단어빈도와 동시링크의 결합을 통한 웹 문서 클러스터링 성능 향상에 관한 연구*

Clustering of Web Document Exploiting with the Union of Term Frequency and Co-link in Hypertext

이교운(Kyo-Woon Lee)** · 이원희(Won-Hee Lee)***
박흠(Heum Park)**** · 김영기(Young-Gi Kim)***** · 권혁철(Hyuk-Chul Kwon)******

〈목 차〉

초 록

이 연구에서는 웹 문서가 갖고 있는 특성, 특히 웹 문서에 포함된 단어 수가 클러스터링 성능에 결정적인 영향을 미친다는 전제 하에, 웹 문서에 포함된 단어 수와 클러스터링 성능과의 관계를 밝힌 다음, 이 부분을 웹 문서의 동시인용 빈도를 이용해 보완할 수 있는 알고리즘을 제시한다. 이 연구에서는 네이버 디렉터리 중 '자연과학' 범주에 포함된 1,449개의 웹 문서를 대상으로 단어기반 클러스터링과 링크기반 클러스터링, 그리고 단어-링크 혼합 클러스터링 기법으로 클러스터링 해 보았으며, 그 결과를 네이버 디렉터리에 초기 할당된 범주와 비교해 보았다.

주제어 : 단어기반 클러스터링, 링크기반 클러스터링, 단어-링크 혼합 클러스터링, 동시링크, 텀 벡터

Abstract

In this paper, we have focused that the number of word in the web document affects definite clustering performance. Our experimental results have clearly shown the relationship between the amounts of word and its impact on clustering performance. We also have presented an algorithm that can be supplemented of the contrast portion through co-links frequency of web documents. Testing bench of this research is 1,449 web documents included on 'Natural science' category among the Naver Directory. We have clustered these objects by term-based clustering, link-based clustering, and hybrid clustering method, and compared the output results with originally allocated category of Naver directory.

Key Words: term-based clustering, link-based clustering, hybrid clustering, Co-link, Term Vector

* 이 논문은 과학기술부(한국과학기술기획평가원)의 국가지정연구실 사업지원으로 이루어진 것임.
** 울산과학대학 컴퓨터정보학부 교수, 부산대학교 한국어정보처리연구실(kwlee@mail.ulsan-c.ac.kr)
*** 부산대학교 한국어정보처리연구실(whlee@pusan.ac.kr)
**** 부산대학교 한국어정보처리연구실(parkheum@pusan.ac.kr)
***** 부산대학교 문헌정보학과 강사, 부산대학교 한국어정보처리연구실 IR팀(ykiki6292@hanafos.com)
****** 부산대학교 전자전기정보컴퓨터공학부 교수(hckwon@pusan.ac.kr)
• 접수일 : 2003. 8. 21 • 최초심사일 : 2003. 9. 3 • 최종심사일 : 2003. 9. 8

Ⅰ. 서 론

1. 연구목적

인터넷 정보자원의 급격한 증대로 이를 효과적으로 조직하여, 검색성능을 향상시키기 위한 다양한 연구와 실험이 수행되고 있다. 특히 클러스터링(문헌 분류)은 가장 원시적인 자료관리 행위이면서, 동시에 고도의 지식과 테크닉을 요구하는 일로서[1] 정보자원의 효과적인 탐색과 이용을 위한 출발점이며, 검색 성능을 향상시킬 수 있는 매우 강력한 수단 중의 하나이다. 정보검색 분야에서 클러스터링은 매우 다양한 방식으로 연구되고 있는데, 크게 단어 유사도에 기반한 단어기반 클러스터링(term-based clustering)과 링크기반 클러스터링(link-based clustering), 그리고 이 둘을 혼합한 혼합형 클러스터링으로 나누어진다.

이 중에서 가장 광범위하게 연구되고 활용되고 있는 클러스터링 기법은 단어기반 클러스터링이다. 이것은 웹 문서에 등장하는 단어 유사도에 기초하여 관련 문서를 군집화 하는 기법으로서, 특히 대규모 웹 문서를 대상으로 할 경우, 자연언어 처리에 기반한 자동색인 기법이 필수적이다. 따라서 단어기반 클러스터링의 경우 그 효과에는 명확한 한계가 존재할 수밖에 없다. 최근에는 온라인 시소러스나 단어의 의미적 관련성을 활용하는 시도도 부분적으로 이루어지고 있지만 역시 그 효과는 상당히 불확실한 편이다.

최근 웹 문서는 텍스트보다는 이미지를 전면에 내세우는 경향이 강하게 나타나고 있으며, 심지어 텍스트마저 이미지로 처리되는 경우도 나타나고 있다. 따라서 실제 웹 문서를 분석해 보면 텍스트가 아예 존재하지 않거나 매우 적은 수의 텍스트로 이루어진 문서도 상당히 많이 나타나고 있다. 따라서 테이블과 같은 웹 문서의 구조정보나[2] 하이퍼링크와 같은 외부 정보원의 부가적 이용을 고려해 볼 수 있다.

이 연구에서는 웹 문서가 갖고 있는 특성, 특히 웹 문서에 포함된 단어 수가 클러스터링 성능에 결정적인 영향을 미친다는 전제 하에, 웹 문서에 포함된 단어 수와 클러스터링 성능과의 관계를 밝힌 다음, 이 부분을 웹 문서의 동시인용 빈도를 통해 보완할 수 있는 알고리즘을 제시한다.

