계층적 문서 클러스터링을 위한 응집식 기법과 분할식 기법의 비교 연구

A Comparative Study on the Agglomerative and Divisive Methods for Hierarchical Document Clustering

이재윤, 경기대학교 {memexlee@kgu.ac.kr}
정진아, 연세대학교 대학원 {jajeong@yonsei.ac.kr}

Lee, Jae-Yun, Kyonggi University
Jeong, Jin-Ah, Graduate School of Yonsei University

계층적 문서 클러스터링에 있어서 실험집단에 따라 응집식 기법과 분할식 기법의 성능이 다르며, 이를 좌우하는 요소는 분류의 깊이, 즉 분류수준이라고 가정하였다. 조금만 나누면 되는 대분류인 경우는 상대적으로 분할식 기법이 유리하고, 조금만 합치면 되는 소분류인 경우에는 응집식 기법이 유리할 것이라고 판단했기 때문이다. 그에 따라 분할식 클러스터링 기법인 양분(Bisecting) K-means 기법과 응집식 기법인 완전연결, 평균연결, WARD기법의 성능을 실험집단이 대분류인 경우와 소분류인 경우의 유사계수를 적용하여 각 기법별 성능을 비교하여 실험집단의 특성에 따른 적합 클러스터링 기법을 찾고자 하였다. 실험결과 응집식 기법과 분할식 기법의 성능 우열에 영향을 미치는 것은 분류수준보다는 변이계수로 측정된 상대적인 군집의 크기 편차인 것으로 나타났다.

1. 서론

클러스터링 기법을 응집식과 분할식으로 구분할 때, 흔히 계층적 클러스터링 기법의 대명사로 간주되는 응집식 기법이 성능이 좋은 것으로 알려져왔다. k-Means 기법으로 대표되는 분할식 기법은 빠른 속도에 비해서 분류 성능이 떨어지므로 클러스터링 관련 서적에서도 간략하게 다루어질 뿐이었다(Anderberg 1973).

2000년에 발표된 양분식(bisecting) k-means 기법은 클러스터를 두 개로 양분하는 작업을 반복하면서 고품질의 계층적 클러스터링을 수행하는 새로운 방식으로 제안되었다(Steinbach et al. 2000). 이 방식을 지속적으로 개발해나간 Y. Zhao와 G. Karypis는 검증결과 문서 클러스터링에 있어서 양분식 k-means 기법으로 대표되는 (반복적) 분할식 클러스터링 기법이 항상 기존의 응집식 기법보다 우수하다고 주장하였다(Zaho & Karypis 2004).

그러나 문서 클러스터링 기법의 성능은 대상 문서집단의 특성에 따라 달라질 수 있어서 어느 클러스터링 기법의 성능이 가장 우수하다고 단정적으로 말하기는 어렵다(정영미 2004).

따라서 이 논문에서는 구성이 상이한 문서집단에 대해서 응집식 혹은 분할식 기법 중 어느 기법의 성능이 좋은 지를 알아보기 위한 실험적 연구를 수행하였다.

응집식과 분할식 기법의 성능 우열을 좌우하는 실험집단의 특성으로는 분류수준을 고려하였다. 계층적 클러스터링은 기계가 수행하는 것이므로 매번의 응집 혹은 분할 판단마다 오류가 생길 수밖에 없다. 계층처리를 위해 응집

이나 분할 판단이 반복되면서 오류는 누적된다. 분할식 기법은 하향식이므로 상대적으로 나누는 횟수가 적은 대분류 상황에서 오류가 적게 누적될 것이다. 반면에 응집식 기법은 상향식이므로 군집의 크기가 작은 소분류 상황이 대분류 상황보다 판단의 횟수가 적어서 오류가 적게 누적될 것이다.

이를 감안하면 소분류인 경우에는 응집식이 상대적으로 유리하고 대분류인 경우는 분할식이 상대적으로 유리하다고 생각할 수 있다.

이를 검증하는 실험을 위해서 이 연구에서는 분류 수준이 상이한 실험집단을 복수로 준비하여 응집식과 분할식의 성능을 비교해보기로 한다.

