대용량 문서학습을 위한 분류기 생성 및 결합방법[1)]

정도헌, 황명권, 성원경
한국과학기술정보연구원 정보기술연구실
e-mail:{heon, mgh, wksung}@kisti.re.kr

A Method for Generating and Combining Classifiers for Large Scale Data

Do-Heon Jeong, Myung-Gwon Hwang, Won-Kyung Sung
Korea Institute of Science and Technology Information

요 약

대용량 데이터 환경에의 적용이 가능한 대용량 학습기반의 자동범주화 기법과 범용적으로 사용할 수 있는 기법은 대량의 정보를 처리해야하는 정보분석 및 정보서비스 환경에 가장 필요한 기술요소라 할 수 있다. 본 논문에서는 대용량의 문서를 단위 컴포넌트로 분할하여 학습하고 이를 동적으로 결합하는 대용량 분류기 생성 기법을 소개하고 자동범주화 성능을 SVM 모델과 비교하여 봄으로써, 본 기술의 활용 가능성을 살펴보도록 한다.

1. 서론

문서분류를 위한 기존의 텍스트 마이닝 기법은 분석대상이 되는 도메인에 의존적일 뿐 아니라, 해당 도메인의 정보변경으로 인해 자질셋 특성이 추가 또는 변경이 되는 경우, 새롭게 변경된 자질정보에 대해 학습결과물인 단위 분류기를 재생성해야 하는 과정이 상시 존재한다.

그러나, 이러한 일련의 과정은 처리해야 하는 대상 데이터가 동적으로 급증하고 이를 효과적으로 분석해야 하는 업무영역에서는 커다란 장애요소가 될 수 있다. 대용량 자원의 효과적인 처리방안의 필요성은 국내외의 수많은 관련연구가 수행되고 있으며, 다양한 방법론이 제안되고 있어 그 중요성을 가늠할 수 있다[1][2][3][4][5].

대량의 논문을 다양한 클러스터링 기법을 결합하여 효과적으로 처리하고자 하는 연구[1]와 바이오 인포매틱스 분야에서 대량의 유전자 정보를 효과적으로 분석하고자 하는 연구[2], 멀티미디어(이미지) 처리를 위한 병렬 데이터 클러스터링 기법과 이를 위한 데이터 축소방법의 연구[3], 그리고, 물류, 주식정보 등 대량의 사실정보 스트림을 효과적으로 처리하고자 하는 분류기법의 연구[4] 등이 과거부터 현재까지 다양하게 진행되었다.

본 논문은 대용량의 문서를 학습함에 있어 자질축소 기법에 의존하지 않고 대량의 문서[2)]를 자유롭게 학습하고 부분적인 자질추가 변경 시에 변경요소만을 효과적으로 반영할 수 있는, 범용적이고 일반적인 분류기의 구조설계 방법에 관한 것이다.

논문의 2장에서는 분류기의 단위 컴포넌트 생성방법과 이를 동적으로 결합하는 방법에 대한 핵심사항을 설명하고, 3장에서 성능측정을 통해 활용 가능성을 살펴본 후, 향후 진행할 연구방향을 간단히 언급하고자 한다.

2. 단위 분류기 생성과 동적 결합방법

자동범주화 기법을 실제 서비스에 응용하고자 할 때, 경우에 따라서는 수백만 건 이상의 정보자원을 학습하고 해석해야 하는 경우가 있다. 한국과학기술정보연구원의 NDSL 서비스의 논문과 특허정보의 보유정보량는 7천만 건 이상에 이르며[3)], 보다 향상된 지능형 정보검색을 위해 논문단위의 분류정보를 부여하고자 노력하고 있다.

일반적으로 효율적인 문서처리를 위해 자질선택 기법을 사용하는데, 이는 정보량의 축소 뿐만 아니라 성능의 향상을 위해서도 필요한 과정으로 알려져 있다. 그러나, 대용량의 문서학습을 하는 과업에서는, 과도한 비율이상으로 자질을 제거하는 과정이 성능에 영향을 끼치게 되므로 자질선택 및 축소기법의 적용 역시 한계가 존재하게 된다.

