자동문헌분류를 위한 대표색인어 추출에 관한 연구

A Study on the Feature Selection for Automatic Document Categorization

황재영, 이응봉, 충남대학교 대학원 문헌정보학과

Hwang, Jae-Young, Lee, Eung-Bong
Graduate School of Library & Information Science, Chungnam National University

인터넷 학술정보자원이 급증하고 있는 가운데 자동문헌분류에 대한 관심과 필요성도 늘어가고 있다. 자동문헌분류에 관한 실험은 전처리 단계인 대표색인어 추출과 추출된 대표색인어의 분류성능 평가 실험으로 구분 할 수 있는데, 본 연구에서는 우선 대표색인어 추출을 위해 다양한 대표색인어(자질) 추출 방법에 따른 색인어 성능평가 실험 및 최적의 대표색인어 개수 선정 실험을 수행하였다.

1 서 론

1.1 연구의 필요성

인터넷상의 정보자원은 기하급수적으로 증가하고 있다. 도서관·정보센터에서는 이용자에게 제공해야 할 정보자원의 범위를 기존의 학술정보자원에서 인터넷상의 학술정보자원으로까지 확대하기에 이르렀다. 따라서 인터넷 학술정보자원에 대해 수집에서부터 조직, 유통, 정보서비스에 이르기까지 다양한 연구가 계속되고 있는데 그 중에서도 자동분류에 대한 관심이 점차 늘어가고 있다. 그러나 아직까지도 이에 대한 연구성과가 텍스트 문서 중심으로 이루어지고 있고, 실제 도서관 정보센터에서도 상용 자동분류시스템 도입에 있어 범주별 대표색인어 추출과 같은 별도의 전처리 과정 없이 그대로 적용하는 수준이라 자동분류 성능은 기대에 미치지 못하고 있다.

자동분류 중 자동 문서 범주화(automatic text categorization)는 미리 정의된 범주(또는 분류체계)에 문서를 자동으로 할당하는 기법과 관련된 연구 분야이다. 이전에는 수작업으로 문서마다 범주를 지정해 주는 방식이 사용되었으나 이 경우에 사람의 노력, 시간, 비용 면에서 심각한 어려움을 초래할 수 있다. 이러한 작업을 자동분류 시스템으로 교체하거나 보조시스템으로 활용하면 비용을 크게 줄일 수 있을 것이다. 그러므로 자동 문서 범주화는 대량의 문서를 효율적으로 관리하고 검색할 수 있게 하는 동시에 방대한 양의 수작업을 감소시키는데 그 목적이 있다. (Yang, 1997)

자동 문서 범주화와 관련된 연구의 필요성과 문제점을 해결하기 위해 본 고에서는 상용 자동분류시스템을 도입하는 과정 중 전처리 단계에서 필요시 되는 대표색인어 선정과 다양한 선정 방법이 분류성능에 어떤 영향을 미치는지에 대해 중점적으로 다루고자 한다.

1.2 연구의 방법

최근의 국내 일부 도서관·정보센터에서는 인터넷 학술정보자원을 이용자에게 제공하는 디렉토리서비스를 시행하고자 자동분류시스템을 도입하여 운용하고 있다. 자동분류시스템의 처리 절차를 간단히 살펴보면 우선 특정 주제분야와 관련된 웹사이트의 URL 정보를 시스템에 등록하고 웹크롤러와 같은 로봇 엔진이 사전에 등록된 웹사이트를 대상으로 변경되거나 새로 생성된 문서를 자동으로 수집한다. 수집된 신규 생성문서는 기존 자동분류시스템에서 설정한 범주(분류체계)별로 문서가 자동으로 할당되게 되는데 할당되는 기본 알고리즘은 주로 범주를 대표할 수 있는 대표 색인어(용어)와 신규로 입력한 문서와의 단어 및 유사도 비교에 의해 할당되게 된다.

본 고에서는 국방무기체계분야 자동분류시스템 구현을 위한 사전 단계로서 국방무기체계분류에 적용할 수 있는 대표색인어를 다양한 방법으로 추출하고 이를 상용 자동분류시스템에 적용하여 그 분류성능을 실험하였다. 이를 위해 현재 국방과학연구소에서 운용중인 국방과학기술 분석정보시스템(InfoBrain)에 구축된 무기체계 관련 문헌을 대상으로 자동분류 실험에 적용할 대표색인어를 추출하였다. 국방과학기술 분석시스템은 무기체계별 주제전문가가 직접 생산 또는 수집한 무기체계 관련 문헌을 "합동무기체계분류표"에 따라 직접 수동으로 문헌을 할당할 수 있는 시스템이기 때문에 무기체계 범주별로 대표 색인어를 추출할 수 있고 자동분류시스템의 분류성능을 평가할 수 있는 실험 컬렉션이라 할 수 있다.

