목차, 책 소개를 이용한 단행본 문서 범주화에 관한 기초연구

A preliminary Study on Text Categorization of Book using Table of Contents and Book Description

도현호, BK21플러스팀, 계명대학교 문헌정보학과, loverdorpo@kmu.ac.kr
이용구, 계명대학교 문헌정보학과, yonggulee@kmu.ac.kr
Hyun-Ho Do, BK21 Plus, Dept. of LIS, Keimyung University
Yong-Gu Lee, Dept. of LIS, Keimyung University

이 연구에서는 도서관의 주요 장서에 해당하는 단행본 도서에 대한 자동 분류를 적용가능한지 알아보고자 하였다. 분류자질로 메타데이터인 서명, 목차, 책 소개를 사용하였으며, 다양한 자질 가중치를 적용하여 581건의 단행본 도서를 통해 kNN 분류기의 분류성능을 파악하였다. 실험 결과 이들 메타데이터를 모두 사용하였을 때 가장 좋은 분류성능을 가져왔으며, 실험문헌집단의 규모가 작은 한계가 있지만 로그 TF를 취한 가중치 방법이 좋은 성능을 가져왔다.

1. 서 론

현대사회는 과학과 기술의 발전으로 많은 양의 정보가 증가하고 있으며, 많아진 정보를 수작업으로 분류하는 것은 어려워지고 있다. 따라서 1960년대부터 분류 알고리즘에 의해 대상물들을 유사한 패턴을 갖는 것 끼리 모아 집단화 하는 자동분류(정영미, 2011, p.166)를 통해서 분류하는 것을 연구하고 있다.

정보가 증가함에 따라 도서관에서도 소장하는 자료의 양도 급격하게 증가하고 있으며, 서명, 저자명, 출판사정보, 책 소개, 책 리뷰 등 자료와 관련된 정보도 증가하고 있다.

그러나 대부분의 도서관에서는 많아지는 자료의 정보들을 자동분류를 통해 활용하거나 이용자에게 제공하지는 않으며, 그에 따른 연구는 미미한 상태이다.

본 연구에서는 자동분류에서 범주정보가 있어 지도학습에 해당하는 범주화 기법을 이용하여 단행본 메타데이터들을 자동분류 하였을 때, 이들 메타데이터에 따른 단행본의 자동분류 성능을 알아보고자 하였다.

2. 선행연구

자동분류의 연구는 50년 동안 이루어졌으며, 문헌정보학 분야에서도 자동분류 연구가 많이 이루어 졌다. 하지만 이를 도서관이과 관련 분야와 결합한 연구는 많지 않았다.

클러스터링 기법을 적용한 방법으로, 노정순(2004)은 문헌정보학분야의 단행본과 학위논문을 대상으로 245필드를 색인하여 DDC 분류와 결과가 유사한 기법을 발견하였다고 주장하였다. Enser(1985)는 서명, 목차, 권말색인의 단어를 이용하여 실험하였다. 도현호와 이용구(2014)는 서명, 목차, 책 소개 메타데이터를 이용한 클러스터링에서 모든 메타데이터 사용이 가장 좋은 성능을 가져왔다.

텍스트 범주화 기법을 적용한 연구인 Pong 등(2007)은 도서관 환경에서 사서의 수작업 분류를 지원하기 위한 방법으로 전자자원을 kNN 분류기와 naive Bayes 분류기를 사용하여 실험하였으며, 분류범주로 LCC 분류기호를 적용하여 분류성능 평가를 수행하였다.

이 연구에서는 서명, 목차, 책 소개를 이용하여 단행본을 범주화를 하였을 때 어떤 메타데이터를 이용하면 좋은 성능을 발휘하는지 알아보고자 한다.

3. 데이터분석

3.1 실험대상 및 방법

이 연구의 데이터 수집을 위해서 대학도서관이 지난 1년간 수서한 단행본을 추출하였으며, 이중에서 사회과학 분야의 한국어로 된 단행본을 실험 데이터로 추출하였다. 추후 자동분류에 필요한 수작업 범주를 부여하기 위해 해당 도서의 분류기호(KDC)를 국립중앙도서관에서 검색하여 KDC 분류기호를 추출하였다. 또한 인터넷 서점 교보문고에서 해당 도서를 검색하여 목차와 책 소개 메타데이터를 수집하였다.

KDC 분류기호, 목차, 책 소개 정보가 모두 있는 단행본 621권을 선정하였고, 이 중 강목이 50건 이상인 다섯 범주(강목 320/256건, 330/108건, 340/61건, 360/87건, 370/69건)를 대상으로 총 581권을 문헌 범주화를 위한 최종 실험 데이터로 선정하였다.