2. 분할식 계층적 클러스터링 기법

분할식 클러스터링 기법인 양분(Bisecting) K-means기법은 비계층적 기법을 반복 사용하는 계층적 클러스터링 기법이다. 클러스터링 과정은 모든 문서들이 소속된 단일 클러스터에서부터 시작하여, 다음과 같은 순서로 이루어진다.

① 한 클러스터를 선택해서 분할한다.

② 기본적인 K-means 알고리즘을 사용한 2개의 하부 클러스터(sub-cluster)를 찾는다.

③ 고정된 횟수 동안 양분하는 단계인, 2단계를 반복하고, 총유사도가 가장 큰 클러스터를 생성하는 분열을 한다. (각 클러스터에 대한, 유사도는 평균 쌍 방식 문서 유사도이고, 모든 클러스터를 합한다.)

④ 1, 2단계 그리고 3단계를 원하는 수의 클러스터에 이를 때까지 반복한다.

G. Karypis가 제작하여 공개한 클러스터링 소프트웨어인 CLUTO(Karypis 2003)에서는 양분식 k-means 클러스터링을 위한 평가함수에 RBR이라는 이름을 붙여서 사용하고 있다. 따라서 이 논문에서도 RBR이라고 약칭하기로 한다.

3. 실험 설계

3.1 적용 기법 및 소프트웨어

비교대상 클러스터링 기법으로는 응집식 기법인 완전연결, 평균연결, WARD 기법과 분할식 기법인 양분 K-means 기법을 선정하였다. 응집식 기법 중에서 단일연결기법은 사전 실험결과 성능이 현저하게 나쁘게 나타났으므로 포함하지 않았다.

실험을 위한 클러스터링 소프트웨어로 CLUTO와 SPSS 통계패키지를 이용하였다. CLUTO는 다양한 종류의 클러스터링 기법들과, 기준 함수(criterion function), 유사도 함수를 설정할 수 있도록 되어 있으며, 이 연구에서는 분할식 기법인 RBR 방식을 실행하는데 사용하였다. SPSS는 응집식 클러스터링 기법을 수행하기 위한 도구로 사용하였다. CLUTO에도 응집식 기법 중 완전연결과 평균연결 기법을 선택할 수 있는 옵션이 있지만, 분할식 기법을 제안한 연구자가 제작한 것이므로 객관성을 유지하기 위해서 이를 사용하지 않았다. 또한 CLUTO에는 WARD 기법이 구현되어 있지 않은 점도 고려한 결과이다.

각 기법에 대해 문서간 유사도 함수로 코사인 유사계수를 사용하여 문서 X, Y간 유사도를 측정하여 유사계수 행렬을 생성하고 이로부터 클러스터 집합을 생성하였다.

각 기법의 클러스터링 성능 평가를 위한 척도로는 WACS(정영미, 이재윤 2001)를 사용하였다.

3.2 실험 데이터

이 연구에서 원하는 분류 수준을 가진 분류 실험용 문서집단을 여러 개 구성하기 위해서 정보검색 실험용 문서집단을 활용하였다. 정보검색 실험용 문서집단은 질의마다 적합문서가 정해져 있으므로, 각 질의를 하나의 주제범주로 보고 질의에 대한 적합문서를 주제범주에

소속한 문서로 간주하였다. 이와 같이 검색 질의와 적합문서로 분류범주와 소속문서 집단을 구축한 것은, 최근 문서 클러스터링 연구가 검색결과문서의 클러스터링을 주요 과제로 하고 있는 점도 감안한 것이다.

실험집단 중 MQ와 MQS는 Medline 실험집단에서 추출한 것이다. Medline 실험집단은 의학논문 초록 1,033개로 구성되었고, 이중에서 696개의 문서가 30개의 질의에 적합한 것으로 나뉘어져 있다. 30개 질의 중에서 적합문서 수가 많은 순서대로 8개 질의를 대분류 범주로, 이 질의들의 적합문서 276개를 소속문서로 추출하여 분류실험집단 MQ를 구성하였다. 이와 별도로 적합문서 수가 이보다 적은 11개의 질의를 소분류 범주로, 이 질의들의 적합문서 157개를 소속문서로 추출하여 분류실험집단 MQS를 구성하였다.