2 선행연구

2.1 자동분류의 개념

문헌분류란 유사한 내용의 문헌들을 모아 집단화하는 작업이다. 전통적으로 문헌분류는 미리 만들어 놓은 분류표에 의거하여 수작업으로 수행하여 왔으나 1960년대에 들어서 자동분류의 개념이 발전하기 시작하였다. 문헌의 자동분류는 기본적으로 두 가지 부류로 구분된다. 첫째는 분류작업 이전에 분류체계가 만들어져 있는 상태에서 각 문헌을 가장 적합한 클레스에 배정함으로써 문헌을 집단화하는 것으로서 자동분류(automatic classification)가 있으며, 둘째는 사전 분류체계 없이 문헌간의 유사성에 근거하여 유사한 문헌들의 집단을 형성하는 클러스터링(clustering)이 있다. 전자를 주로 사전 분류체계에 의한 자동분류라 하는데 이는 기존 분류체계를 그대로

또는 수정하여 이용하거나 아니면 실험문서집단을 이용하여 경험적으로 작성한 분류체계를 이용하며, 이 분류체계를 구성하는 각 카테고리로 문헌들을 자동적으로 분류하는 것을 말한다.(정영미, 1993)

전산학분야에는 사전 분류체계에 의한 문헌 자동분류를 주로 "자동 문서 범주화"라고 하며 다음과 같은 개념으로 사용한다.

자동 문서 분류는 각 범주의 특징을 표현하는 범주특성 벡터와 입력 문서의 특징을 표현하는 입력문서 벡터 사이에 유사도를 계산하여 유사도 값에 따라 입력 문서의 범주를 할당하는 것이다. 즉, 자동분류 시스템은 문서에서 출현한 용어들에 대해 각 범주를 판별하는데 기여하는 정도에 따라 가중치를 부여한다. 입력 문서에 대해서도 문서에 출현한 용어들을 추출하고 해당 문서에 대한 가중치를 계산하여 문서 벡터를 구한다. 입력 문서 벡터와 각 범주들의 특성 벡터 사이에 유사도 계산에 의해 입력 문서의 범주를 결정한다. (이경찬, 2002)

2.2 국내 · 외 선행연구

자동 분류에 관한 연구 가운데 대표 색인어 또는 자질 추출과 관련된 선행 연구 논문을 살펴보면, 김지숙(2002)은 대량의 문서를 자동으로 분류하기 위하여 비감독 학습 기법에 의한 카테고리별 대표 색인어 추출 방법을 제안하였다. 제안된 방법은 사전에 문서를 분류하지 않고 대표 색인어 추출을 위해 데이터마이닝 기법 중의 하나인 연관 규칙 탐사 알고리즘을 이용하였다.

진훈(2001)은 텍스트 형태로 존재하는 문서가 특정 범주에 속하는지를 판별하는데 있어서 그 문서를 표현하고 있는 특징을 어떻게 선택할 것인가와 얼마나 선택할 것인가가 미치는 영향을 실험을 통하여 측정하였다. 즉, 실험을 통하여 특징 선택 방법이 분류 성능에 미치는 영향을 알아보고자 하였고, 특징의 개수와 분류 성능과의 관계 그리고 범주의 개수와 특징의 개수와의 관계를 규명하고자 하였다.

Y. Yang(1997)은 문서 범주화의 통계적 학습에 있어 다양한 자질 선정 방법을 비교 연구하였다. 그의 연구는 Documentation Frequency (DF), Information Gain(IG), Term Strength(TS) X^2-test (CHI), Mutual Information(MI)등 5가지 자질선정 기법의 성능을 실험하였는데, 그중 IG와 CHI가 가장 효과적인 것으로 나타났다.

