기계학습을 기반으로 한 인터넷 학술문서의 효과적 자동분류에 관한 연구

The Study on the Effective Automatic Classification of Internet Document Using the Machine Learning

노 영 희(Young-Hee Noh)*

<목 차>

1. 서론
2. 이론적 배경
   2.1 선행연구
   2.2 인터넷 디렉토리 서비스 현황
   2.3 kNN분류기
3. 실험 설계
   3.1 디렉토리 구조 형성
   3.2 실험 환경
4. 실험 결과 분석
   4.1 평가방법 및 평가척도
   4.2 실험결과 분석
5. 요약 및 결론

초 록

본 연구에서는 kNN분류기를 이용한 범주화 방법에 대한 성능 실험을 하였다. kNN분류기와 같은 대부분의 예제기반 자동 분류기법은 학습문서집단의 자질을 축소하게 되는데 자질을 몇 퍼센트 축소함으로써 높은 성능을 얻을 수 있는지를 알아보고자 하였다. 또한, kNN분류기는 학습문서집단에서 검증문서와 가장 유사한 k개의 학습문서를 찾아야 하는데, 이때 가장 적합한 k값은 얼마인지를 실험을 통하여 검증하여 보고자 하였다.

주제어 : 자동 분류기법, 범주화, kNN분류기

Abstract

This study experimented the performance of categorization methods using the kNN classifier. Most sample based automatic text categorization techniques like the kNN classifier reduces the feature set of the training documents. We sought to find out which percentage reductions in the feature set would result in high performances. In addition, the kNN classifier has to find the k number of training documents most similar to the test documents in the training documents. We sought to verify the most appropriate k value through experiments.

Key Words : Automatic Text Categorization Techniques, kNN Classifier

* 이화여대 국제정보센터 실장(trustme@irs4u.net)

· 접수일 : 2001. 8. 16 · 최초심사일 : 2001. 9. 9 · 최종심사일 : 2001. 9. 15

1. 서 론

인터넷 학술문서를 효율적으로 분류하여 제공하고자 하는 연구 노력이 있어 왔으며, 이러한 연구 결과를 상용화된 포털 사이트, 즉 인터넷디렉토리시스템에 적용하고자 하는 노력이 있다.

인터넷 학술문서를 체계화된 구조로 제공하기 위해 먼저, 디렉토리 체계를 구성해야 하는데 야후나 알타비스타, 심마니, 네이버 등은 일반적인 학문분류를 엄격히 따르기보다는 웹 로봇에 의해 수집된 문서 또는 사이트 관리자가 등록요구를 한 문서의 특성을 파악하여 그 문서가 분류될 적절한 주제범주를 새로 생성하는 식으로 유지되고 있다.

그러나 인터넷을 하나의 거대한 도서관이라고 보고 인터넷 문서도 도서관의 자료처럼 학문분류에 따라 엄격하게 분류되어야 한다는 주장과 연구가 있다. biz/ed나 Blue Web'n Browse by Subject Area, PICK 사이트의 경우는 DDC 분류에 따라 경제학 관련 인터넷 문서를 분류하였고, Nordic WWW 페이지는 UDC 분류체계에 따라 그 문서들을 분류[1]하고 있다. 이 중에서 가장 많이 사용되고 있는 것은 DDC 분류인 것으로 McKiernan[2]의 연구보고서에 나타나 있다.

또한 이러한 분류체계에 인터넷 문서를 실제로 분류하는 한 방법으로 특정 분류표를 사용할 경우, 각 분류표에 나타난 상관색인을 사용하여 인터넷 문서와 상관색인과의 유사도를 산출하여 가장 높은 유사도를 갖는 주제범주로 분류해 주는 방법이 있다. 또 다른 방법으로는 각 분류의 주제명을 각 범주의 대표단어로 추가함으로써 새로 들어온 인터넷 문서와 범주의 대표단어를 비교하여 유사도가 높은 범주에 분류해 주는 방법이 있다.

이처럼 완전 통계기반 문서 자동 분류 방법 외에 최근에는 기계학습을 이용한 문서 할당 방법이 연구되고 있다. 기계학습 기반 문서 범주화(text categorization) 방법에는 규칙기반 방법과 귀납적 학습방법이 있으며, 규칙기반 방법은 지식공학과 범주화 규칙에 대한 지식베이스를 이용하고, 귀납적 학습방법은 수동으로 구축된 학습집단을 대상으로 귀납 학습에 의해 자동으로 범주화시킨다. 귀납적 학습방법에 기반한 방법은 미리 주어진 예제들의 유사성을 이용하여 일반화 과정을 수행하고 가설을 생성한 다음 새로운 예제에 대한 범주를 예측하는 방법으로서, 분류기의 구축과 갱신, 개개인의 관심분야에 대한 범주 생성이 용이하고 업무에 따른 정확률과 재현율 조정이 가능하다는 장점이 있다[3].

---

1) Anders Ard and, Traugott Koch, "Automatic Classification of WAIS databases", 1994. [1997. 3. 7]. <http://www.ub2.lu.se/autoclass.html>.

2) Gerry McKiernan, "Beyond Bookmarks : Schemes for organizing the Web", 1999. [1999.2.4]. <http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm>.

기계학습을 이용해 하나의 문서를 특정 범주로 할당 해 주는 방법으로는 다중회귀모형(multivariate regression model), 최근접 문서기반 분류기(k-Nearest Neighbor classifier: kNN), 확률적 베이지언 모형(probabilistic Bayesian model), 결정트리(decision tree), 신경망(neural network) 기반 모형, SVM(Support Vector Machine: 지지벡터기) 등이 있으며, 특히 kNN분류기는 여러 문서 범주화 기법 중에서 그 알고리즘이 비교적 간단하면서도 성능도 우수한 것으로 알려지고 있으며[4], 위에서 언급한 귀납적 학습방법 중의 하나라고 할 수 있다.

본 연구에서는 kNN분류기를 이용한 범주화 방법에 대한 성능 실험을 하였다. kNN분류기는 학습문서로부터 자질을 축소하고 역시 자질이 축소된 검증문서와 가장 유사한 k개의 학습문서를 찾아 그 문서들이 가장 많이 할당된 범주에 검증문서를 할당해 주는 기법이다. 대부분의 예제기반 자동 분류기법은 자질을 축소하게 되는데 자질을 몇 퍼센트로 축소함으로써 높은 성능을 얻을 수 있는지를 알아보고자 하였다.

또한 kNN분류기는 학습문서집단에서 검증문서와 가장 유사한 k개의 학습문서를 찾아야 하는데, 이 때 가장 적합한 k값은 얼마인지를 실험을 통하여 검증하여 보고자 하였다.

2. 이론적 배경

2.1 선행연구

인터넷디렉토리시스템과 관련된 연구 및 시스템을 보면, 주제기반 디렉토리시스템으로는 Yahoo, Planet Earth, 및 NCSA Meta-Index 등이 있으며, 이러한 시스템들은 브라우즈하기에 편리하게 정보자료를 재조직하고 있다. 그러나 정보자료가 급격하게 증가함에 따라 모든 디렉토리를 브라우즈하고 검색하는데 어려움을 갖게 되었고 신속하게 인터넷 문서를 분류하여 서비스하는데도 어려움이 있다. 비교적 잘 조직된 디렉토리 구조를 가지고 있다고 평가되고 있는 야후는 수작업으로 문서를 분류하는 내용기반 분류시스템이다.

Nordic WAIS/World Wide Web Project는 수작업 분류가 갖는 단점, 즉 표준분류체계를

3) M. A. Hearst et al., "Support vector machines", *IEEE Intelligent Systems*, Vol. 13, No. 4(1998), pp. 18-28.

4) Y. Yang, and Xin Liu. "A re-examination of text categorization methods", *Porceedings of ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR)*, (1999), pp. 42-49, [online]. [cited 2001.03.30]. <http://www.cs.cmu.edu/~yiming>.