CQ와 CQS는 컴퓨터과학 분야 초록으로 구성된 CACM 실험집단에서 추출한 것이다. 적합문서 수가 많은 순서대로 8개 질의를 대분류 범주로, 이 질의들의 적합문서 263개를 소속문서로 추출하여 분류실험집단 CQ를 구성하였다. 이와 별도로 적합문서 수가 이보다 적은 11개의 질의를 소분류 범주로, 이 질의들의 적합문서 49개를 소속문서로 추출하여 분류실험집단 CQS를 구성하였다. 원래 CACM 실험집단에서 추출한 적합문서들 중에는 둘 이상의 질의에 적합한 복수주제 문서가 있었으나 배타적 클러스터링 실험을 위해서 제거하고 CQ와 CQS를 구성하였다.

HQ와 HQS는 HANTEC 실험집단에서 추출한 것이다. HANTEC에는 약 4만 여건의 문서들을 대상으로 질의 30여개에 대한 12만 건의 적합성 판정이 이루어져있다. 적합문서 수가 많은 순서대로 5개 질의를 대분류 범주로, 이 질의들의 적합문서 351개를 소속문서로 추출하여 분류실험집단 HQ를 구성하였다. 이와 별도로 적합문서 수가 이보다 적은 11개의 질의를 소분류 범주로, 이 질의들의 적합문서 104개를 소속문서로 추출하여 분류실험집단 HQS를 구성하였다. 6개 실험집단의 구성은 <표 1>과 같다.

<표 1> 실험집단의 구성

실험집단	범주 수	문서 수	범주당 문서 수	범주크기의 변이계수*
CQ	8	263	32.9	0.331
MQ	8	276	34.5	0.150
HQ	5	351	70.2	0.367
CQS	11	49	4.5	0.485
MQS	11	157	14.3	0.186
HQS	11	104	9.5	0.576

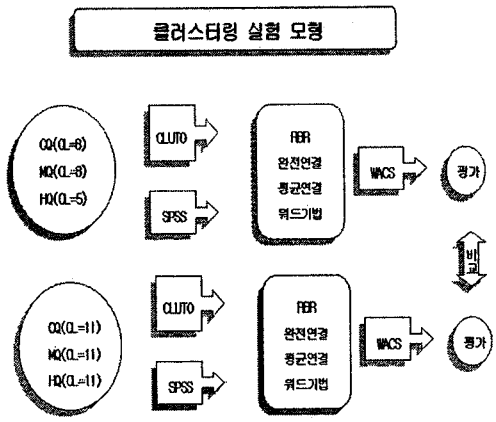
* 변이계수는 표준편차를 평균으로 나눈 값

구축된 분류실험집단의 특성을 알아보기 위해서 군집내 문서 간 유사도와 타 군집에 속한 문서와의 유사도를 각각 평균한 결과를 <표 2>에 제시하였다. 실험집단의 군집내 문서 간 유사도가 높고 타 군집 소속 문서와의 유사도가 낮을수록 분류가 잘 될 것이다. 군집내 문서 간 유사도는 MQ가 가장 높으며 CQ와 HQ은 비슷하다. 타 군집 소속 문서와의 유사도는 CQ가 가장 높으며 MQ가 이보다 조금 낮고 HQ는 매우 낮다. 결국 두 평균 유사도간의 차이는 MQ가 가장 크며 HQ가 그 다음이고 CQ가 가장 작다. 이로 미루어볼 때 분류 문제로서는 CQ가 가장 어렵고 MQ가 가장 쉬운 집단이라고 할 수 있다.