3 자동 문헌 범주화 실험

3.1 실험 컬렉션 및 자동분류시스템

실험 문서 집단은 현재 국방과학연구소 인트라넷에서 운용되고 있는 국방과학기술분석정보시스템 내의 국방 무기체계 데이터베이스로 하였다. 이 데이터베이스는 무기체계별 주제 전문가가 직접 생산하거나 수집한 문헌을 기 마련된 "합동무기체계분류표" 기준에 따라 시스템에서 직접 수동으로 문헌을 할당할 수 있도록 되어있기 때문에 무기체계 문헌에 대해서는 가장 정

확한 분류가 이루어지고 있다고 할 수 있다. 따라서 이 문헌을 실험 문서 집단으로 채택함으로써 정확한 무기체계 분야 대표 색인어 추출이 가능하고 이를 자동분류시스템에 적용했을 경우 분류성능의 적합성을 쉽게 평가할 수 있는 것으로 판단된다.

본 고에서는 자동분류시스템을 잉크토미사의 CCE(Contents Classification Engine)를 적용하였다.

3.2 실험설계 및 내용

다양한 대표 색인어 추출 방법 및 가중치 부여 방법이 자동분류시스템의 성능 평가에 영향을 줄 수 있다는 가정 하에 다음과 같이 실험설계를 하였다.

첫째, 무기체계별 주제 전문가가 "합동무기체계분류표"에 의해 수동으로 분류한 무기체계 관련 문헌을 국방과학기술 분석정보시스템 (InfoBrain)으로부터 총 2,000건의 실험문서를 추출하였다. 대표색인어 추출을 위해 10개 무기체계 분류항목별로 학습문서 150건, 분류성능 평가 문서 50건 즉, 각 분류항목 당 200건의 문서를 추출하였다.

둘째, 추출된 문헌의 메타데이터 중 본문의 내용을 가장 잘 표현한다고 할 수 있는 제목과 초록 정보만을 다시 추출하였으며, 이를 다시 분류항목별로 형태소 분석을 하였다. 본 고에서 적용한 형태소 분석기는 현재 공개 테스트가 가능한 한글 형태소 분석기, HAM(강승식, 2002)을 사용하였으며 형태소 분석의 성능을 높이기 위해 스태밍, 불용어 제거, 국방과학기술 한글시소러스 사전을 대입하였다.

셋째, 형태소 분석기를 통해 추출된 용어(색인어)는 별도로 개발한 "용어빈도 및 가중치 계산 알고리즘" 프로그램에 의해 색인어를 저장하고, 용어별 출현빈도 및 문헌빈도를 계산하도록 하였다.

넷째, 각 범주별로 추출된 색인어는 용어빈도(TF)와 용어빈도*역문헌빈도(TF*IDF) 값에 따라 가중치 값을 계산하고 이를 중요도 순으로 정렬하였다.

다섯째, 각 범주별 최적의 대표색인어 개수를 정하기 위해 동일 실험문헌 집단에서 주제 전문가가 기 부여한 키워드집합과 비교하는 실험을 하였다. 이 실험에서는 TF 및 TF*IDF로부터 추출한 색인어를 상위 몇 퍼센트 적용했을 때 키워드집합을 최대한 만족할 수 있는가를 알아보는 실험으로서 대표색인어 개수를 정하는데 유용하다.

여섯째, 다양한 자질 가중치 계산 기법에 따른 색인어 추출과 대표 색인어 개수 선정 실험을 한 후 최종적으로 추출된 대표 색인어를 자동분류시스템에 등록하고 자동분류를 시행하였다.

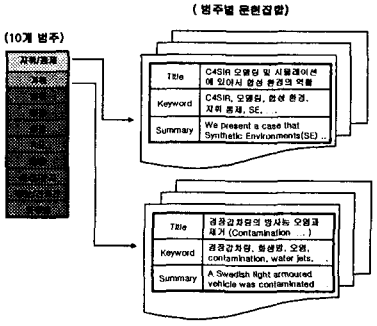
일곱째, 분류항목별 분류 성능 테스트 문서 50건에 대해 자동분류를 실시하고 분류성능을 평가하였다. 평가방법은 기존의 국방과학기술 분석정보시스템의 무기체계별 수동 분류 결과를 정답으로 추정하고 이와 비교한 정확률, 재현율을 평가하였다.

3.3 색인어 추출 및 가중치 계산

"합동무기체계분류표"에 의해 분류된 실험문서집단을 〈표 1〉과 같이 선정하고 선정된 개별 문헌에 대해 Title(제목), Keyword(주제어) Summary(초록)를 대상으로 〈그림 1〉 과 같은 형식으로 추출하였다.

