크립터를 부여하도록 하였다. 즉, 10개 학습집합으로 학습된 각 분류기들에 대하여, 검증집합을 구성하고 있는 2004년(1년)의 논문들에 대한 특정 디스크립터의 부여 여부를 이원 판정(yes, no)으로 결정하도록 하였다.

<그림 3>은 디스크립터 자동부여에서 학습집합의 크기가 증가함에 따라 3개 분류기

의 성능이 어떻게 변화하는 가를 마이크로 평균 F_1 척도로 나타낸 것이다.



〈그림 3〉 학습집합의 크기에 따른 3개 분류기의 성능(마이크로 평균 F_1)

학습 데이터의 증가에 따른 3개 분류기의 성능 변화를 마이크로 평균 F_1 척도로 살펴본 〈그림 3〉에서 발견된 사실은 다음과 같다.

첫째, 전반적으로 SVM과 VPT가 NB보다 좋은 성능을 나타냈다. 특히, 최근 1년(2003년)의 학습집합에서 지난 10년 간(2003년~1994년)의 학습집합으로 학습 데이터가 증가함에 따라 앞의 두 분류기와 NB 간의 성능 차이가 점점 더 커지는 현상(최대 0.10~0.11)을 보였다.

둘째, SVM과 VPT는 학습 데이터의 증가에 따른 성능 향상이 뚜렷하게 나타났으며(SVM: 최대 0.10, VPT: 최대 0.12), 특히, SVM은 보다 안정된 경향의 성능 향상을 보여주는 동시에 최근 10년간의 학습 데이터 모두를 사용하였을 때 가장 좋은 결과를 보여주었다. 따라서 SVM을 사용한 디스크립터 자동부여에서 이전 10년까지는 모든 데이터를 학습집합으로 사용하는 것이 가장 좋은 결과를 가져온다고 볼 수 있다. 그러나

최근 10년 이전의 데이터 전체를 사용하는 경우에 더 좋은 결과를 산출할 것인가에 대해서는 더 많은 실험 및 연구가 필요할 것이다. 예를 들면, JD의 경우 1960년대에 창간되었으므로 전체 데이터는 본 실험에서의 분류 시점인 2004년을 기준으로 한다면 약 40년간의 데이터를 말한다. 2004년에 발표된 논문들에 디스크립터를 부여하는 상황에서, 이전 40년 동안의 논문 전체를 학습데이터로 사용한 결과가 좋을 것인가는 의문의 여지가 있다.

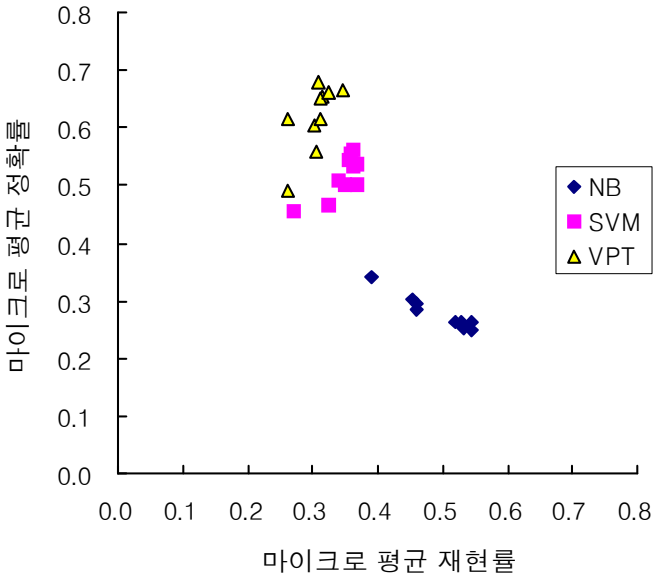
셋째, SVM과 VPT와는 대조적으로 NB는 학습 데이터의 증가에 따른 성능 향상이 없고 오히려 미세하게 낮아지는 현상을 보였다. 그러나 이러한 낮은 수치(-0.02)로는 분류 성능이 감소한 것으로 보기 보다는 학습 데이터의 증가에 따른 영향을 NB가 거의 받지 않는 것으로 보는 것이 타당할 것으로 생각된다. 주목할 점은 NB가 최소한의 학습 데이터(최근 1년: 0.36)로도 NB의 전체 성능 측면에서 가장 좋은 성능을 보여주었다는 사실이다. 반면, SVM이나 VPT는 최소한 최근 2년 이상의 학습집합을 사용하는 모든 경우에 NB와 동등하거나 훨씬 더 좋은 성능을 보여주었다.

