메타데이터를 활용한 조사자료의 문서범주화에 관한 연구

An Exploratory Study on Survey Data Categorization using DDI metadata

박자현, 연세대학교 문헌정보학과, parkja04@ksdc.re.kr
송민, 연세대학교 문헌정보학과, min.song@yonsei.ac.kr
Ja-Hyun, Park, Dept. of Library & Information Science, Yonsei University
Min, Song, Dept. of Library & Information Science, Yonsei University

본 연구는 DDI 메타데이터를 활용하여 귀납적 학습모델(supervised learning model)의 문서범주화 실험을 수행함으로써 조사자료의 체계적이고 효율적인 분류작업을 설계하는데 그 목적이 있다. 구체적으로 조사자료의 DDI 메타데이터를 대상으로 단순 TF 가중치, TF-IDF 가중치, Okapi TF 가중치에 따른 나이브 베이즈(Naive Bayes), kNN(k nearest neighbor), 결정트리(Decision tree) 분류기의 성능비교 실험을 하였다. 그 결과, 나이브 베이즈가 가장 좋은 성능을 보였으며, 단순 TF 가중치와 TF-IDF 가중치는 나이브 베이즈, kNN, 결정트리 분류기에서 동일한 성능을 보였으나, Okapi TF 가중치의 경우 나이브 베이즈에서 가장 좋은 성능을 보였다.

1. 서론

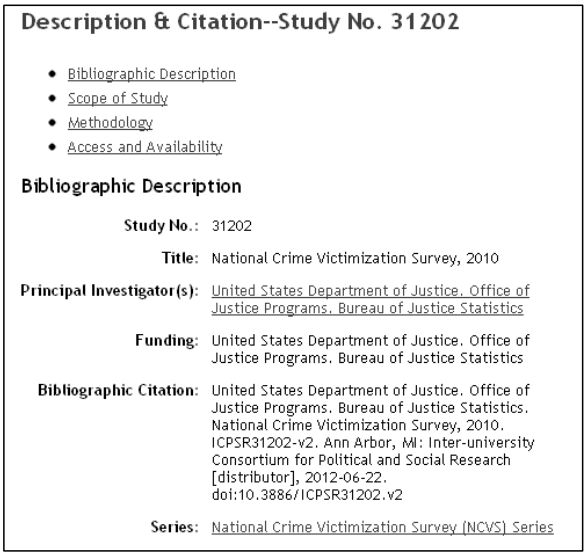
문서범주화는 문서의 내용을 바탕으로 미리 정의된 범주를 하나 이상 문서에 부여하는 것이며, 수작업으로 문서를 분류하는 데 소요되는 시간과 노력, 비용을 감소시키기 위하여 자동 문서범주화 방법이 등장하였다(Yang, and Pedersen 1997). 자동 문서범주화 방법은 효율적인 정보의 조직 및 검색을 가능하게 할 뿐만 아니라 최근에는 문서 필터링이나 스팸메일 필터링, 뉴스 기상 주제선정, 웹 문서 분류 등의 응용분야로 확대·적용되고 있다.

본 연구에서 다루고자 하는 조사자료 메타데이터를 활용한 문서범주화는 일종의 웹 문서 분류에 해당한다. 사회조사자료를 전문적으로 아카이브하여 웹 DB형태로 제공하고 있는 ICPSR (Inter-university Consortium for Political and Social Research)은 조사자료마다 메타데이터 표준 스키마인 DDI(Data Documentation Initiative)를 적용한 요약정보를 제공하고 있다. 그러나 국제적인 표준 적용이 권장되는 조사자료 메타데이터의 기술과 달리, 표준화된 분류체계와 분류표는 부재한 실정이다. 이에 따라 국내 데이터 아카이브의 경우, 해당 조사자료의 메타데이터를 참조하여 분류자가 임의로 분류하고 있는 실정이며, 조사자료의 양적 증가로 인해 수작업 분류에 상당한 시간과 인력이 투입되어야 한다. 본 연구는 DDI 메타데이터를 활용하여 귀납적 학습모델의 문서범주화 실험을 수행함으로써 조사자료의 체계적이고 효율적인 분류작업을 설계하는데 그 목적이 있다.