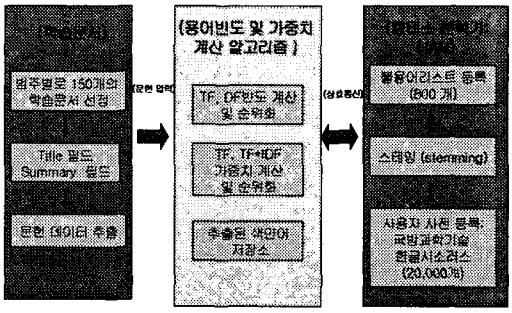
〈표 1〉 범주별 실험대상 문서 선정

대분류	중분류	분류 코드	학습 문서 (개)	테스트 문서 (개)	계
지휘/통제	자동화체계	01010000	150	50	200
기동	전차	02010000	150	50	200
화력	탄약	03040000	150	50	200
항공	무인항공기	04030000	150	50	200
함정	잠수함	05020000	150	50	200
유도	대지 유도무기	06010000	150	50	200
방공	대공 유도무기	07010000	150	50	200
통신/전자	전자 광학장비	08060000	150	50	200
정보/전자전	전자전장비	09020000	150	50	200
화생방	화생방 장비/물자	10010000	150	50	200
계			1,500	500	2,000



〈그림 1〉 문헌별로 추출된 메타데이터 구조

범주별 150건의 학습문서는 한글형태소 분석기 HAM에 의해 형태소분석이 이루어지며 형태소 분석결과 추출된 색인어에 대해서는 별도로 개발한 "용어빈도 및 가중치 계산 알고리즘" 프로그램에 의해 계산되도록 하였다. 〈그림 2〉참조

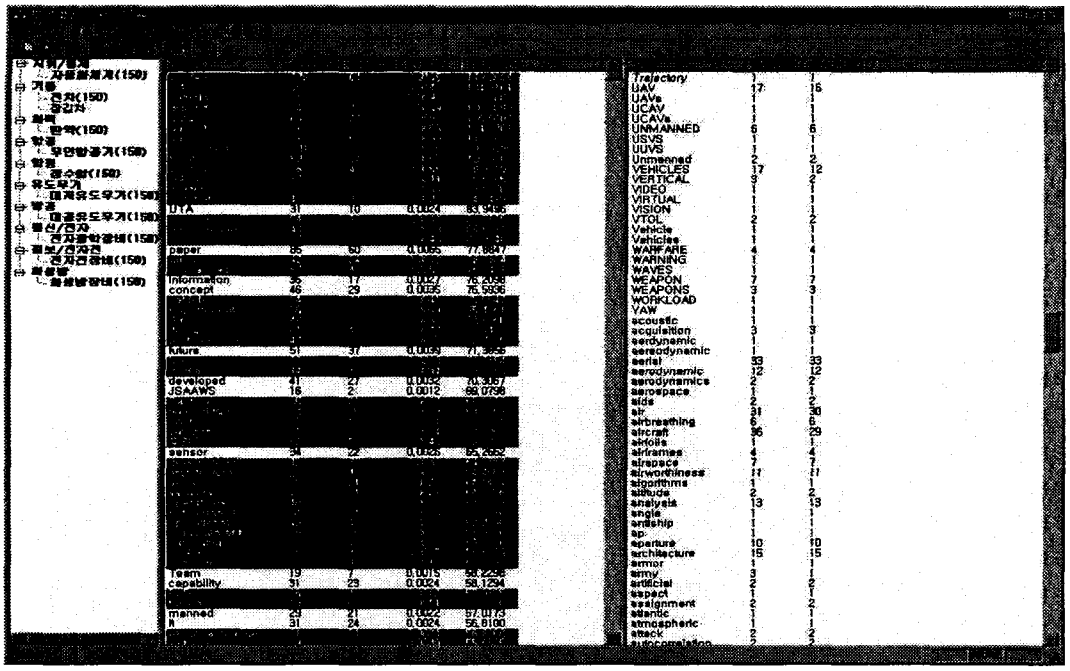


〈그림 2〉 형태소분석과 용어빈도 & 가중치 계산 알고리즘 모델

"용어빈도 및 가중치 계산 알고리즘"에 적용한 문헌당 용어빈도(TF)의 가중치 값과 용어빈도*역문헌빈도(TF*IDF)에 적용한 수식은 다음과 같다.