<표 2> 실험집단의 군집내·외간 유사도 비교

실험집단 / 평균유사도 측정 대상	CQ	HQ	MQ
① 동일 군집내 문서간	0.0663	0.0642	0.0956
② 타 군집 소속 문서간	0.0184	0.0045	0.0169
① - ②	0.0479	0.0596	0.0787

이상의 전체 실험과정을 정리해서 <그림 1>에 제시하였다.



<그림 1> 세부 실험 단계

4. 실험 결과 및 분석

4.1 대분류 실험집단의 실험결과

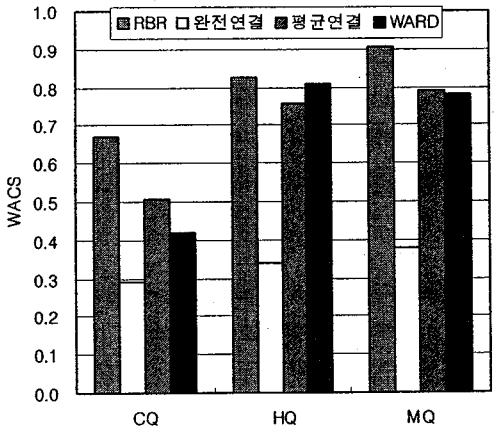
실험집단, 용어가중치 할당방식, 적용기법에 따른 대분류 실험결과는 <표 3>과 같다. 이 표에서 용어가중치 적용공식을 나타내는 두 자리 표기 중에서 앞 자리가 n이면 TF가중치를, l이면 로그TF 가중치를 적용한 것이고, 뒷 자리의 n과 t는 각각 역문헌빈도의 미적용과 적용을 나타낸 것이다.

<표 3> 기법별 대분류 성능 비교

		RBR	완전연결	평균연결	WARD
MQ	nn	**0.9065**	0.3450	**0.7905**	0.6383
	nt	0.8749	**0.3771**	0.6195	0.6686
	ln	0.8999	0.3190	0.7501	**0.7828**
	lt	0.8813	0.3251	0.5792	0.7198
		RBR	완전연결	평균연결	WARD
HQ	nn	0.8221	0.3227	0.7256	0.5607
	nt	0.6756	0.3201	0.6436	0.6263
	ln	**0.8277**	0.3407	**0.7557**	**0.8087**
	lt	0.7563	**0.3408**	0.5906	0.7558
		RBR	완전연결	평균연결	WARD
CQ	nn	**0.6685**	0.2829	0.4366	0.3801
	nt	0.6010	0.2753	0.4391	0.4091
	ln	0.6218	**0.2919**	0.3945	**0.4197**
	lt	0.6048	0.2855	**0.5085**	0.3584

실험집단 MQ에서 분할식 RBR의 최고성능은 0.9065, 평균연결의 최고성능은 0.7805로 분할식 RBR기법이 월등히 성능이 뛰어남을 알 수 있다. HQ에서는 분할식 RBR의 WACS성능이 0.8277로 평균연결의 0.7557보다 좋은 것으로 나타났다. CQ에서는 분할식 RBR의 최고성능은 0.6685이었으며, 평균연결의 최고성능은 0.5085로 분할식 RBR의 최고성능이 더 좋은 것으로 나타났다.

대분류 실험집단의 분할식 RBR과 응집식 완전연결, 평균연결, WARD기법의 성능평가를 종합하면, 분할식 RBR의 성능이 응집식 기법보다 월등히 좋게 나타났다. 각 기법별로 가중치설정에 상관없이 최고성능만 그림으로 비교하면 <그림 2>와 같다.

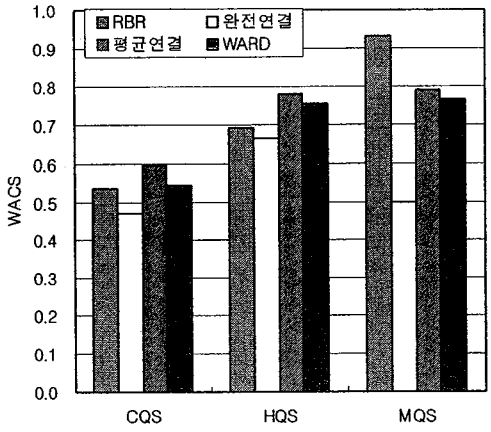


<그림 2> 기법별 대분류 최고성능 비교

결론적으로 대분류 실험집단인 CQ, HQ, MQ에서는 하향식으로 판단 횟수가 적은 분할식 RBR 기법의 성능이 좋은 것으로 나타났다. 반대로 상향식으로 진행하여 대분류 문제에서 판단 횟수가 많은 응집식 기법인 완전연결, 평균연결, WARD기법은 분할식보다 좋지 않은 것으로 나타났다.

