자동 도서분류를 위한 확장된 나이브베이지안 알고리즘

김성수*, 정현준**, 백두권***†
*고려대학교 컴퓨터정보통신대학원 소프트웨어공학과
**고려대학교 정보통신대학 컴퓨터·전파통신공학과
***고려대학교 융합소프트웨어전문대학원
e-mail : sskimm@korea.ac.kr, junghj85@gmail.com, baikdk@korea.ac.kr

An Extended Naive Bayesian Algorithm for Automatic Book Classification

Sung-Soo Kim*, Hyun-Jun Jung**, Doo-Kwon Baik***†
*Dept. of Software Engineering, Korea University
**Dept. of Computer and Radio Communications Engineering, Korea University
***Graduate School of Convergence IT, Korea University

요 약

국내 공공도서관에서는 잘못 분류된 도서의 서가(bookshelf) 배치로 인해 이용자의 불편과 해당 도서관의 도서분류체계와의 불일치 등으로 도서관리에 어려움을 겪고 있다. 또한 자동 도서분류를 위한 기계학습 등 다양한 알고리즘의 연구가 진행되어 왔으나 적은 학습데이터에서의 분류효과 향상에 한계가 있었다. 이에 이 연구에서는 KORMARC(Korea Machine Readable Cataloging) 의 색인어(키워드) 정보를 결합한 확장된 나이브베이지안 알고리즘을 제안하였다. 색인어 정보는 일반적으로 도서검색시스템에서 검색 효과를 높이기 위해 이용되고 있으며 실제 공공도서관에서의 실험을 통해 도서량이 적은 경우에 보다 높은 분류효과를 얻을 수 있음을 실험 평가하였다.

1. 서론

국내 대부분의 공공도서관에서는 미국과 유럽에서 널리 사용되고 있는 듀이십진분류법(Dewey Decimal Classification) 을 우리나라 실정에 맞추어서 만든 한국십진분류법(Korean Decimal Classification, KDC) 을 도서분류체계로 이용하고 있다. 도서분류체계의 궁극적 목적은 자료의 체계적 배치를 통한 이용자의 편의성을 향상시키고 같은 자료가 같은 분류번호에 배정이다[1]. 도서관에서 분류업무의 주체가 도서납품업체인 경우가 많아 해당 도서관의 분류체계와의 불일치 등으로 사서의 검수가 필요하다[2]. (표 1) 은 동일한 ISBN(International Standard Book Number) 의 도서가 도서관에 따라 서로 다른 도서분류체계를 나타내고 있다. 한편 방대한 자료를 수작업으로 분류하기가 불가능한 상황으로 자동 문서분류를 위한 기계학습 알고리즘의 다양한 적용방법과 성능개선에 대한 연구가 진행되고 있다.

도서분류에 있어 기계학습 알고리즘은 해당 도서관의 도서분류체계를 기계학습을 통하여 유지할 수 있는 장점이 있다. 하지만 학습데이터(도서량) 가 적을 경우 정확도가 떨어지는 단점으로 국내 공공도서관의 약 48% 인 2 만권이하의 작은(소규모) 도서관에서의 적용에 문제점이 있다[3].

(표 1) 도서분류체계 현황

도서관	총도서수	동일 ISBN 도서수	도서분류체계가 다른 도서수
E 구립도서관	249,765	135,767	15,977 (11.7%)
K 교육청도서관	339,767		

이 논문에서는 기계학습을 이용한 자동 도서분류 연구를 통해서 기계학습의 실용적 한계점을 극복하고 보다 정확한 분류서비스를 수행할 수 있는 확장된 나이브베이지안 알고리즘을 제안한다. 제안방법은 도서분류의 특성인자를 적용한 확장된 나이브베이지안 알고리즘을 제안하고 다양한 도서량의 실험을 통해 국내 공공도서관에서의 도서분류의 성능 향상 방안을 제시한다.

이 논문의 구성은 다음과 같다. 제 2 장에서는 관련 연구에 대해 언급하고 제 3 장은 이 논문에서 제안하는 알고리즘을 정의한다. 제 4 장은 제안한 알고리즘의 실험과 정량적 평가를 기술한다. 제 5 장은 결론 및 향후 연구에 대해 기술한다.