첫째, 용어빈도(TF: Term Frequency)란 문헌내 용어의 출현빈도가 문헌에 대한 용어의 중요성을 나타낼 수 있는 것으로 용어 가중치 값 부여에 있어 가장 단순하면서도 보편적인 방법이다. 본 고에서는 하나의 범주안에 나타나는 모든 용어의 출현빈도를 조사하고 이를 가중치로 계산하였으며 이때 적용한 수식은 다음과 같다.

$$W_{i,c} = \frac{freq(t)}{\sum_{i} freq(t_i)}$$



〈그림 3〉 용어빈도 및 가중치 계산 결과

범주 C에 대한 용어 t의 가중치 $W_{t,c}$는 용어 t가 범주 C를 대표하는 정도를 반영하기 위하여 용어 t의 빈도수를 범주 C를 기술하는 모든 용어들의 빈도수 합으로 나눈 값으로 계산하였다.

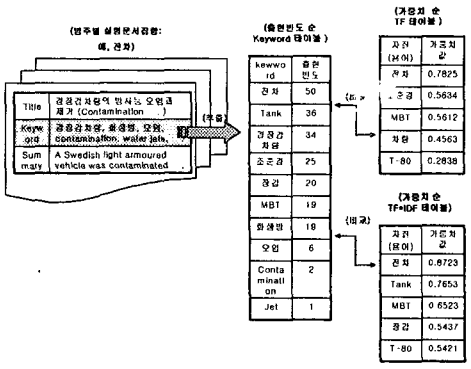
둘째, 용어빈도(TF)*역문헌빈도(IDF: Inverse Document Frequency)는 용어빈도(TF)값이 가지는 단점을 보완하기 위한 것으로 하나의 용어가 여러 문헌에서 여러번 나타날수록 대표성이 떨어진다는 가정에 기인한 것이다. 즉 실제로 많은 문서에서 그 문서를 대표하는 단어는 빈도가 그리 높지 않게 발생한다는 것이다. 본 고에서는 하나의 범주안에 나타나는 모든 용어의 용어빈도와 역문헌빈도를 곱한 값을 대상으로 가중치를 계산하고 이를 중요도순으로 정렬하였으며, 적용한 수식은 다음과 같다.

$$w_{ij} = tf_{ij} \log\left(\frac{N}{df_i}\right)$$

〈그림 3〉은 용어빈도 및 가중치 계산 알고리즘에 의해 출력된 결과화면이다.

3.4 각 범주별 대표색인어 개수 선정 실험

CCE 자동분류시스템에 등록할 대표색인어 선정을 위해 각 범주별로 추출된 색인어를 대상으로 최적 개수 선정 실험을 하였다. 이 실험은 범주별로 TF 및 TF*IDF 가중치 값으로부터 추출한 색인어와 동일



〈그림 4〉 색인어집합과 키워드집합의 비교 모델

실험문서집단에서 주제별 전문가가 기 추출한 키워드필드 집합과의 비교실험으로서 상위 몇 퍼센트 적용 시 키워드필드 집합을 최대한 만족할 수 있는가를 알아보는 실험이다. 이 실험의 전제는 주제별 전문가가 기 부여한 키워드필드 집합이 정답이라는 가정하에서 수행된다. 〈그림4〉 참조.