2. 조사자료의 DDI 메타데이터

본 연구에서 조사자료 문서범주화에 활용하는 DDI 메타데이터는 크게 4가지 영역으로 구분할 수 있다. 'Bibliographic Description'은 제목, 조사자와 같은 일반 서지사항을 기술하

며, ‘Scope of Study'는 초록, 주제용어, 조사지역, 조사내용 등 연구내용에 대한 기술이다. 또한 ‘Methodology'는 샘플링, 가중치, 응답률 등 조사의 방법론을 기술하며, ‘Access and Availability'는 이용자의 자료접근에 대한 안내이다. ‘Methodology', ‘Access and Availability'는 자료의 내용과 무관한 기술적 요소(technical element)로 문서범주화에 적합하지 않은 정보이다. 따라서 본 연구의 분석대상에서 제외하였다. DDI 메타데이터의 예시는 <그림 1>과 같다.


Description & Citation--Study No. 31202

- Bibliographic Description
- Scope of Study
- Methodology
- Access and Availability

Bibliographic Description

Study No.: 31202

Title: National Crime Victimization Survey, 2010

Principal Investigator(s): United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics

Funding: United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics

Bibliographic Citation: United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. National Crime Victimization Survey, 2010. ICPSR31202-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2012-06-22. doi:10.3886/ICPSR31202.v2

Series: National Crime Victimization Survey (NCVS) Series

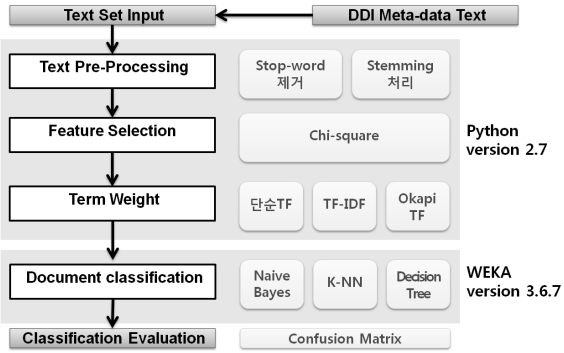
<그림 1> DDI 메타데이터 예시

3. 조사자료의 문서범주화 실험

3.1 실험내용 및 방법

본 실험대상인 DDI 메타데이터 텍스트 450건은 ICPSR의 ‘Topic classifications' 중 'Education', 'Economic Behavior and Attitudes', 'Health Care and Facilities'에 분류된 조사자료의 요약정보를 각 150건씩 수집한 것이다. 위의 세 가지 범주로 국한한 것은 다른 범주의 경우, 외연이 명확하지 않아 중복되는 자료가 많으며, 이에 따라 문서범주화를 체계적으로 수행하기에 적합하지 않다고 판단하였기 때문이다.

실험문헌집단의 수집은 2012년 5월 5일부터 2012년 6월 1일까지 수행하였으며, 수집방법은 ICPSR 조사자료의 상세정보 페이지를 복사하여 텍스트 파일로 저장하였다. 수집한 실험문헌집단은 <그림 2>와 같은 절차로 문서범주화 실험을 수행하였다.



<그림 2> 실험절차 개요도

문서범주화 실험에서 모든 단어를 문서 표현에 사용할 경우 방대한 문서벡터가 형성되어 학습의 효율성이 떨어지고 많은 학습시간이 소요되게 된다. 이에 따라 색인어 및 자질의 선정을 위하여 전처리 작업(text pre-processing)을 수행해야 하는데, 불용어(stop-word) 제거와 어간추출과정(stemming)이 포함된다.

불용어 제거는 텍스트를 구성하는 각 단어들을 분리한 후 먼저 불용어 리스트와 대조하여 비주제어를 제거하는 것이다. 이 때 제거되는 불용어들은 대개 고빈도 기능어들이다.

본 연구에서는 'the', 'of', 'and', 'about'과 같은 일반적인 불용어 리스트에 조사자료에 자주 등장하는 조사자료 불용어들을 추가하여 빈도 높은 기능어 및 주제어로서 가치 없는 기타 고빈도어를 제거하였다. 조사자료 불용어 리스트는 <표 1>과 같다.

어간추출과정은 불용어가 제거된 나머지 단어들의 어간을 추출하여 단어의 어원을 복구하는 작업을 수행한다. 본 연구에서는 가장 널

리 활용되고 있는 'Porter Stemming Algorithm'을 이용하여 단어들의 어간을 추출하고 단어의 어원을 복구하는 작업을 수행하였다. 어간추출 과정 완료 후의 단어집합은 <그림 3>와 같다.