<표 4> 기법별 소분류 성능 비교

		RBR	완전연결	평균연결	WARD
MQS	nn	0.9180	0.4413	**0.7889**	0.5943
	nt	0.9067	0.4136	0.7608	0.6319
	ln	0.9286	**0.4993**	0.7281	**0.7631**
	lt	**0.9299**	0.4708	0.6841	0.7110
		RBR	완전연결	평균연결	WARD
HQS	nn	**0.6916**	0.5545	0.6741	0.5233
	nt	0.6682	0.5020	**0.7798**	0.5529
	ln	0.6636	**0.6645**	0.7539	0.6816
	lt	0.6879	0.5205	0.6803	**0.7539**
		RBR	완전연결	평균연결	WARD
CQS	nn	0.4875	0.3799	**0.5951**	0.4862
	nt	**0.5377**	0.4243	0.5505	**0.5444**
	ln	0.4903	0.4567	0.5657	0.4782
	lt	0.4626	**0.4689**	0.4000	0.5057



<그림 3> 기법별 소분류 최고성능 비교

4.2 소분류 실험집단의 실험결과

실험집단, 용어가중치 할당방식, 적용기법에 따른 소분류 실험결과는 <표 4>와 같다.

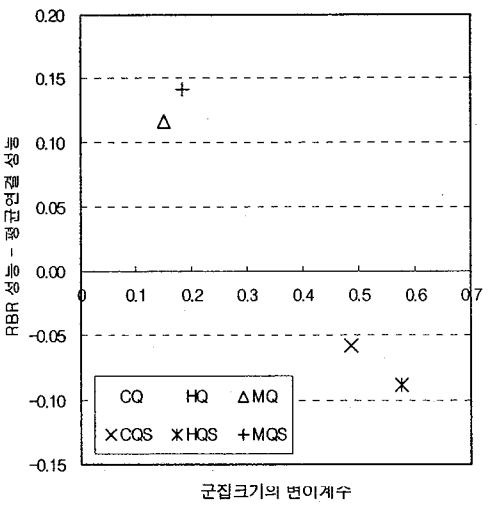
실험집단 HQS에서 RBR의 최고성능은 0.6916이며, 평균연결의 최고성능은 0.7798로 응집식인 평균연결 기법의 성능이 분할식인 RBR 기법보다 좋은 것으로 나타났다. 이는 같은 HANTEC에서 추출하여 구성한 대분류 집단인 HQ와는 다른 결과이다.

실험집단 CQS에서 가장 좋은 기법은 평균연결로서 0.5951이었고, RBR은 0.5377로 평균연결과 WARD기법보다 낮은 성능을 보였다. 역시 같은 CACM 집단에서 추출하여 구성한 CQ와는 다른 결과이다.

실험집단 MQS에 대한 분할식 RBR과 응집식 기법간의 성능을 비교한 결과, 다른 두 집단과 달리 대분류인 MQ에서처럼 분할식인 RBR의 성능이 응집식 기법인 완전연결, 평균연결, WARD기법보다 월등히 좋은 것으로 나타났다.

가중치할당 방식에 상관없이 기법별 소분류 최고성능을 그림으로 비교하면 <그림 3>과 같다. CQS와 HQS에서는 응집식인 평균연결기법과 WARD 기법이 분할식인 RBR 기법보다 좋은 반면에 MQS에서는 RBR 기법이 가장 좋은 것으로 나타났다.