이 논문은 2014 년도 정부(미래창조과학부) 의 재원으로 한국연구재단-차세대정보・컴퓨팅기술개발사업의 지원을 받아 수행된 연구임(No.2012M3C4A7033346).
† 교신저자

2. 관련 연구

나이브베이지안은 베이즈 이론(Bayes' Rule) 에 기초한 확률모델의 기계학습 알고리즘이다[4]. 확률모델의 기본형식은 문서에 속한 자질(feature) 이 클래스에 속할 확률을 계산하여 가장 높은 확률 값의 클래스를 찾아내는 방법으로 클래스에 속할 확률은 문서에 출현하는 자질의 빈도로 나타내는 것을 기본으로 한다[5]. 자질이란 문서 내에 존재하는 모든 용어에서 불용어와 일반적으로 자주 쓰이는 단어를 제외한 색인이 가능한 용어를 말한다. 하지만 단순한 자질의 빈도만으로 클래스를 결정하기 부족한 점이 발생하여 많은 연구에서 자질 선별에 대한 다양한 연구가 진행되고 있다.

확률기법을 이용한 자동 문서분류시스템에서는 유즈넷의 7,224 개 학습집단과 3,106 개 평가문서의 나이브베이지안 뉴스그룹 분류 실험에서 자질의 수가 가장 많은 8 만에서 약 88%의 정확도를 나타냈다[5]. 기존의 단순한 용어 빈도에 의존하던 방식에서 역범주 빈도 가중치, 역문헌 빈도 가중치, 용어 가중치 등 다양한 가중치를 적용하여 역문헌 빈도 가중치 실험에서 좋은 결과를 나타냈지만 2 만이하의 자질 실험에서는 평균 82%의 낮은 성능을 보인다.

베이지안 학습을 이용한 문서의 자동분류 연구에서는 유즈넷의 20 개 범주별로 1,000 개 학습문서의 뉴스그룹 분류 실험으로 약 77% 의 정확도를 나타냈다[6]. 이 연구의 특징은 학습문서에 3 번 이상의 같은 단어가 나타난 특성단어의 수를 중요한 요인으로 설정하였으나, 문서명에 의한 단어 빈도만으로는 확률모델 알고리즘의 한계를 보인다.

선행 연구에서는 분류효과를 얻기 위해 통계적 기법을 이용한 단순 빈도(term frequency) 와 역문헌 빈도(inverted document frequency) 등 다양한 가중치 부여 방법을 주로 이용하였다. 하지만 학습데이터의 양에 따라 단순 빈도 정보만을 이용한 방법은 성능 및 분류효과에 한계가 있다. 자동 도서분류에 있어서는 사용자가 요구하는 질의정보만을 검색해주는 정보검색 시스템과는 달리 명확한 분류 성능을 제시해야 하며 특히 확률기반 알고리즘의 단점인 자질의 수가 적을 경우의 낮은 성능을 극복하려는 방안이 요구되었다.

이 논문에서는 도서분류의 성능 향상을 위한 방법으로 일반적으로 도서검색시스템에서 색인어로 이용되고 있는 키워드 정보를 도서분류의 특성인자로 인식하고 도서명과 해당 도서의 키워드를 결합한 확장된 나이브베이지안 알고리즘을 제안한다.

3. 확장된 나이브베이지안 알고리즘

이 장에서는 색인어 키워드정보를 이용한 확장된 나이브베이지안 알고리즘을 제안한다.

국립중앙도서관에서는 도서관 업무의 기계화가 급속하게 진전되고 있는 상황에서 컴퓨터가 목록으로 작성된 레코드(cataloging record) 를 정보로 읽어 해석할 수 있도록 하기 위해 1980 년 처음으로 한국문헌자동화목록(KORMARC) 을 개발했다. 이후 수정과 보완하여 1993 년 KS(국가표준) 로 제정하여 전국 도서관이 표준화된 방법으로 문헌정보 데이터베이스를 구축하였다[7]. KORMAC 에는 도서목록 작성 규칙에 의해 (표 2) 의 0XX~9XX 태그(tag) 로 구성되며, (그림 1) 의 653 태그에는 해당 도서의 키워드 정보를 포함하고 있다. 예를 들어 학위논문의 경우 저자가 제시한 “핵심어(키워드)” 등이 기술되어 있다. 653 태그정보는 공공도서관 표준자료관리시스템(KOLAS) 의 도서검색용으로 생성된 IDX_BO_TBL 에 존재하고 있다.

