지식 분류의 자동화를 위한 클러스터링 모형 연구*

Development of a Clustering Model for Automatic Knowledge Classification

정영미(Young-Mee Chung)**, 이재윤(Jae-Yun Lee)***

초 록

본 연구에서는 문헌을 기반으로 한 지식의 자동분류를 위해 최적의 클러스터링 모형을 제시하고자 하였다. 클러스터링 실험을 위해서 신문기사 실험집단과 학술논문 초록 실험집단을 구축하였고, 분류 성능 평가 척도인 WACS를 개발하였다. 분류자질로 사용한 용어의 집합은 다양한 자질 축소 기준을 적용하여 생성하였으며, 다양한 용어 가중치를 사용하였다. 유사계수 공식으로는 코사인 계수와 자카드 계수를 적용하였으며, 클러스터링 알고리즘으로는 비계층적 기법인 완전연결 기법과 계층적 기법인 K-means 기법을 각각 사용하였다. 실험 결과 신문기사 원문 집단에서의 성능이 좋았으며, 완전연결 기법의 성능이 K-means 기법보다 높게 나타났다. 역문헌빈도의 적용은 완전연결 클러스터링에서는 긍정적인 효과가 나타났으나, K-means 클러스터링에서는 그렇지 못했다. 분류자질은 전체의 7.66%만 사용하였을 경우에도 성능 저하가 크지 않았으며, K-means 클러스터링에서는 오히려 성능 향상 효과가 있었다.

ABSTRACT

The purpose of this study is to develop a document clustering model for automatic classification of knowledge. Two test collections of newspaper article texts and journal article abstracts are built for the clustering experiment. Various feature reduction criteria as well as term weighting methods are applied to the term sets of the test collections, and cosine and Jaccard coefficients are used as similarity measures. The performances of complete linkage and K-means clustering algorithms are compared using different feature selection methods and various term weights. It was found that complete linkage clustering outperforms K-means algorithm and feature reduction up to almost 10% of the total feature sets does not lower the performance of document clustering to any significant extent.

키워드 : 자동 분류, 문헌 클러스터링, 클러스터링 모형, 용어가중치, 자질 축소
automatic classification, document clustering, clustering model, term weighting method, feature reduction

---

* 이 논문은 1999년도 한국학술진흥재단의 연구비에 의하여 지원되었음(KRF-99-041-C00531)

** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)

*** 연세대학교 문헌정보학과 강사(memexlee@lis.yonsei.ac.kr)

■ 논문 접수일 : 2001년 5월 21일

■ 게재 확정일 : 2001년 6월 5일

1 서 론

지식기반 사회에서는 넘쳐나는 정보를 선별하고 가공하여 유용한 지식을 획득하는 작업이 대단히 중요하다. 이런 작업에는 색인, 검색, 필터링, 요약, 시각화 등이 포함되며 이들은 모두 인간의 인식 행위의 기본이 되는 분류에 기반하고 있다. 지식 분류의 도구로서의 클러스터링은 1960년대 후반에 용어 분류와 문헌 분류에 대한 응용 연구로 시작되었으며, 1990년대에 들어서 컴퓨터 처리 능력의 향상과 접근 가능한 정보의 폭증에 힘입어 클러스터링에 대한 관심이 다시 커지고 있다. 특히 최근에는 지식 분류와 시각화 분야에서 많은 응용이 시도되고 있다.

클러스터링 기법은 접근방법에 있어서 대상 물간 유사도를 이용한 계층적 클러스터링 기법과 자기발견적 클러스터링 기법의 두 종류가 있는데, 최근 문헌이나 용어의 클러스터링에 주로 사용되는 것은 계층적 기법이다.

어떤 대상을 분류하기 위한 계층적 클러스터링 모형은 대상 항목의 선정, 분류자질의 빈도 행렬 작성, 유사계수의 적용, 클러스터 생성 기법의 적용 등 여러 단계로 구성된다. 각 단계마다 다양한 경우의 수가 있으므로 클러스터링 결과도 적용한 모형에 따라서 달라지게 된다. 즉 동일한 대상에 대해서 상이한 유사계수를 적용하거나 상이한 클러스터 생성 기법을 적용하면 클러스터링 결과도 다르게 나타난다. 또한 동일한 유사계수와 클러스터 생성 기법을 적용하더라도 클러스터링 대상이 무엇인가에 따라 클러스터링 성능이 달라질 수 있다. 예를 들어 동일한 말뭉치를 대상으로 얻어진 용어-문헌 행렬에서 유클리드 거리와 같은 거리계수를 유사계수로 사용할 경우, 행렬을 구성하는 각 값의 특성 때문에 용어 클러스터링에서는 별다른 문제가 없으나 문헌 클러스터링에서는 유효한 결과를 얻지 못할 수 있다.

이와 같이 클러스터링은 상황에 따라서 대단히 많은 경우의 수가 생기게 되는 데도 불구하고 지식 분류와 시각화를 비롯한 최근의 여러 연구에서는 대부분 대상 항목의 특성에 따른 적절한 클러스터링 모형에 대한 비판적인 검토가 결여되어 있다. 특히 지식의 구성 단위가 되는 용어와 문헌에 대한 클러스터링에 있어서 이런 사례가 많은데, 이는 클러스터링을 용어와 문헌의 분류에 적용하면서 각기 다른 응용 영역에 대한 모형을 구분하지 않은 것도 한 가지 원인으로 판단된다. 어떤 경우에는 대부분의 클러스터링 기법이 결과적으로 별다른 차이가 없다는 전제하에 상대적으로 단순한 기법만을 적용하기도 한다.

따라서 클러스터링 모형을 구성하는 각 요소의 특징을 실험을 통해 파악하고, 한국어 문헌 분류에 적합한 클러스터링 모형을 찾아내는 연구가 필요하다. 본 연구에서는 자동분류 실험용 문헌집단을 구축한 다음, 분류자질과 유사계수 및 클러스터 생성 기법이 분류 결과에 미치는 영향을 파악함으로써 문헌기반 지식 분류를 위한 최적의 클러스터링 모형을 개발하는 것을 목적으로 한다.

2 문헌 클러스터링 관련 연구

문헌 클러스터링에 대한 초기 연구는 클러스

터 파일을 대상으로 검색 실험을 하기 위한 것이 대부분이었다. 이후 유사한 환경에서 검색 성능을 향상시킬 목적으로 여러 클러스터링 알고리즘을 검토하는 연구가 이어졌다.

1990년대에 들어와서 문헌 클러스터링은 검색 대상 문헌들을 클러스터 파일로 조직하기 위한 목적보다는 데이터베이스 또는 검색 결과의 브라우징이나 분류 자체를 목적으로 하는 연구들이 증가하였다. 예를 들어 Scatter/Gather(Cutting et al. 1992 ; Cutting, Karger, and Pedersen 1993)에서는 데이터베이스의 내용을 이용자가 능동적으로 브라우징함으로써 적합문헌을 찾아낼 수 있도록 하기 위해 클러스터링 기법을 이용하였다. 즉, 유사한 주제의 문헌들을 소집단으로 클러스터링한 다음 각 클러스터의 요약정보를 이용자에게 제공함으로써 몇 개의 선택된 클러스터만을 대상으로 다시 클러스터링 작업을 반복하여 정보요구에 가장 적합한 클러스터를 생성하도록 하였다.

문헌 클러스터링은 검색 작업 이전에 데이터베이스의 내용을 시각적으로 보여 주기 위한 연구에서 더 발전하여 일단 검색한 결과를 시각적으로 보여 주거나 브라우징할 수 있도록 하기 위한 연구로 발전하였다. 또한 최근 들어 검색된 문헌들의 클러스터링을 통해 일차 검색 결과를 자동으로 정련하여 보여 줌으로써 검색 성능을 향상시키기 위한 연구들이 나타나고 있다.

자동 분류를 위한 연구로는 OCLC의 SCORPION 프로젝트와 스탠포드대학의 SONIA 프로젝트가 있다. SCORPION 프로젝트는 전자 정보자원의 자동색인과 편목을 실험하는 연구로서 특히 듀이십진분류법(DDC)과 같은 전통적인 분류체계를 적용하는 데 초점을 두고 있다. 이 프로젝트에서는 지도학습 방식으로 전자 정보자원에 DDC 분류번호를 부여하며, 일단 분류된 문헌집단을 비계층적 클러스터링 기법인 싱글패스 클러스터링 기법을 이용하여 상세 분류하는 연구를 진행하였다.

SONIA(Service for Organizing Networked Information Autonomously) 프로젝트는 클러스터링 기법과 기계학습 기법을 이용하여 전자문헌을 자동으로 분류하는 실용적인 시스템을 개발하는 것을 목적으로 하였다(Sahami, Yusufali, and Baldonado 1998). 이용자 프로파일을 사용할 경우에는 지도학습 방식의 자동분류를 수행하며, 프로파일이 없는 경우에는 계층적 클러스터링 기법과 비계층적 클러스터링 기법을 적용하여 문헌을 분류하였다.

이와 같이 대량의 문헌을 분류하기 위해서 효율적인 문헌 클러스터링 모형을 검토하거나 새롭게 개발하려는 연구가 최근에 매우 활발하게 수행되고 있다.

3 실험 설계

3.1 문헌 클러스터링 모형

지식의 자동분류를 위한 클러스터링 모형은 분류 대상물의 선정, 분류자질의 선정, 유사계수의 선정, 클러스터링 알고리즘의 선정 등의 여러 측면으로 구성된다. 본 연구에서는 다음과 같은 요소를 변수로 하는 클러스터링 모형을 설계하여 각 모형의 성능을 평가하고, 또한 모형을 구성하는 각 변수가 문헌기반 지식 분

류 결과에 미치는 영향을 살펴보고자 한다.

(1) 분류대상물의 유형

분류 대상물인 문헌은 학술적 문헌과 비학술적 문헌의 두 가지 유형으로 구분하여 신문기사 실험집단(KFCM-CL1020)과 학술논문 실험집단(KTSET-990)을 구축하였고, 각 실험집단을 대상으로 하는 실험을 통해 문헌 유형의 차이가 클러스터링 결과에 미치는 영향을 파악한다.

(2) 분류자질

문헌 분류에는 흔히 문헌에 포함된 단어를 분류자질로 사용하게 된다. 이때 한 문헌의 모든 단어를 분류자질로 이용할 경우가 반드시 가장 좋은 성능을 보인다고 보장할 수 없으며, 잡음 정보를 배제하기 위한 일정한 기준이 필요하다. 분류자질의 선정은 처리 효율 측면에서도 자동분류에 소요되는 시간을 줄이는 효과가 크다. 선정기준으로는 문헌빈도, 장서빈도, 문헌내 단어빈도, 역문헌빈도 등을 적용하여, 상이한 분류자질 선정기준이 문헌기반 지식 분류 결과에 미치는 영향을 파악한다. 각 선정기준에 대해서 분류자질 집합의 크기를 단계적으로 변화시켜 자동분류 성능을 평가한다. 또한 다양한 용어 가중치를 사용하여 클러스터링 결과를 비교함으로써 가장 효과적인 가중치 공식을 발견하고자 한다.