2. 관련연구

정보검색에 있어서 클러스터링은 다양한 방법으로 연구되어 왔다. 웹 정보검색의 경우

1) 최정태, 양재한, 도태현, 문헌분류의 이론과 실제(부산대학교 출판부, 1998P, p.i.
2) 정성원, 이원희, 김영기, 권혁철, "웹 문서 중 의미 있는 표의 추출", 한글 및 한국어 정보처리, 제14집 (2002. 10), pp.332~339.

클러스터링은 주제별 커뮤티니의 확인[3]이나 웹 구조의 추출[4], 그리고 중복 문서의 발견[5], 적합성 피드백을 통한 질의확장[6] 등의 수단으로 연구되어 왔다. 단어기반 클러스터링은 각 문헌에 포함된 단어 유사도에 근거하여 전체 문헌집단에 대한 조직으로 나아가고 있으며, 최근 인터넷 환경에서의 정보검색은 링크기반 클러스터링 알고리즘과 검색결과의 동적 클러스터링 양쪽을 포괄하는 연구경향을 보이고 있다.

Ⅱ. 연구방법

1. 분석 대상 선정

이 연구에서는 웹 문서에 포함된 단어 수와 클러스터링 성능과의 관계를 분석하기 위한 실험 대상으로 검색 포탈 엔진인 네이버 디렉터리(http://www.naver.com) 중에서 하나를 선정하였다. 14개의 네이버 기본 디렉터리 중에서 '학문, 과학' 영역을 선정하였으며, 이 중에서도 그 하위영역인 '자연과학'을 선택하였다. '자연과학' 분야의 경우 웹 문서가 풍부하고 디렉터리 구조도 상대적으로 명확하기 때문에 분석 대상으로 적절하다고 판단하였다. '자연과학' 분야는 다시 16개의 하위 디렉터리로 나누어지는데 이 중에 상위 디렉터리가 겹치지 않는 11개의 카테고리가 최종 실험대상으로 선정되었다. 예를 들어 '자연과학/지구과학/해양학/해양법'에 속하는 문서가 '사회, 문화/법/해양법'의 문서에도 분류되어 있기 때문에 이번 실험대상에서는 제외하기로 하였다.

최종적으로 분석대상으로 선정된 범주와 대상 웹 문서 수는 다음과 같다.

---

3) Kumar, S. R., Raghavan, P., Rajagopalan, S., & Tomkins, A. "Trawling the Web for emerging cyber-communities", *Proceedings of the 8th WWW Conference.* 1999. ; Mukherjea, S, "Organizing topic-specific Web information", *Proceedings of the 11th ACM Conference on Hypertext.*(2000), pp.133~141. : Mukherjea, S, "WTMS: a system for collecting and analyzing topic-specific Web information", *Proceedings of the 9th International World Wide Web Conference*(2000), pp.457~471.

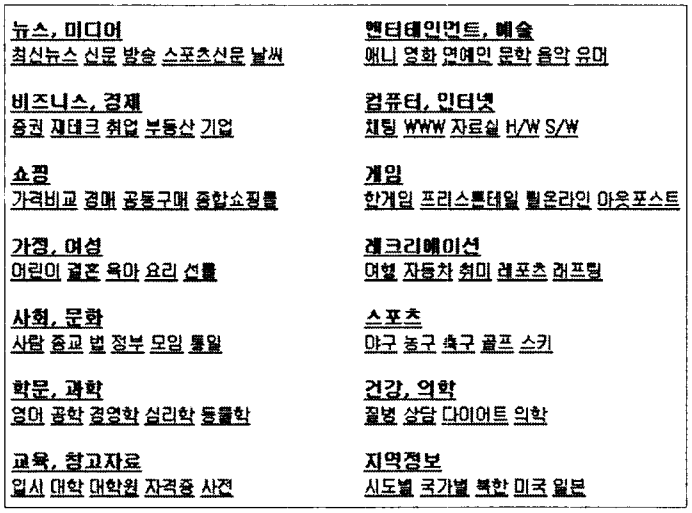
4) Larson R. R., "Bibliometrics of the World Wide Web: An Exploratory Analysis of the Intellectual Structure of Cyberspace", Proceedings of the 1996 American Society for Information Science Annual Meeting. 1996. Pirolli, P., Schank, P., Hearst, M., & Diehl, C., "Scatter/ Gather browsing communicates the topic structure of a very large text collection", *Proceedings of the Conference on Human Factors in Computing Systems*(1996), pp.213-220.

5) Broder, A. Z., Glassman, S. C., Manasse, M. S., & Zweig, G., "Syntactic clustering of the Web", *Proceedings of the 6th International WWW Conference*(1997), pp.391-404.

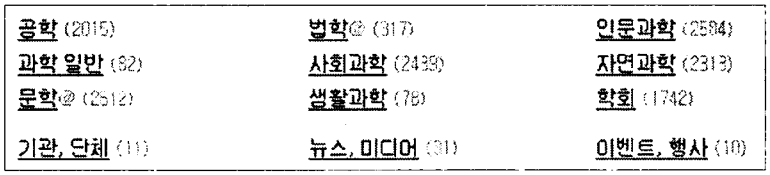
6) Chang, C. H. & Hsu, C. C., "Integrating query expansion and conceptual relevance feedback for personalized Web information retrieval", *Proceedings of the 7th International WWW Conference*(1998).