(표 2) 태그 구성도

Tag No.	내 용
0XX	제어정보
1XX	기본표목
2XX	서명과서명관련사항
3XX	형태사항
4XX	총서사항
5XX	주기사항
6XX	주제명 부출표목
7XX	부출표목, 연관저록
8XX	총서명 부출표목
9XX	자관용 필드

MARC 정보 리스트

tag	ind	contents
001		KMO200241690
005		20140110173916
008		020531s2013 ulk 000af kor
020		a9788970121277(2)g03830:c\12000
020	1	a9788970121734(세트)
040		a111042c111042
041	1	akorhjpn
056		a833.625
090		a833.6bㅁ666ㅌc2
245	00	a태엽 감는 새.n2;p예언하는 새 편/d무라카미 하루키 지음;e윤성원 옮김
246	11	aWind-up bird chronicle
246	19	aねじまき鳥クロニクル
260		a서울:b문학사상,c2013
300		a383 p.;c20 cm
500		a무라카미 하루키의 한자명은 '村上春樹'임
653		a태엽a새a삶a의미a일본소설
700	1	a무라카미 하루키
700	1	a윤성원
740	2	a예언하는 새 편
900	10	a촌상춘수
900	10	aMurakami, Haruki
950	0	b\12000
049	0	lEM0000006565v2lEM0000006566v2c2lEM0000127836v2c3

(그림 1) 마크 데이터 예제

기존방법에서는 도서명을 구성하는 단어의 단순 빈도에 의존하여 도서분류명을 결정함에 따라 분류정확도 향상에 한계가 있다. 제안방법은 해당 도서의 핵심어인 키워드 정보가 결합되므로 분류정확도 향상에 도움을 줄 수 있다.

기존의 도서명에 의한 나이브베이지안 계산식은 분류하고자 하는 도서의 분류 가능한 클래스(class) 들 가운데 가능성이 가장 높은 확률의 클래스를 찾는 것으로 클래스(c) 가 다수이고, 도서명이 n 개의 단어 (w_1,w_2,…, w_n) 일 경우 가장 가능성이 높은 클래스의 확률은 식 (1) 과 같다[8].

$$C_{MAP} = \underset{c \in C}{\text{Arg max}}\ P(c) \prod_{i=1}^{n} P(w_i \mid c) \quad \text{- (1)}$$

도서명과 키워드를 결합한 제안방법의 계산식은 식 (1) 의 나이브베이지안 알고리즘을 기본형식으로 도서명이 n 개 단어(w_1,w_2,…, w_n) 이고, 키워드가 m 개의 단어(kw_1,kw_2,…, kw_m) 일 경우의 가장 가능성이 높은 클래스의 확률은 식 (2) 와 같다.

$$C_{MAP} = \underset{c \in C}{\text{Arg max}}\ P(c) \{ \prod_{i=1}^{n} P(w_i \mid c) + \prod_{j=1}^{m} P(kw_j \mid c) \} \quad \text{- (2)}$$

이 연구의 확장된 나이브베이지안 알고리즘은 기존의 도서명에 의한 도서분류 방법에서 키워드 정보를 결합하여 분류하고자 하는 클래스의 확률을 높이데 있다. 따라서 보다 향상된 분류효과를 높이기 위해 도서의 저자(author), 출판사(publisher) 정보 등을 결합하여 확장이 가능하다.

4. 실험 및 평가

4.1 실험방법

제안 알고리즘의 도서분류 실험은 2014 년 1 월부터 2 월까지 서울시 소재의 E 구립도서관에서 진행되었다. (표 3) 은 E 구립도서관의 KDC 분류별 소장도서 현황이다.

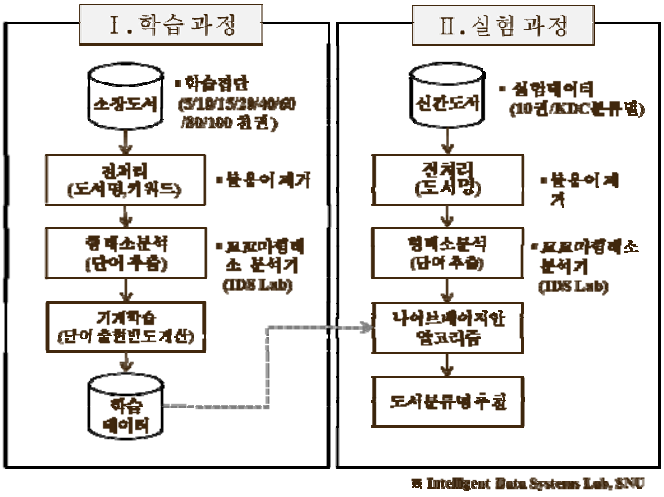
(표 3) E 구립도서관의 소장도서 collection

총류	철학	종교	사회과학	언어
9,926	11,338	7,396	42,693	16,598
자연과학	기술과학	예술	문학	역사
19,646	16,549	12,141	93,177	20,301

(총 249,765 권)

기계학습 확률모델 알고리즘은 사전확률(prior probability) 에 의존하여 새로 입력되는 문서에 사후확률(posterior probability) 을 수행한다. (그림 2) 의 학습과정과 실험과정으로 이루어진다[9].