(3) 유사계수

문헌과 용어 클러스터링에 사용되고 있는 유사계수로는 코사인 계수(cosine coefficient), 자카드 계수(Jaccard coefficient), 피어슨 상관계수(Pearson's correlation coefficient), 상호정보량(Mutual information), 율의 Y(Yule's Y), 카이제곱 통계량(x^2 statistic), 우도비(likelihood ratio) 등이 있다. 본 연구에서는 일반적으로 문헌이나 용어 클러스터링에서 많이 사용되고 있으며, 여러 연구(Willet 1983 ; Kim and Choi 1999)에서 유용한 척도로 밝혀진 코사인 계수와 자카드 계수를 적용한다.

(4) 클러스터링 알고리즘

클러스터링 기법은 계층적 기법과 비계층적 기법으로 구분된다. 일반적으로 클러스터링 성능에 있어서는 계층적 기법이 비계층적 기법에 비해 우수하지만 처리 시간에 있어서는 비계층적 기법이 훨씬 효율적인 것으로 나타나 있다. 계층적 기법으로는 단일연결(single linkage), 완전연결(complete linkage), 그룹평균연결(group average linkage), 워드 기법(Ward's method) 등이 있으며, 비계층적 기법으로는 싱글패스(single pass), K-means, EM (expectation maximization) 알고리즘 등이 있다. 본 연구에서는 계층적 기법 가운데 사전 실험을 통해 가장 성능이 우수한 것으로 나타난 완전연결 기법을 선택하며, 비계층적 기법으로는 일반적으로 많이 사용되고 있는 K-means 기법을 선택하여 클러스터링 성능을 비교한다.

3.2 실험용 문헌집단

3.2.1 KFCM-CL1020

KFCM은 국제 및 경제분야의 신문기사 말

〈표 1〉 KFCM-CL1020의 기사당 통계

	색인어 수	색인어 종수	문헌길이
최대	1, 141	636	8,714
최소	15	9	87
평균	187.76	123.02	1,531.48

뭉치이다. 여기에는 국내 여러 일간지의 1992년 기사가 포함되어 있는데 본 연구에서는 이 가운데 3개 신문의 4월, 7월, 10월 기사 중에서 각 달의 1일부터 4일까지의 기사 340건씩을 모아 모두 1,020건의 실험집단 KFCM-CL1020을 구성하였다. 각 기사를 한국어 자동색인기 HAM으로 색인한 결과 전체 색인어 수는 191,513개, 색인어 종수는 40,008개였으며, 기사당 평균 색인어 수는 188개, 평균 색인어 종수는 123개, 평균 문헌 길이는 1,531바이트로 나타났다〈표 1 참조〉.

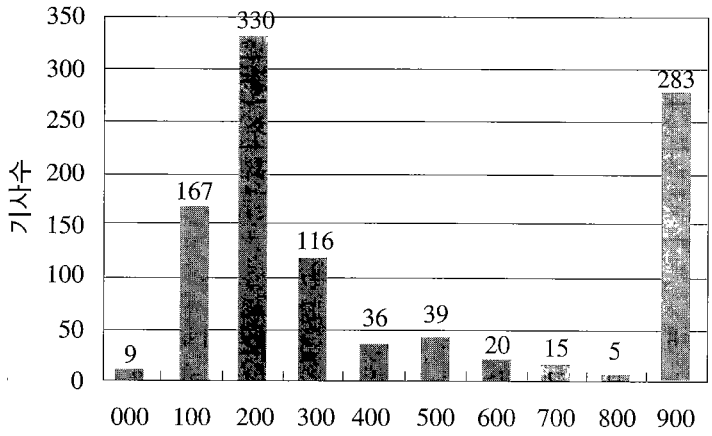
1,020건의 기사에 대해서 1992년도판 〈전국언론사 기사자료 표준분류표〉를 이용하여 분류번호를 부여하였다. 이 분류표는 십진분류표로서 000번대부터 900번대까지의 대분류 항목의 주제는 총류, 정치, 경제, 산업, 사회, 사건, 문화, 과학, 스포츠, 국제로 되어 있다. 분류 결과 기사 1,020건의 주제 분포는 '경제'가 330건으로 가장 많고 '국제'가 283건, '정치'가 167건, '산업'이 116건으로 나타나 대다수의 기사(896건)가 4개의 주제에 관련되는 것으로 나타났다.

그림 1은 기사 1,020건에 대한 분류 결과로서 각 대분류 항목별 기사 수를 보여 준다. 또한 분류 결과 실제로 부여된 소분류 항목의 총 수는 360개이며, 1건씩의 기사가 분류된 항목이 가장 많아 173개이고, 가장 많은 30건의 기사가 분류된 항목이 2개로 나타났다.

분류 수준에 따른 클러스터링 성능을 검토하기 위해서 360개의 소항목을 결합하여 모두 네 계층의 계층적인 분류체계를 만들었다. 가장 상위의 1계층에는 4개의 주제범주, 2계층에는 19개의 주제범주, 3계층에는 39개의 주제범주, 4-계층에는 114개의 주제범주를 생성하였다.

3.2.2 KTSET-990

정보과학 분야 학술논문 초록으로 구성된 KTSET 1.0은 ACM Computing Classification System에 따른 분류번호가 할당되어 있다. KTSET의 1-1000번 사이의 문헌 중 분류번호가 없는 6건과 영문초록으로만 된 4건을 제외한 990건으로 실험집합 KTSET-990을 구성하였다. KTSET의 경우에는 영어 단어와 외래어가 많이 포함되어 있으므로 한-영 표기의 차이나 외래어 표기의 차이를 해결하기 위해서, HAM으로 자동색인을 한 다음 컴퓨터 용어 사전을 이용하여 전거제어를 하였다. 예를 들어 시스템/시스팀/system은 시스템으로, 데이터/데이타/data는 데이터로 통일하였다. 전거제어 후의 전체 색인어 수는 69,885개, 색인어 종수



〈그림 1〉 KFCM-CL1020의 주제별 기사 분포

는 10,624개였으며, 기사당 평균 색인어 수는 71개, 평균 색인어 종수는 43개, 평균 문헌 길이는 626 바이트로 나타났다〈표 2 참조〉.

KTSET-990의 각 문헌은 복수 분류가 되어 있는데 이를 모두 수용하였다. 단, 분류 수준은 3단계까지만 인정하되, 한 문헌에 할당된 복수 분류번호가 H.3과 H.3.2와 같이 계층적인 주제인 경우에는 더 하위 주제범주로 통일하였다. 분류 수준에 따른 클러스터링 성능을 검토하기 위해서 3단계 수준 이외에 2단계 수준의 분류도 함께 검토하였다. 그 결과 KTSET-990의 분류에 사용된 범주의 수는 3단계 수준에서 231개, 2단계 수준에서 61개였고, 문헌당 평균 범주 수는 3단계 수준에서 1.76개, 2단계 수준에서 1.62개였다. 한 범주에 포함된 문헌의 수는 3단계에서 평균 7.55건, 2단계에서 평균 26.30건이었다〈표 3 참조〉.

3.3 클러스터링 성능의 평가척도

클러스터링 성능의 평가는 정보 검색이나 텍스트 범주화 기법의 성능 평가에 비해 어려운 점이 있다. 정보 검색과 텍스트 범주화의 평가에서 흔히 사용되는 정확률과 재현율은 각 문헌에 대한 적합 질의나 적합 범주가 미리 판정이 되어 있는 상태이므로 객관적이고 절대적인 평가가 가능하다. 그러나 클러스터링의 경우에는 생성된 클러스터가 어느 범주에 해당하는지, 또는 특정 문헌이 어느 범주로 자동 분류되었는지를 판정하기가 어렵다. 따라서 동일한

〈표 2〉 KTSET-990의 기사당 통계

	색인어 수	색인어 종수	문헌길이
최대	243	126	1,620
최소	19	13	173
평균	70.71	43.05	626.12

〈표 3〉 KTSET-990의 분류 수준별 통계

분류 수준	문헌당 평균 범주 수	범주수	범주당 문헌 수			단일 문헌 범주의 수
			평균	최대	최소	
3계층	1.76	231	7.55	66	1	68
2계층	1.62	61	26.30	185	1	6

환경에서 상대적인 평가를 하는 것이 현실적인 방안이다. 여기서 동일한 환경이란 클러스터 생성과 관련된 파라미터를 일치시키며 동일한 평가척도를 적용하는 것을 말한다.

본 연구에서는 클러스터링 성능을 단일척도로 평가하기 위해서 가중 평균 클러스터 유사도인 *WACS(Weighted Average Cluster Similarity)* 척도를 고안하였다. *WACS* 척도에서는 먼저 각 수작업 분류 범주와 자동 생성된 각 클러스터와의 유사도를 다이스 계수 공식을 적용하여 산출한 후 각 클러스터 C_j의 성능을 계산한다. 즉 클러스터 C_j에 속한 문헌이 하나 이상 속한 수작업 분류 범주를 모두 찾아서 유사도를 계산한 다음 일치하는 문헌의 수를 반영하여 가중 평균을 산출한다.

$$Sim(M_i, C_j) = \frac{2|M_i \cap C_j|}{|M_i|+|C_j|}$$

$$WACS(C_j) = \sum_{i=1}^{m} \frac{Sim(M_i,C_j)\,|M_i \cap C_j|}{|C_j|} = \frac{1}{|C_j|}\sum_{i=1}^{m} \frac{2|M_i \cap C_j|^2}{|M_i|\ |C_j|}$$

클러스터링 기법의 전체 성능$WACS(C)$는 각 클러스터에 대한 $WACS(C_j)$를 모두 합한 후 클러스터 크기를 반영한 가중 평균을 구하여 산출한다.

범주와 클러스터 사이의 유사도 산출을 위해 코사인 계수가 아닌 다이스 계수를 사용한 이유는 코사인 계수 공식이 분모에 벡터의 길이를 곱하도록 되어 있기 때문에 긴 벡터의 경우

$$WACS(C) = \frac{1}{D}\sum_{j=1}^{n} WACS(C_j)|C_j| = \frac{1}{D}\sum_{j=1}^{n}\sum_{i=1}^{m} \frac{2|M_i \cap C_j|^2}{|M_i|+|C_j|}$$

분모가 커지므로 상대적으로 작은 값을 가지게 되는 경향이 있기 때문이다(Salton and McGill 1983, 203). 즉, 클러스터를 크게 분할한 경우가 작게 분할한 경우에 비해서 낮은 값을 가지게 된다. 이보다 더 큰 문제는 두 클러스터 사이의 크기의 편차가 큰 경우에 곱한 결과가 작아지므로, 평가 기준이 되는 수작업 분류 범주의 수가 자동 생성한 클러스터의 수와 큰 차이가 날수록 좋은 성능으로 평가될 여지가 많다는 점이다.

4 용어 가중치에 따른 클러스터링 성능 평가

정보검색에서 문헌을 표현하는 색인어에 적절한 가중치를 부여하는 문제는 오래 동안 연구되어 온 과제이다(Salton and Buckley 1988). 특히 최근에는 전문검색이 활발해지면

서 가중치 공식의 중요성이 더욱 강조되고 있으며, 문헌 자동분류에 있어서도 분류자질이 사용되는 용어에 적절한 가중치를 부여할 필요가 있다. 그러나 정보검색에서 좋은 성능을 보이는 가중치 공식이 자동분류에서도 좋은 성능을 보인다고 볼 수는 없다. 따라서 본 연구에서는 검색 환경에서 제안된 기존의 가중치 공식들이 자동분류 환경에서 어떤 성능을 보이는지를 실험을 통해 파악하고자 한다.