따르지 못하거나 웹 문서의 동적 특성을 반영하지 못하는 점 등을 극복하기 위해 WWW과 WAIS의 강점을 통합 개발하였으며 이 프로젝트는 ALIS 도서관 시스템을 WWW에 통합하고 WAIS 데이터베이스를 자동분류하였으며, Nordic WWW 페이지를 자동 색인한 후 UDC 분류체계에 따라 그 문서들을 자동 분류하였다[5]. 이 프로젝트의 결과로서 나온 주제그룹들이 WWW-WAIS 게이트웨이를 통해 직접적으로 탐색되거나 정보자료가 잘 조직되어 있는 WWW 하이퍼텍스트 시스템을 사용하여 브라우즈 할 수 있는 등 자료에 대한 다양한 접근 방법을 제시하고 있다.

HyPursuit은 탐색과 브라우징을 위한 것으로서, 특정 정보자료 데이터베이스를 구조화하기 위해 하이퍼텍스트 문서를 클러스터화한 계층적 네트웍 탐색엔진이다. 이 시스템의 클러스터링 알고리즘은 하이퍼링크 구조와 문서내용에 나타난 의미정보(semantic information)를 기반으로 한다. HyPursuit은 문서들을 그룹화하기 위해 LC 주제명목록표와 같은 원리를 기반으로 다중의 공존하는 클러스터 계층을 형성한다[6]. HyPursuit의 요약기능은 클러스터 내용을 요약하여 광범위한 질의처리를 지원한다. 하이퍼텍스트 클러스터링을 기반으로 한 이 기능은 의미 있고 광범위한 클러스터 계층을 구축하는데 사용될 수 있다.

한편, 인터넷 문서를 디렉토리 구조로 서비스함에 있어 DDC 분류체계를 따르는 것이 적합하다는 주장이 많이 나오고 있다. Svenonious[7]는 온라인 정보검색시스템에 있어서 분류체계는 재현율과 정확률을 향상시키고 브라우징을 가능하게 하며, 언어간 변환을 위한 수단에 기여한다고 주장하고 있다.

Markey와 Demeyer[8]는 DDC 온라인 프로젝트를 수행한 후 최종이용자의 주제접근, 브라우징, 그리고 배열에 있어서 도서관 분류체계는 매우 유용하다고 주장하고 있다. 즉, DDC 분류체계를 따름으로써 주제접근이 향상되고 DDC에 출현한 색인용어로 목록 이용자들이 추가적으로 관련 항목을 검색할 수 있게 되었다는 것이다.

Dahlberg[9]는 네트웍환경에서의 분류이론 적용에 관한 연구에서 LCC와 DDC를 선정하여 분류체계의 적합성 여부를 분석하였으며, 분류이론의 적용은 각 주제의 분석과 추적 및 계층

---

5) Anders Ard and Traugott Koch, "Automatic Classification of WAIS databases", 1994.

6) Ron Weiss et al., "HyPursuit: a Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering", *Proceedings of the Seventh ACM Conference on Hypertext*, Washington, DC, March, 1996.

7) Elaine Svenonius, "Use of classification in online retrieval", *Library Resources and Technical Services*, Vol. 27, No. 1(1983), pp. 76-80.

8) K. Markey, and A. N. Demeyer. Dewey Decimal Classification Online Project : Evaluation of a Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog. Dublin, Ohio : OCLC Online Ocmputer Library Center, Inc., Office of Research, 1986.

9) Ingtraut Dahlberg, "The future of classification in libraries and networks, a theoretical point of view", *Cataloging & Classification Quarterly*, Vol. 21, No. 2(1995), pp. 23-36.

구분에 유용하고, 기존에 분류되어 있는 자료의 네트웍상 재조직에도 필요하다고 하였다.

Vizine-Goetz[10]는 DDC와 LCC의 분류체계를 탐색엔진 Yahoo의 분류체계와 비교, 분석하였는데 각 항목들의 용어와 구성에 표현된 수를 조사하여 항목간의 균형성과 적절성에 대해 논의하였다. Vizine-Goetz[11]의 또 다른 연구에서는 OCLC에서 이러한 분류체계를 적용한 사례를 레코드 형식과 구축방법에 대한 소개와 함께 제시하였다.

한편, 국내연구로 이명희[12]는 주제별 디렉토리 시스템인 Yahoo와 키워드 검색엔진인 AltaVista의 검색효율에 대한 비교연구를 수행하였으며, 김영보[13]는 컴퓨터와 인터넷을 중심으로 인터넷 탐색엔진에 적용될 수 있는 분류체계의 모형을 구축하였다.

최재황[14]은 DDC 분류체계를 이용하여 학술분야 인터넷 정보자원을 검색할 주제별 디렉토리 검색엔진을 설계하였으며, 최희윤[15]은 문헌분류체계와 인터넷 기반 분류체계의 계층구조와 접근방법을 구조적 측면과 검색사례를 통하여 조사하고 이에 대한 비교분석을 통해 인터넷 환경에 적합한 분류체계의 구성방안을 제시한 바 있다.

2.2 인터넷 디렉토리 서비스 현황

인터넷 정보자료에 대하여 특정 형태의 분류체계를 적용하여 주제별 접근이 가능하게 한 디렉토리시스템으로 심마니, 정보탐정, 까치네, 네이버, Yahoo, Magellan, Infoseek 등등을 들 수 있다. 인터넷 학술정보에 중점을 두고 조사한 사례로는 최재황[16]의 논문을 들 수 있는데, 그는 DDC 분류체계로 인터넷 정보를 분류하여 제공하고 있는 관련 사이트를 제시하였다. McKiernan[17]은 다양한 분류체계를 이용하여 인터넷 학술정보를 재조직하여 제공하고 있는 사이트들을 조사하였다. 그에 의하면, 알파벳순으로 분류한 사이트는 2개, 숫자(Numeric)에

10) Diane Vizine-Goetz, "Using library classification schemes for internet resources", *Proceedings of the OCLC Internet Cataloging Colloquium 1996.* [1999.2.5] <http://www.oclc.org/oclc/man/colloq/vg.htm>.
11) Diane Vizine-Goetz, "Classification Research at OCLC", 1996. [1999.2.4] <http://www.oclc.org/oclc/research/publications/review96/class.htm>.
12) 이명희. "네트웍 데이터베이스에서의 주제별 디렉토리와 키워드 탐색엔진의 탐색효율에 관한 탐색적 연구", ≪한국문헌정보학회지≫ 3권 2호(1997), pp. 177-197.
13) 김영보, "인터넷 탐색엔진의 분류체계에 관한 연구 : 컴퓨터, 인터넷 분야를 중심으로". 성균관대학교 대학원 석사학위논문, 1997.
14) 최재황, "인터넷 학술정보자원의 디렉토리 서비스 설계에 있어서 DDC 분류체계의 활용에 관한 연구", ≪정보관리학회지≫ 15권, 2호(1998), pp. 47-67.
15) 최희윤, "인터넷 정보서비스의 분류체계에 대한 비교연구 : 물리학을 중심으로", ≪정보관리학회지≫ 15권 3호(1998), pp. 45-72.
16) 최재황, "인터넷 학술정보자원의 디렉토리 서비스 설계에 있어서 DDC 분류체계의 활용에 관한 연구", (1998), pp. 47-67.
17) Gerry McKiernan, "Beyond Bookmarks : Schemes for organizing the Web", 1999.

의한 분류사이트는 1개, DDC 분류체계를 따른 사이트는 20개, 그리고 UDC 분류체계를 따른 사이트는 3개 등으로 조사하고 있다.

이와 같이 인터넷 정보자원을 DDC 또는 UDC 등 기타의 분류체계로 분류하여 제공하고 있는 사이트가 적지 않지만 특정 주제분야로 특성화시켜 그 주제분야에 대해서 전문화된 서비스를 제공하고 있는 사이트는 많지 않다.

주제별 디렉토리 서비스를 제공하고 있는 사이트 중 특정 주제분야를 전문적으로 분류하여 제공하고 있는 사이트가 있는데, biz/ed는 경영학교육을, Blue Web'n Browse by Subject Area는 교육분야를, Expanding Universe는 천문학 분야의 인터넷 정보자원을, 그리고 PICK은 문헌정보학의 관심분야를 전문 주제항목으로 선정하여 제공하고 있다.