학습과정에서는 E 구립도서관의 소장도서 약 25 만권중에서 8 개 학습집단(5 천권, 1 만권, 1.5 만권, 2 만권, 4 만권, 6 만권, 8 만권, 10 만권) 으로 구성하였다. 학습집단별 도서는 도서명과 키워드의 전처리과정과 형태소분석(꼬꼬마형태소분석기[10] 사용) 으로 자질을 추출한다. 추출된 자질은 기계학습에 의해 (표 4) 의 학습데이터를 생성한다. C 는 분류하고자 하는 c 개의 클래스이고, w_{cn} 은 c 번째 클래스에 해당하는 n 번째 도서명의 자질의 빈도(혹은 가중치), kw_{cm} 은 c 번째 클래스에 해당하는 m 번째 키워드의 자질의 빈도(혹은 가중치) 이다.



(그림 2) 도서분류 실험과정

(표 4) 학습데이터 구조

클래스	도서명				키워드			
	w_1	w_2	…	w_n	kw_1	kw_2	…	kw_m
C_1	w_{11}	w_{12}	…	w_{1n}	kw_{11}	kw_{12}	…	kw_{1m}
C_2	w_{21}	w_{22}	…	w_{2n}	kw_{21}	kw_{22}	…	Kw_{2m}
:	:	:	:	:	:	:	:	:
Cc	w_{c1}	w_{c2}	…	w_{cn}	kw_{c1}	kw_{c2}	…	kw_{cm}

도서명과 키워드의 자질은 제안 알고리즘의 분류효과에 미치는 영향을 알아보기 위해 (표 5) 와 같이 도서명 자질의 빈도 가중치(α) 와 키워드 자질의 빈도 가중치(β) 를 적용한다.

(표 5) 가중치 적용방법

실험	가중치	설명
E1	α=0 , β=1	키워드만을 적용한 실험
E2	α=1 , β=0	도서명만을 적용한 기존방식의 실험
E3	α=1 , β=0.1~1	키워드의 가중치를 단계적으로 증가한 실험

실험과정에서는 E 구립도서관의 2013 년 10 월부터 12 월까지 구입한 약 4,800 여권(비도서 제외) 의 신간도서 중에서 KDC 분류별 10 권씩 총 100 권을 실험데이터로 사용한다. 실험데이터의 도서명 전처리과정과 형태소분석을 걸쳐 학습과정에서 생성된 학습데이터(사전확률) 를 이용하여 실험데이터의 도서분류명(사후확률) 을 추천한다. 제안 알고리즘의 실험환경은 다음과 같다.

- 서버컴퓨터 : CPU Intel® CORE™ i7-3930k 3.20GHz, 16GB RAM, Windows7 64Bit
- 데이터베이스 : ORACLE 10g (공공도서관 표준자료관리 시스템, KOLASⅢ)
- 개발툴 : JDK 1.7

성능평가 방법은 신간도서의 실제 분류명과 추천된 도서분류명의 정분류율(correct classification rate) 을

이용하였으며, 식(3) 과 같다.

$$\text{정분류율} = \frac{\text{정분류된 신간도서의 수}}{\text{총 신간도서의 수}} \quad \text{- (3)}$$

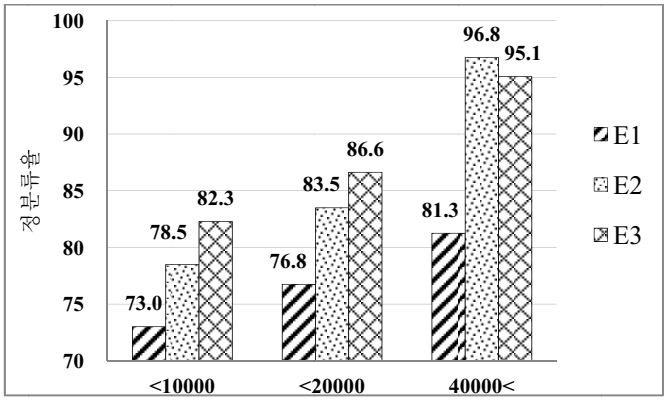
4.2 평가

제안 알고리즘의 가중치 실험 결과는 (표 6) 과 같다.