다른 측면에서, 학습집합의 크기에 따른 3개 분류기의 성능을 개별 디스크립터를 중심으로 살펴보았다. 각 분류기에 대하여 학습집합 크기별로 성능이 가장 높은 것은 SVM과 VPT의 경우 지난 10년을 모두 학습집합으로 사용한 것(10t1)이었고, NB의 경우에는 가장 최근의 1년을 학습집합으로 사용한 것(1t1)이었다. 따라서 이들 세 가지 경우에 전체 분류기의 성능에 가장 큰 영향을 준 디스크립터를 찾아보았다. 그 결과 디스크립터별 정확률, 재현율, F_1값이 모두 0으로 산출되어 3개 분류기의 성능에 공통적으로 크게 부정적인 영향을 준 디스크립터들은 모두 6개('digital libraries', 'online databases', 'library and information science', 'information science', 'medicine', 'survey')였다. 이 중에서 너무 광범위한 주제('library and information science', 'information science', 'medicine')이거나 주제로서 부적합한 디스크립터('survey')는 텍스트 범주화에서 '잘못 분류된 범주(mislabeled categories)'의 경우와 같이 '잘못 부여된 디스크립터(misassigned descriptors)'로 간주할 수 있다(Zhang and Yang 2003). 그러나 나머지 두 개의 디스크립터('digital libraries', 'online databases')의 경우에는 해당 디스크립터로 제대로 부여한 경우가 전혀 없는 원인을 기존 주제의 변화로 인하여 해당 디스크립터가 그 주제를 표현할 수 없게 되는 등 다른 측면에서 찾아야 할 것이다(김판준 2005).

4.3 분류기별 특성에 따른 디스크립터 자동부여

학술지 논문에 디스크립터를 자동부여하는 목적으로 기계학습 기반 분류기를 적용하

는 경우를 전제로 본 연구에서 사용한 3개 분류기의 특성을 살펴보았다. <그림 4>는 학습 데이터의 증가에 따른 3개 분류기의 성능을 정확률과 재현율의 두 가지 측면에서 살펴보기 위하여 x축을 마이크로 평균 재현율, y축을 마이크로 평균 정확률로 나타낸 것이다.



〈그림 4〉 학습집합의 크기에 따른 3개 분류기의 재현율과 정확률

<그림 4>에서 알 수 있는 사실은 다음과 같다. 첫째, SVM과 VPT는 정확률에서 강한 경향을 보이지만, NB는 재현율에서 더 나은 경향을 나타냈다. 둘째, VPT는 정확률에서 가장 높은 경향을 보이지만 상대적으로 재현율이 크게 떨어지는 문제가 있다. 대조적으로 NB는 재현율에서는 가장 좋은 성능을 보이지만 정확률에서 SVM이나 VPT의 최소 성능보다도 훨씬 낮은 문제가 있다. 셋째, 안정된 성능을 시사하는 정확률과 재현율의 trade-off 측면에서는 일관성 있게 정확률과 재현율이 일정 수준 이상으로 균형을 이루고 있는 SVM이 가장 좋은 경향을 보여주었다.