〈표 1〉 범주별 분석대상 문서

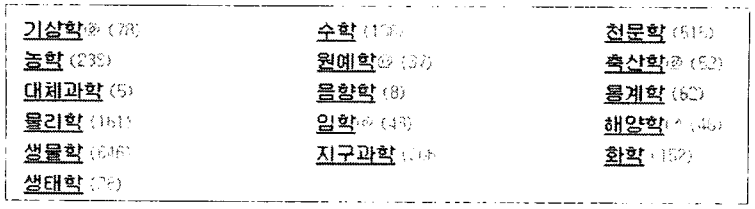
범 주	문서 수	문서번호
농학(Aggr)	145	1-145
대체과학(Alt)	4	146-149
물리학(Phys)	102	150-251
생물학(Bio)	426	252-677
생태학(Ecol)	24	678-701
수학(Math)	102	702-803
음향학(Acou)	5	804-808
지구과학(Earth)	149	809-957
천문학(Astro)	323	958-1280
통계학(Stat)	56	1281-1336
화학(Chem)	113	1337-1449
계	1,449	1-1449



〈그림 1〉네이버 최상위 범주 (http://www.naver.com)



〈그림 2〉네이버 '학문과학' 범주
(http://dir.naver.com/Science and Technology/)



〈그림 3〉 네이버 '학문과학' 내 '자연과학' 범주
(http://dir.naver.com/Science and Technology/Science/)

2. 단어기반 클러스터링(term-based clustering)

실험 대상 문서를 단어기반 클러스터링을 위해 문서-단어빈도 행렬(document-term frequency matrix)을 작성하였다.

클러스터링 작업과 클러스터링 성능 분석을 위해서는 미네소타 대학 컴퓨터 과학과(University of Minnesota, Department of Computer Science)에서 2002년 8월에 배포한 클러스터링 시스템(clustering Toolkit)인 Cluto2.1 (http://www-users.cs.umn.edu/~karypis/cluto/)을 사용하였다. Cluto 시스템은 문서 클러스터링을 위해 하향식(분할식; partitional), 상향식(집적식; agglomerative), 그래프 분할식(graph partitional) 패러다임 등과 같은 다양한 클러스터링 알고리즘과 함께 각 알고리즘에서 사용될 수 있는 일곱 가지의 클러스터 평가함수(criterion function)를 제공하여 사용자가 선택할 수 있도록 해 주고 있다. Cluto 시스템에서 제공해 주고 있는 기본적인 평가함수 일곱 가지는 다음과 같다.

그 외에도 상향식 클러스터링 알고리즘에 사용되는 전통적인 클러스터 평가함수인 단일링크(single-link), 완전링크(complete-link), UPGMA 등도 함께 제공해 주고 있다. Cluto 시스템에서 제공되는 유사도 계산 방법으로는 코사인(cosine function), 상관계수(correlation coefficient), 역 유클리디안 거리(inverse Euclidean distance), 확장 자카드 계수(extended Jaccard coefficient) 등이 있다. Cluto 시스템은 클러스터링 결과를 매우 다양한 형태의 그래프나 도표, 트리 형식으로 제시할 수 있는 옵션도 제공하고 있다.

실험대상 문서에 대한 문서-단어 빈도행렬을 Cluto 시스템으로 클러스터링 한 다음 그 결과를 네이버 디렉터리에 초기 할당된 범주와 비교하였으며, 그 일치 여부를 클러스터링 성능으로 간주하였다. 이와 함께 보편적인 클러스터링 성능평가 척도로 사용되고 있는 복잡도(Entropy)와 순정도(Purity)도 함께 제시해 보았다.

〈표 2〉 CLUTO clustering 평가함수. *k*: 전체 클러스터 수, *S*: 클러스터 된 전체 문서 수, *Si*: *i*번째 클러스터에 할당된 문서 집합, *ni*: *i*번째 클러스터에 할당된 문서 수, *v*, *u*: 문서, *sim*(*v*, *u*): 두 문서간의 유사도

Criterion Function	Optimazition Function
$\mathcal{I}_1$	maximize $\sum_{i=1}^{k} \frac{1}{n_i} \left(\sum_{v,u \in S_i} \text{sim}(v,u) \right)$
$\mathcal{I}_2$	maximize $\sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} \text{sim}(v,u)}$
$\mathcal{E}_1$	minimize $\sum_{i=1}^{k} n_i \frac{\sum_{v \in S_i, u \in S} \text{sim}(v,u)}{\sqrt{\sum_{v,u \in S_i} \text{sim}(v,u)}}$
$\mathcal{G}_1$	minimize $\sum_{i=1}^{k} \frac{\sum_{v \in S_i, u \in S} \text{sim}(v,u)}{\sum_{v,u \in S_i} \text{sim}(v,u)}$
$\mathcal{G}_1'$	minimize $\sum_{i=1}^{k} n_i^2 \frac{\sum_{v \in S_i, u \in S} \text{sim}(v,u)}{\sum_{v,u \in S_i} \text{sim}(v,u)}$
$\mathcal{H}_1$	maximize $\frac{\mathcal{I}_1}{\mathcal{E}_1}$
$\mathcal{H}_2$	maximize $\frac{\mathcal{I}_2}{\mathcal{E}_1}$

3. 링크기반 클러스터링(link-based clustering)

웹 문서는 각종 정보원들이 링크를 통해 서로 연결되어 있는데, 링크 기반 클러스터링은 하이퍼링크에 포함된 정보를 이용하여 관련문서 군집화시키는 것으로서, 하이퍼링크가 두 문서 간에 의미적 연관성을 갖고 있을 것이라는 것을 전제로 하고 있다. 이러한 연관성은 패스(path)의 길이에 반비례하며, 링크 수에 비례한다고 볼 수 있다.