4.1 용어 가중치 공식

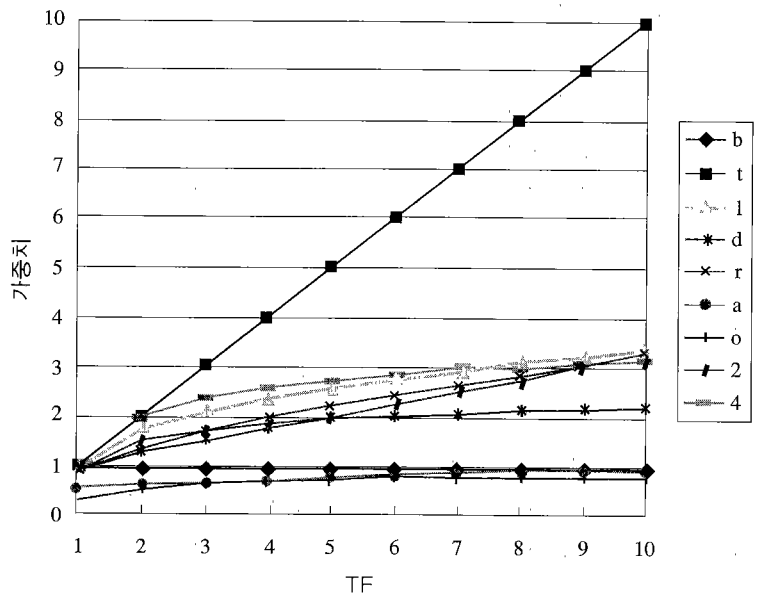
용어 가중치를 구성하는 요소는 단어빈도(TF), 역문헌빈도(IDF), 문헌길이 정규화(normalization)의 세 가지이다. 즉, 단어빈도만을 사용하는 경우, 단어빈도와 역문헌빈도를 함께 사용하는 경우, 그리고 앞의 두 경우의 가중치를 정규화시키는 경우 등이 실험 대상이 된다. 본 실험에서는 다양한 단어빈도 공식과 정규화 공식을 적용하였으며, 모든 경우 역문헌빈도로는 $log_2 N/DF$ 값을 사용하였다. 각 가중치 공식은 세 가지 구성요소를 나타내기 위하여 세 자리 기호로 표시하였다. 즉, tin은 단순 단어빈도(t=tf)와 역문헌빈도(i=idf)로 구성되며 정규화는 적용하지 않은 경우(n=no normalization)를 표시한다.

단어빈도는 문헌 내 출현여부만을 반영하는 이진값이나 출현빈도 자체를 가중치로 사용할 수 있으며, 이외에 다양한 공식이 제안되어 있다. 〈표 4〉는 본 실험에서 사용한 단어빈도 가중치 공식이다.

표 4의 공식 가운데 마지막에 나와 있는 〈더블로그 2 TF〉와 〈루트직선 TF〉는 본 실험에서 제안한 공식으로서 전자는 단어빈도가 0, 1, 2일때는 가중치도 0, 1, 2가 되고 단어빈도가 3 이상일 때에는 가중치가 로그 곡선을 따

〈표 4〉 단어 빈도 가중치 공식

이 름	공 식	기 호
이진 TF	$TF = 1(\text{if } tf > 0), 0$	b
단순 TF	$TF = tf$	t
로그 TF	$TF = 1 + \log(tf)$	l
더블로그 TF	$TF = 1 + \log(1 + \log(tf))$	d
루트 TF	$TF = \sqrt{tf}$	r
보정 TF	$TF = (1 - w) + w \times \frac{tf}{max\_tf}$	a
오카피 TF	$TF = \frac{tf}{2 + tf}$	o
더블로그 2 TF	$TF = 1 + \log_2(1 + \log_2(tf))$	2
루트직선 TF	$TF = \frac{tf + 3}{4}$	4



〈그림 2〉 TF에 따른 각 공식값의 변화

르도록 고안한 것이며, 후자는 예비실험에서 좋은 성능을 보인 〈루트 TF〉 공식과 가중치 선의 기울기가 유사하도록 고안한 직선 공식이다. 위의 공식 가운데 〈보정 TF〉는 문헌길이로 정규화시킨 것과 마찬가지이므로 다른 단어빈도 가중치와는 차이가 있다.

이상의 여러 가지 단어빈도 가중치 공식은 결국 그림 2와 같이 가중치 값에 의해 단순 단어빈도를 어느 범위로 축소시키는 가를 의미한다. 그림에서 보는 바와 같이 가중치 값을 나타내는 선의 기울기는 〈이진 TF〉가 0으로 가장 낮고 〈단순 TF〉가 1로 가장 높다. 다른 공식들은 모두 〈이진 TF〉와 〈단순 TF〉 사이에서 각기 다른 기울기를 표현한 것으로 볼 수 있다.

각 공식별로 실험문헌에 출현한 최저빈도어와 최고빈도어에 대한 가중치 값의 비율을 살펴보았다. KFCM-CL 1,020건의 최고빈도어의 평균은 약 9.6이며, 가장 높은 값은 69인 것으로 나타났다. 따라서 tf=1일 때와 tf=10일 때, 그리고 tf=1일 때와 tf=69일 때의 가중치 값의 비율을 계산한 결과가 표 5에 나와 있다.

〈표 5〉를 보면 이진 TF(b)와 단순 TF(t)를 제외하였을 때 가중치 값의 비율이 높게 나타난 공식은 로그 TF(l), 루트 TF(r), 루트직선 TF(4)로서 이들 공식은 고빈도어에 대한 가중치 값이 상대적으로 높아 저빈도어와 고빈도어

〈표 5〉 공식별 저빈도어와 고빈도어의 가중치 비율

	b	a	d	o	2	r	4	l	t
w(tf=10)/w(tf=1)	1.00	1.82	2.20	2.50	3.11	3.16	3.25	3.30	10.00
w(tf=69)/w(tf=1)	1.00	1.97	2.66	2.92	3.83	8.31	18.00	5.23	69.00

의 식별력이 다른 공식에 비해 우수할 수 있다는 것을 보여 준다.

전문검색에서는 길이가 긴 문헌일수록 각 단어의 출현빈도가 높고 출현하는 단어의 종류가 많기 때문에 짧은 문헌에 비해서 검색될 확률이 높다. 이는 클러스터링에서도 마찬가지여서 길이가 긴 문헌일수록 다른 문헌과의 유사도가 상대적으로 높아질 여지가 있으므로 정규화가 필요하다. 〈표 6〉은 본 연구에서 적용한 문헌길이를 반영한 정규화 공식이다. 각 공식을 표현한 기호 조합에서 가운데 ? 기호는 역문헌빈도의 적용여부를 나타낸다. 마지막 기호는 정규화 유형을 나타내는 것이며 정규화를 시키지 않을 때는 기호 n이 사용된다.

용어 가중치와 관련된 선행연구들을 살펴본 결과 문헌 클러스터링에서 단어빈도 가중치 공식으로는 〈단순 TF〉보다는 기울기가 낮은 〈로그 TF〉나 〈루트 TF〉 공식, 심지어는 〈이진 TF〉 공식이 좋은 경우도 있는 것으로 나타났다(Cutting et al. 1992 ; Schütze and Silverstein 1997 ; Sahami, Yusufali, and Baldonado 1998 ; Wong and Fu 2000). 한편 역문헌빈도는 〈단순 TF〉와 결합될 경우에 부정적인 것으로 나타나 있다.

본 연구에서는 이와 같은 선행연구의 주장을 검토한 결과 다음의 질문을 가지고 용어 가중치

〈표 6〉 문헌길이 정규화

명 칭	공 식	기호	기호조합
코사인 정규화	$TW = \frac{w}{\sqrt{\sum w_i^2}}$	c	t?c
최대 TF 정규화	$TW = (1-w) + w \times \frac{tf}{max\_tf}$	a	t?a(=a?n)
바이트길이 정규화	$TW = \frac{tf}{2 \times (1-b+b \times \frac{document\ length}{average\ document\ length}) + tf}$	o	t?o
피벗 유니크 정규화	$TW = \frac{\frac{1+\log(tf)}{1+\log(average\ tf)}}{0.8+0.2 \times \frac{unique\ tf}{average\ unique\ tf}}$	u	u?u
피벗 바이트길이 정규화	$TW = \frac{tf}{0.8+0.2 \times \frac{length\ of\ document\ (in\ bytes)}{average\ unique\ of\ document\ (in\ bytes)}}$	b	t?b l?b d?b
로그 정규화	$TW = \frac{1+\log(tf)}{1+\log(total\ tf)}$	l	l?l
단어빈도 정규화	$TW = \frac{tf}{0.5+1.5 \times \frac{total\ tf}{average\ total\ tf} + tf}$	w	t?w
피벗 단어빈도 정규화	$TW = \frac{tf}{0.8+0.2 \times \frac{total\ tf}{average\ total\ tf}}$	x	t?x l?x d?x
루트 정규화	$TW = \sqrt{\frac{tf}{\sum tf}}$	r	r?r

에 따른 문헌 클러스터링 실험을 수행하였다.

첫째, 단어빈도 가중치 공식에서 단순 TF 값을 누그러뜨리는 정도, 즉 그림 2에서의 곡선의 기울기는 클러스터링 성능에 어떤 영향을 미치는가? 과연 단순 TF 공식은 기울기가 낮은 공식에 비해서 반드시 낮은 성능을 보이는가?

둘째, 역문헌빈도의 적용은 클러스터링 성능에 어떤 영향을 미치는가?

셋째, 문헌길이 정규화는 클러스터링 성능에 어떤 영향을 미치는가? 긍정적인 영향을 미친다면 어느 공식이 가장 좋은 성능을 보이는가?

4.2 완전연결 기법을 사용한 용어 가중치 실험

4.2.1 단어빈도/역문헌빈도 가중치 실험

완전연결 클러스터링에서 클러스터 수를 400개부터 200개까지 25개씩 줄이면서 9단계에서 클러스터를 생성한 결과를 수작업 분류 범주 360개 및 39개와 각각 비교해 보았다. 이 실험은 소분류와 대분류 수준에서 단어빈도 가중치 공식의 성능을 비교하기 위한 것으로서 유사계수 공식으로는 코사인 계수와 자카드 계수를 사용하였다.