이를 보다 세부적으로 규명하기 위해 학습집합의 크기에 따른 3개 분류기의 성능을 각각 마이크로 평균 재현률과 마이크로 평균 정확률로 나타낸 결과를 각각 <그림 5>

와 <그림 6>으로 제시하였다. 여기서 각 분류기의 특성을 보다 세부적으로 파악할 수 있다. 먼저, SVM은 재현율과 정확률 모두에서 일정 수준 이상의 좋은 성능을 나타내면서, 일관성 있게 성능이 향상되는 경향을 보이고 있다. VPT는 정확률 면에서 가장 강하지만 재현율은 상대적으로 약한 경향을 보이고 있다. 이에 반하여, NB는 전반적으로 재현율면에서 가장 높지만 재현율의 향상과 함께 정확률이 크게 떨어지는 경향을 보이고 있다. 이러한 사실들을 종합하면 분류기의 특성에 따른 디스크립터 자동부여는 다음과 같이 고려해 볼 수 있다.

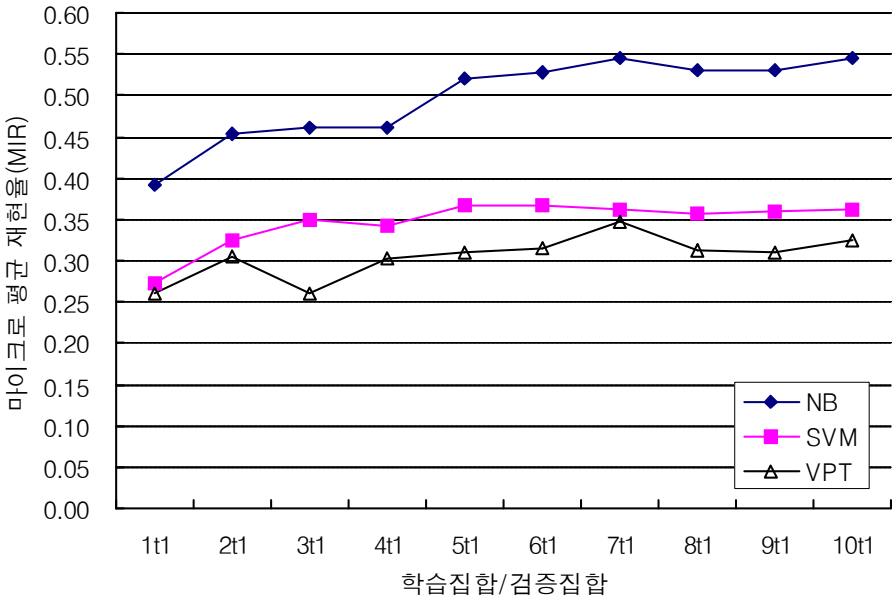
첫째, 최소 2년~10년의 학습데이터가 있는 경우에는 SVM과 VPT를 사용하는 것이 좋을 것이다. 또한, 디스크립터 부여에서 다소 엄격한[4] 경향을 보이는 SVM과 VPT는 색인작업이 색인전문가의 개입이 없이 완전 자동으로 이루어질 경우에 더 적합할 것이다(Lauser and Hotho 2003). 또한, 부여한 디스크립터의 정확성이 크게 강조되는 경우에는 VPT가, 보다 안정적인 성능을 위해서는 SVM이 더 선호될 수 있다.

둘째, 학습데이터가 거의 없는 상황(1년 이하)에서는 학습집합의 크기에 영향을 적게 받는 NB를 우선적으로 고려해 볼 수 있다. 또한, 디스크립터 부여 판정에서 다소 너그러운[5] 경향을 보이는 NB는 상대적으로 많은 수의 후보 디스크립터를 제공하여 색인전문가의 색인작업을 지원할 수 있는 상호작용적인 시스템에 적합하다고 할 수 있다(Sebastiani 2002; Yang 1997).