웹 문서의 링크는 우선 문서 내 링크(intra-document link)와 문서 간 링크(inter-document link)로 나눌 수 있으며, 나가는 링크(out-link)와 들어오는 링크(in-link)로로 나눌 수도 있다. 나가는 링크가 많은 사이트는 hubness가 높으며, 들어오는 링크가 많은 사이트는 authority가 높다고 말한다.[7]

들어오는 링크 수는 AltaVista와 Google, Alltheweb 등의 "link:url" 검색을 통해, 그리고 자기인용(문서 내 링크)을 제외한 링크 수는 AltaVista나 Alltheweb 등의 "link:url and not url", 또는 Google의 결과 내 재검색 방법 등을 통해 알 수 있다. 나가는 링크(out

7) R. K. Belew, *Finding Out About: A Cognitive perspective on search engine technology and the WWW.* Cambridge University Press, 2000. p.196.

link)의 경우 웹 문서의 파싱을 통해 그 개수를 쉽게 구할 수 있지만, 들어오는 링크(in link)는 해당 분야의 전체 코퍼스(corpus)를 직접 갖고 있어야만 구할 수 있다.

이 연구에서는 연구대상이 된 웹 문서들을 대상으로 각 두 개의 문서를 동시에 링크하고 있는 사이트 수를 조사하기 위해 Alltheweb의 "Advanced Search"를 이용하였다. 사용된 검색식은 다음과 같다.

LINK:url_A AND LINK:url_B

이러한 방법으로 웹 사이트들의 쌍을 동시에 링크하고 있는 웹 문헌들의 검색 건수를 구하여 동시인용빈도 행렬을 작성하였다.

다음으로 웹 문서들 간의 관련성의 정도, 즉 상대적인 유사성과 비유사성을 나타내기 위해 동시인용 빈도는 새로운 척도로 변형될 필요가 있다. 일반적으로 키워드를 통한 두 문헌간의 관계를 나타내기 위해서는 단어출현빈도(Term Frequency, TF)와 역문헌빈도(Inverse Document Frequency, IDF)를 이용한 TFIDF(TF × IDF)를 이용한 단어벡터(word vector)와 문서 헤드의 거리비교(distance comparison)가 사용된다. 또한 TFIDF를 변형하여 CCIDF(Common Citation × Inverse Document Frequency) 알고리즘을 고려해 볼 수도 있을 것이다. 이 실험에서는 CCIDF를 이용하여 웹 문서들 간의 상대적인 유사도와 비유사도를 구했으며, 실험 환경은 단어기반 클러스터링과 같은 Cluto 시스템을 사용하였다.

Cluto 시스템을 통해 클러스터링 결과를 덴드로그램(dendrogram)으로 나타내었으며, 이를 단어기반 클러스터링 결과와 비교하였다.

4. 혼합 클러스터링(hybrid clustering)

단어기반 클러스터링 기법과 링크기반 클러스터링 기법을 혼합하기 위해 각 단어와 링크에 가중치를 주는 다양한 알고리즘이 개발되어 있다. 이 실험에서는 단어기반 클러스터링 기법으로 클러스터링이 잘 되지 않는 문서들의 공통된 특성을 추출한 다음, 이런 문서들에 한해서 링크기반 클러스터링을 하는 방법을 고려하였다. 이와 더불어 단어기반 클러스터링에서 각 문서의 자질(feature)로 규정된 각 문서의 색인어에 동시링크 수를 또 하나의 자질로 추가하는 방법도 함께 사용하였다.

Ⅲ. 실험결과

1. 단어기반 클러스터링

분석대상이 된 문서 전체를 대상으로 각 문서별 단어출현빈도 행렬을 작성하였다. 분석대상이 된 문서는 모두 1,449개이며, 등장한 총 단어 수는 293,596개, 문서 당 평균 포함된 단어 수는 202.6개로 나타났다. 파싱을 통한 전체 색인어 수는 33,253개로 나타났다. 그러나 클러스터링 성능을 높이기 위해 이러한 색인어 중에서 불용어와 숫자, 전자우편 주소, 한 글자 색인어 등을 제거하였다. 또한 실험대상이 자연과학 분야이기 때문에 이 범위를 넘어서는 '학문, 대학, 학회' 등과 같이 클러스터링에 부정적인 영향을 미칠 것으로 판단되는 일부 색인어도 제거하였다. 이리하여 최종적으로 선정된 색인어 수는 17,223개였다. 이 실험에서 사용된 문서-단어 행렬은 1,449×17,223이다.

크게 상향식 클러스터링과 하향식 클러스터링 기법에 앞에서 제시한 열두 가지 평가함수(criterion function)를 적용하였으며, 유사도 측정으로는 코사인(cosine)과 상관계수(correlation coefficient)를 별도로 적용해 보았다. 여기에 칼럼 모델(column model)로 상관계수 유사도 방식과 역문헌 빈도(idf)를 각각 적용하였다. 그 결과를 보편적인 클러스터링 성능 평가 척도인 클러스터링 복잡도(entropy)와 순정도(purity)로 비교해 보면 다음 표와 같다.