〈표 7〉은 실험집단 KFCM-CL에서 360범주를 기준으로 한 완전연결 클러스터링의 성능

〈표 7〉 가중치 공식별 완전연결 클러스터링 성능 비교

(k=400부터 200까지 9단계 평균/360범주 기준)

가중치	유사계수	b	a	d	o	2	r	4	l	t
TF	코사인	0.52776	0.54489	0.53806	0.53971	0.54002	0.54557	**0.54614**	0.54287	0.52495
	자카드	0.51832	0.54528	0.52875	0.53181	0.53419	**0.54801**	0.53321	0.53827	0.51833
TF · IDF	코사인	0.53609	0.55549	0.56221	0.56592	0.55749	**0.57022**	0.56729	0.56521	0.55315
	자카드	0.53738	0.55484	0.55239	0.55439	0.54801	0.55826	**0.56130**	0.55558	0.55261

*굵은 글씨는 여러 가중치 공식 중 최대값
*밑줄은 단순 TF 관련 가중치보다 높은 경우

〈표 8〉 가중치 공식별 완전연결 클러스터링 성능 비교

(k=400부터 200까지 9단계 평균/39범주 기준)

가중치	유사계수	b	a	d	o	2	r	4	l	t
TF	코사인	0.17787	0.18817	0.18469	0.18544	0.18506	0.18732	**0.18915**	0.18747	0.18636
	자카드	0.16812	**0.18617**	0.17704	0.17923	0.18227	0.18601	0.18127	0.18135	0.17918
TF · IDF	코사인	0.18189	0.19092	0.19539	0.19691	0.19408	0.19970	**0.20067**	0.19786	0.19998
	자카드	0.18068	0.19130	0.18937	0.19217	0.18869	0.19514	0.19541	0.19262	**0.19559**

*굵은 글씨는 여러 가중치 공식 중 최대값
*밑줄은 단순 TF 관련 가중치보다 높은 경우

을 나타내고 있는데 모든 가중치 공식이 *WACS* 값 0.5 이상인 좋은 성능을 보이고 있다. 이 실험에서 단어빈도만을 사용하였을 경우에 비해 역문헌빈도를 함께 사용한 가중치 공식이 더 높은 성능을 가져왔으며, 또한 단순 TF와 이진 TF에 비해 다른 모든 가중치 공식이 더 높은 성능을 나타내고 있다. 유사계수 공식에 있어서는 코사인 계수를 사용한 경우가 자카드 계수를 사용한 경우보다 나은 성능을 보인다.

〈표 8〉은 39범주를 기준으로 했을 때의 완전연결 클러스터링의 성능을 보여 준다. 39범주를 기준으로 한 성능은 360범주를 기준으로 하였을 때와는 다른 양상을 보이고 있다. 평균 성능은 *WACS* 값 0.2 미만으로서 상당히 낮은 성능을 보인다. 유사계수에 따른 차이도 현저하지가 않으며, 단순 TF의 성능이 다른 가중치 공식과 별 차이가 없는 것으로 나타났다. 360범주를 기준으로 하였을 때보다 폭은 좁지만 여전히 역문헌빈도를 적용하였을 경우가 성능이 좋게 나타났다.

〈표 9〉와 〈표 10〉은 실험집단 KTSET을 대상으로 동일한 실험을 수행한 결과를 보여 준다. 완전연결 클러스터링 결과는 각각 수작업 분류 결과인 231범주와 61범주를 기준으로 하여 평가하였다. KTSET을 대상으로 한 실험에서는 역문헌빈도를 적용한 가중치가 단어빈도만 사용한 가중치에 비해 다소 성능이 높기는

〈표 9〉 가중치 공식별 완전연결 클러스터링 성능 비교

(k=400부터 200까지 9단계 평균/231범주 기준)

가중치	유사계수	b	a	d	o	2	r	4	l	t
TF	코사인	0.34700	0.37108	0.36813	0.37391	0.37778	**0.37912**	0.36952	0.37857	0.36649
	자카드	0.34656	0.37137	0.36738	0.37611	0.38239	0.38424	0.37138	0.38517	0.36667
TF · IDF	코사인	0.36870	0.37873	0.38545	0.38174	0.37711	0.38700	0.38381	0.38765	**0.39053**
	자카드	0.36239	0.37773	0.38104	0.38433	0.37729	0.38519	0.38069	0.37973	**0.38988**

*굵은 글씨는 여러 가중치 공식 중 최대값
*밑줄은 단순 TF 관련 가중치보다 높은 경우

〈표 10〉 가중치 공식별 완전연결 클러스터링 성능 비교

(k=400부터 200까지 9단계 평균/61범주 기준)

가중치	유사계수	b	a	d	o	2	r	4	l	t
TF	코사인	0.15694	0.16650	0.16791	0.16882	**0.17341**	0.16978	0.16748	0.16955	0.16445
	자카드	0.15517	0.16645	0.16769	0.16961	**0.17487**	0.17276	0.16677	0.17168	0.16376
TF · IDF	코사인	0.16159	0.16765	0.17328	0.16961	0.16787	0.17301	0.17127	0.17373	**0.17480**
	자카드	0.15978	0.16941	0.16844	0.17011	0.16945	0.17209	0.17108	0.16784	**0.17361**

*굵은 글씨는 여러 가중치 공식 중 최대값
*밑줄은 단순 TF 관련 가중치보다 높은 경우

〈표 11〉 문헌길이 정규화 공식별 완전연결 클러스터링 성능 비교

(k=400부터 200까지 9단계 평균/360범주 기준)

가중치	b?n	t?n	t?c	t?a	t?o	u?u	t?b	l?b	d?b	l?l	t?w	t?x	l?x	d?x	r?r
TF	0.51832	0.51833	0.52495	0.54528	0.51893	0.54027	0.52451	0.53308	0.53168	0.54152	0.53601	0.52724	0.53984	0.53751	0.54557
TF·IDF	0.53738	0.55261	0.54742	0.55484	0.53645	0.56620	0.54742	0.54091	0.53879	0.56369	0.55947	0.55011	0.56235	0.55786	0.56886

*굵은 글씨는 여러 가중치 공식 중 최대값

*밑줄은 '단순 TF+비정규화' (t?n)보다 높은 경우

하지만 KFCM-CL에 비해 차이가 크지 않다. 자카드 계수를 사용하였을 경우에는 오히려 단어빈도만을 사용한 경우의 성능이 더 나은 가중치 공식(로그 TF, 더블로그 TF, 루트 TF)을 볼 수 있다.

4.2.2 문헌길이 정규화 실험

문헌길이 정규화 공식은 비교 대상인 두 문헌 벡터에 모두 적용할 경우에는 코사인 계수에서는 대부분 효과가 나타나지 않는다. 이는 코사인 계수에서는 분자와 분모에서 정규화 항이 소거되기 때문이다. 따라서 정규화 실험은 자카드 계수만을 사용하여 수행하였으며, 실험 집단으로는 KFCM-CL을 사용하였다.

완전연결 클러스터링에서 클러스터 수를 400개부터 200개까지 25개씩 줄이면서 9단계로 클러스터링하여 수작업 분류결과(범주 360개)를 기준으로 성능을 평가한 결과는 표 11와 같다. 〈표 11〉에서 가중치 공식은 세 자리 기호로 표현되는데 예를 들어 〈t?c〉는 〈단순 TF, IDF 적용여부, 코사인 정규화〉를 나타낸다. 〈표 12〉에서 〈b?n〉과 〈t?n〉은 정규화하지 않은 이진 TF와 단순 TF 적용 가중치로서 정규화한 가중치의 성능을 평가하기 위해 베이스라인 가중치로 사용하였다.

전반적으로 역문헌빈도를 적용하지 않은 가중치들은 단순 TF(tnn)와 비교하였을 때 문헌길이 정규화를 하였을 경우에 성능이 약간씩 향상된 것을 볼 수 있다. 그러나 역문헌빈도를 적용했을 경우(tin)에는 정규화 결과가 오히려 성능이 낮은 경우도 많이 나타났다. 특히 바이트길이 정규화(t?o)나 피벗바이트길이 정규화(t?b, l?b, d?b)는 모두 성능이 나쁘게 나타났으며, 바이트길이 대신 단어 수로 정규화한 경우(t?x, l?x, d?x)가 상대적으로 나은 성능을 보였다.

정규화 결과 성능이 가장 좋은 경우는 역문헌빈도의 적용 여부에 상관없이 루트 정규화(r?r)였으며, 역문헌빈도를 적용한 경우에는 피벗유니크 정규화(u?u)와 로그 정규화(l?l)가 그 다음으로 나타났다. 이 세 공식의 공통점은 모두 분자와 분모항 모두를 일정한 방식으로 처리했다는 것이다. 즉, 분자는 기본적인 TF 가중치, 분모는 단어 총 빈도나 종수(total number of tokens or types)를 쓰고, 분자와 분모 각각에 대해서 루트나 로그를 취하거나 평균으로 나누는 방법을 동일하게 적용한 것이다. 루트정규화 공식이 이중에서도 가장 좋은 것은 원래 루트 TF 공식의 성능이 좋은 영향인 것으로 판단된다. 전체적으로 볼 때, 정보검색에서 좋은 효과를 보이는 문헌길이 정규화 공식에 비해서 단순한 로그정규화(l?l)나 루트정규화(r?r) 공식이 클러스터링에서 더 좋은 성능을 보였다.

4.3 K-means 기법을 사용한 용어 가중치 실험

비계층적 클러스터링에서 용어 가중치에 따른 성능의 차이를 보기 위하여 K-means 클러스터링 기법을 사용하여 앞에서의 완전연결 기법과 같은 순서로 실험을 진행하였다. 완전연결 클러스터링에서는 분류자질을 축소하지 않고 모든 단어를 자질로 사용하였으나 K-means에서는 k값이 커짐에 따라 클러스터링 속도가 저하되는 것을 고려하여 5장에 나오는 자질 축소 실험 결과 장서빈도가 10이상인 단어만을 자질로 사용하였다.

K-means 클러스터링은 실험집단 KFCM-CL만을 대상으로 수행되었다. 먼저 k=200으로 하여 클러스터링한 결과를 수작업 분류결과인 360범주와 비교하여 평가하였고, 다시 k=39로 하여 수작업 분류결과인 39범주와 비교하여 평가하였다. 각 실험에서 k개의 종자문서를 첫 번째 문서에서 시작하여 n/k 간격으로 선정하고 다시 두 번째 문서에서 시작하여 n/k 간격으로 2회 선정하였다. 따라서 두 집단의 종자문서를 가지고 각각 클러스터링한 결과를 평균내어 성능을 평가하였다.

k=200에서의 실험결과는 〈표 12〉에서와 같이 단어빈도만을 사용한 가중치와 역문헌빈도를 적용한 가중치 사이에 뚜렷한 성능 차이를 보이지 않고 있다. k=39에서의 실험결과

〈표 12〉 가중치 공식별 K-means 클러스터링 성능 비교 : KFCM-CL의 경우

(CW1 기준 7.66% 자질집합 사용/k=200/360범주 기준)

가중치	유사계수	b	a	d	o	2	r	4	l	t
TF	코사인	0.37140	0.38909	0.39821	0.40304	**0.40520**	0.40369	0.39735	0.40419	0.39945
	자카드	0.36835	0.39426	0.39822	0.40404	0.40471	0.40885	0.40717	**0.41384**	0.39838
TF · IDF	코사인	0.37992	0.39457	0.39995	0.40186	0.40328	0.40888	0.40748	**0.41255**	0.40601
	자카드	0.38490	0.39864	0.40145	0.40133	0.40444	**0.41299**	0.41246	0.40602	0.40416

*굵은 글씨는 여러 가중치 공식 중 최대값
*밑줄은 단순 TF 관련 가중치보다 높은 경우

〈표 13〉 가중치 공식별 K-means 클러스터링 성능 비교 : KFCM-CL의 경우

(CW1 기준 7.66% 자질결합 사용/k=39/39범주 기준)

가중치	유사계수	b	a	d	o	2	r	4	l	t
TF	코사인	0.20435	0.23091	0.22581	0.23224	0.23374	0.23657	**0.24640**	0.23937	0.23695
	자카드	0.20967	0.22812	0.22910	0.23797	0.23264	0.23936	**0.24251**	0.23344	0.24013
TF · IDF	코사인	0.19532	0.21294	0.21509	0.21733	0.21744	0.22458	0.22809	0.21997	**0.23142**
	자카드	0.19817	0.21707	0.21856	0.22278	0.22332	0.22075	**0.23308**	0.22998	0.22111

*굵은 글씨는 여러 가중치 공식 중 최대값
*밑줄은 단순 TF 관련 가중치보다 높은 경우

는 〈표 13〉에 나와 있는데 모든 경우 단어빈도만을 사용한 가중치가 역문헌빈도까지 고려한 가중치에 비해 성능이 우수한 것으로 나타났다.