실험 E1 은 도서명을 제외한 키워드만으로 실험한 결과로 전체 실험에서 가장 낮은 분류정확도를 나타내어 키워드만으로는 도서분류에 한계가 있다. 실험 E2 는 기존방식의 도서명만으로 실험한 결과로 4 만권 이상(40000<) 에서 평균 96.8% 와 10 만권에서 98%의 높은 정확도를 나타냈다. 실험 E3 은 키워드 가중치를 도서명의 가중치와 동일한 수준까지 단계적으로 실험한 결과로 (그림 3) 의 2 만권이하(<20000) 에서 기존방식의 실험 E2 보다 평균 3.1% 향상되었고, 1 만권이하(<10000) 에서 3.8% 향상되었다. 이는 제안 알고리즘이 적은 도서량에서 기존방식보다 우수한 것으로 평가되었다.

(표 6) 실험별 결과표

Data Sets / Method	5,000	10,000	15,000	20,000	40,000	60,000	80,000	100,000
E1	70%	76%	79%	82%	84%	80%	79%	82%
E2 (기존방식)	76%	81%	88%	89%	96%	97%	96%	98%
E3	79%	85.6%	91.2%	90.7%	94.8%	95.2%	94.2%	96.1%
정확도 향상률	+3%	+4.6%	+3.2%	+1.7%	-1.2%	-1.8%	-1.8%	-1.9%



(그림 3) 실험별 성능비교

5. 결론

이 논문에서는 제안 알고리즘의 도서분류 실험을 통하여 확률기반 알고리즘의 도서분류 효과는 학습 도서의 자질보다는 학습 도서량에 크게 의존되고 있음을 알 수 있었다. 제안 알고리즘의 4 만권 이상(40000<) 실험에서는 기존방식만으로도 높은 분류정확도를 나타내어 도서분류의 특성인자(키워드) 가 분류효과에 크게 도움을 주지 못하고 있다. 이는 실험 대상의 공공도서관이 이미 체계화된 도서분류체계를 유지하고 있어 도서분류 정확도가 높게 나타나고 있음을 추정케도 한다. 하지만 2 만권이하(<20000) 실험에서는 기존방식보다 평균 3.5% 향상되었다. 이는 제안 알고리즘이 도서량이 적은 작은(소규모) 도서관에서 기존방식보다 높은 분류효과를 얻을 수 있음을 실험 평가하였다.

이 연구에서는 공공도서관의 실제 소장도서를 실험데이터로 이용하였으나, 일개 공공도서관 에서의 한정된 실험이며 또 다른 도서 환경에서의 일관성을 단정하기는 어렵다. 따라서 좀더 다양한 도서 환경에서의 검증이 필요하다.

참고문헌

[1] 윤희윤, “공공도서관 분류오류의 실증적 분석과 대안” 한국도서관·정보학회논문지, Vol.34, No.1, 2003.

[2] 최인성, “학교도서관 자료분류의 문제점 및 개선방안”, 계명대학교 교육대학원 논문집, 2010.

[3] 도서관정보정책기획단, “한국도서관연감 2012” 문화체육관광부·한국도서관협회, 2012.

[4] D.Michie,D.J. Spiegelhalter,C.C. Tayer, “Machine Learning, Neural and Statistical Classification” Ellis Horwood, 1994.

[5] 이경찬, “확률기법을 이용한 자동 문서분류시스템” 국민대학교대학원 전산과학학과 논문집, 2004.

[6] 김진상,신양규, “베이지안 학습을 이용한 문서의 자동분류” 한국데이터정보과학회논문지, Vol.11, No.1, 2000.

[7] 국립중앙도서관, ”한국문헌자동화목록형식(KORMARC)” 국립중앙도서관, 2005.

[8] 김인철,조수선, “북 마크 자동 분류를 위한 학습 에이전트” 정보처리학회논문지, Vol.8-B, No.5, 2001.

[9] 박찬정,김기용,성동수,이건배, “KNN 알고리즘을 이용한 특허문서의 다중 IPC 자동분류” 정보과학회논문지, pp.1502-1504, 2013.

[10] 이동주,연종흠,황인범,이상구, “꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구” 정보과학회논문지, Vol.16, No.11, 2010.