결과가 다른 MQS 실험집단의 특징은 <표 1>에서 보듯이 군집크기의 변이계수가 낮다는 점이다. 즉 MQS는 군집의 크기가 상대적으로 고르게 구성된 실험집단이다. 각 실험집단마다 분할식 RBR 기법과 응집식 평균연결 기법의 성능 차이와, 실험집단 군집크기의 변이계수를 <그림 4>에 함께 나타내 보았다.



<그림 4> 실험집단별 변이계수와 기법간 성능 차이

<그림 4>를 보면 실험집단내 군집크기의 편차가 상대적으로 클 수록(변이계수가 높을 수록) RBR 기법이 평균연결 기법보다 성능이 나쁜 것을 볼 수 있다.

이 관찰 결과를 통계적으로 검증하기 위해서 분할식인 RBR 기법에서 응집식인 평균연결 기법의 성능을 뺀 값과, 실험대상 문서집단의 평균군집크기 및 군집크기 변이계수와의 상관을 분석해보았다. 피어슨 상관계수를 산출한 결과는 <표 5>와 같다.

<표 5> 군집크기 및 변이계수와 성능차이의 상관 분석

	RBR성능 - 평균연결성능
평균군집크기	0.347
군집크기 변이계수	-0.811

분석 결과 평균군집크기는 두 기법간 성능차이와 통계적으로 상관이 유의하지 않은데 반해서 군집크기의 변이계수는 상관계수가 -0.811로 99% 유의수준에서 상관이 있는 것으로 나타났다. 즉, 군집크기의 편차가 평균 크기에 비해서 심하면 심할수록 응집식인 평균연결기법의 성능이 RBR에 비해서 좋다는 의미이다.

군집크기의 편차가 상대적으로 심하면 분할식 기법인 RBR이 불리한 이유는, 전체를 단계적으로 양분해나가는 특성상 크게 차이나는 분할은 발생하지 않기 때문이라고 생각된다.

5. 결론

계층적 문서 클러스터링에 있어서 실험집단에 따라 응집식 기법과 분할식 기법의 성능이 다르며, 이를 좌우하는 요소는 분류의 깊이, 즉 분류수준이라고 가정하고 이를 확인하는 실험을 수행하였다. 분할식 클러스터링 기법인 양분(Bisecting) K-means기법과 응집식 기법인 완전연결, 평균연결, WARD기법의 성능을 실험집단이 대분류인 경우와 소분류인 경우로 나누어 비교하였다.

실험결과 대분류 문제에 대해서는 응집식 기법보다 분할식 기법이 항상 좋은 성능을 보였다. 소분류 문제에 있어서는 이와 달리 응집식 기법이 세 실험집단 중 두 집단에서 분할식 기법보다 좋은 클러스터링 성능을 보였다. 분석결과 분할식 기법의 성능 우열에 영향을 미치는 것은 분류수준이나 군집의 크기 자체보다는 변이계수로 측정된 상대적인 군집의 크기 편차인 것으로 나타났다.

결국 클러스터링 알고리즘은 문서집단의 특성에 따라 성능이 달라질 수 있으므로 이를 고려한 적용이 필요할 것이다.

참고문헌

정영미. 2004. 정보검색을 위한 클러스터링 기법의 성능 비교 연구. 『지원 윤구호 박사 정년기념논문집』, 463-481.

정영미, 이재윤. 2001. 지식 분류의 자동화를 위한 클러스터링 모형 연구. 『정보관리학회지』, 18(2): 203-230.

Anderberg, Michael R. 1973. *Cluster analysis for Applications*. Academic Press.

Karypis, George. 2003. *CLUTO: A Clustering Tookit*, Release 2.1.1. Department of Computer Science, University of Minnesota.

Steinbach, Michael, George Karypis, and Vipin Kumar. 2000. "A comparison of document clustering techniques." *Proceedings of the TextMining Workshop, KDD 2000.*

Zhao, Ying, and George Karypis. 2004. "Empirical and theoretical comparisons of selected criterion functions for document clustering." *Machine Learning*, 55(3): 311-331.