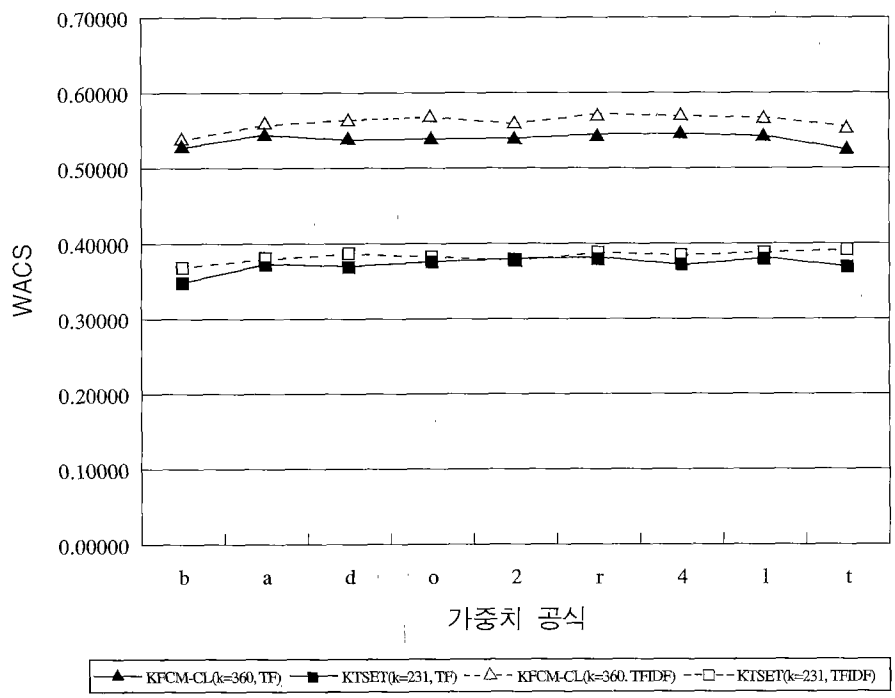
4.4 용어 가중치 실험 결과 분석

다른 특성을 가진 2개의 실험집단을 대상으로 수행한 용어 가중치 실험에서 신문기사 원

〈표 14〉 실험집단별 용어 가중치 실험 결과 비교

(완전연결 기법/소분류기준/코사인 계수)

가중치	유사계수	b	a	d	o	2	r	4	l	t
TF	KFCM-CL (k=360)	0.52776	0.54489	0.53806	0.53971	0.54002	0.54557	0.54614	0.54287	0.52495
	KTSET (k=231)	0.34700	0.37108	0.36813	0.37391	0.37778	0.37912	0.36952	0.37857	0.36649
TF · IDF	KFCM-CL (k=360)	0.53609	0.55549	0.56221	0.56592	0.55749	0.57022	0.56729	0.56521	0.55315
	KTSET (k=231)	0.36870	0.37873	0.38545	0.38174	0.37711	0.38700	0.38381	0.38765	0.39053

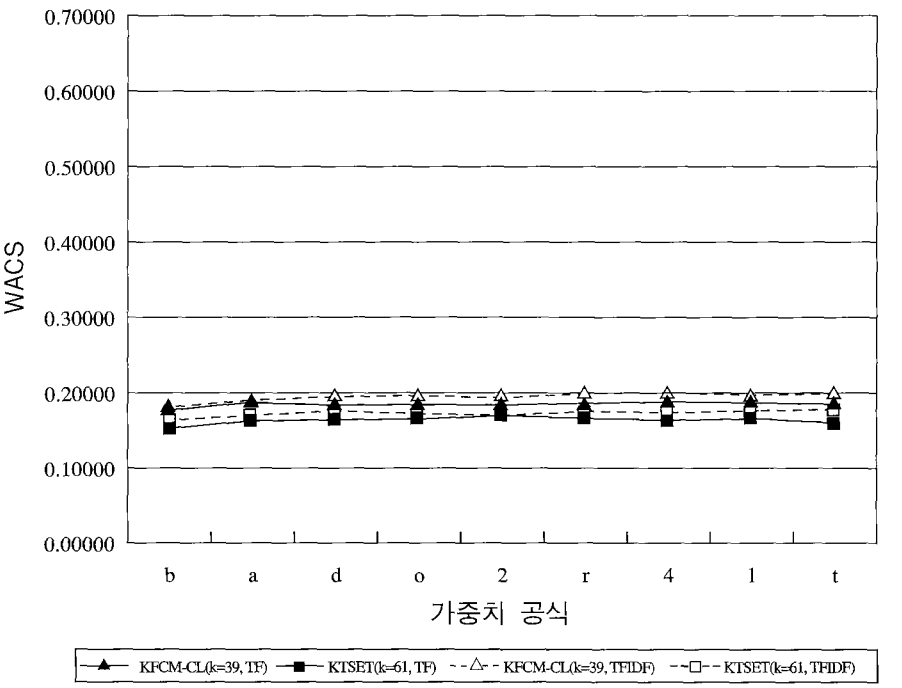


〈그림 3〉 실험집단별 용어 가중치 실험 결과 비교
(완전연결 기법/소분류기준/코사인 계수)

〈표 15〉 실험집단별 용어 가중치 실험 결과 비교

(완전연결 기법/대분류기준/코사인 계수)

가중치	유사계수	b	a	d	o	2	r	4	l	t
TF	KFCM-CL (k=39)	0.17787	0.18817	0.18469	0.18544	0.18506	0.18732	0.18915	0.18747	0.18636
	KTSET (k=61)	0.15694	0.16650	0.16791	0.16882	0.17341	0.16978	0.16748	0.16955	0.16445
TF · IDF	KFCM-CL (k=39)	0.18189	0.19092	0.19539	0.19691	0.19408	0.19970	0.20067	0.19786	0.19998
	KTSET (k=61)	0.16159	0.16765	0.17328	0.16961	0.16787	0.17301	0.17127	0.17373	0.17480



〈그림 4〉 실험집단별 용어 가중치 실험 결과 비교
(완전연결 기법/소분류기준/코사인 계수)

문을 수록한 KFCM-CL과 학술논문 초록을 수록한 KTSET은 상이한 실험결과를 보이고 있다. 두 실험집단의 차이를 한 눈에 보기 위하여 앞의 실험결과를 비교하여 〈표 14-15〉, 〈그림 3-4〉에 정리하였다. 각 그림에서 Y-축의 *WACS*는 0에서 1사이의 값을 갖지만, 용어 가중치에 따른 클러스터링 성능의 차이가 눈에 잘 띄도록 값의 범위를 임의로 표시하였다. 〈표

14〉와 〈그림 3〉은 소분류를 기준으로 하였을 때 두 실험집단에서의 클러스터링 성능을 나타내며, 또한 단어빈도 가중치와 역문헌빈도를 적용한 가중치에 따른 성능의 차이를 보여 준다. 반면 〈표 16〉과 〈그림 4〉는 대분류를 기준으로 하였을 때의 성능을 비교한 것이다. 그러나 두 실험집단이 상이한 문서들을 포함하고 있으며, 또한 다른 분류체계를 사용하여 수작업 분류를 하였기 때문에 클러스터링 결과의 차이를 일반화할 수는 없다. 다만 두 상이한 실험집단에서도 가중치의 효과가 유사한 경향을 보이는가를 평가할 수 있다.

다음은 용어 가중치 실험결과 발견된 중요한 사실을 요약한 것이다.

(1) 단어빈도만을 사용한 가중치와 역문헌빈도를 함께 사용한 가중치의 실험에서 KFCM-CL에서의 클러스터링 성능이 KTSET에 비해 월등히 좋은 것으로 나타났다. 또한 수작업 대분류 결과와 비교하였을 때보다 소분류 결과와 비교하였을 때 두 실험집단에서의 성능 차이가 더욱 뚜렷하게 나타났다.

(2) 두 상이한 실험집단에서 모두 TF · IDF 가중치를 사용하였을 경우 클러스터링 성능이 더 우수하게 나타났다.

(3) KFCM-CL에서 완전연결 기법으로 소분류(클러스터 수 400~200) 클러스터링할 경우에는 역문헌빈도를 적용한 가중치가 수작업 소분류(360범주) 결과 및 대분류(39범주) 결과와 비교하였을 때에 모두 뚜렷한 성능 향상 효과를 보였다.

(4) KTSET에서 완전연결 기법으로 소분류(클러스터 수 400~200) 클러스터링할 경우에는 수작업 소분류(231범주) 결과와의 비교에서는 역문헌빈도의 적용이 대부분 성능 향상 효과를 보였지만, 수작업 대분류(61범주) 결과와의 비교에서는 자카드 계수 사용시 성능 저하를 나타내는 경우가 있었다.

(5) KFCM-CL에서 K-means 기법으로 소분류(k=200) 클러스터링할 경우에는 역문헌빈도의 적용이 대부분 미미한 성능 향상 효과를 보였으며, 오히려 성능이 저하된 경우도 여러 가중치 공식에서 나타났다.

(6) KFCM-CL에서 K-means 기법으로 대분류(k=39) 클러스터링할 경우에는 역문헌빈도의 적용이 모든 경우에 뚜렷한 성능 저하를 가져왔다.