〈표 3〉 단어기반 하향식 클러스터링 성능

평가함수	역문헌빈도		상관계수	
	entropy	purity	entropy	purity
I1	0.501	0.586	0.457	0.604
I2	0.363	0.708	0.352	0.723
E1	0.356	0.699	0.374	0.712
G1	0.426	0.675	0.374	0.705
G1P	0.367	0.705	0.398	0.652
H1	0.394	0.676	0.427	0.619
H2	0.382	0.684	0.397	0.666

〈표 4〉 단어기반 상향식 클러스터링 성능
(문서 유사도: cos, 칼럼(단어) 유사도 idf)

평가함수	entropy	purity
I1	0.602	0.507
I2	0.483	0.587
E1	0.539	0.546
G1	0.810	0.306
G1P	0.541	0.557
H1	0.485	0.596
H2	0.533	0.546
단일링크	0.812	0.309
단일링크(가중치)	0.812	0.309
완전링크	0.741	0.354
완전링크(가중치)	0.746	0.345
UPGMA	0.811	0.309

클러스터링은 복잡도가 낮으면 낮을수록, 순정도가 높으면 높을수록 그 성능이 높다고 볼 수 있다. 따라서 이러한 결과들을 종합해 보면 본 연구의 실험 대상 문서의 경우 클러스터링 방식은 상향식(agglomerative) 보다는 하향식(partitional)이, 문서 유사도의 경우 상관계수보다는 코사인이, 단어 유사도의 경우는 역문헌빈도(idf)보다는 상관계수가 더 나은 클러스터링 성능을 보였다. 또한 사용된 평가함수 중에서는 대체적으로 I2, H2 등의 경우에 성능이 높은 것으로 나타났다. 전체적으로 보아 실험대상 문서의 경우 가장 클러스터링 성능이 좋은 것으로는 문서유사도에 코사인, 칼럼 유사도에 상관계수를 사용한 하향식 클러스터링으로 그 중에서도 I2의 경우에 가장 높은 성능을 보였다. 따라서 이 실험에서는 대상 웹 문서에 클러스터링 방식으로 하향식을, 문서유사도에는 코사인을, 단어 유사도에는 상관계수를, 평가함수로는 I2를 사용하여 20개의 클러스터로 클러스터링 하였다. 각 클러스터별 클러스터링 결과는 다음 표와 같다.