학습문서의 수나 자질의 수에 대한 고려는 분류기의 생성효율과 관련이 있다. 최적화된 문서와 자질을 이용할

1) 관련특허: 분류기의 동적 결합에 의한 대용량 분류기 자동 생성 시스템 및 방법 (출원번호:10-2010-0099164, 출원일:2010-10-12)

2) 여기서 대량의 문서라 함은, 이론적으로는 학습데이터의 용량 제한이 없음을 의미한다.

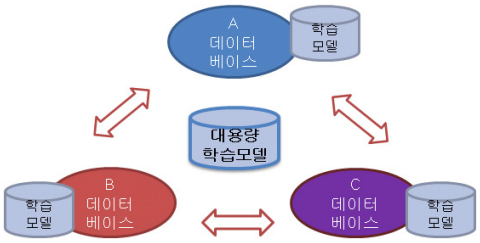
3) http://www.ndsl.kr 의 자원소개에 따르면, 해외논문 약 5천4백만 건, 국내논문 약 72만 건, 국내외 특허 약 2천2백만 건을 서비스하고 있다.

때 빠른 처리와 함께 높은 성능을 낼 수 있기 때문에 이와 관련한 여러 기법을 연구하고 적용하고 있지만, 정보량을 축소하는 보다 근본적인 이유는 분류기 생성에 소요되는 시간과 실제 메모리 점유의 문제로 인해 대량의 문서를 학습할 수 없는 제한점이 존재하기 때문이다.

문서의 기계학습에 되도록 많은 문서와 많은 자질을 이용하면 성능이 좋아진다는 기본 가설에는 변함이 없으므로, 본 기술을 통해 대용량 정보의 기계학습 시 고민해야 하는 정보의 차원축소 문제라는 제약사항을 해소할 수 있는 방법을 제공하고자 한다. 즉, 이 기법은 정보를 축소하는 것이 아니라, 실제로 대용량의 매트릭스를 생성하는 것과 작은 용량의 매트릭스를 다수 생성하여 동적으로 결합하는 두 가지 분류기 생성방법에 있어 학습결과의 수치상 차이가 전혀 없도록 하는 방법이다.

(그림 1)은 데이터베이스별로 여러 개별 분류기를 조합하는 예시이며, 개별 데이터베이스도 구성문서의 수가 많을 경우, 여러 개의 분할된 복수의 분류기로 구성하여 동적으로 결합하여 최종 분류기를 생성할 수 있다

.



(그림 1) 대용량 분류기 생성 개념

2-1. 단위 분류기의 생성과정

단위 분류기의 생성을 위해 아래와 같은 전처리 과정을 포함한 일련의 과정을 거친다.

(1) 자질 추출

자질을 추출하기 위해 아래의 두가지 타입을 고려할 수 있다. 타이틀, 초록 등으로부터 정보를 추출하는 경우에는 스테밍(영문) 또는 형태소분석(한글)을 거쳐 자질집합을 생성한 후 자질축소의 과정을 고려하는 것이 좋다. 또한, 전체문서 집합에서 저빈도(CF=1) 자질은 제거한다.[4)]

① 키워드, 디스크립터

: 논문 저자의 키워드 필드나 통제어휘인 디스크립터 필드를 이용한다.

② 용어 추출(Info Extraction)

: 타이틀, 초록 등의 비구조적인 정보로부터 명사구를 포함한 주요 정보를 추출한다.

(2) 문헌별 자질정보 추출 및 생성

: 문헌를 구성하는 개별 자질에 범주코드를 부여한다.

*주요 생성필드 : 문헌고유ID, 자질, 범주코드

4) 일반적으로 CF=1인 자질 종수가 전체 중 약 40-60%까지 차지함

(3) 자질 특성 매트릭스 생성[5)]

: 최종 자질 벡터를 연산하기 위한 매트릭스 정보를 생성하여 DB나 바이너리 파일로 적재한다.