국내의 경우, 특정 주제분야로 제한시켜 주제별 디렉토리시스템을 개발해야하고 DDC나 UDC 등 전문화된 분류체계를 따라 인터넷 학술정보자원을 관리하고 제공해야 한다는 연구와 주장은 있지만 실제로 특정 분류체계를 따르거나 특정 주제분야로 전문화시켜 주제별 디렉토리시스템을 제공하고 있는 기관은 거의 없는 실정이다.

본 절에서는 인터넷 정보자료에 대한 주제별 디렉토리시스템 중 특정 주제분야로 전문화시켜 인터넷 정보자료를 제공하고 있는 시스템을 중심으로 각 시스템의 분류체계, 분류방법 및 검색기법 등에 대해 간단히 기술하고자 한다.

2.2.1 Blue Web'n Browse by Subject Area

이 시스템은 교육분야의 인터넷 정보자원을 서비스하고 있다. DDC 분류체계를 따라 디렉토리시스템을 구축하였지만 DDC 분류순이 아니라 알파벳순으로 디렉토리를 나열하고 있고 DDC 분류번호는 알파벳순 주제항목의 괄호 안에 표시하고 있다[18].

또한 총 12개의 대분류와 104개의 소분류로 구분하고 있으며 웹자원의 자료유형을 7 종으로 구분하여 제공하고 있다. 인터넷 문서를 "Web based Tutorials", "Web based Activities", "Web Based Projects", "Unit & Lesson Plans", "Hotlists", "Other Resources", "Referece & Tools"로 구분하여 동일한 주제분야 중에서도 자료의 유형별 접근을 허용함으로써 이용자가 원하는 자료로 신속하게 접근할 수 있도록 하고 있다.

Blue Web'n이 지원하고 있는 검색방법도 매우 다양하다. 기본적으로 디렉토리 체계에 따라 해당 디렉토리를 따라 내려가면서 원하는 주제를 검색할 수 도 있지만 다양한 검색조건을 줌으로써 신속하게 원하는 문서를 검색할 수 있는 기능도 제공하고 있다. 먼저 인터넷 문서를 이용하고 하는 사람의 연령별로 어린이, 초등학교, 중학교, 고등학교, 대학, 성인 구분할

18) "Blue Web'n, Browse by Subject Area", 1999. [1999.07.15] <http://www.kn.pacbell.com/wired/bluewebn/categories.html>.

수 있게 하고 자료유형을 구분하게 한다. 또한 주제영역과 DDC 분류번호에 의한 제한검색이 가능하며 기타 키워드 검색도 지원하고 있다.

2.2.2 biz/ed

biz/ed는 경영학 교육(bussiness education)을 위한 인터넷 서비스이다. DDC 분류를 이용하여 인터넷 자료 목록 시스템을 구축하였으며 DDC 분류번호 330과 650이 주류를 이루고 있다[19]. 이용자는 DDC 분류번호에 의해 해당 문서에 접근하는 것이 아니라 알파벳순으로 나열된 주제분야별로 접근할 수 있다.

그러나 DDC 분류체계를 따라 디렉토리 구조로 체계화시켜 인터넷 자료를 제공한다기보다는 DDC의 H, 細目을 무시하고 경영학 관련 분류에 해당하는 모든 주제항목을 상위 항목으로 분류하고 있고 하위항목은 존재하지 않는 구조이며 단일층으로 구성되어 있다. 즉, 분류번호 중 332(Financial Economics)와 332.1(Banks)은 분류표상에서 서로 다른 등급으로 332.1이 332의 하위 분류체계에 속하지만 biz/ed 시스템은 이것을 동일한 등급으로 분류하여 알파벳순으로 배열하기 때문에 이 두 항목은 인터넷 목록표상에서 흩어지게 된다. 현재 총 54개의 분류를 알파벳순으로 배열하고 있다.

biz/ed가 제공하는 검색기법으로 불리언 검색기법과 키워드 검색기법이 있으며 그 외에 자료의 성격별로 구분하여 접근할 수 있도록 하는 기능도 제공하고 있다.

2.2.3 Expanding Universe

메트로폴리탄 토론토 참고 도서관(Metropolitan Toronto Reference Library)에 의해 개발된 Expanding Universe 디렉토리시스템은 천문학 분야의 인터넷 정보를 제공하고 있다. DDC 분류체계를 따르고 있으며 520에서 525 분류번호 사이에 선택된 모든 세부분류를 알파벳순으로 배열하여 주제의 알파벳순 목록을 통하여 접근할 수 있도록 하고 있다[20].

Expanding Universe는 DDC 분류번호 중 520, 522, 523, 525의 각 목(目)아래에 세목(細目) 분류를 하고 있다. 즉 4개의 목 아래에 15개의 세목이 있으며 각 세목아래는 다시 분류되어 DDC의 분류체계를 엄격하게 따르고 있다.

이 시스템은 알파벳순에 의한 접근과 분류체계에 따른 접근을 허용하고 있으며 각 디렉토리에 분류된 문서에 대한 직접적인 검색방법은 제공하고 있지 않다.

---

19) "biz/ed: Internet Catalogue", 1999 [1999.07.16] <http://bized.ac.uk/listserv/listhome.htm>.

20) "Expanding Universe: A Classified Search Tool for Amateur Astronomy", 1999. [1999.07.19]. <http://www.mtrl.toronto.on.ca/centres/bsd/astronomy/index.html>.

2.2.4. PICK

PICK 시스템은 University of Wales Aberystwyth의 Thomas Parry 도서관에 의해 개발되었으며 문헌정보학분야 인터넷 정보자원을 제공하고 있다. DDC 분류번호의 020에서 028이 주류를 이루고 있으며 기타 관련분야를 주제항목으로 선정하여 DDC 분류순으로 제공하고 있다[21].

주제항목으로 39개의 주분류를 제공하고 있으나 하위 디렉토리는 존재하지 않는다. 검색방법으로 주제분야에 따른 접근이 가능하고 검색기법으로 불논리 검색과 키워드 검색을 지원하고 있다.

2.3 kNN분류기

2.3.1 kNN분류기의 원리

kNN(k-Nearest Neighbor classification)은 지난 4년 동안 패턴인식분야에서 집중적으로 연구되어온 통계적 접근방법으로 잘 알려져 있다[22]. 그 이후 kNN은 문서 범주화에 응용되었다[23].

kNN 알고리즘은 다른 기계학습 기반 자동분류 알고리즘에 비해 비교적 간단하다. 새로이 분류될 입력문서가 있을 때, 시스템은 학습문서집단 중에서 k개의 최근접 문서를 찾아낸다. 그리고 k개의 최근접 문서들이 할당된 범주정보를 이용하여 후보 범주에 가중치를 부여할 수 있다. 즉, 입력문서와 각 근접문서와의 유사도는 이웃문서가 속한 범주의 가중치가 되는 것이다. 만약 k개의 최근접 문서 중 여러 개가 하나의 범주에 분류되어 있다면 여러 개의 근

21) "PICK: Quality Internet Resources in Library and Information Science", 1999. [1999.07.20]. <http://www.aber.ac.uk/~tplwww/e/contents.html>.

22) Belur V. Dasarathy. *Nearest Neighbor(NN) Norms: NN Patern Classification Techniuges* McGraw-Hill Computer Science Series. Las Alamitos, California : IEEE Computer Society Press, 1991.

23) B. Masand, G. Linoff, and D. Waltz. "Classifying news stories using memory based reseonin", *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, (1992), pp. 59-64; Y. Yang, "Expert network : effective and efficient learning from human decisions in text categorization and retrieval", *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, (1994), pp. 11-21; Makato Iwayama and Takenobu Tokunaga. "Cluster-based text categorization: a comparison of category search strategies", *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, (1995), pp. 273-281.

접 문서들의 가중치가 그 범주에 모두 더해지며, 그 결과로서 나온 가중치의 합은 입력문서에 대한 그 범주의 유사도로 사용될 수 있는 것이다. 후보 범주의 가중치를 정렬하여 입력문서를 후보 범주 중 하나에 최종 분류할 수 있다.