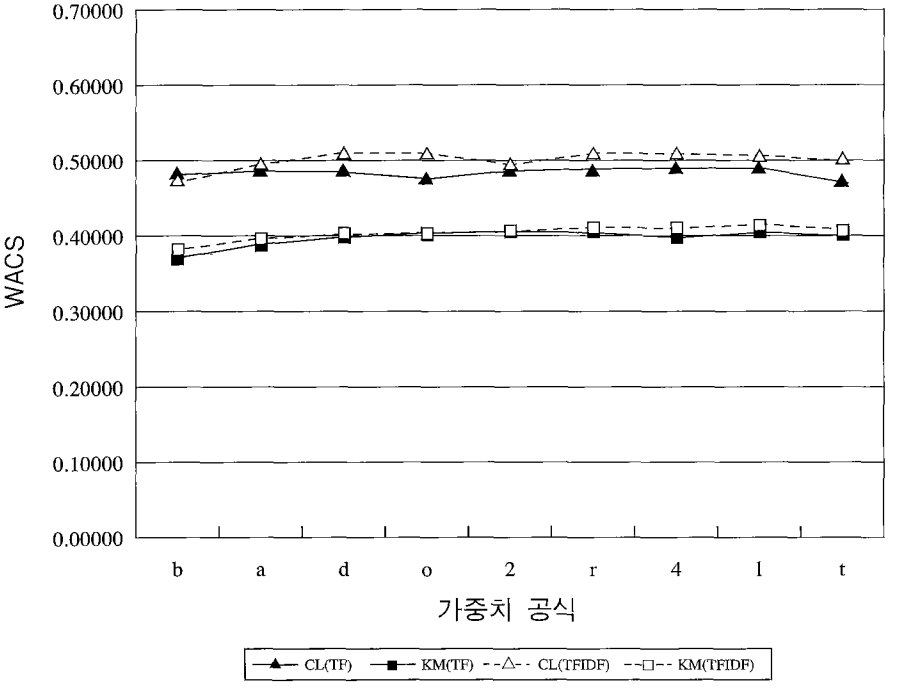
결론적으로 역문헌빈도의 적용은 완전연결 기법을 사용한 소분류 클러스터링에서는 뚜렷한 성능 향상을 가져왔으나 대분류에서는 오히려 성능 저하를 가져오는 경우가 있었다. 또한 K-means 기법을 사용한 경우에는 두 경우 모두 효과가 없는 것으로 나타났다. 이는 분류 수준의 특성상 고빈도어의 역할이 큰 대분류에서는 역문헌빈도를 적용하지 않는 것이 좋다는 것을 시사한다. 그러나 상세하게 소분류를 해야 하는 상황에서는 반대의 경우가 되므로 역문헌빈도의 적용 여부는 분류 수준을 고려하여 결정해야 할 것이다. 특히 계층적 클러스터링에서는 분류 수준에 따라서 역문헌빈도를 유동적으로 반영했을 때 성능을 최적화시킬 수 있을 것으로 보인다.

앞의 실험 결과만을 가지고는 완전연결 기법

〈표 16〉 클러스터링 기법별 시험 결과 비교

(KFCM-CL/소분류기준/코사인계수/k=200)

가중치	클러스터링 기 법	b	a	d	o	2	r	4	l	t
TF	완전연결	0.48038	0.48514	0.48429	0.47537	0.48524	0.48791	0.48868	0.48842	0.46964
	K-means	0.37140	0.38909	0.39821	0.40304	0.40520	0.40369	0.39735	0.40419	0.39945
TF · IDF	완전연결	0.47151	0.49503	0.51011	0.50950	0.49492	0.50969	0.50859	0.50656	0.49594
	K-means	0.37992	0.39457	0.39995	0.40186	0.40328	0.40888	0.40748	0.41255	0.40601



〈그림 5〉 클러스터링 기법별 실험결과비교
(KFCM-CL/소분류기준/코사인 계수/k=200)

(CL)과 K-means 기법(KM)의 성능을 직접 비교하기는 어렵다. 따라서 KFCM-CL 실험 집단에서 생성되는 클러스터의 수(k)를 200으로 지정하고 코사인 계수를 사용하여 각각 실험한 결과를 수작업 소분류(360범주) 결과와 비교하여 보았다. 〈표 16〉과 〈그림 5〉에서 보는 바와 같이 완전연결 기법의 성능이 K-means 기법보다 높은 성능을 보이고 있으며, 완전연결 기법에서만 역문헌빈도를 적용한 가중치의 효과가 뚜렷하게 나타나고 있다.

5 자질 축소에 따른 성능 평가

5.1 자질 축소의 필요성

기계학습 기반 자동분류에서 자질 축소(feature reduction) 또는 자질 집합 축소는 처리 시간의 단축과 분류 성능의 향상이라는 두 가지 목적을 가지고 있다. 지도학습(supervised learning) 방식인 범주화 기법에서는 다양한 자질 축소 방법이 연구되고 있으며, 자질 축소가 성능 향상을 위한 중요한 요소로 간주되고 있다. 그러나 범주정보가 제공되지 않는 비지도학습(unsupervised learning) 방식인 클러스터링에서는 자질축소에 대한 연구가 상대적으로 많지 않다(Liu and Motoda 1998, 182).

Dash 등(Dash et al. 1997)은 특정 자질 제거 전의 집합내 엔트로피와 제거 후의 엔트로피의 차이로 자질의 중요도를 계산하는 방법을 사용하였다. Schütze와 Silverstein (1997)은 LSI(latent semantic indexing)를 통해서 자질 차원을 축소한 결과 전체 자질을 모두 사용한 것보다 클러스터링 성능이 약간 향상되었음을 보고하였다. 한편 Vaithyanathan과 Dom(1999 ; 1999a)은 K-means 클러스터링에서 LSI를 사용한 결과 불용어를 제거한 전체 자질을 사용한 것에 비해 약간 낮은 성능을 얻었으며, 분포 클러스터링(distributional clustering)을 이용해서 자질 집합을 20% 정도로 축소한 경우에는 다소 향상된 결과를 얻었다. 그러나 이 방식들은 계산량이 지나치게 많아서 처리 시간의 단축이라는 자질 축소의 근본 목적을 달성하기는 어렵다.

따라서 복잡한 계산을 필요로 하지 않는 간단한 자질 선택 기준을 검토할 필요가 있다. 문헌 클러스터링에서 사용되는 자질 축소 방법은 벡터를 처리하는 문헌 단위의 축소와 행렬을 처리하는 장서 단위의 축소로 구분할 수 있다. 문헌 단위의 축소를 '절단(truncation)' 방법 또는 '지역적(local)' 방법이라고 하며, 장서 단위의 축소를 '차원 축소(dimension reduction)' 또는 '전역적(global)' 방법이라고도 말한다(Schütze and Silverstein 1997). 그러나 어느 경우에나 축소 작업 이전에 불용어 제거 과정을 거치게 된다.

문헌 단위의 축소 방법은 각 문헌 벡터나 센트로이드 벡터의 길이를 일정하게 만드는 방식이다. Larsen과 Aone(1999)은 자질의 가중치를 기준으로 각 문헌을 상위 n개 단어로 표현하였다. 이들이 사용한 가중치 공식은 앞에서 실험한 가중치 공식 가운데 tnn, tin, lnn, rnn의 네 가지인데, 문헌의 수가 많은 경우에는 tin을 적용한 경우가 가장 좋은 성능을 보이고 나머지 세 공식의 차이는 미미한 것으로 보고하였다. 이들은 또한 문헌당 자질의 수를 5에서 500까지 변화시켜서 실험한 결과, 문헌 벡터를 축소할수록 성능은 저하되는 결과를 얻었지만, 처리 시간 측면에서 n을 25로 하는 것이 무난한 것으로 판단하였다.

문헌 단위의 자질 축소 방법은 자질의 중요도를 문헌이라는 좁은 관점에서 평가하게 되므로 장서 단위의 자질 축소 방법에 비해서 불리한 면이 있다. 비록 역문헌빈도를 이용해서 장서 단위의 정보를 어느 정도 반영할 수는 있지만, 전체적인 측면에서 자질의 중요도를 정밀하게 반영하지는 못한다. 또한 각 문헌 벡터의

자질 수를 동일하게 만들어 줌으로써 실제 문헌의 크기 차이를 무시해버리는 것도 단점이 될 수 있다.

반면 장서 단위의 자질 축소는 역문헌빈도를 기준으로 하여 저빈도어를 제거하거나 일부 고빈도어를 함께 제거한다. 지도학습 방식의 문헌 범주화에서도 자질 집합의 축소 기준으로 문헌빈도(DF)를 사용하는 단순한 방법이 다른 복잡한 기준에 비해서 뒤떨어지지 않는 성능을 보이는 것으로 보고되어 있다(Yang and Pederson 1997). SONIA 프로젝트(Sahami, Yusufali, and Baldonado 1998)에서는 장서빈도(CF)가 2 이하거나 1000 이상인 단어를 제거한 후 남은 단어를 대상으로 문헌빈도를 기준으로 엔트로피를 계산하여 값이 낮은 15%를 다시 제거하는 방법을 사용하였다.

Wong과 Fu(2000)는 클러스터당 평균 문헌 수에 해당하는 DF 값을 기준으로 위와 아래로 같은 범위를 적절히 잡는 방법을 사용하였다. 예를 들면 1000개 문헌을 대상으로 20개 클러스터를 생성할 경우 DF=50을 기준으로 경험적으로 상, 하위 3씩 범위를 늘려서 DF가 47 이상, 53 이하인 단어를 선택하는 방법이다. 이는 DF를 기준으로 저빈도어와 고빈도어를 동시에 제거하는 방법인데 기준점을 최저빈도와 최고빈도에 두는 것이 아니라 중간빈도어에 두는 점이 특징이다. 이 방법은 문헌이 주제별로 고르게 구성되지 않은 집합일 경우에는 좋은 성능을 기대하기가 어렵다.

본 연구에서는 간단한 자질 집합 축소 기준으로 문헌빈도 이외에 장서빈도, 그리고 문헌빈도와 장서빈도의 복합 기준을 검토해보았다. 문헌빈도 이외에 장서빈도를 함께 검토하는 이유는, 여러 문헌에 고르게 출현하는 단어보다는 일부의 문헌에 집중적으로 출현하는 단어가 문헌 클러스터의 식별에 도움이 될 것이라고 가정했기 때문이다.

5.2 자질 축소 기준

본 연구의 자질 축소 실험은 KFCM-CL을 대상으로 수행되었다. 자질 축소 기준으로는 기본적으로 문헌빈도(DF)와 장서빈도(CF)를 사용하였고, 단순빈도인 DF, CF, TF를 활용한 다양한 복합기준을 고안하여 사용하였다. 단, DF를 기준으로 할 경우 각 빈도별 단어집합의 크기가 CF를 기준으로 할 경우와 다르므로 동일한 DF 값을 가진 단어들은 임의로 자모순으로 제거하여 자질 집합의 크기를 CF 기준과 맞추었다.

뒤에서 설명될 자질 축소 실험 결과 모든 복합기준이 비슷한 성능을 보였기 때문에 기본적인 CF/DF 기준과 비교할 필요가 있을 때는 공식의 성격이 분명한 CW1과 CW6의 두 기준만을 검토하였다. CW1은 일반적인 정보검색에서의 TF · IDF 공식과 유사하지만, 문헌 하나가 아닌 장서 전체를 대상으로 하므로 TF 대신 로그 CF를 적용하였고, CW6은 문헌 벡터의 연산에서 특정 자질이 미치는 전반적인 영향을 추정하기 위해서 각 문헌에서 특정 자질이 가지는 TF · IDF 가중치를 합하였다. 다음은 자질 축소 실험에서 사용한 복합기준 공식이다.

$$CW1 = \log_2 \frac{N+1}{DF} \times \log_2(CF+1)$$

$$CW2 = \frac{CF}{DF}$$

$$CW3 = \ln(CF) \times \frac{CF}{DF}$$

$$CW4 = \ln(\ln(CF)+1) \times \frac{CF}{DF}$$

$$CW5 = \sum(tf \times \log_2 \frac{N}{DF})$$

$$CW6 = \sum(\sqrt{tf} \times \log_2 \frac{N}{DF})$$

KFCM-CL 문헌 1020건에 포함된 단어 종수는 40,008개이며, 〈표 17〉은 CF를 축소 기준으로 하여 저빈도어를 제거한 결과를 보여준다.