실험 데이터를 대상으로 울산대학교 UTagger (http://nlplab.ulsan.ac.kr)를 이용하여 한국어 형태소 분석을 하였고, 통계용 패키지가 포함된 R 프로그래밍 언어를 이용하여 명사를 추출하여 문헌×용어 행렬을 구축하였다.

문헌×용어 행렬을 이용하여 다수의 선행연구를 통하여 가중치로는 단순빈도(TF), 이진 TF(binTF), 로그 TF(logTF), 로그 TF*IDF (logTFIDF)를 적용하였다. 또한 kNN 기법은 R언어의 class 패키지의 knn 분류기를 사용하여 자동분류를 수행하였다. 이 분류기는 유사도로 유클리디언(Euclidean) 거리계수를 사용한다. 또한 범주 결정은 가장 많은 이웃문헌들이 속한 범주를 선택하는 방법(majority vote)을 쓴다.

분류기 성능평가로 매크로 평균 정확률 및 매크로 평균 재현율 기법과, 이들의 단일가척도인 매크로 평균 F_1 척도를 사용하였다.

3.2 실험결과 분석

실험 데이터의 특성을 살펴보기 위해 실험문헌의 메타데이터에 따른 출현한 분류 자질 수와 출현빈도를 정리하면, <표 1>과 같다.

<표 1> 실험문헌의 메타데이터 자질 수와 출현빈도

	서명	목차	책 소개	서명+목차	서명+책소개	목차+책소개	전체
자질 수	1,351	13,073	4,728	13,208	4,852	14,053	14,098
출현빈도	3,061	109,523	25,026	112,584	28,087	134,549	137,610

서명, 목차, 책 소개와 같은 단일 메타데이터에서는 목차가 가장 많은 분류 자질을 가지며, 출현빈도도 가장 많은 수가 출현하였으며, 이들 단일 메타데이터를 결합한 메타데이터에서는 전체가 14,098개로 가장 많았다.

이 연구에서는 텍스트 문헌에 많이 사용하는 분류기인 kNN 분류기를 적용하였다. 이 분류기는 새로 입력된 문헌(검증문헌)을, 분류기 구축시에 적용된 학습문헌과 유사도를 비

교하여 가장 유사도가 높은 k개의 최근접 이웃문헌을 찾아내고 이들이 가진 범주 정보를 이용하여 검증문헌의 범주를 결정한다. 특히 이때 k의 값에 따라 분류기의 성능이 달라지므로 최적의 값을 가져오는 k값을 미리 결정해야 한다.

사전 실험을 위해 가중치 방법은 4가지(TF, binTF, logTF, logTFIDF)를 적용하였다. 메타데이터의 경우 결합 여부에 따라 단일 3가지(서명, 목차, 책 소개), 일부 결합 3가지(서명+목차, 서명+책 소개, 목차+책 소개), 그리고 모두 결합한 전체 등 총 7가지로 조합하였다. 이들 28개 사례를 대상으로 오분류율을 적용하였다. 오분류율은 수작업 범주정보와 분류기의 자동분류 결과를 비교하여 일치하지 않는 정도를 비율로 나타낸 값이다. 이는 잘못 분류된 것을 말하기에 작은 값이 좋은 성능을 뜻한다.

k값을 결정하기 위한 사전 실험으로, 28개의 경우에 해당하는 오분류율을 분석하였으며, 그 중 최소값에 해당하는 k값의 수를 계산하여, <표2>를 얻었다. 여기에는 동률의 최소값을 갖는 경우가 8건이 포함되어 있다.

<표 2> k 값에 따른 오분류율 최소값 빈도

k	1	2	3	4	5이상	전체
최소값 빈도	**16**	4	12	4	0	36

실험 결과 581건의 사회과학 분야의 단행본 도서를 수작업 범주인 KDC 분류기호에 따른 kNN 분류기의 분류 성능을 계산하였다. 성능 평가 방법으로 매크로 평균 정확률, 매크로 평균 재현율, 매크로 평균 F1 척도를 사용하였으며, 그 결과 각각 <표 3>, <표 4> 그리고 <표 5>를 얻었다.

<표 3>의 매크로 평균 정확률의 성능을 살펴보면, 가장 좋은 성능을 가져온 경우는 logTFIDF에서 서명+목차로 0.7665이며, 다음으로 목차+책소개(0.6752), 전체(0.6732) 순으로 나타났다. 단일 메타데이터만을 살펴보면, 서명의 매크로 평균 정확률이 큰 차이로 더 좋은 성능을 보이는 것을 알 수 있다. 가중치 측면에서는 평균적으로 logTFIDF가 0.6211로 가장 좋은 성능을 보였으며, logTF(0.6016), TF(0.5692) 순으로 나타났다. 다만 단일 메타데이터에서 자질의 수가 적은 TF의 서명이 제일 좋은 성능을 보였으며, 자질의 수가 많아지는 결합 메타데이터에서는 logTFIDF의 서명+목차가 좋은 성능을 보였다는 부분을 눈여겨 볼 필요가 있다.