*주요 생성필드 : 자질고유ID, 자질, 범주코드, TP, TN, FP, FN, CF, IDF 등 (표 1)

(표 1) 자질-범주간 출현관계 분할표

	범주 c_j 소속	범주 c_j 미소속
자질 f_i 출현	TP	TN
자질 f_i 미출현	FP	FN

2-2. 단위 분류기 결합을 통한 대용량 분류기 생성

대용량 분류기 생성의 핵심은 단위 분류기 생성 프레임워크 단계 중 3단계에서 생성된 자질특성 매트릭스를 결합하는 방법을 이용해 분류기의 동적결합의 수행하는 것이다. 단위 분류기는 일반적으로 도메인별로 생성될 수 있으나, 학습할 대상문헌이 많은 도메인의 경우에는 적당한 크기로 자유롭게 생성하여 필요한 경우 동적으로 결합해 거대한 매트릭스를 재생산할 수 있다.

(1) 매트릭스 동적결합 수행

① 우선 복수개의 결합 대상 '자질특성 매트릭스'를 메모리에 상주하여, 모든 매트릭스에 출현한 자질값의 고유한(distinct) 전체 셋을 만든다.

② 개별 자질에 결합 대상 매트릭스들을 참조하여 정보를 가져온다. 이때, 자질이 모든 자질특성 매트릭스에서 출현하지 않으므로 자질의 개수, 전체 문헌의 수 등 각 매트릭스의 통합정보를 동적으로 산출하여 TP, TN, FP, FN과 IDF, CF 등 주요 정보를 재계산한다.[6)]

(2) 개별 자질에 대한 주제-가중치 벡터를 생성

: 통합된 자질 특성 매트릭스로부터 거리계수 및 Cosine, LOR 등 유사척도를 이용해 최종 투표분류기에 적합한 자질 벡터형태를 생성하여 DB나 바이너리 파일로 적재한다.

LogTF*IDF*Cosine 계수를 이용한 자질벡터는 아래와 같이 표현이 가능하다[5].

$$vs(f_i, c_j)=(1+\log tf)*\log(N/df)*\cos(f_i, c_j)$$

(3) 문헌범주화 수행

: 통합 매트릭스에서 생성된 자질벡터를 이용해 투표형 분류기법으로 분류를 수행한다. 자질값 투표형 분류기(Feature-value Voting Classifier: FVC)는 여러 관련 연구를 통해 수행되었다[5][6][7]. 생성된 자질 벡터를

5) 개별 단위분류기를 생성하기 위한 핵심정보 매트릭스를 본 논문에서는 '자질 특성 매트릭스'라 칭한다.

6) 이 과정은 10만 건씩 학습된 10개의 분류기를 결합한 통합매트릭스 생성결과와 100만 건 전체를 한번에 학습한 분류기 매트릭스 내의 개별 파라미터요소의 수치가 정확히 일치함을 의미한다.

메모리에 상주한 후, 대량의 입력문헌에 대해 고속의 다원분류를 수행하여 입력문서를 분류한다.

최종 생성된 분류기는 최종계산된 벡터의 데이터량이 상대적으로 많지 않아 메모리 상주용량이 적기 때문에 자질 종수의 제한이 없으며 각 가중치의 선형결합을 실시하므로 자질 종수의 증가에 따른 속도저하도 거의 없는 고속의 분류기이다.

3. 실험

3-1. 데이터 콜렉션

KISTI의 NDSL 과학기술정보 통합서비스의 해외학술논문 데이터 중 2006년도 초록 733,930 건을 추출하고, 42개 세부 주제분야를 10개의 대분류 주제분류 체계로 재설정하여, 최종 597,671 건의 초록데이터를 실험에 사용하였다(표 2).

스페이스와 구두점을 기준으로 토크나이징을 하고, 불용어 사전을 이용하여 기본적인 불용어를 1차 제거한 후 스테밍 처리를 하였다. 자질축소는 실시하지 않았으며, 전체 문서집단에서 1번 출현한 자질(CF=1)은 사전 제거하였다.