전체 자질의 7.66% 수준으로 자질을 축소할 경우 축소된 자질 집합에 의해 각 문헌 벡터를 표현하였을 때 평균적인 길이를 계산한 결과는 〈표 18〉와 같다. 이 표에서 보는 바와 같이 복합기준 사용시 더 적은 수의 단어로 한 문헌을 표현하게 됨을 알 수 있다. 이는 복합기준을 적용할 경우 일부 고빈도어도 제거되기 때문인 것으로 보인다.

〈표 17〉 저빈도어를 제거한 자질 집합의 크기

기준	단어 종수	비율
전체	40,008	100.00%
CF〉1	15,237	38.08%
CF〉2	9,784	24.46%
CF〉3	7,281	18.20%
CF〉4	5,968	14.92%
CF〉9	3,064	7.66%

각 기준에 따라서 축소된 자질 집합이 얼마나 다른지를 알아보기 위해서 7.66%(3,064개)로 축소된 자질집합을 비교해 보았다. 그 결과 〈표 19〉에서와 같이 CF와 DF를 기준으로 축소된 집합은 서로 87%가 일치하였으며, 복합기준인 CW1과 CW6에 의해 축소된 집합은 서로 86%가 일치하였다. 반면에 CW1과 CW6을 기준으로 축소된 집합은 DF 기준 축소 집합과는 각각 70%, 61%만이 일치하였으며, CF 기준 축소 집합과는 이보다 조금 높은

〈표 18〉 7.66%(3,064)개로 축소된 자질 집합의 문헌당 단어 수 평균

	전체	CF	DF	CW1	CW6
문헌당 단어 종수 평균	123.02	67.23	67.53	52.17	58.15
문헌당 총 단어수 평균	187.76	121.90	117.79	96.59	113.81

〈표 19〉 7.66%(3,064)개로 축소된 자질 집합의 일치율 비교

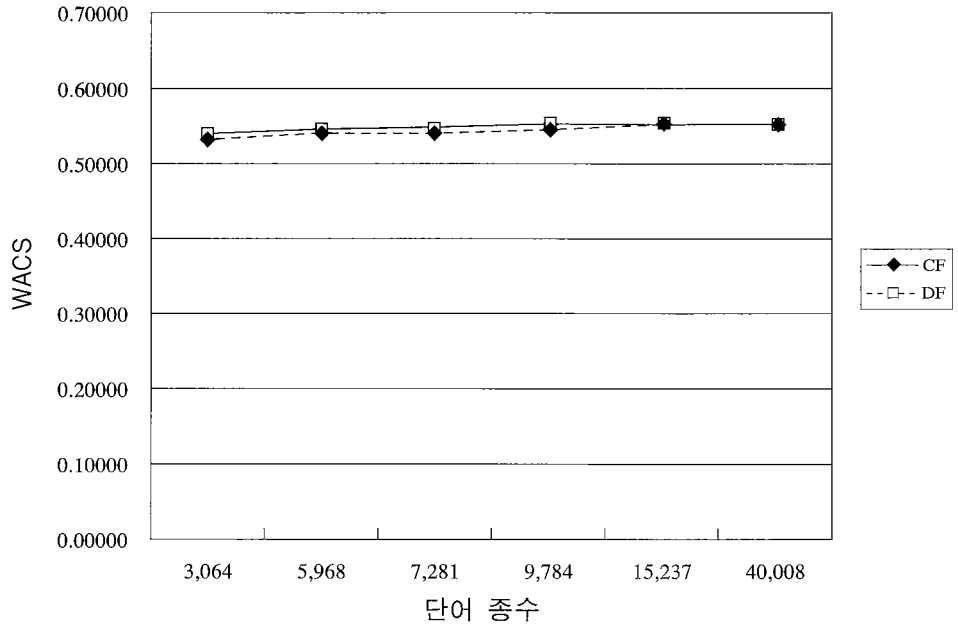
	CF	DF	CW1	CW6
CF	100%(3,064)	87%(2,653)	84%(2,565)	74%(2,262)
DF	87%(2,653)	100%(3,064)	70%(2,154)	61%(1,863)
CW1	84%(2,565)	70%(2,154)	100%(3,064)	86%(2,638)
CW6	74%(2,262)	61%(1,863)	86%(2,638)	100%(3,064)

84%, 74%의 일치율을 보였다. 기존에 많이 적용되고 있는 DF를 기준으로 축소된 자질 집합은 복합기준에 의한 축소 집합과는 1/3 정도가 불일치한다는 것을 알 수가 있다. 고빈도어의 제거 여부를 살펴보면 CW1을 기준으로 할 경우 저빈도어만 제거되는 것이 아니라 고빈도어 중에서도 최상위 고빈도어 41개(전체의 0.1%)가 제거되었다. 반면에 CW6의 경우에는 최상위 고빈도어가 제거되지 않으며 제거되는 단어중 DF가 가장 큰 것은 57인 경우(DF 순 237위)였다.

〈표 20〉 CF 기준 자질 집합 축소에 따른 완전연결 클러스터링의 성능 변화

(tim 가중치/클러스터 수 400-200개 9단계 평균/360범주 기준)

자질 축소 비율 (단어 종수)		100.00% (40,008)	38.08% (15,237)	24.46% (9,784)	18.20% (7,281)	14.92% (5,968)	7.66% (3,064)
코사인	CF 기준	0.55315	0.55265	0.55350	0.54952	0.54710	0.54138
	DF 기준	0.55315	0.55265	0.54616	0.54167	0.54186	0.53386
자카드	CF 기준	0.55261	0.55086	0.55252	0.55145	0.54639	0.53945
	DF 기준	0.55261	0.55086	0.55075	0.54123	0.54404	0.53546



〈그림 6〉 자질 집합 축소에 따른 완전연결 기법의 성능 변화 : CF와 DF 비교
(코사인 계수/클러스터 수 400-200개 9단계 평균/360범주 기준)

5.3 완전연결 기법을 사용한 자질 축소 실험

KFCM-CL 신문기사 1,020건을 대상으로 완전연결 기법에 의한 클러스터링을 수행하여 자질 축소에 따른 성능 차이를 살펴보았다. 클러스터링 성능은 360개 수작업 분류 범주를 기준으로 하되, 용어 가중치 실험에서와 마찬가지로 자동생성 클러스터의 수를 400개부터 200개까지 25개 간격으로 줄이면서 9단계로 생성하여 각 단계의 성능을 평균내어 평가하였다.

5.3.1 CF/DF를 기준으로 자질 집합을 축소한 경우

CF와 DF 기준 평가에서는 TF · IDF 가중치(tin)를 적용한 결과를 비교하였다. 실험 결과 CF 기준 자질 축소 집합의 성능이 DF 기준 자질 축소 집합의 성능과 거의 같거나 약간 높게 나타났으며, 자질 집합의 축소율이나 유사계수에 상관없이 모두 같은 결과가 나타났다. 전반적으로는 자질을 축소할 경우 모든 자질을 사용한 경우에 비해 약간 못 미치는 성능이 나타났으나 그 차이는 미미했다. 예를 들어 자질을 7.66% 수준으로 축소하였을 경우의 성능이 0.541로서 전체 자질을 사용하였을 때의 0.553에 비해 거의 차이가 없는 것을 볼 수 있다. 〈표 20〉과 〈그림 6〉은 CF/DF를 기준으로 축소한 자질 집합에서의 클러스터링 성능을 요약한 것이다.

5.3.2 복합기준으로 자질 집합을 축소한 경우

〈표 21〉은 CF와 DF, 또는 TF와 DF를 함께 사용한 복합기준을 적용하여 자질 집합을 각각 14.92%(5,968 단어)와 7.66%(3,064)로 축소한 후 클러스터링한 결과를 보여 준다. 복합기준들의 성능은 서로 별 차이가 없으며, CF와 DF의 단일 기준과도 비슷한 성능을 보여준다. 〈그림 7〉은 자질 축소를 하지 않았을 경우를 상한선으로 하여 각 축소 기준에 따른 클러스터링 성능을 보여 준다.

5.4 K-means 기법을 사용한 자질 축소 실험

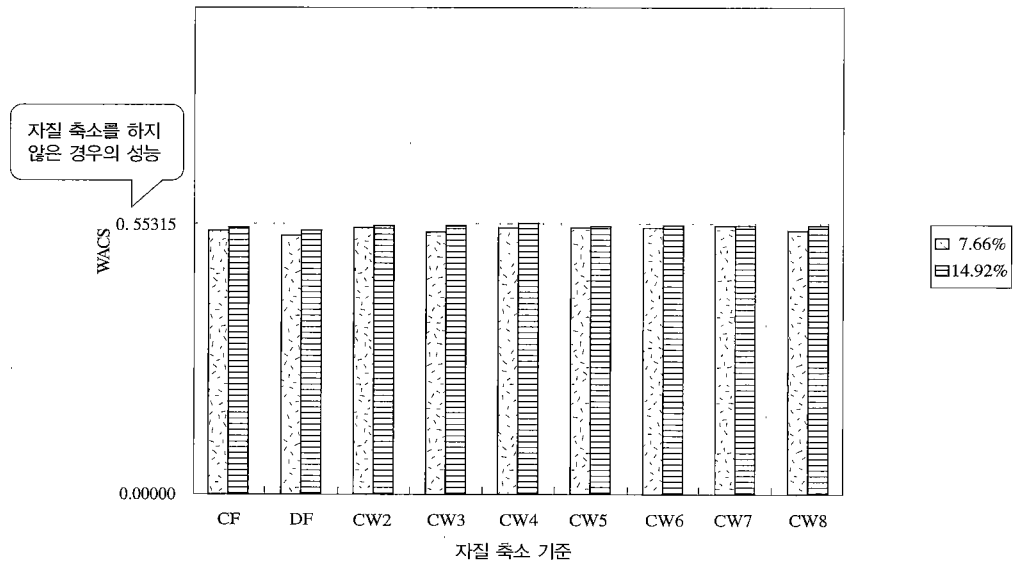
K-means 클러스터링에서는 용어 가중치 실험에서와 마찬가지로 분류 수준을 대분류와 소분류로 구분하였다. 대분류 실험에서는 k=39로 하여 클러스터링한 결과를 수작업 분류 범주 39개와 비교하였고, 소분류 실험에서는 k=200으로 하여 클러스터링한 결과를 수작업 분류 범주 360개와 비교하였다. 초기 클러스터 중심문헌은 문헌 번호에 따라서 간격을 k/N씩 일정하게 건너뛰면서 선정하되 1번 문헌부터 선정한 경우와 2번 문헌부터 선정한 경우로 2회 실험하여 평균을 산출하였다.