또한 정렬된 후보 범주에 특정 기준치를 적용하여 기준치 이상의 범주들에 복수 할당될 수 있도록 할 수 있으며, 문서 분류를 위한 kNN 공식은 다음과 같다. k개의 최근접 문서들 각각의 유사도를 학습문서 D_j가 해당 범주 C_k에 속할 조건확률, $P(C_k, D_j)$과 곱하고 이들을 모두 합산하여 각 범주별로 적합성 점수, $rel(C_k \mid D_x)$을 구한다. 이 중에서 적합성 점수가 가장 높은 상위 범주에 새로운 입력문서를 할당한다[24].

$$rel(C_k \mid D_x) \approx \sum_{D_j \in kNN} sim(D_x, D_j) \times P(C_k, D_j) \qquad \text{<공식 1>}$$

위 식에서 $sim(D_x, D_j)$와 $P(C_k, D_j)$는 다음 공식 a와 b로 상세하게 풀어 쓸 수 있다.

선정된 자질들을 벡터로 표현하기 위해 $tf \cdot idf$를 가중치로 사용하였으며 학습문서 D_j와 입력문서 D_x의 유사도를 산출하기 위해서 다음과 같이 코싸인 유사계수공식을 이용한다.

$$W'(D_x, D_j) = \frac{\sum_{i=1}^{n} t_{xk} \times t_{jk}}{\sqrt{\sum_{i=1}^{n} (t_{xk})^2 \times \sum_{i=1}^{n} (t_{jk})^2}} \qquad \text{<공식 a>}$$

$$P(C_k, D_j) \approx \frac{\text{범주 } C_k \text{가문서 } D_j \text{에할당된빈도}}{\text{학습집단에서문서 } D_j \text{가출현한빈도}} \qquad \text{<공식 b>}$$

위 공식은 입력문서 i와 최근접 문서 j간의 유사도를 측정하는 공식이다. t_{ik}는 입력문서 i내의 용어 k의 가중치를 나타내며, d_{jk}는 최근접 문서 j내의 용어 k의 가중치를 나타낸다. 이러한 매칭함수에 의해 산출된 유사도 순으로 정렬되며 가장 높은 순위의 범주로 문서를 할당한다.

---

24) Y. Yang, "Expert network : effective and efficient learning from human decisions in text categorization and retrieval", *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, (1994), pp. 11-21.

2.3.2 kNN분류기에 대한 선행연구

kNN분류기를 이용한 선행연구를 살펴보면, 먼저 Yang[25]은 다양한 문서 범주화 기법들을 비교하는 실험에서 kNN 분류기의 성능을 비교하였는데, 이 실험에서 kNN 분류기는 Widrow-Hoff, 신경망 기반 모형, LLSF 매핑 모형과 함께 실험대상 11개의 기법 중 높은 성능을 보여 주었다. 특히, kNN은 분류공간이 수백개의 범주를 갖는 수준에서 수천개의 범주를 갖는 수준으로 증가하더라도 무리없이 진행됨으로써 MEDLINE 전체 주제 범주를 처리할 만한 유일한 기계학습방법으로 평가되었다.

Yang의 또 다른 연구[26]에서는 12 종류의 문서범주화 기법의 성능을 직, 간접적으로 비교했는데, 전반적으로 kNN, LLSF, 그리고 신경망 기법이 가장 높은 성능을 보여주는 것으로 실험결과 나타났으며, 나이브 베이즈 기법을 제외한 다른 기계학습 알고리즘들도 비교적 높은 성능을 보여 주는 것으로 나타났다.

Yang과 Liu[27]는 1999년에 kNN 분류기를 포함한 다섯 개의 분류기의 성능을 비교한 바 있다. 비교대상 분류기는 SVM(Support Vector Machine: 지지벡터기), 최근접 문서기반 분류기(k-Nearest Neighbor classifier: kNN), 신경망(neural network; NNet) 기반 모형, the Linear Least-squares Fit (LLSF) mapping, 나이브 베이즈(Naive Bayes; NB) 분류기이다. 이 연구의 실험 결과, 범주가 충분히 일반적이면서 범주 당 적합한 학습문서의 수가 10개 이하로 적을 경우에 SVM, kNN 및 LLSF가 NNet나 NB보다 훨씬 높은 성능을 보여주는 것으로 나타났다.

이영숙과 정영미는 kNN기법에서 보편적으로 사용되는 범주 할당 방법을 응용하여 k개의 유사문서 중 최상위 및 상위 M개 문서에 가중치를 부여하는 방법들을 고안하고자 하는 실험과 함께 k값의 변화에 따른 이들의 성능을 비교하는 실험을 수행한 바 있다[28].

---

25) Yiming Yang, *An Evaluation of Statistical Approaches to Text Categorization.* Computer Science Technical Report CMU-CS-97-127, School of Computer Science, Carnegie Mellon University, 1997. <http://reports-archive.adm.cs.cmu.edu/anon/1997/abstracts/97-127.html>.

26) Y. Yang, "An evaluation of statistical approaches to text categorization", *Journal of Information Retrieval*, Vol. 1 No. 1-2(1999), pp. 67-88.

27) Y. Yang, and Xin Liu. "A re-examination of text categorization methods", *Porceedings of ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR)*, (1999), pp. 42-49.

28) 이영숙, 정영미, "KNN 분류기의 범주할당 방법 비교 실험", 《정보관리학회 학술대회 논문집》, (2000), pp. 37-40.

3. 실험설계

3.1 디렉토리 구조 형성

인터넷 정보자료에 대하여 특정 분류체계를 적용함으로써 주제별 접근이 가능하도록 하도록 하는 시스템이 있다. 앞에서 언급했듯이 McKiernan의 조사에 의하면, 대부분의 시스템이 일반 도서관이 채택하고 있는 분류체계를 따르고 있으며 그 중에서 DDC 분류체계가 가장 많이 사용되고 있는 것으로 분석되었다.

본 연구에서는 인터넷 디렉토리의 구조를 설계함에 있어서 가장 많이 사용되고 있는 DDC 분류체계를 따르고 있으며 디렉토리명은 DDC 분류의 주제명으로 하였다. DDC는 모든 대상을 0에서 9까지 10구분하는 십진식 분류법으로서 인류가 기록한 모든 지식을 10개의 주류(main classes)아래 모으고, 각 主類는 다시 10개의 綱(divisions)아래 모으며, 각 강은 다시 10의 目(sections)아래 모은다. 본 연구는 이와 같은 분류체계를 따르되 주류 중 "Social Sciences(300)"분야의 "Economics(330)"만을 대상으로 하고 있다. 이렇게 해서 구축된 디렉토리 체계 내 총 주제범주는 757개였다.

3.2 실험 환경

1) 실험문서집단 구성

실험문서집단의 구성을 위해 웹 로봇 에이전트를 사용하여 인터넷에서 정보자료를 수집하였다. 웹 로봇 에이전트가 처음으로 방문해야 할 초기 URL은 경제학 관련 기관 및 연구소로 주었다. 문서가 수집된 후 디렉토리체계 내의 특정 범주로 분류해 주어야 하는데, 이를 위해 일단 자동분류기법을 사용하여 자동으로 문서들을 특정 범주로 분류되게 하고 인간 분류전문가의 수작업 확인과정을 거쳐 완전한 실험문서집단을 구성하였다. 자동분류를 위해 사용한 방법은 디렉토리 구조를 형성할 때 만들어진 각 범주 대표단어와 인터넷 문서내의 색인어와의 코싸인 유사도 가중치를 활용하는 것이다.

코싸인 유사도 공식을 사용할 때 문서가 각 범주에 할당되는 조건으로 특정 기준치 이상의 유사도를 갖는 디렉토리시스템 내의 특정 범주로 문서를 분류해 줄 수 있도록 하였는데, 여기에서는 0.1로 하였다. 또 다른 기준치를 주어 특정 기준치 이상이면 하나의 문서가 한 개 이상의 범주에 분류되게 함으로써 다중분류를 허용할 수 있으나 본 연구에서는 다중분류

를 허용하지 않고 유사도가 가장 높은 범주로 문서가 분류될 수 있도록 하였다.

자동분류과정과 수작업 확인 과정을 거친 후 총 범주 757개 중 문서가 하나 이상 분류된 범주 수는 386개였다. 실험문서집단은 다시 학습문서집단과 검증문서집단으로 나뉘는데 학습문서집단과 검증문서집단의 비율은 7:3이 되도록 하였다. 각 범주에 분류된 문서들 중 30% 정도를 검증문서집단으로 뺀 결과, 남아 있는 학습문서집단은 총 1,888건이었고 검증문서집단은 624건으로 실제로는 7.4:2.6의 비율로 분리되었다.