〈표 2〉 TF 색인어 집합과 키워드집합과의 비교

범주(중분류)	입력 문헌수	추출 색인 어수	키워드 수	상위5%색인어			상위10%색인어			상위20%색인어			상위50%색인어			상위100%색인어		
				입력수	적중수	적중율	입력수	적중수	적중율	입력수	적중수	적중율	입력수	적중수	적중율	입력수	적중수	적중율
자동화체계	150	4591	1153	229	153	67%	459	260	57%	918	419	46%	2295	655	29%	4591	915	20%
전차	150	4265	1346	213	143	67%	426	237	56%	853	391	46%	9132	698	33%	4265	1047	25%
탄약	150	3989	1080	199	126	63%	398	225	57%	797	361	45%	1994	646	32%	3989	952	24%
무인항공기	150	3594	769	179	104	58%	359	171	48%	718	270	38%	1797	456	25%	3594	623	17%
잠수함	150	4247	905	212	121	57%	424	205	48%	849	326	38%	2123	566	27%	4247	830	20%
대지유도무기	150	2597	744	129	82	84%	259	143	55%	519	226	44%	1298	382	29%	2597	554	21%
대공유도무기	150	2386	549	119	65	55%	238	108	45%	477	180	38%	1193	287	24%	2386	436	18%
전자광학장비	150	3731	567	186	103	55%	373	156	42%	746	232	31%	1865	382	20%	3731	566	15%
전자전장비	150	3533	763	176	109	62%	353	168	48%	706	268	38%	1766	450	25%	3533	654	19%
화생방장비	150	4360	1425	218	145	67%	436	242	56%	872	413	47%	2180	733	34%	4360	1113	26%
계	1500	37293	9301	1860	1151	62%	3725	1915	51%	7455	3086	41%	18643	5255	26%	37293	7690	21%
평 균	150	3729.3	9301.1	186	115.1	62%	372.5	191.5	51%	745.5	308.6	41%	1864.3	525.5	28%	3729.3	769	21%

〈표 3〉 TF*IDF 색인어 집합과 키워드집합과의 비교

범주(중분류)	입력 문헌수	추출 색인 어수	키워드 수	상위5%색인어			상위10%색인어			상위20%색인어			상위50%색인어			상위100%색인어		
				입력수	적중수	적중율	입력수	적중수	적중율	입력수	적중수	적중율	입력수	적중수	적중율	입력수	적중수	적중율
자동화체계	150	4591	1153	229	161	70%	459	261	57%	918	415	45%	2295	655	29%	4591	915	20%
전차	150	4265	1346	213	144	68%	426	241	57%	853	409	48%	2132	698	33%	4265	1047	25%
탄약	150	3989	1080	199	132	66%	398	237	60%	797	367	46%	1994	646	32%	3989	952	24%
무인항공기	150	3594	769	179	110	61%	359	166	46%	718	275	38%	1797	454	25%	3594	623	17%
잠수함	150	4247	905	212	122	58%	423	206	49%	849	333	39%	2123	566	27%	4247	830	20%
대지유도무기	150	2597	744	129	83	64%	259	148	57%	519	223	43%	1298	382	29%	2597	554	21%
대공유도무기	150	2386	549	119	65	55%	238	111	47%	477	191	40%	1193	287	24%	2386	436	18%
전자광학장비	150	3731	567	186	103	55%	373	168	50%	746	243	33%	1865	382	20%	3731	566	15%
전자전장비	150	3533	763	176	113	64%	353	172	49%	706	271	38%	1766	450	25%	3533	654	19%
화생방장비	150	4360	1425	218	147	67%	436	237	54%	872	406	47%	2180	733	34%	4360	1113	26%
계	1500	37293	9301	1860	1180	63%	3725	1947	52%	7455	3133	42%	18643	5253	28%	37293	7690	21%
평 균	150	3729.3	9301.1	186.0	118	63%	372.5	194.7	52%	745.5	313.3	42%	1864.3	525.3	28%	3729.3	769.0	21%

〈표 4〉 tf*log(N/df) 알고리즘 적용시 추출된 대표색인어 집합(상위 25개)

순위	지휘통제체계	전차	탄약	무인항공기	잠수함	지상유도무기	해공유도무기	전자통제장비	전자전장비	화생방장비
1	information	Army	data	system	submarine	missile	시스템	systems	pulse	chemical
2	Army	tank	projectile	UAV	추진	순항	시험	thermal	system	기술
3	decision	system	missile	vehicle	data	미사일	미사일	design	power	입자
4	systems	전차	weapon	ground	system	미사	방어	optical	systems	system
5	command	armor	system	data	잠수함	시험	Defense	system	EW	탐지
6	data	ammunition	model	aircraft	잠수	개발	미국	IR	propagation	Army
7	information	tanks	explosive	air	vehicle	방어	개발	lens	electromagn	Biological
8	control	propellant	IM	unmanned	연료	Missile	유동	infrared	radar	protective
9	software	gun	systems	flight	design	시스템	지대공	temperature	microwave	시험
10	Joint	training	자료	control	시스템	비행	미사	laser	시스템	Chemical
11	System	장갑	results	UAVs	sonar	test	missile	zoom	speech	NBC
12	Force	model	tests	systems	force	설계	대한	target	전자전	개발
13	technology	개발	control	combat	AUV	Block	사	FLIR	field	detection
14	interoperabili	육군	substrate	UTA	UUV	대한	Missile	performanc	radio	작용제
15	Command	차량	using	missions	submarines	순항미사	발사	night	recognition	System
16	support	자료	developmen	technologie	model	경로	해군	개발	developmen	provide
17	system	기술	developed	mission	fluid	System	추진	imaging	HPM	agents
18	operational	systems	Army	paper	payload	기술	로켓	aircraft	data	화학
19	operations	Tank	fiber	target	Navy	미국	방공	image	target	대한
20	tactical	simulation	propellant	Unmanned	underwater	로켓	도입	time	processing	분석
21	development	전투	target	information	vessel	사	러시아	method	techniques	mask
22	network	performanc	process	concept	missile	엔진	순항	paper	coupling	Detection
23	paper	시스템	GPS	vehicles	개발	재돌입체	사용	Laser	aircraft	생물학
24	Control	tests	based	requirement	Submarine	탄도	방공시스템	레이저	In	deg
25	future	developmen	erosion	Air	전지	Tomahawk	MEADS	using	It	체계