〈표 22〉와 〈그림 8〉에서 보는 바와 같이 K-means 클러스터링 성능은 자질을 축소한 경우가 전체 자질을 사용하였을 때 보다 대부분 성능 향상을 가져왔다. 특히 k=200인 소분류보다 k=39인 대분류의 경우에 성능 향상 효과가 뚜렷하게 나타났다. 축소 기준별 비교에서는 큰 차이는 없지만 대분류의 경우에는 DF, 소분류의 경우에는 CW6이 가장 좋은 성능을 보였다. DF는 특이하게도 대분류에서는 가장 좋은 성능을 보였지만, 소분류에서는 가장 낮은 성능을 보였다. 이는 용어 가중치로 tin을 사용하여 DF가 낮은 단어의 특성을 이

〈표 21〉 자질 집합 축소 기준에 따른 완전 연결 클러스터링의 성능 비교

(tin 가중치/클러스터 수 400-200개 9단계 평균/360범주 기준)

유사계수	자질집합 크기	전체	CF	DF	CW1	CW2	CW3	CW4	CW5	CW6	CW7
코사인	14.92%	0.55315	0.54710	0.54186	0.54921	0.54805	0.55088	0.54985	0.54500	0.54772	0.54591
	7.66%	0.55315	0.54138	0.53386	0.54550	0.53857	0.54546	0.54684	0.54442	0.54956	0.54051
자카드	14.92%	0.55261	0.54639	0.54404	0.54660	0.53530	0.54718	0.54682	0.54471	0.54511	0.54980
	7.66%	0.55261	0.53945	0.53546	0.54251	0.52506	0.54468	0.54159	0.53665	0.54414	0.53992



〈그림 7〉 자질 집합 축소 기준에 따른 완전연결 클러스터링의 성능 차이
(코사인 계수 적용/TF · IDF가중치/클러스터 수 400-200 9단계 평균/360범주 기준)

미 반영하였는데 다시 DF 기준에 의한 자질 축소에서 DF가 낮은 단어를 우선적으로 제거하게 되므로 저빈도어의 기여도가 큰 소분류에서 성능의 저하를 가져온 것으로 분석된다.

5.5 자질 축소 실험 결과 분석

완전연결 기법에서 CF와 DF를 기준으로 자질 집합을 축소한 결과 클러스터링 성능은 전체 자질을 사용하는 경우와 큰 차이가 없었으며, 두 기준을 비교하였을 경우에도 별 차이가 나타나지 않았다. 복합기준을 사용한 경우에도 CF와 DF만을 사용한 경우와 유사한 결과를 보여 주었다.

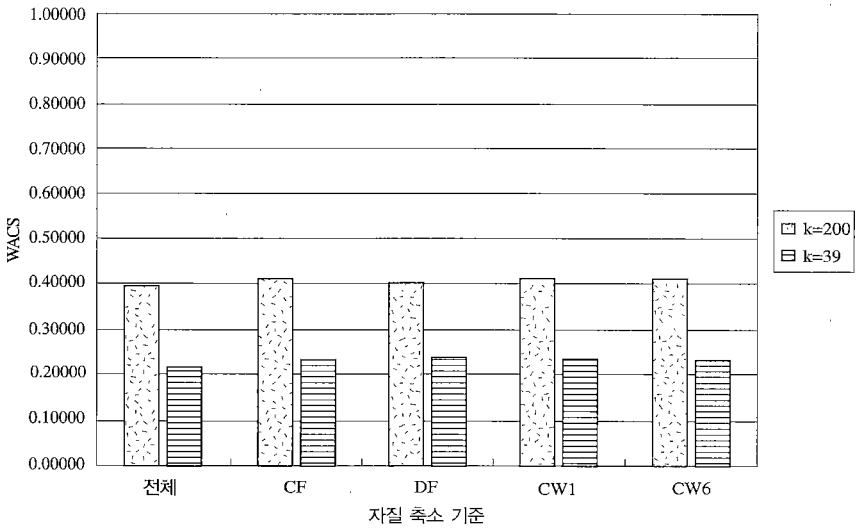
그러나 K-means 클러스터링에서는 자질 축소 결과 성능 개선 효과가 확실하게 나타

〈표 22〉 자질 집합 축소 기준에 따른 K-means 클러스터링 성능

(7.66% 자질 집합/소분류 기준)

	유사계수	전체 (40,008)	CF (3,064)	DF (3,064)	CW1 (3,064)	CW6 (3,064)
k=200	코사인	0.39666	0.40817	0.40122	0.40601	**0.41318**
	자카드	0.40102	0.40625	0.39715	0.40416	**0.40874**
k=39	코사인	0.21403	0.23164	**0.23498**	0.23142	0.23119
	자카드	0.21004	0.22832	**0.23797**	0.22111	0.23265

*굵은 글씨는 전체를 제외한 네 경우 중 최대값
*밑줄은 전체보다 높은 경우



〈그림 8〉 자질 집합 축소 기준에 따른 K-means 클러스터링 성능
(7.66% 자질 집합/코사인 계수/소분류 기준)

났으며, 소분류에서 보다는 대분류에서 성능이 뚜렷하게 향상되었다. 이는 자질 축소 결과 제거된 저빈도어는 작은 크기의 클러스터를 발견하는 데에는 긍정적인 기여를 하지만, 큰 크기의 클러스터를 발견하는 데에는 별다른 기여를 하지 못하거나 오히려 부정적인 영향을 미치기 때문인 것으로 추측된다.

결과적으로 문헌 클러스터링에서도 범주화에서와 마찬가지로 분류자질의 축소를 통해 처리 속도를 개선하면서 동시에 자동분류 성능을 전체 자질을 사용하는 경우와 비슷하게 유지하거나 오히려 향상시킬 수 있음을 알 수 있다.

6 결 론

본 연구에서는 문헌을 기반으로 한 지식의 자동분류를 위해 최적의 클러스터링 모형을 제시하고자 하였다. 이를 위해 클러스터링 모형의 구성 요소인 분류 대상물, 분류자질, 유사계수, 클러스터링 알고리즘에 따라 클러스터링 성능이 어떻게 달라지는가를 평가하는 실험을 수행하였다. 분류 대상물로는 각각 신문기사 원문과 학술지 기사 초록으로 구성되는 실험집단을 구축하였고, 분류자질로 사용한 용어의 집합은 다양한 자질 축소 기준을 적용하여 생성하였으며, 용어의 가중치로는 다양한 가중치 공식을 사용하였다. 유사계수 공식으로는 코사인 계수와 자카드 계수를 적용하였으며, 클러스터링 알고리즘으로는 비계층적 기법인 완전연결 기법과 계층적 기법인 K-means 기법을 각각 사용하였다.

본 연구의 실험결과를 요약하면 다음과 같다.

첫째, 용어 가중치에 따른 클러스터링 성능 평가 실험에서 신문기사 원문으로 구성된 실험집단에서의 클러스터링 성능이 훨씬 높게 나타났다. 그러나 이는 비교 대상이 되는 수작업 분류 결과에 의해서도 영향을 받으므로 일반화하기는 힘들다.

둘째, 용어 가중치별 성능 평가에서는 완전연결 클러스터링에서는 단어빈도만 사용한 경우보다 단어빈도와 역문헌빈도를 함께 사용한 경우의 성능이 더 좋게 나타났다. 반면 K-means 클러스터링에서는 역문헌빈도의 적용이 별다른 성능 향상을 가져오지 않는 것으로 나타났다.

셋째, 분류자질의 축소에 따른 클러스터링 성능 평가 실험에서는 자질을 7.66%까지 축소하였는데도 전체 자질을 사용하였을 경우에 비해 별다른 성능 차이가 나타나지 않았다. 오히려 K-means 클러스터링에서는 성능 향상 효과도 눈에 띄게 나타났다.

넷째, 계층적 기법인 완전연결 기법과 비계층적 기법인 K-means 기법의 성능을 비교한 결과 완전연결 기법의 성능이 높게 나타났다.

본 연구에서는 선행연구들에서와 마찬가지로 계층적 기법이 비계층적 기법보다 좋은 클러스터링 성능을 보였으며, 자질 축소에 있어서도 10% 수준까지 축소하여도 전체자질을 사용하였을 경우에 비해 약간의 성능 저하가 나타났을 뿐이며 오히려 성능이 더 향상되는 효과도 나타났다. 그러나 용어 가중치에 따른 성능은 클러스터링 기법에 따라 다소 차이가 나타났다. 결론적으로 자동분류의 응용 영역과 상황에 따라서 적절한 클러스터링 기법, 용어 가중치 공식, 자질 축소 수준 등을 선정함으로써 클러스터링 모형을 최적화하는 것이 바람직하다.

참 고 문 헌

Association for Computing Machinery. 1991. *ACM Computing Classification System.* [online]. 〈http : // www.acm.org/ class/〉.

Cutting, D.R., Karger, D.R., Pedersen, J.O., and Tukey, J.W. 1992. "Scatter/Gather : a cluster-based approach to browsing large document collections." In *Proceedings of the Fifteenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 318-329).

Cutting, D.R., Karger, D.R., and Pedersen, J.O. 1993. "Constant interaction-time scatter/gather browsing of very large document collections." In *Proceedings of the Sixteenth Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 126-134).

Kim, M. and Choi, K. 1999. "A comparison of collocation-based similarity measures in query expansion." *Information Processing & Management*, 35(1) : 19-30.

Larsen, B., and Aone, C. 1999. "Fast and effective text mining using linear-time document clustering." In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp.16-22).

Liu, H., and Motoda, H. 1998. *Feature Selection for Knowledge Discovery and Data Mining.* Boston : Kluwer Academic Publishers.

Sahami, M., Yusufali, S., and Baldonado, M.Q.W. 1998. "SONIA : a service for organizing networked information autonomously." In *Proceedings of the Third ACM Conference on Digital Libraries* (pp. 200-209).

Schütze, H., and Silverstein, C. 1997. "A comparison of projections for efficient document clustering." In *Proceedings of the Twentieth Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 74-81).

Silverstein, C., and Pedersen, J. 1997. "Almost-constant-time clustering of arbitrary corpus subsets." In *Proceedings of the Twentieth Annual ACM SIGIR Conference*

on Research and Development in Information Retrieval" (pp. 60-66).

Salton, G., and Buckley, C. 1988. "Term-weighting approaches in automatic text retrieval." *Information Processing & Management*, 24(5) : 513-523.

Scorpion Project Homepage. 〈http : //orc.rsch.oclc.org : 6109/〉.

Vaithyanathan, S., and Dom, B. 1999. "Generalized model selection for unsupervised learning in high dimensions." In *Proceedings of the Neural Information Processing Systems 1999*. [online]. 〈http : //www. almaden. ibm.com/cs/k53 /papers/nips99.ps〉.

Vaithyanathan, S., and Dom, B. 1999a. "Model selection in unsupervised learning with applications to document clustering." In Proceedings of the 16th *International Conference on Machine Learning* (pp. 423-433). 〈http : //www. almaden. ibm. com/cs/k53/ irpapers/dom. ps〉.

Willett, P. 1983. "Similarity coefficients and weighting functions for automatic document classification : an empirical comparison." *International Classification*, 10(3) : 138-142.

Wong, Wai Chiu, and Fu, A. 2000. "Incremental document clustering for Web page classification." In *Proceedings of the IEEE 2000 International Conference on Information Society in the 21st Century : Emerging Technologies and New Challenges* (IS2000). [online]. 〈http : //www.cs.cuhk. hk/ ~adafu/Pub/IS2000full.ps〉.