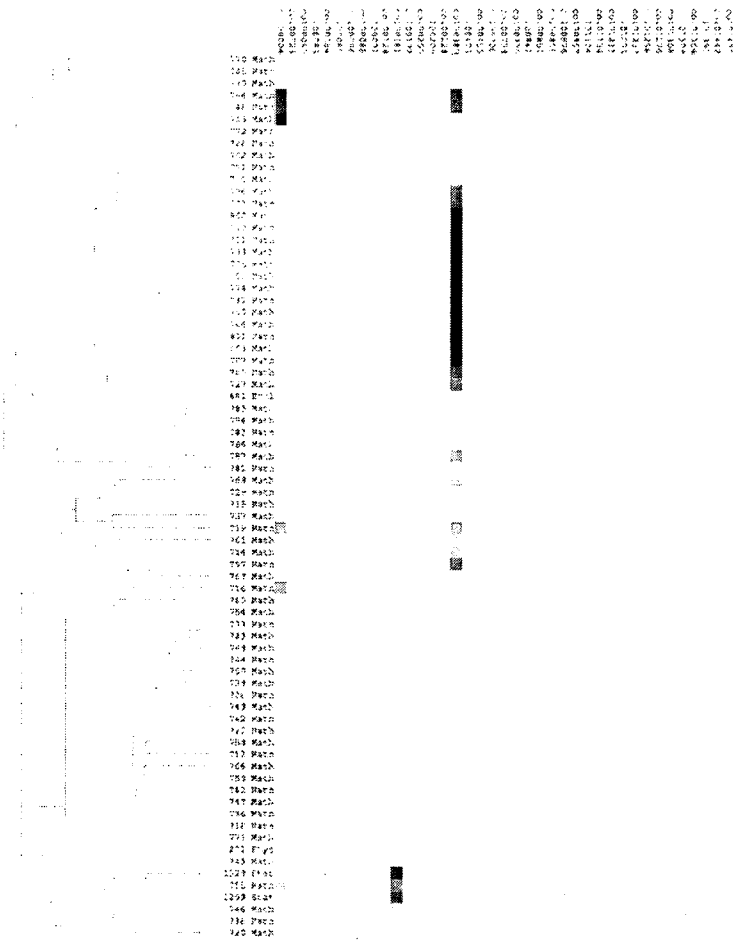
〈표 5〉 단어기반 클러스터링 결과

20-way clustering: [I2=3.92e+002] [1068 of 1449], Entropy: 0.352, Purity: 0.723

cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty	Aggr	Alt	Phys	Bio	Ecol	Math	Acou	Eart	Astr	Stat	Chem
0	67	+0.388	+0.112	+0.008	+0.004	0.184	0.910	0	0	61	0	0	1	1	2	1	0	1
1	37	+0.307	+0.156	+0.005	+0.003	0.311	0.784	0	0	0	4	0	2	0	0	2	29	0
2	38	+0.215	+0.095	+0.009	+0.004	0.215	0.789	8	0	0	30	0	0	0	0	0	0	0
3	58	+0.203	+0.087	+0.009	+0.006	0.206	0.897	0	1	1	1	0	0	0	0	52	1	2
4	49	+0.193	+0.075	+0.004	+0.002	0.112	0.939	0	0	1	0	0	0	0	2	46	0	0
5	43	+0.196	+0.088	+0.011	+0.006	0.468	0.628	7	0	0	27	5	0	0	2	0	0	2
6	74	+0.179	+0.072	+0.006	+0.004	0.111	0.946	0	0	1	0	1	70	0	0	0	2	0
7	62	+0.181	+0.082	+0.009	+0.005	0.134	0.919	4	0	0	57	0	0	0	1	0	0	0
8	60	+0.164	+0.083	+0.006	+0.004	0.172	0.900	1	0	0	4	0	0	0	0	0	1	54
9	17	+0.156	+0.053	+0.015	+0.010	0.578	0.471	0	0	8	0	0	3	2	3	1	0	0
10	36	+0.117	+0.066	+0.004	+0.003	0.538	0.472	3	0	0	10	0	0	0	17	5	0	1
11	56	+0.103	+0.050	+0.008	+0.005	0.612	0.429	5	0	0	11	12	0	0	24	2	0	2
12	39	+0.099	+0.041	+0.008	+0.006	0.307	0.821	1	0	0	1	0	1	0	1	32	0	3
13	50	+0.098	+0.045	+0.009	+0.006	0.337	0.760	0	1	7	0	0	1	0	3	38	0	0
14	67	+0.088	+0.058	+0.005	+0.003	0.485	0.582	0	0	0	11	0	3	0	39	12	0	2
15	59	+0.075	+0.049	+0.004	+0.003	0.669	0.407	12	0	2	24	2	0	0	1	11	1	6
16	56	+0.077	+0.036	+0.009	+0.005	0.398	0.750	5	0	0	42	0	1	0	4	1	1	2
17	71	+0.064	+0.043	+0.004	+0.003	0.298	0.831	3	1	1	59	0	0	0	5	1	0	1
18	73	+0.063	+0.027	+0.007	+0.005	0.469	0.521	38	0	0	24	1	0	0	1	1	0	8
19	56	+0.039	+0.019	+0.004	+0.003	0.666	0.446	25	0	1	10	0	1	1	5	10	1	2

이 표에서 cid, Size, ISim, ISdev, ESim, ESdev는 각각 클러스터 ID와 각 클러스터에 포함된 문서 수, 클러스터 내부 유사도와 표준편차, 클러스터 외부 유사도와 표준편차를 나타내며, 각 클러스터별 복잡도와 순정도를 보여주고 있다. 전체적으로 보아 1,449개의 문서 중에서 1,068개의 문서가 클러스터링 되었으며, 클러스터로 묶이지 않은 381개의 문서에는 '단어개수 0'인 문서 350개가 포함되어 있다. 클러스터별로는 클러스터 ID 6, 4, 7, 8, 0의 경우 매우 높은 성능을 보여 주고 있으며, 주제별로는 농학, 물리학, 수학, 통계학, 화학 등의 분야가 클러스터링이 잘 수행되었다. 반면 클러스터 ID 15, 11, 9, 10의 경우 매우 낮은 클러스터링 성능을 보여주고 있으며, 생물학, 지구과학, 천문학 등의 분야는 여러 클러스터로 흩어져 클러스터링 되었다.

한편 클러스터링의 결과를 각 문서에 대한 단어별 출현빈도와 덴드로그램으로 표현해 보면 다음과 같다.



〈그림 4〉 단어기반 클러스터링 결과 중 〈클러스터 6〉

이 그림은 <클러스터 6>의 한 부분으로서, 가로 줄은 각각 하나의 문서를 나타내며, 문서번호와 해당 주제범주가 나타나 있다. 한편 세로 줄은 각각 하나의 색인어에 대응되며, 그림 내부의 점은 해당 문서에 특정 색인어가 나타난 빈도를 보여준다. 그리고 가장 왼쪽의 덴드로그램은 각 문서가 군집화 하는 과정을 보여주고 있다. 이 그림에서 클러스터 6의 경우 대부분이 수학에 속하는 문서들이 클러스터링 되었지만, 생태학에 속하는 681번 문서라든가 물리학의 221번, 통계학의 1329번과 1299번 문서 등이 함께 이 클러스터에 포함되어 있음을 알 수 있다. 이러한 문서들은 잘못 클러스터링 된 사례들로서, 이러한 사례들의 공통된 특징을 추출해 보았다.

2. 단어 수와 클러스터링 성능

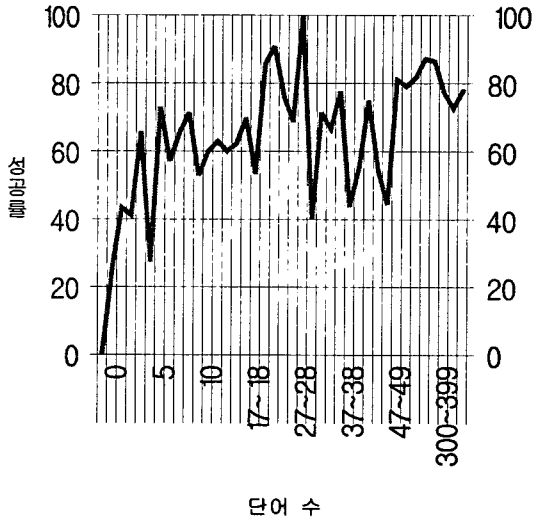
단어기반 클러스터링의 경우 색인어 자체에 직접적인 영향을 받을 수밖에 없다. 특히 대규모 웹 문서를 대상으로 한 색인의 경우 자연언어 처리에 기반한 자동색인 기법이 필수적인데, 이 경우, 웹 문서에 포함된 단어 수가 매우 적거나 이미지만으로 되어 있는 웹 문서의 경우 클러스터링 자체가 불가능하게 된다. 특히 이번 실험의 경우에도 실험 대상 문서 1,449개 중에서 350개(약 24%)가 '단어개수 0'으로 나타났으며, 클러스터링 성능이 현저하게 떨어지는 단어 수 다섯 개 이하의 문서가 489개로 전체 실험 대상의 33.7%에 달하였다. 이런 문제를 해결하기 위해 일반적으로 대상 웹 문서의 링크를 따라 링크된 외부 정보원을 이용하여 추가로 색인어를 추출하는 것이 보편적이지만, 그럴 경우 실험 대상 문서 수가 웹 문서에 포함된 평균 링크 개수만큼 증가하기 때문에 계산량이 엄청나게 늘어나게 되어 그 효율은 오히려 떨어지게 된다.