3.5 자동분류 성능 실험

각 범주별 대표 색인어 개수 선정 실험에 의해 추출된 최종 대표 색인어를 잉크토미사의 CCE 자동분류시스템에 범주별로 등록한 후 각 대표 색인어가 자동분류성능에 어떤 영향을 미치는가를 실험하였다. 실험을 위한 입력문서는 동일 실험문서집단에서 기 선정한 테스트문서 50건으로 하였으며 자동분류 실행 후 분류성능은 정확률과 재현율로서 평가하였다. 단 이 실험의 전제는 국방과학기술 분석정보시스템(Info-Brain)에서 주제별 전문가가 수동으로 할당한 범주가 정답이라는 전제하에서 수행된다.

4 실험결과 분석

4.1 색인어 추출 및 가중치 계산결과 분석

학습문서 1,500건(각 범주별 150건)에 대해 형태소를 분석한 결과 총 37,293개의 색인어가 추출되었으며, 동일한 방법으로 실험문서집단 내 주제별 전문가가 부여한 키워드필드의 형태소를 분석한 결과 총 9,301개의 키워드가 추출되었다. 각 범주별 TF 가중치 값 및 TF*IDF 가중치 값에 따른 추출 색인어 수와 키워드 수의 현황은 〈표 2, 표 3〉과 같다.

또한 각 범주별로 추출된 색인어들은 TF 가중치 값과 TF*IDF 가중치 값에 따라 중요도 순으로 순위화 할 수 있으며 이렇게 하여 추출된 대표색인어 결과는 〈표 4〉와 같다.

4.2 색인어집합과 키워드집합의 비교결과 분석

각 범주별 대표색인어 개수 선정 실험을 위해 추출된 키워드집합과 색인어집합을 상호 비교하였다. 키워드집합이 고정된 상

태에서 색인어집합을 중요도 순으로 상위 5%, 10%, 20%, 50%, 100%를 각 범주별로 대입한 결과는 〈표 2, 표 3〉과 같다. 자동화체계의 경우 상위 5%의 색인어(229개)를 키워드집합(1,153개)과 비교한 결과 153개(적중률 66.8%)가 적중한 것으로 나타났다.

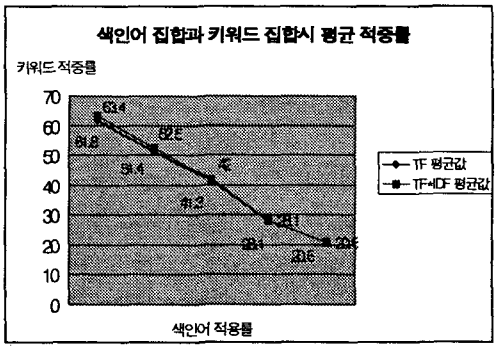
동일한 방법으로 상위 10%의 색인어(459개)의 경우 260개가 적중(56.6%), 상위 20%(918개)의 경우 419개가 적중(45.6%), 상위 50%(2,295개)의 경우 655개가 적중(28.5%), 전체 100% (4,591개)를 대입한 경우 915개가 적중(19.9%)하였다. 각 범주별로 중요도 순으로 대입한 색인어 입력수와 적중수에 대한 결과는 〈표 2, 표 3〉과 같으며 각 범주별로 TF값과 TF*IDF 값에 따른 평균 적중률 분석결과는 〈그림 5〉와 같다.