실험문서집단의 통계적 특성으로 전체 문서 수, 총 범주 수 및 문서 당 범주 수 등에 관한 것은 표 1과 같다.

<표 1> 실험문서집단의 실험구성요소

내 용		갯 수
전체 문서 수		2,512
총 범주 수		757
문서 당 범주 수		1
범주 당 문서 수	평균	3.32
	최대	184
	최소	0
학습문서집단		1,888
검증문서집단		624

2) 자질 축소

여기에서 자질 축소란 하나의 문서를 특정 범주로 분류해 줄 때 문서 내에 출현한 모든 색인어를 자동분류에 이용하는 것이 아니라 문헌빈도(Document Frequency)를 이용하여 문헌빈도 순위 20~30%에 해당하는 색인어만을 자동분류의 자질로 사용하는 것을 말한다.

본 연구에서는 학습문서 및 검증문서의 자질을 각각 문헌빈도의 20%, 30%, 50%, 100%로 각각 추출하고 문서 내 색인어 중 몇 퍼센트를 자질로 하였을 때 가장 높은 성능을 보여주는지 비교하였다.

3) 입력문서에 대한 k개의 최근접 문서 선정

각 범주에 속한 실험문서를 범주별로 벡터형태로 표현하고 검증문서를 하나씩 입력하여 특정 범주로 분류되는 과정을 거치게 되는데 이 때, 하나의 검증문서를 특정 범주로 분류하기 위해 k개의 최근접 문서를 산출해야 한다. k개의 최근접 문서가 산출되면 최근접 문서가 가장 많이 출현한 범주로 검증문서를 분류해 줄 수 있다.

본 연구에서는 k값을 얼마로 하였을 때 가장 높은 성능을 보여 주는지 알아보기 위해 k값

을 각각 10, 20, 30으로 하여 실험하였으며 k개의 최근접 문서 산출을 위해 사용한 공식은 코싸인 유사계수 공식이다.

또한 k개의 최근접 문서를 기반으로 입력문서를 특정 범주로 할당 할 때 사용한 공식은 2장의 공식 1이다. 대부분의 시스템은 보통 복수 범주를 허용하지만 여기에서는 가장 유사한 하나의 범주로 분류되게 하고 복수 범주 분류는 허용하지 않았다.

4. 실험 결과 분석

4.1 평가방법 및 평가척도

kNN분류기에 다양한 기준을 적용하여 구축된 디렉토리시스템을 비교 평가하기 위해 각 범주에 대한 분류결과를 표현하는 표 2의 2×2 분할표를 이용하였다. 아래 표에서 보이는 적합문서와 부적합문서에 대한 판정은 DDC 분류체계에 따라 구성된 각 디렉토리의 주제범주로 분류된 문서를 인간 주제전문가가 수작업으로 각 범주에 분류된 문서의 적합성을 평가한 것이다. 즉, 각 범주에 할당된 문서를 적합문서와 부적합문서로 평가하여 구분하고 부적합문서는 실제로 분류되어야 할 범주를 찾음으로써 각 범주에 할당되지 못한 적합문서를 산출할 수 있다.

a : 특정 범주에 분류된 문서 중 적합 문서의 수
b : 특정 범주에 분류된 문서 중 부적합 문서의 수
c : 특정 범주에 분류되지 않은 적합 문서의 수
d : 특정 범주에 분류되지 않은 부적합 문서

<표 2> 2×2 분할표

	적합문서	부적합문서
범주에 할당	a	b
범주에 할당되지 못함	c	d

위 표를 이용하여 정확률(precision)과 재현율(recall), 정확도(accuracy), 오류율(error), 그

리고 부적합률(fallout)을 구할 수 있다.

$$\text{정확률}(r) = \frac{a}{a+b}$$

$$\text{재현율}(p) = \frac{a}{a+c}$$

$$\text{정확도}(a) = \frac{a+d}{a+b+c+d} \qquad (n = a+b+c+d \quad)$$

$$\text{오류율}(e) = \frac{b+c}{a+b+c+d} \qquad (n = a+b+c+d \quad)$$

$$\text{부적합률}(f) = \frac{b}{b+d}$$

일반적으로 사용되는 척도로 van Rijsbergen[29]에 의해 처음으로 도입된 F척도(F-measure)가 있으며[30], 반비례 관계에 있는 정확률과 재현율을 하나의 값으로 나타나기 위하여 β값을 1로 하는 F_1척도를 사용하였다.

$$F_\beta(r,p) = \frac{(\beta^2+1)2pr}{\beta^2 p + r} \qquad F_1(r,p) = \frac{2pr}{p+r}$$

또한, 분산도와 계층별 분산도를 평가 척도로 이용하였는데, 정확률은 디렉토리시스템 내에 있는 범주에 분류된 문서를 평가하여 각 범주에 적합하게 분류된 문서의 비율을 산출하는 것인 반면, 분산도는 웹 로봇에 의해 수집된 문서가 디렉토리시스템의 특정 범주에 집중되지 않고 각 범주에 얼마나 잘 분산되어 분류되어 있는지를 측정하는 것이다. 분산도 산출 공식은 아래와 같다.

$$\text{분산도} = \frac{\text{문서가 분류된 범주의 수}}{\text{전체 범주의 수}} \qquad \text{<공식 3>}$$

29) C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.

30) I. Moulinier, G. Raskinis, and J. Ganascia. "Text Categorization: a Symbolic Approach", In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval*, 1996 ; David D. Lewis, et al. "Training Algorithms for Linear Text Classifiers", *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, (1996), pp. 298-306 ; William W. Cohen and Yoram Singer, "Context-Sensitive Learning Methods for Text Categorization", *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, (1996), pp. 307-315.

또한, 계층별 집중도는 전체 범주에 분류된 문서 중 각 계층 범주에 얼마나 집중되어 분류되어 있는지를 평가하는 것이다. 계층별 집중도를 산출하는 공식은 다음과 같다.

$$\text{계층별집중도} = \frac{\text{각 계층 범주에 분류된 문서 수}}{\text{전체 범주에 분류된 문서 수}} \qquad \text{<공식 4>}$$

4.2 실험결과 분석

본 실험에서는 kNN분류기에 다양한 조건을 부여하여 최적의 조합을 발견하고자 하였다. 표 3은 자질축소 비율과 k개의 최근접 문서에 대한 조건을 조합하여 12가지 경우를 만들었으며, 실험결과에 대한 분석은 이 조합을 기준으로 하였다.

<표 3> kNN분류기의 조건 조합

자질축소 조건(%)	k개의 최근접 문서 조건(개)
20	10
20	20
20	30
30	10
30	20
30	30
50	10
50	20
50	30
100	10
100	20
100	30