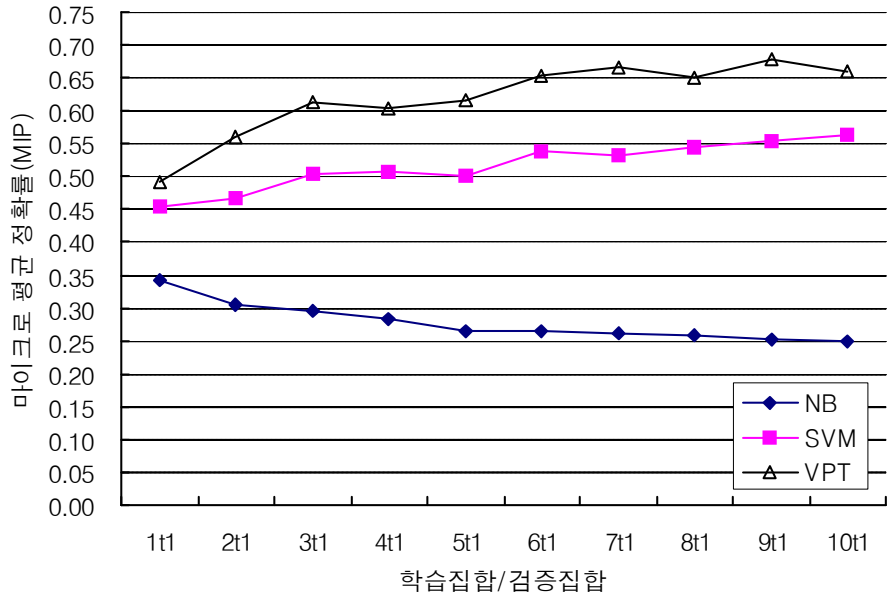
---

4) '엄격한'의 의미는 SVM과 VPT가 특정 디스크립터의 부여 여부를 결정할 때 상대적으로 소극적이라는 것이다. 즉, 학습 데이터의 증가와 함께 해당 디스크립터를 부여하는 문헌의 수도 늘어나는데, 그 수가 상대적으로 작은 대신 비교적 정확한 판정을 내리는 경향을 보인다.

5) '너그러운'의 의미는 NB가 특정 디스크립터의 부여 여부를 결정할 때 상대적으로 적극적이라는 것이다. 즉, 학습집합의 크기가 커질수록 SVM이나 VPT에 비해 훨씬 더 많은 수의 문헌에 디스크립터를 부여하지만, 이러한 판정이 틀린 경우가 큰 폭으로 증가하는 경향을 보인다.



〈그림 5〉 학습집합의 크기에 따른 3개 분류기의 마이크로 평균 재현율



〈그림 6〉 학습집합의 크기에 따른 3개 분류기의 마이크로 평균 정확률

5 결론

색인은 사용되는 용어의 통제 여부에 따라 자연언어색인과 통제언어색인으로 구분할 수 있지만, 지금까지의 자동색인은 대개 문헌에 출현한 단어를 그대로 추출하는 전자를 주된 관심의 대상으로 하고 있다. 학술정보서비스를 제공하는 온라인 데이터베이스는 자동색인을 통한 자연언어색인과 색인전문가에 의한 통제언어색인을 함께 제공하고 있는데, 이러한 두 가지 유형의 색인은 서로 다른 측면에서 문헌의 내용을 표현하는 필수적인 요소로서 상호보완적으로 사용된다. 그러나 색인전문가가 처리해야 하는 문헌의 양이 크게 증가함에 따라 종전과 같이 전적으로 수작업에 의존하는 통제언어색인은 현실적으로 불가능한 문제가 되고 있어, 이를 대상으로 하는 자동색인의 필요성이 점차 커지고 있다. 특히, 온라인 데이터베이스는 통제언어색인을 위해 기존의 통제어휘집을 사용하거나 데이터베이스의 필수적인 데이터 요소로서 디스크립터를 부여하여 유지하고 있으므로, 이를 기반으로 통제어휘 자동색인을 모색하는 것은 보다 실용적인 대안이 될 수 있을 것이다.

본 연구는 통제언어색인에서 색인전문가의 색인작업을 일부 대체 또는 지원할 수 있는 방안으로 기계학습 접근법의 적용 가능성을 실험을 통해 살펴보았다. 즉, 학술지 논문을 대상으로 디스크립터를 자동부여하기 위한 목적으로 정보학 분야의 핵심 학술지를 선정하여 1994년부터 2004년까지 11년간의 수록 논문을 대상으로 문헌집단을 구성하였고, 이전 10년 동안의 논문들을 학습집합(1,666/1994년~2003년), 최근 1년의 문헌들을 검증집합(192/2004년)으로 구분하여 여러 분류기의 성능을 비교하는 실험을 수행하였다.