분석결과, 추출된 색인어 상위 5% 사용시 가장 높은 상대적 적중률을 보였으며 10%, 20%, 50%, 100%와 같이 색인어를 많이 사용할수록 키워드집합에 대한 상대적 적중률은 낮아짐을 알 수 있다. 즉 특정 문헌에서 추출된 색인어 중 대표색인어의 적정 개수를 선정하기 위해서는 상위 5%만을 대표색인어로 사용해도 좋은 결과 값을 가져온다는 것을 알 수 있다. 그리고 상위 50%이후의 TF와 TF*IDF값을 비교하면 적중수(적중률)가 이 거의 동일하게 나타나는데 이는 평균적으로 상위 색인어 순위가 18,643번 이후의 용어는 카테고리내에서 한 번만 출현하고, 문헌빈도(DF)도 한 번만 출현한 용어이기 때문이다.

4.3 TF와 TF*IDF의 비교결과 분석

TF 가중치 값과 TF*IDF 가중치 값에 의한 색인어 집합을 키워드집합과 비교한 결과 평균적으로 TF*IDF 가중치 값에 의한 색인어 추출이 상대적으로 만족할만한 결과 값을 보였다. TF 가중치 값에 의한 평균 결과 값을 살펴보면, 상위 5%의 색인어(186개) 적용시 115개(적중률 61.8%)가 적중하였으며, 이후 10% 색인어(51.4% 적중), 20% 색인어(41.3% 적중), 50% 색인어(28.1% 적중), 100% 색인어(20.6% 적중)의 상대적 적중률을 보였다.

TF*IDF의 경우 상위 5%의 색인어(186개) 적용시 118개(적중률 63.4%)가 적중하였으며, 이후 10% 색인어(52.5% 적중), 20% 색인어(42.0% 적중), 50% 색인어(28.1% 적중), 100% 색인어(20.6% 적중)의 상대적 적중률을 보였다. 〈그림 5〉 참조.



〈그림 5〉 색인어집합과 키워드집합 비교시 평균 적중률

5 결 론

본 고는 상용 자동분류시스템 도입과정에서 필요시 되는 대표색인어 추출에 관한 연구이다. 국방무기체계 분류항목별 대표색인어(자질) 추출을 위해 TF 가중치 값과 TF*IDF 가중치 값에 의해 추출된 색인어를 주제별 전문가가 부여한 키워드집합과 비교한 결과 TF*IDF 가중치 값에 의해 추출된 색인어가 더 우수한 것으로 나타났다. 또한 대표색인어 개수 선정 실험에서는 키워드집합을 고정한 상태에서 추출된 색인어를 상위 5%, 10%, 20%, 50%, 100%로 나누어 비교 실험한 결과 상대적으로 상위 5%의 대표색인어가 가장 높은 적중률을 보였다. 즉, 특정 문서집단에서 상위 5%의 대표색인어만을 자동분류시스템에 적용해도 충분히 우수한 성능을 발휘하는 것으로 나타났다.

향후 다양한 방법으로 추출된 대표색인어가 자동분류성능에는 어떤 영향을 줄 수 있는가에 대한 계속적인 연구가 이루어지길 기대해 본다.

참 고 문 헌

김지숙, 등저. 2002, "효율적인 문서 자동분류를 위한 대표 색인어 추출 기법", 정보기술과 데이터베이스저널, 8(1), pp. 117~128

강승식, 2002. "HAM:한국어 형태소 분석 라이브러리", http://ham.hansung.ac.kr/ham

이경찬, 강승식. 2002, "범주 대표어의 가중치 계산 방식에 의한 자동 문서 분류 시스템", 2002년도 한국정보과학회 봄 학술대회논문집, 29(1), pp. 475

이경찬, 강승식. 2003, "자질 중요도 계산 기법에 의한 자동 문서 범주화", 2003년도 한국정보과학회 봄 학술발표논문집, 30(1), pp. 538

정영미, 1993. 『정보검색론』, 구미무역(주) 출판부

진훈, 김인철. 2001, "문서분류를 위한 특징 선택", 2001년도 한국정보과학회 봄 학술발표논문집, 28(1), pp. 262~264

Yang, Y and Pedersen, J.O. 1997, "A Comparative Study on Feature Selection in Text Categorization", Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97), pp. 412~420