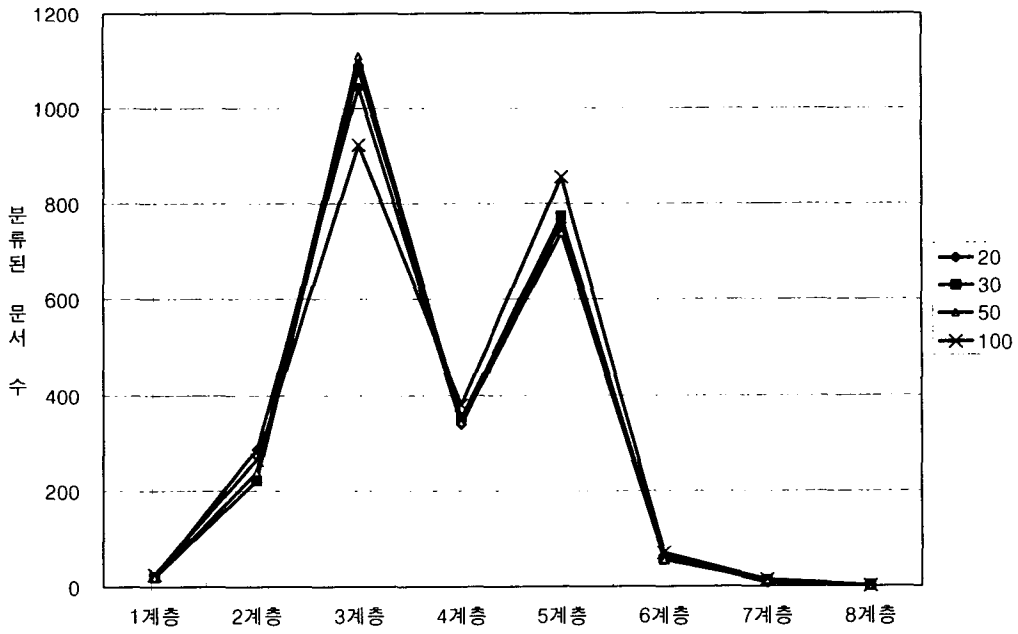
4.2.1 계층별 분류 문서 수

본 실험에서는 학습문서집단에 나타난 각 문서의 색인어 중 유사도 측정에 사용할 자질의 비율과 k개의 최근접 문서의 수를 변화시켰을 때에 각 계층별로 분류되는 문서의 수 및 분산도의 변화를 알아보고자 하였다. 표 4는 조건변화에 따른 계층별 분류문서의 수이고 그림 1은 이를 그림으로 표현한 것이다. 그림 1에서 보듯이 자질의 비율이나 k값의 변화에 관계없이 3계층, 5계층, 4계층 순으로 문서가 많이 분류되는 것으로 나타났다. 즉, 전체 범주에 분류된 문서 중 각 계층 범주에 얼마나 집중되어 분류되어 있는지를 분석하였는데, 계층별 집중도는 3계층, 4계층, 5계층이 높은 것으로 나타났으며, 자질의 비율 및 k값의 변화에 많은

영향을 받지 않는 것으로 나타났다.

<표 4> 조건변화에 따른 계층별 문서의 집중도

계층 / 조건	1계층	2계층	3계층	4계층	5계층	6계층	7계층	8계층	평 균
20%/10위	21	300	931	386	801	65	12	1	314.63
20%/20위	20	287	1043	339	758	62	7	1	314.63
20%/30위	20	279	1070	339	743	54	12	1	314.75
30%/10위	20	303	961	364	797	55	9	4	314.13
30%/20위	20	222	1081	353	773	56	12	1	314.75
30%/30위	20	231	1087	319	791	56	12	1	314.63
50%/10위	20	325	950	355	784	71	12	1	314.75
50%/20위	20	239	1109	343	738	56	12	1	314.75
50%/30위	20	233	1086	328	782	56	12	1	314.75
100%/10위	25	265	923	380	855	70	14	1	316.63
100%/20위	25	265	923	380	855	70	14	1	316.63
100%/30위	25	265	923	380	855	70	14	1	316.63
평 균	21.33	267.83	1,007.25	355.50	794.33	61.75	11.83	1.25	315.14



<그림 1> 조건변화에 따른 계층별 문서의 집중도

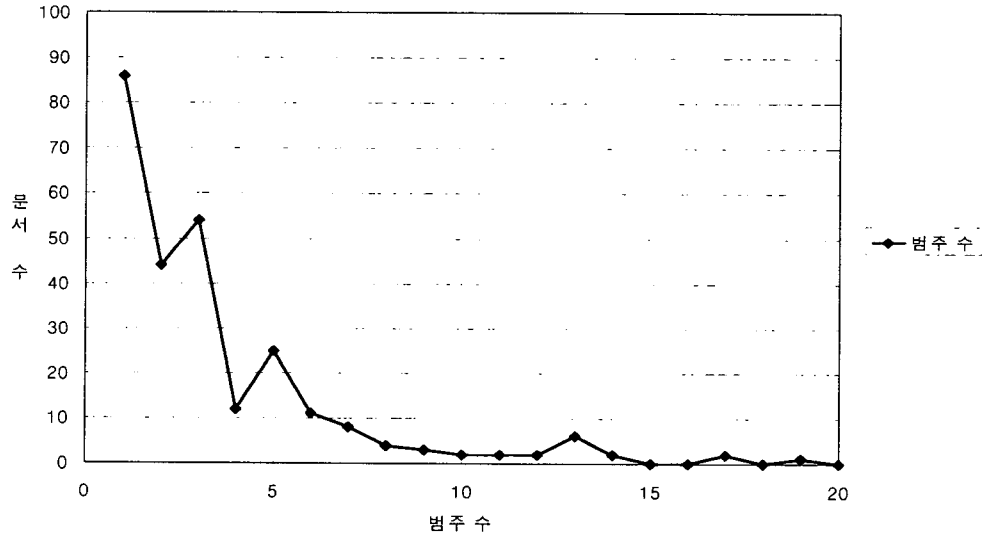
4.2.2 범주 분포도

문서가 각 범주에 분류된 정도, 즉 범주 분포도를 분석하기 위해 자질을 20%, k값을 10으로 한 경우의 조건을 택하였으며, 표 5와 그림 2가 여기에 해당한다. 표와 그림에서 보듯이 문서 1개를 포함하고 있는 범주 수는 86개로 문서가 분류된 총 범주 수, 283개의 30%를 차지하고 있는 것으로 나타났으며 20개까지 분류된 범주 수는 총 264개로 문서가 하나라도 분류된 범주의 93.28%를 차지하고 있어, 하나의 범주에 문서가 집중되어 분류되는 현상은 발생하지 않는다는 사실을 알 수 있다.

또한, 웹 로봇에 의해 수집된 문서가 디렉토리 시스템 내 각 범주에 분산되어 있는 정도를 측정하는 분산도는 37.38%로 나타났다. 즉, 디렉토리 시스템 내 전체 범주 수는 757개로 이 범주 중 문서가 하나라도 분류된 범주는 283개로 나타났다. 전체 문서의 수가 2500건 정도밖에 되지 않는다는 사실을 감안하면, 비교적 높은 분포도를 보여주고 있는 것을 알 수 있다.

<표 5> 범주 분포도

문서 수	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	총
범주 수	86	44	54	12	25	11	8	4	3	2	2	2	6	2	0	0	2	0	1	0	264



<그림 2> 범주 분포도

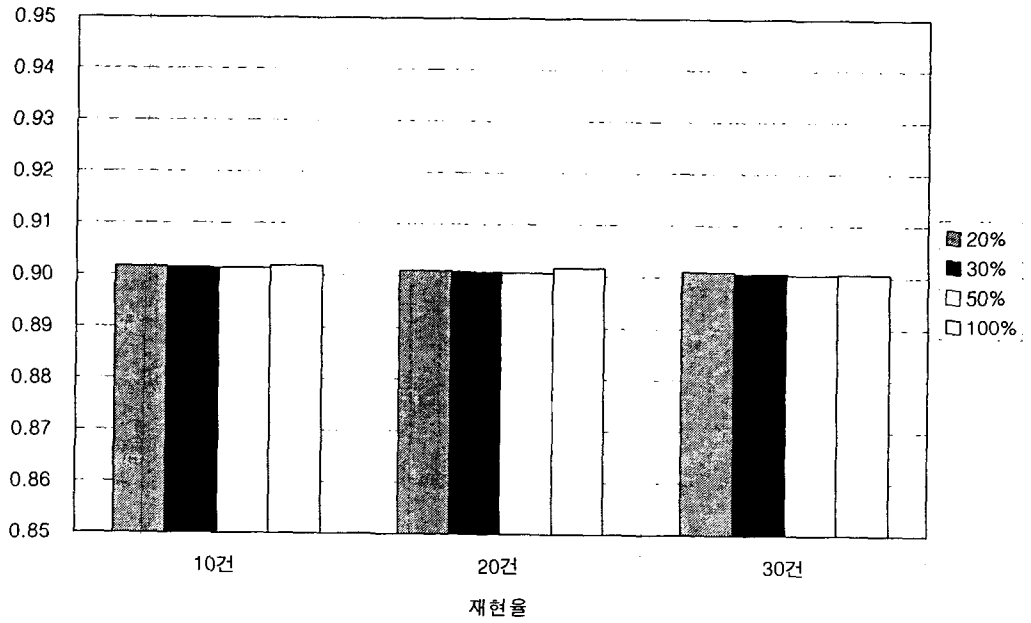
4.2.3 분할표를 이용한 성능 평가

2x2 분할표를 이용하여 정확률과 재현율, 정확도, 오류율, 그리고 부적합률을 모두 구할 수 있다.