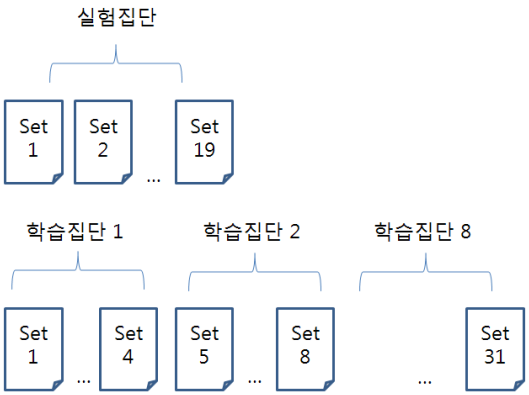
(표 2) 주제범주 설정과 세부 분야별 문헌 수

주제범주	세부분야	문헌수	합계 (비율)
ECO	경영	7,787	30,827 (5.16%)
	경제	20,776	
	무역	2,264	
CHE	화학	56,353	77,489 (12.97%)
	화학공업	21,136	
PHI	물리	56,724	96,669 (16.17%)
	응용물리	39,945	
BIO	생명과학	82,921	137,393 (22.99%)
	식물	32,263	
	동물	22,209	
INF	수학	22,589	33,835 (5.66%)
	전산/정보	11,246	
MED	약학/치료학	18,446	150,949 (25.26%)
	내과	93,661	
	외과	30,604	
	산부인과	8,238	
CON	건설/건축/토목	5,407	9,130 (1.53%)
	도시/환경	3,723	
POL	정치	6,633	12,470 (2.09%)
	행정	2,613	
	법률	3,224	
ART	예술	4,819	13,738 (2.30%)
	역사/지리	4,233	
	철학	1,859	
	종교	2,312	
	문학	515	
GEO	지학/지질	22,072	35,171 (5.88%)
	농업	13,099	
합계		**597,671**	**100%**

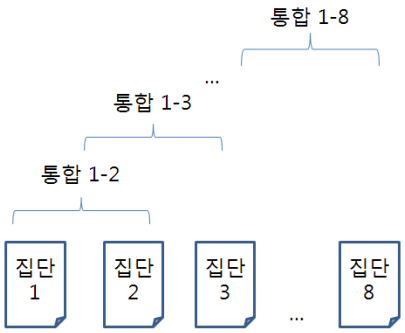
3-2. 단계별 분류기 생성

주제분야 별로 50개 그룹으로 분할하여 실험 셋을 구축하고, 무작위 19개 셋을 실험 집단으로 구성하고 21개 셋을 학습용으로 사용한다.

우선 4개 셋을 편의상 1개 그룹단위로 묶어 '자질 특성 매트릭스'를 만들고(그림 2), 상기 기술한 분류기 동적 결합방법을 이용해 총 8개의 개별 매트릭스를 순차적으로 통합하여 학습 환경을 구성한다(그림 3).



(그림 2) 실험집단과 학습집단 구성



(그림 3) 단계별 통합 분류기 생성

3-3. 실험결과

분류기의 성능을 측정하기 위해 성능이 좋은 것으로 알려진 SVM의 성능을 Baseline으로 비교실험을 수행하였다. SVMlight 공개 소프트웨어[7]를 실험에 사용하였고, 문서측 자질선정은 실시하지 않았다. SVM은 이전 관련연구를 참고로 오류 패널티 값인 C=1000으로 결정하였다[8].

실험결과 SVM은 입력문헌이 20만 건 수준인 학습 셋인 "set 01-16" 부터는 메모리 허용한계 초과에러가 발생하고 있다[8]. 또한 본 연구에서 사용한 로그승산비(Log Odd Ratio) 기반의 FV-LOR_t_AM 가중치 모델[9][9]은 최종 학습 셋인 "set 01-31" 지점에서 SVM의 최고성능을 앞지르고 있음을 확인할 수 있었다. 실험결과를 통해 문

7) http://svmlight.joachims.org/
8) 단, 시스템 물리메모리 2GB에서 실험하였음
9) FV-LOR_t_AM에서 _t는 LOR 측정결과의 음수를 0으로 보정한 모델이며, _AM은 Mengle과 Goharian(2009)이 제안한 Ambiguity Measure 기법을 사용한 것임.