<표 1> 조사자료 불용어 리스트

analysis	data	national	series
analyze	dataset	period	source
analyzing	datasets	population	statistical
archive	date	principal	subject
bibliography	design	project	summary
citation	format	questionnaire	survey
code	geographic	record	terms
codebook	investigator	research	time
codebooks	level	response	title
collection	levels	result	universe
collections	measure	sample	variable
coverage	method	scope	variables

abus	academ	access	achiev	acr	adai	administ	administr	admiss
admit	adult	ag	agenc	aggreg	agricultur	aid	alameda	alcohol
altern	ama	ambulatori		america	amount	andersen	annual	
antebellum		argentina	arriv	ascii	aspir	assess	asset	assist
associ	atack	athlet	attend	attitud	automobil	background		bank
baselin	bateman	bed	began	bilingu	bill	blood	bolivia	bond
british	budget	build	bureau	busi	cancer	capit	cardiovascular	
care	career	carnegi	cash	censu	center	centuri	certif	
characterist		check	child	childcar	childhood		children	china
chronic	cigarett	citizen	citizenship		class	classroom		clinic
cognit	cohort	collect	colleg	colombia	columbia	commod	commun	
communic		comorbid	compani	compet	complet	comprehens		compris
concept	condit	conduct	confidenti	congest	consent	consum	contracept	
control	counti	cours	cport	creation	criteria	croatia	curriculum	
custom	dc	deana	death	defin	degre	deliveri	democrat	
demograph		dental	depart	depositor	depress	dept	develop	diabet
diagnos	diagnosi	diagnost	diet	disabl	discharg	disciplin	diseas	disord
dispos	distribut	district	divers	doctor	dollar	domest	dropout	drug
durabl	earli	econom	economi	ecuador	educ	eighteenth		elderli
elementari		emerg	engag	enhanc	enrol	entrepreneur		
entrepreneuri		environ	epidemiolog		episod	equiti	estat	europ
evalu	exact	exactli	exam	examin	exercis	experi	explor	export
extant	facil	factor	faculti	fall	farm	field	file	financ
financi	finland	firm	fiscal	fish	fit	flexibl	fluctuat	focus
follow	foreign	foundat	francisco	frequenc	freshman	freshmen	function	fund
furnish	gavin	gener	global	goal	good	governor	grade	grader
graduat	gross	group	growth	guid	health	healthcar	hear	heart
hegi	height	high	higher	highest	hispan	histor	hiv	hmo

<그림 3> 어간 추출과정 후 단어집합 예시

문서범주화 실험은 전처리 작업 후, 학습문서에 나타나는 다양한 단어(혹은 구)들 중에서 범주화 학습에 유용하게 사용될 만한 단어를 선정하는 자질선정(feature selection)의 과정을 거치게 된다.

자질선정 기준에 따른 실험 결과에서 카이제곱 통계량(chi-square statistics)과 함께 문헌빈도(document frequency)가 좋은 분류 성능을 나타냈다(Yang, and Pedersen 1997). 그러나 문헌빈도는 적은 문헌빈도를 갖는 용어가 정보량이 많다는 정보검색의 기본가정에 상치되기 때문에 본 연구의 자질선정에서 문헌빈도를 제외하였다. 반면 카이제곱 통계량은 일반적으로 자질선정 기준을 비교 평가하는 실험에서 높은 성능을 보일 뿐만 아니라 과학적 방법으로 통계적 유의성을 판단할 수 있으므로 문서범주화에 적합한 방법이다. 따라서 본 연구에서는 자질선정 기준으로 카이제곱 통계량을 사용하였다.

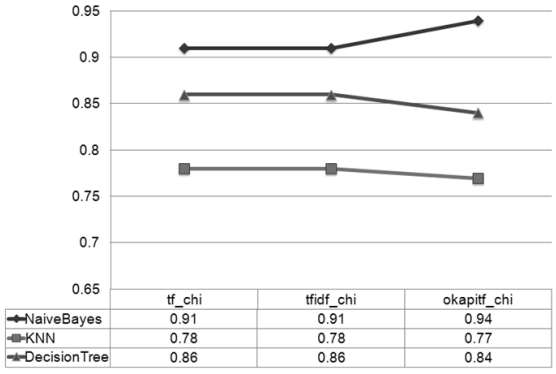
단어가중치(term weight)는 문서범주화에서 색인 과정에 이루어지는 작업으로 문서를 표현하는 단어들을 선별한 후 수치화된 점수를 매기는 방법이다. 선정된 자질을 사용하여 어떻게 문서를 더 잘 표현할 것인가에 대한 단계이며 문서의 표현은 문서범주화 분류기의 전체적인 성능에 큰 영향을 미치므로 각 문서를 학습에 적합한 형태로 표현해야 한다(이용훈 2010). 따라서 본 연구는 단순 TF 가중치와 함께, 문서범주화 분야에서 가장 많이 활용되고 있는 TF -IDF 가중치와 TF를 문헌 길이로 정규화한 Okapi TF 가중치를 사용하여 각 성능을 비교·평가하였다.