실험에서는 분류기의 성능에 가장 큰 영향을 주는 요인들로 자질 선정과 학습집합의 크기라는 두 가지 측면에서 3개 분류기(NB, SVM, VPT)의 성능을 비교한 다음, 실험에 사용된 3개 분류기의 특성에 따른 실제 디스크립터 자동부여에서의 고려사항을 살펴보았다.

먼저, 실험을 통해 발견된 사실은 다음과 같다. 첫째, 학습 데이터를 구성하기 위한 자질 선정 측면에서는 많은 선행연구들과 마찬가지로 카이제곱 통계량(CHI) 기준이나 고빈도 선호 자질 선정 기준(COS, GSS, JAC)을 적용하여 자질집합을 축소한 다음, SVM으로 학습한 결과가 가장 좋은 성능을 보였다. 둘째, 이전 몇 년간의 문헌으로 학습한 경우에 가장 효과적인가를 알아보기 위한 학습집합의 크기 측면에서 SVM과 VPT는 학습에 사용된 데이터의 양이 증가할수록 지속적으로 성능이 향상되는 경향을 보였다. 반면, SVM과 VPT 보다 다소 낮은 성능의 NB는 학습집합의 크기에 따른 영향을 거의 받지 않는 것으로 나타났다. 전반적으로는 자질 선정과 학습집합의 크기 양 측면에서 3개 분류기 중 SVM이 가장 안정적이면서도 높은 성능 수준을 유지하였으며, 학습 데이터의 증가에 따라 일관성 있게 성능이 향상되는 현상을 보여주었다.

다음으로, 디스크립터 자동부여에의 실제 적용을 위한 차원에서 실험을 통해 밝혀진 각 분류기의 특성에 따른 고려사항은 다음과 같다. 첫째, 데이터베이스 환경에서 학습 데이터

가 어느 정도 확보되어 있는 경우(2년 이상)에는 SVM이나 VPT의 사용을 우선적으로 고려할 수 있는데, 일반적인 경우에는 SVM이 가장 안정된 성능을 보장할 수 있을 것이다. 반면, 디스크립터 부여의 정확성이 크게 강조될 경우에는 대안적으로 VPT를 고려해 볼 수 있다. 특히, 디스크립터 부여작업이 완전 자동으로 이루어지는 상황에서 SVM과 VPT는 틀린 디스크립터 부여를 최소한으로 줄이는 효과를 기대할 수 있다. 둘째, 이용 가능한 학습 데이터가 거의 없는 경우(1년 이하)에는 NB의 사용을 우선적으로 고려할 수 있다. 또한, 색인전문가의 반자동 색인작업을 지원하기 위한 후보 디스크립터 리스트의 생성에서 NB는 상대적으로 더 많은 대안적인 디스크립터를 제공할 수 있을 것이다.

참 고 문 헌

김판준. 2005. 『새로운 주제 탐지를 통한 지식 구조 갱신에 관한 연구』. 박사학위논문, 연세대학교 대학원, 문헌정보학과.

윤구호. 1999. 『색인·초록』. 서울: 도서관협회.

이재윤. 2005. 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 관한 연구. 『문헌정보학회지』, 39(2): 123-146.

이재윤. 2005. 문헌간 유사도를 이용한 SVM 분류기의 문헌분류성능 향상에 관한 연구. 『정보관리학회지』, 22(3): 261-287.

정영미. 2005. 『정보검색연구』. 서울: 구미무역(주) 출판부.

Borko, H. and Myma Bernick. 1963. "Automatic Document Classification." *JACM,* 10(2): 151-162.

Chang, Jeffrey. 2000. *Using the MeSH Hierarchy to Index Bioinformatics Articles*. CS224N/Ling237 Final Projects 2000, Stanford University.

Chung, Y., W. M. Pottenger, and B. R. Schatz. 1998. "Automatic subject indexing using an associative neural network." *Proceedings of the 3rd ACM international Conference on Digital Libraries(DL '98)*, ACM Press, 59-68.