<표 4>의 매크로 평균 재현율의 성능을 살펴보면, logTF의 목차+책소개가 0.5805로 가장 좋았으며, 근소한 차이로 전체(0.5793), TF의 전체(0.5518) 순으로 나타났다. 대체로 전체 메타데이터를 사용하는 것이 더 좋은 성능을 보였으며, 단일 메타데이터에서는 서명이 더 좋은 성능을 보였다. 가중치 측면에서는 logTF와 TF가 좋은 성능을 보였다.

일반적으로 정확률과 재현율은 상반된 결과를 가져오는 경향이 있기 때문에 단일 척도가 필요하며, 이 연구에서도 매크로 평균 F_1 척도를 적용하였다. <표 5>를 보면, 단일 척도를 사용한 성능평가에서는 logTF를 적용하고 전체 메타데이터를 사용한 경우가 0.5966으로 가장 좋았으며, logTF의 목차+책소개(0.5874), TF의 전체(0.5744) 순으로 나타났다. 메타데이터 측면에서는 주로 전체를 사용하는 방법이 좋은 성능을 보였으며, 목차+책소개와 서명 순으로 나타났다. 가중치 측면에서는 logTF와 TF가 좋은 성능을 보였다.

4. 결 론

이 연구에서는 도서관에서 수집하는 단행본 도서의 메타데이터 중에 서명을 추출하고 이

<표 3> 매크로 평균 정확률

가중치	서명	목차	책 소개	서명+목차	서명+책소개	목차+책소개	전체	평균
TF	0.6523	0.4800	0.5693	0.5397	0.5632	0.5808	0.5991	0.5692
binTF	0.6371	0.5313	0.4673	0.5449	0.4928	0.4885	0.5056	0.5239
logTF	0.6479	0.5676	0.5454	0.6463	0.5948	0.5945	0.6149	0.6016
logTFIDF	0.5858	0.6149	0.4557	**0.7665**	0.5767	0.6752	0.6732	0.6211

<표 4> 매크로 평균 재현율

가중치	서명	목차	책 소개	서명+목차	서명+책소개	목차+책소개	전체	평균
TF	0.4947	0.4406	0.5274	0.4713	0.5136	0.5552	0.5518	0.5078
binTF	0.4577	0.3340	0.4101	0.3893	0.3304	0.4380	0.4479	0.4011
logTF	0.4899	0.4860	0.4794	0.4826	0.4625	**0.5805**	0.5793	0.5086
logTFIDF	0.4979	0.3568	0.3141	0.3812	0.3072	0.3964	0.4460	0.3856

<표 5> 매크로 평균 F_1

가중치	서명	목차	책 소개	서명+목차	서명+책소개	목차+책소개	전체	평균
TF	0.5627	0.4594	0.5475	0.5032	0.5372	0.5677	0.5745	0.5360
binTF	0.5327	0.4102	0.4368	0.4541	0.3956	0.4619	0.4750	0.4523
logTF	0.5579	0.5237	0.5103	0.5526	0.5204	0.5874	**0.5966**	0.5498
logTFIDF	0.5383	0.4516	0.3719	0.5092	0.4009	0.4995	0.5365	0.4725

에 도서 관련 기관에서 목차와 책 소개를 수집하여 이들의 조합을 통한 자동분류 성능을 알아보고 도서의 자동 분류에 적용가능성을 알아보았다.

실험결과를 정리하면, 서명을 비롯하여 목차와 책 소개를 모두 포함한 자동분류가 가장 좋은 성능을 보여 이들의 이용가능성을 확인할 수 있었으며, 가중치 측면에서는 로그를 취한 단어빈도가 좋은 성능을 보였다. 다만 이 연구의 경우 일반적인 자동분류 실험보다 적은 실험집단을 적용하여 추가 보완할 필요성이 있다.

참고문헌

노정순 (2004). OPAC에서 자동분류 열람을 위한 계층 클러스터링 연구. 정보관리학회지, 21(1), 93-116.

도현호, 이용구 (2014). 목차와 책 소개 정보를 이용한 단행본 클러스터링에 관한 기초연구. 2014년도 한국도서관·정보학회 하계학술발표대회, 175-180.

정영미 (2011). 정보검색연구 (증보판). 서울: 연세대학교 출판문화원.

Enser, P. G. B. (1985). Automatic classification of book material represented by back of book index. *Journal of Documentation*, 41(3), 135-155.

Pong, J., et al. (2007). A comparative study of two automatic document classification methods in a library setting. *Journal of Information Science*, 34(2), 213-230.