한편 웹 문서에 등장하는 단어 수와 클러스터링 성능과의 관계는 다음 표와 같으며, 이를 단어 수에 대한 성능 곡선으로 나타내면 다음 그림과 같다.

〈표 6〉 웹 문서 내 단어 수와 클러스터링 성능과의 관계

단어개수	문서 수	성공	실패	성공률(%)	단어개수	문서 수	성공	실패	성공률(%)
0	350	0	350	0.0	27~28	16	11	5	68.8
1	8	2	6	25.0	29~30	13	13	0	100.0
2	56	23	33	41.1	31~32	10	4	6	40.0
3	20	7	13	35.0	33~34	7	5	2	71.4
4	44	27	13	61.4	35~36	15	10	5	66.7
5	11	3	8	27.3	37~38	9	7	2	77.8
6	52	38	14	73.1	39~40	16	7	9	43.8
7	7	4	3	57.1	41~42	9	5	4	55.6
8	29	19	10	65.5	43~44	12	9	3	75.0
9	14	10	4	71.4	45~46	11	6	5	54.5
10	17	9	8	52.9	47~49	9	4	5	44.4
11	5	3	2	60.0	50~99	143	116	27	81.1
12	27	17	10	63.0	100~149	125	99	26	79.2
13~14	15	9	6	60.0	150~199	82	67	15	81.7
15~16	32	20	12	62.5	200~299	78	68	10	87.2
17~18	20	14	6	70.0	300~399	37	32	5	86.5
19~20	15	8	7	53.3	400~499	22	17	5	77.3
21~22	14	12	2	85.7	500~599	22	16	6	72.7
23~24	11	10	1	90.9	600 이상	45	32	13	77.8
25~26	21	16	5	76.2	계	1,449	779	666	53.8

위의 표와 그림에 따르면 단어기반 클러스터링 성능은 문서에 포함된 단어의 수와 밀접한 연관을 맺고 있음을 보여주고 있다. 단어 수에 대한 클러스터링 성능 곡선은 단어 수가 특정한 개수까지는 급격하게 늘어나다가 어느 순간부터는 단어 수에 관계없이 비슷한 성능을 보여주고 있다. 특히 전체 실험대상 문서 1,499개의 약 1/3에 달하는 489개의 문서가 다섯 개 이하의 단어를 포함하고 있으며, 이들 중 62개 문서만이 클러스터링 전의 범주와 일치하게 클러스터링 되어 약 12.7%의 클러스터링 성공률을 보여줌으로써, 전체 클러스터링 성능에 결정적인 영향을 미치고 있다. 또한 전체 문서를 대상으로 한 클러스터링 성공률은 53.8%이지만, '단어개수 0'인 문서 350개를 제외하면 성공률이 70.9%로 급격하게 상승한다. 다음으로 단어개수 5까지를 차례로 제외해 나가면 1% 정도씩 성공률이 증가해 나가다가 6개 이상부터는 75~6%대로 수렴하게 된다.



〈그림 5〉 단어 개수와 클러스터링 성능과의 관계

따라서 '단어개수 5' 이하인 문서에 대해서는 별도의 클러스터링 알고리즘의 개발이 필요하게 되며, 이 연구에서는 두 문서를 동시에 링크하는 웹 문서의 수를 두 문서에 동시에 등장하는 단어수와 같은 비중의 자질로 간주하는 방법을 고려하게 된 것이다.

3. 링크기반 클러스터링

분석대상이 된 문서 전체를 대상으로 각 문서별 1,449×1,449의 동시링크빈도 행렬을 작성하였다. 이를 Cluto 시스템을 통해 앞의 단어기반 클러스터링과 같은 여러 가지 방법으로 클러스터링 해 본 결과는 다음 표와 같다.

이러한 결과들을 종합해 보면 본 연구의 실험 대상 문서의 경우 링크기반 클러스터링은 단어기반 클러스터링과 비슷한 결과를 보였다. 다만 단어 유사도의 경우는 반대로 상관계수보다는 역문헌빈도(idf)가 더 나은 클러스터링 성능을 보였다. 최종적으로 클러스터링 한 결과는 다음 표와 같다.

〈표 7〉 링크기반 하향식 클러스터링 성능

평가함수	역문헌빈도		상관계수	
	entropy	purity	entropy	purity
I1	0.319	0.695	0.338	0.667
I2	**0.263**	**0.743**	0.266	0.744
E1	0.278	0.725	0.296	0.711
G1	0.265	0.755	0.290	0.738
G1P	0.297	0.707	0.313	0.692
H1	0.280	0.721	0.301	0.713
H2	0.269	0.741	0.297	0.711