Freund, Yoav and Robert E. Schapire. 1998. "Large Margin Classification Using the Perceptron Algorithms." *Proceedings of the 11th Annual Conference on Computer Learning Theory*, ACM Press, 209-217.

Humprey, Susanne M. 1999. "Automatic indexing of Documents from Journal Descriptors: A Preliminary Investigation." *JASIS*, 50(8): 661-674.

Joachims, Thorsten. 1998. "Text Categorization with Support Vector Machines:

Learning with Many Relevant Features." *Proceedings of the 10th European Conference on Machine Learning*, 137-142.

Joachims, Thorsten. 2001. *Learning to Classify Text Using Support Vector Machines.* Boston: Kluwer Academic Publishers.

John, George H. and Pat Langley. 1995. "Estimating Continuous Distributions in Bayesian Classifiers." *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, San Mateo, 338-345.

Lan, Man et al. 2005. "A Comprehensive Comparative Study on Term Weighting Schemes for Text Categorization with Support Vector Machines." *Proceedings of the 14th International Conference on World Wide Web, WWW(Special Interest Tracks and Posters)*, Chiba, Japan, May 10-14, 1032-1033.

Lauser, B. and Andreas Hotho. 2003. "Automatic Multi-Label Subject Indexing in a Multilingual Environment." *Proceedings of the 7th European Conference in Research and Adavanced Technology for Digital Libraries(ECDL '03)*, 140-151.

Lewis, D. D. et al. 1996. "Training Algorithms for Linear Text Classfiers." *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR '96)*, 298-306.

Liang, Chun-Yan et al. 2006. "Dictionary-based Text Categorization of Chemical Web Pages." *IPM,* 42(4): 1017-1029.

Lewis, D. D. et al. 1996. "Training Algorithms for Linear Text Classifiers." *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR '96)*, 298-306.

Moens, Marie-Francine. 2000. *Automatic Indexing and Abstracting of Document Texts.* The Kluwer International Series on Information Retrieval. Boston: Kluwer Academic Publishers.

Platt, John. 1999. "Fast Training of Support Vector Machines using Sequential Minimal Optimization." In *Advances in Neural Information Processing Systems 11,* by Kearns, M. S., S. A. Solla, and D. A. Cohn. MIT Press.

Plaunt, C. and Barbara A. N. 1998. "An association-based Method for automatic indexing with a controlled vocabulary." *JASIS*, 49(10): 888-902.

Rogati, M. and Y. Yang. 2002. "High-Performing Feature Selection for Text Classification." *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management*, 659-661

Ruiz, Miguel E. and Padmini Srinivasan. 1999. "Combining Machine Learning and Hierarchical Indexing Structures for Text Categorization." To appear in *Advances in Classification Research Vol. 10: Proceedings of the 10th ASIS SIG/CR Classification Research Workshop,* Washington D.C. [cited 2006.1.27]. <http://informatics.buffalo.edu/faculty/ruiz/publications/sigcr%5F10>

Ruiz, Miguel E. and Padmini Srinivasan. 2002. "Hierarchical Text Categorization Using Neural Networks." *Information Retrieval,* 5(10): 87-118.

Sebastiani, Fabrizio. 2002. "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, 34(1): 1-47.

Tzeras, Kostas and Stephan Hartmann. 1993. "Automatic indexing based on Bayesian Inference Network." *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR '93)*, 22-34.

Yang, Y. 1999. "An Evaluation of Statistical Approaches to Text Categorization". *Information Retrieval*, 1: 69-90.

Yang, Y. and Jan O. Pedersen. 1997. "A Comparative Study on Feature Selection in Text Categorization." *Proceedings of the 14th International Conference on Machine Learning(ICML '97)*, 412-420.

Yang, Y. and Xin Liu. 1999. "A Re-examination for Text Categorization Methods." *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval('SIGIR 99)*, 42-49.

Zhang, J. and Y. Yang. 2003. "Robustness of Regularized Linear Classification Methods in Text Categorization." *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR '03)*, 190-197.