(표 3) SVM과 FV-LOR_t_AM의 성능비교(마이크로 정확률)

학습문서 집합	set 01-04	set 01-08	set 01-12	set 01-16	set 01-20	set 01-24	set 01-28	set 01-31
학습문서 수	47,811	95,651	**143,471**	**191,257**	239,043	286,826	334,612	**370,418**
자질 수 (CF>1)	113,221	166,596	211,017	249,150	283,792	315,952	346,250	367,254
FV-LOR_t_AM (%)	75.224	75.909	76.163	76.401	76.557	76.706	76.797	**76.910**
SVM (%)	75.089	76.299	**76.854**	**메모리초과**	-	-	-	-

헌학습량이 많아짐에 따라 일정 수준까지는 지속적으로 성능을 높일 수 있을 것으로 기대한다.

4. 결론 및 향후 연구

본 논문에서는 기존의 자동범주화 모델에서 제한점으로 존재하였던 대용량 문서의 처리와 학습결과의 재사용 문제를 해결함으로써 학습대상 문서의 추가 변경시 전체 데이터를 반복적으로 처리해야 하는 텍스트 마이닝의 취약점을 보완한 대용량 기반의 동적 분류기 생성방안을 제시하였다. 또한 성능이 우수한 것으로 알려진 SVM과의 베이스라인 성능 비교를 통하여 새롭게 제안한 모델의 가능성을 검증하였다.

향후, 성능을 향상시킬 수 있는 자질 가중치 요소를 모델에 추가할 예정이며, 다양한 유형의 데이터를 기반으로 기존의 분류모델과의 비교검증을 추가로 수행하여 일반화된 성능검증 결과를 도출할 예정이다.

참고문헌

[1] Liu, X., S. Yu, F. Janssens, W. Glanzel, Y. Moreau, and B. D. Moor. 2010. "Weighted Hybrid Clustering by Combining Text Mining and Bibliometrics on a Large-Scale Jorunal Database." *Journal of The American Society for Information Science and Technology,* 61(6):1105-1119.

[2] Picardi, E., F. Mignone, and G. Pesole. 2009. "EasyCluster: a fast and efficient gene-oriented clustering tool for large-scale transcriptome data." *BMC Bioinformatics*, 10(Suppl 6):S10.

[3] Judd, D., P. K. McKinley, and A. K. Jain. 1998. "Large-Sclae Parallel Data Clustering." I*EEE Transanstions On Pattern Analysis and Machine Intelligence,* 20(8):871-876.

[4] 이창환. 정인철. 권영식. 2010. 속성 값 빈도 기반의 전문가 다수결 분류기. 『정보과학회논문지:데이터베이스』, 37(4):177-184.

[5] 정도헌. 2010. 최대 개념강도 인지기법을 이용한 데이터베이스 자동선택 방법에 관한 연구. 『정보관리학회지』, 27(3):265-281.

[6] Ko, Y., and J. Seo. 2004. "Using the feature projection technique based on a normalized voting method for text classification." *Information Processing and Management*. 40(2): 191-208.

[7] 이재윤. 2005. 문서측 자질선정을 이용한 고속 문서분류기의 성능향상에 관한 연구. 『정보관리연구』, 36(4):51-69.

[8] 정영미. 임혜영. 2000. SVM 분류기를 이용한 문서 범주화 연구. 『정보관리학회지』, 17(4):229-248.

[9] Mengle, S.S.R. and Goharian, N. 2009. "Ambiguity measure feature-selection algorithm." *Journal of The American Society for Information Science and Technology*. 60(5):1037-1050.