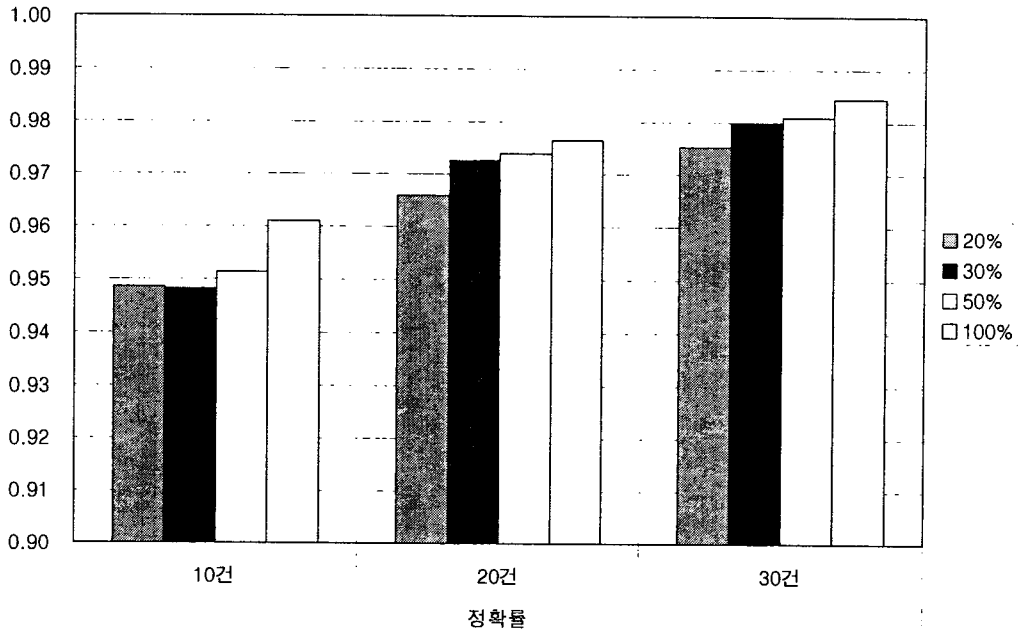
표 6은 자질축소 비율과 k개의 최근접 문서에 대한 조건을 다양하게 조합한 조건별 재현율과 정확률이고 그림 3과 그림 4는 이를 그림으로 표현한 것이다. 표 6에서 보듯이 평균 재현율은 k값이 10건, 20건, 30건일 때, 각각 0.9015, 0.9009, 0.9006으로 나타났고, 평균 정확률은 각각 0.9523, 0.9722, 0.9802로 나타났다. 이와 같이 평균 재현율 성능이 90%를 넘고 평균 정확률 성능은 95%를 넘고 있어서 기계학습기반 자동 분류기법을 이용하면 매우 높은 성능을 얻을 수 있음을 알 수 있다.

<표 6> k값 및 자질축소 비율에 따른 재현율 및 정확률

k값 / 자질	재현율				정확률			
	10건	20건	30건	평 균	10건	20건	30건	평 균
20%	0.9015	0.9010	0.9009	0.9011	0.9486	0.9658	0.9752	0.9632
30%	0.9014	0.9008	0.9007	0.9010	0.9482	0.9726	0.9800	0.9669
50%	0.9013	0.9005	0.9005	0.9008	0.9513	0.9739	0.9811	0.9688
100%	0.9018	0.9015	0.9005	0.9013	0.9610	0.9765	0.9845	0.9740
평균	0.9015	0.9009	0.9006	0.9010	0.9523	0.9722	0.9802	0.9682



<그림 3> k값 및 자질축소 비율에 따른 재현율

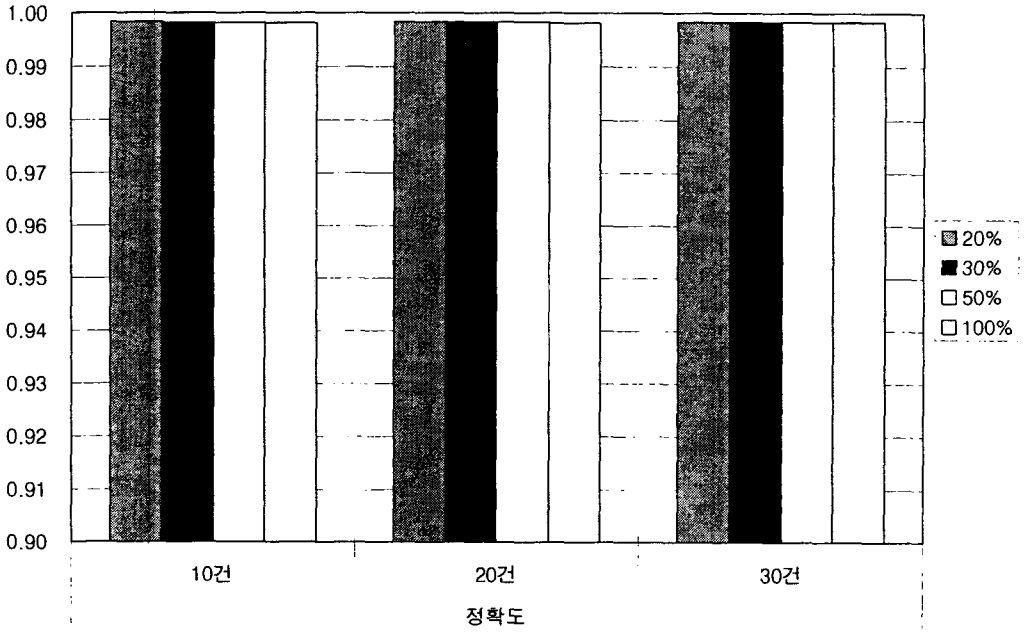


<그림 4> k값 및 자질축소 비율에 따른 정확률

정확도는 각 범주에 분류된 전체 문서 중 각 범주에 분류되어야 할 문서는 정확히 분류되고 배제되어야 할 문서는 정확히 배제된 정도를 측정하는 것이다. 표 7은 k값 및 자질축소 비율에 따른 정확도이고 그림 5는 이를 그림으로 표현한 것으로서, 표에서 보듯이 정확도는 자질을 20%로 축소하였을 때 가장 높았고, 분류될 범주 선택시 사용할 문서 수 즉, k값을 20건이나 30건으로 할 때 좀 더 높다고 할 수 있지만 다른 조건과의 비교에서 0.001밖에 차이가 나지 않는 것으로 나타났다. 전체적인 시스템의 정확도는 0.9984%로 매우 높은 것으로 나타났다.

<표 7> k값 및 자질축소 비율에 따른 정확도

k값 / 자질	정확도			
	10건	20건	30건	평 균
20%	0.9984	0.9984	0.9984	0.9984
30%	0.9983	0.9984	0.9984	0.9984
50%	0.9983	0.9984	0.9984	0.9984
100%	0.9983	0.9983	0.9984	0.9983
평균	0.9983	0.9984	0.9984	0.9984



<그림 5> k값 및 자질축소 비율에 따른 정확도

위와 같이 정확도가 매우 높게 나타난 반면에 오류율과 부적합률은 평균적으로 각각 0.0017과 0.0008로 극히 낮게 나타나고 있어 인간 분류전문가의 분류수준에 이른다는 것을 알 수 있다. k값 및 자질축소 비율 변화에 따른 오류율 및 부적합률의 변화는 표 8과 같다.