불용어 제거와 어간추출과정을 포함한 전처리 작업, 자질선정, 단어가중치부여는 모두 파이썬(python) version 2.7을 통해 수행하였다.

문서범주화 실험(document classification)은 단순 TF 가중치, TF-IDF 가중치, Okapi TF 가중치에 따른 나이브 베이즈, kNN , 결정트리 분류기의 성능비교 실험을 하였다. 학습은 k-fold cross validation 방법을 사용하였고 본 연구에서는 10-fold로 실험하였다. k-fold cross validation은 전체 집합을 k개로 나눈 뒤 하나를 다른 것들과 비교하여 전체의 평가 지표로 만들기 위해 평균을 구하며, 이를 이용해 실험 문헌집단의 검증을 수행할 수 있다(김민희, 권영식 2011). 또한 분류기의 도구는 웨카(WEKA) version 3.6.7을 사용하였다. 마지막으로 분류기의 성능평가는 다범주 분류임을 감안하여 혼동행렬(confusion matrix)을 사용하였다.

3.2 실험결과 분석

본 연구는 단순 TF 가중치, TF-IDF 가중치, Okapi TF 가중치에 따른 나이브 베이즈, kNN , 결정트리 분류기의 성능비교 실험을 하였다. 실험결과를 요약하면 <그림 4>와 같다.



	tf_chi	tfidf_chi	okapitf_chi
NaiveBayes	0.91	0.91	0.94
KNN	0.78	0.78	0.77
DecisionTree	0.86	0.86	0.84

<그림 4> 실험결과 요약

본 연구에서는 나이브베이즈가 가장 좋은 성능을 보였다. 단순 TF 가중치와 TF-IDF 가중치는 나이브 베이즈, kNN, 결정트리 분류기에서 서로 동일한 성능을 보였으나, Okapi TF 가중치의 경우 나이브 베이즈에서 가장 좋은 성능을 보였다. 보편적으로 좋은 성능을 보인다고 알려진 kNN의 경우, 조사자료의 문서범주화에서는 가장 낮은 성능을 보였다.

4. 결론

본 연구는 DDI 메타데이터를 활용하여 귀납적 학습모델의 문서범주화 실험을 수행함으로써 조사자료의 체계적이고 효율적인 분류작업을 설계하는데 그 목적이 있다. 구체적으로 단순 TF 가중치, TF -IDF 가중치, Okapi TF 가중치에 따른 나이브 베이즈(Naive Bayes), kNN (k nearest neighbor), 결정트리(Decision tree) 분류기의 성능비교 실험을 하였다. 연구의 실험결과는 다음과 같다.

첫째, 귀납적 학습 모델의 분류기 중 나이브 베이즈가 kNN, 결정트리보다 더 좋은 성능을 보였다. 보편적으로 좋은 성능을 보인다고 알려진 kNN의 경우, 조사자료 문서범주화에서는 가장 낮은 성능을 보였다.

둘째, 단순 TF 가중치와 TF-IDF 가중치는 나이브 베이즈, kNN, 결정트리 분류기에서 서로 동일한 성능을 보였으나, Okapi TF 가중치의 경우 나이브 베이즈에서 가장 좋은 성능을 보였다.

본 연구의 내용을 발전시켜 다음과 같은 후속 연구를 기대해 볼 수 있다.

첫째, 본 연구의 분류기와 단어가중치 외에 다른 기법을 적용해보는 실험을 할 수 있을 것이다.

둘째, 본 연구에서 사용된 실험문헌집단보다 많은 텍스트 문서를 사용하여 증명하는 것도 필요할 것이며, 한글 텍스트의 조사자료 메타데이터를 대상으로 적용해볼 수 있다.

셋째, 규칙 기반 모델(rule-based model)의 문서범주화 설계를 통해 수작업 분류를 효과적으로 구현할 수 있는 알고리즘을 구현함으로써 본 연구의 귀납적 학습 모델과 그 성능을 비교해 볼 수 있을 것이다.

참 고 문 헌

정영미. 2005. 정보검색연구. 서울: 구미무역 출판부.

이용훈. 2010. 단어 가중치법 적용을 통한 문서 범주화 기법에 관한 연구. 석사학위논문. 단국대학교 대학원, 전자계산학과 컴퓨터과학전공.

김민희, 권영식. 2011. 문서분류 성능 향상을 위한 단어 가중치 기법에 대한 연구. 대한산업공학회 2011년 추계학술대회 논문집: 1-23.

Yang, Yiming, and J. O. Pedersen. 1997. "A comparative study on feature selection in text categorization." Machine Learning. Proceedings of the Fourteenth International Conference (ICML97): 412-420.