<표 8> k값 및 자질축소 비율에 따른 오류율 및 부적합률

성능 / 자질	오류율				부적합률			
	10건	20건	30건	평 균	10건	20건	30건	평 균
20%	0.0016	0.0016	0.0016	0.0016	0.0008	0.0008	0.0008	0.0008
30%	0.0017	0.0016	0.0016	0.0016	0.0008	0.0008	0.0008	0.0008
50%	0.0017	0.0016	0.0016	0.0016	0.0008	0.0008	0.0008	0.0008
100%	0.0017	0.0017	0.0016	0.0017	0.0008	0.0008	0.0008	0.0008
평균	0.0017	0.0016	0.0016	0.0016	0.0008	0.0008	0.0008	0.0008

5. 요약 및 결론

본 연구에서는 kNN분류기를 이용한 범주화 방법에 대한 성능 실험을 하였다. kNN분류기와 같은 대부분의 예제기반 자동 분류기법은 학습문서집단의 자질을 축소하게 되는데 자질을 몇 퍼센트 축소함으로써 높은 성능을 얻을 수 있는지를 알아보고자 하였고, 학습문서집단에서 검증문서와 가장 유사한 k개의 학습문서를 얼마로 하였을 때 높은 성능을 보여 주는지 실험하였다. 그 실험결과는 다음과 같다.

첫째, 학습문서집단에 나타난 각 문서의 색인어 중 유사도 측정에 사용할 자질의 비율과 k개의 최근접 문서의 수를 변화시키면서 각 계층별로 분류되는 문서 수 및 분산도를 측정한 결과, 자질 축소 비율이나 k값의 변화에 관계없이 3계층, 5계층, 4계층 순으로 문서가 많이 분류되는 것으로 나타났다.

둘째, 문서가 각 범주에 분류된 정도, 즉 범주 분포도를 측정할 결과 하나의 범주에 문서가 집중되어 분류되는 현상은 발생하지 않는다는 사실을 알 수 있었다.

셋째, 자질축소 비율과 k개의 최근접 문서에 대한 조건을 다양하게 조합한 조건별 재현율과 정확률은 평균적으로 각각 90%, 95% 이상의 높은 성능을 보여 주었다.

넷째, 정확도는 자질을 20%로 하였을 때 가장 높았고, 분류될 범주 선택시 사용할 문서 수는 20건이나 30건으로 할 때 좀 더 높게 나타났다. 반면에 오류율과 부적합률은 평균적으로 각각 0.0017과 0.0008로 극히 낮게 나타나고 있어 인간전문가의 분류수준에 이른다는 것을 알 수 있었다.

위와 같이 여러 가지 범주화 기법 중 kNN분류기를 이용하여 그 시스템 성능을 측정한 결과 재현율과 정확률은 90% 이상의 높은 성능을 보여주고 있는 것으로 나타나고 있다. 연구결과로 알 수 있는 사실은 첫째, 웹 문서의 기계학습기반 자동분류를 위해 실험문서집단을 학습문서집단과 검증문서집단으로 구분하고 학습문서집단에 속한 문서의 자질을 축소하게 되는데, 이 때 자질 축소의 비율을 20%, 30%, 50%, 100%으로 한 모든 경우에 비슷한 성능을 보여 주었다. 오히려 20%에서 비교적 높은 성능을 보여 주었고 다음으로 30%, 50%, 100%순으로 나타났다. 둘째, 입력문서와 가장 유사한 k개의 최근접 문서를 분석하여 문서가 가장 많이 출현한 주제범주를 택함에 있어, 상위 10건, 20건, 30건까지의 문서를 대상으로 분석한 결과 상위 30건까지의 문서들을 분석해서 문서가 가장 많이 출현한 범주를 택하는 것이 가장 높은 성능을 보여 주었고 20건, 10건 순으로 나타났다.

참 고 문 헌

김영보. "인터넷 탐색엔진의 분류체계에 관한 연구 : 컴퓨터, 인터넷 분야를 중심으로", 성균관대학교 대학원 석사학위논문, 1997.

이명희. "네트웍 데이터베이스에서의 주제별 디렉토리와 키워드 탐색엔진의 탐색효율에 관한 탐색적 연구", ≪한국문헌정보학회지≫ 3권, 2호(1997). pp. 177-197.

이영숙, 정영미. "KNN 분류기의 범주할당 방법 비교 실험", ≪정보관리학회 학술대회 논문집≫, (2000). pp. 37-40.

최재황. "인터넷 학술정보자원의 디렉토리 서비스 설계에 있어서 DDC 분류체계의 활용에 관한 연구", ≪정보관리학회지≫ 15권, 2호(1998). pp. 47-67.

최희윤. "인터넷 정보서비스의 분류체계에 대한 비교연구 : 물리학을 중심으로", ≪정보관리학회지≫ 15권, 3호(1998). pp. 45-72.

"Blue Web'n, Browse by Subject Area". 1999. [1999.07.15] <http://www.kn.pacbell.com/wired/bluewebn/categories.html>.

"biz/ed: Internet Catalogue". 1999 {1999.07.16] <http://bized.ac.uk/listserv/listhome.htm>.

"Expanding Universe: A Classified Search Tool for Amateur Astronomy". 1999. [1999.07.19]. <http://www.mtrl.toronto.on.ca/centres/bsd/astronomy/index.html>.

"PICK: Quality Internet Resources in Library and Information Science". 1999. [1999.07.20]. <http://www.aber.ac.uk/~tplwww/e/contents.html>.

Ard, Anders and Koch, Traugott, "Automatic Classification of WAIS databases", 1994. [1997. 3. 7]. <http://www.ub2.lu.se/autoclass.html>.

Belur V. Dasarathy. *Nearest Neighbor(NN) Norms: NN Patern Classification Techniuqes,* McGraw-Hill Computer Science Series. Las Alamitos, California : IEEE Computer Society Press, 1991.

Cohen, William W. and Yoram Singer. Context-Sensitive Learning Methods for Text Categorization", *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR),* 1996. pp. 307-315.

Dahlberg, Ingtraut. "The Future of Classification in Libraries and Networks, a Theoretical Point of View". *Cataloging & Classification Quarterly*, Vol. 21, No. 2 (1995). pp. 23-36.

Hearst, M.A. et al. "Support vector machines", *IEEE Intelligent Systems,* Vol. 13, No. 4 (1998). pp. 18-28.

Iwayama, Makato, and Takenobu Tokunaga. "Cluster-based Text Categorization: a Comparison of Category Search Strategies", *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR),* 1995. pp. 273-281.

Lewis, David D., et al. "Training algorithms for linear text classifiers", *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR),* 1996. pp. 298-306.

Markey, K., and A. N. Demeyer. *Dewey Decimal Classification Online Project : Evaluation of a Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog.* Dublin, Ohio : OCLC Online Ocmputer Library Center, Inc., Office of Research, 1986.

McKierman, Gerry. "Beyond Bookmarks : Schemes for Organizing the Web". 1999. [1999.2.4]. <http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm>.

Masand, B., G. Linoff, and D. Waltz. 1992. "Classifying News Stories Using Memory Based Reseonin", *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR),* 1992, pp. 59-64.

Moulinier, I., G. Raskinis, and J. Ganascia. "Text Categorization: a Symbolic Approach", In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval,* 1996.

Svenonius, Elaine. "Use of Classification in Online Retrieval". *Library Resources and Technical Services,* Vol. 27, No. 1(1983). pp. 76-80.

van Rijsbergen, C. J. *Informatioh Retrieval.* London: Butterworths, 1979.

Vizine-Goetz, Diane. "Classification Research at OCLC". 1996. [1999.2.4] <http://www.oclc.org/oclc/research/publications/review96/class.htm>.

Weiss, Ron, et al. "HyPursuit: a Hierarchical Network Search Engine That Exploits Content-Link Hypertext Clustering", *Proceedings of the Seventh ACM Conference on Hypertext,* Washington, DC, March, 1996.

Yang, Y. "Expert Network : Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval", *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR),* 1994. pp. 11-21.

Yang, Y. *An Evaluation of Statistical Approaches to Text Categorization.* Computer

Science Technical Report CMU-CS-97-127, School of Computer Science, Carnegie Mellon University, 1997.
<http://reports-archive.adm.cs.cmu.edu/anon/1997/abstracts/97-127.html>.

Yang, Y. "An Evaluation of Statistical Approaches to Text Categorization", *Journal of Information Retrieval,* Vol. 1, No. 1-2(1999). pp. 67-88.

Yang, Y., and Xin Liu. "A Re-examination of Text Categorization Methods". *Porceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR),* (1999). pp. 42-49, [online]. [cited 2001.03.30].
<http://www.cs.cmu.edu/~yiming>.