〈표 8〉 링크기반 클러스터링 결과

20-way clustering: [I2=7.75e+002] [1154 of 1449], Entropy: 0.263, Purity: 0.743

cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty	Aggr	Alt	Phys	Bio	Ecol	Math	Acou	Eart	Astr	Stat	Chem
0	314	+0.956	+0.101	+0.010	+0.003	0.478	0.385	121	4	97	92	0	0	0	0	0	0	0
1	14	+0.903	+0.165	+0.117	+0.020	0.107	0.929	13	0	0	0	0	0	0	0	1	0	0
2	22	+0.684	+0.102	+0.019	+0.011	0.000	1.000	0	0	0	0	0	0	0	0	22	0	0
3	51	+0.540	+0.221	+0.004	+0.002	0.080	0.961	0	0	0	1	0	0	0	0	0	49	1
4	50	+0.464	+0.205	+0.010	+0.009	0.183	0.840	0	0	0	8	0	0	0	0	42	0	0
5	42	+0.463	+0.161	+0.013	+0.014	0.094	0.952	0	0	0	1	0	0	0	1	40	0	0
6	22	+0.446	+0.183	+0.012	+0.008	0.000	1.000	0	0	0	22	0	0	0	0	0	0	0
7	25	+0.438	+0.195	+0.008	+0.009	0.183	0.840	0	0	0	4	0	0	0	21	0	0	0
8	71	+0.399	+0.197	+0.006	+0.004	0.115	0.944	0	0	0	1	0	67	0	1	0	0	2
9	53	+0.406	+0.182	+0.014	+0.007	0.238	0.830	0	0	0	7	0	1	0	0	0	1	44
10	48	+0.405	+0.155	+0.025	+0.012	0.084	0.958	0	0	0	1	0	0	0	1	46	0	0
11	51	+0.263	+0.126	+0.020	+0.013	0.161	0.902	0	0	0	46	3	0	0	2	0	0	0
12	57	+0.259	+0.104	+0.023	+0.013	0.142	0.912	0	0	0	52	4	0	0	0	0	0	1
13	41	+0.244	+0.102	+0.013	+0.008	0.095	0.951	0	0	0	0	0	1	0	0	1	0	39
14	49	+0.229	+0.086	+0.016	+0.014	0.071	0.959	0	0	0	2	0	0	0	0	47	0	0
15	46	+0.205	+0.073	+0.011	+0.008	0.087	0.957	0	0	0	0	0	0	0	44	1	0	1
16	44	+0.182	+0.072	+0.016	+0.012	0.432	0.659	0	0	0	8	3	0	0	29	1	0	3
17	35	+0.171	+0.077	+0.014	+0.010	0.091	0.943	0	0	0	33	2	0	0	0	0	0	0
18	68	+0.135	+0.063	+0.020	+0.015	0.523	0.618	0	0	0	42	8	3	1	8	5	0	1
19	51	+0.121	+0.073	+0.009	+0.008	0.390	0.745	0	0	0	38	0	1	3	5	1	0	3

이 표에서 클러스터링 성능을 주제별로 보면 앞의 단어기반 클러스터링과 거의 유사하게 나타났다. 전체적으로 보아 1,449개의 문서 중에서 1,154개의 문서가 클러스터링 되었으며, 클러스터로 묶이지 않은 295개의 문서에는 '동시링크 빈도 0'인 문서 293개가 포함

되어 있다. 클러스터 ID별로는 클러스터 'ID 0, 16, 18, 19'를 제외하면 그 성능이 월등하게 향상되어 나타났다.

단어기반 클러스터링과 링크기반 클러스터링 성능을 단순 비교해 보면, 각각 복잡도와 순정도가 0.352, 0.723과 0.263, 0.743으로, 링크기반 클러스터링의 성능이 더 나은 것으로 나타났다. 그 결정적인 요인이 된 것은 단어기반 클러스터링으로는 불가능했던 '단어 개수 0'인 문서도 링크기반 클러스터링에서는 클러스터링이 가능했기 때문이다. 그러나 링크기반 클러스터링을 통해서도 '동시링크 개수 0'인 문서 293개를 비롯해 동시링크 개수가 매우 적은 문서에 대한 클러스터링 누락과 실패에 대한 문제가 여전히 남게 된다. 이러한 문제가 위의 표에서 <클러스터 0>으로 나타나 있으며, 전체 클러스터링 성능에 큰 영향을 주고 있다.

따라서 웹 문서에 포함된 단어 수가 일정한 개수 이하인 문서에만 동시링크 빈도를 적용하거나, 단어출현빈도에 동시링크 빈도를 단어출현 빈도와 같은 자질로 추가할 필요가 있다.

4. 단어-링크 혼합 클러스터링

분석대상이 된 문서 전체를 대상으로 각 문서별 단어출현빈도 행렬에 동시링크 행렬을 추가하여 1,499×(17,223+1,499) 행렬을 작성하였다. 앞의 1,499는 실험대상 문서 수이며, 17,223은 사용된 색인어 수, 그리고 뒤의 1,499는 두 문서의 동시링크 빈도를 나타낸다. 이를 Cluto 시스템을 통해 앞의 실험과 같은 여러 가지 방법으로 클러스터링 해 본 결과는 다음 표와 같다.

〈표 9〉 단어-링크 혼합 클러스터링 성능(문서 유사도: 코사인, 칼럼 유사도: idf)

평가함수	역문헌빈도		상관계수	
	entropy	purity	entropy	purity
I1	0.400	0.599	0.392	0.616
I2	**0.260**	**0.748**	0.271	0.745
E1	0.293	0.702	0.309	0.687
G1	0.277	0.727	0.283	0.747
G1P	0.301	0.707	0.318	0.693
H1	0.298	0.709	0.318	0.682
H2	0.293	0.704	0.299	0.701

단어-링크 혼합 클러스터링 역시 본 연구의 실험 대상 문서의 경우 클러스터링 방식은 상향식(agglomerative) 보다는 하향식(paritional)이, 문서 유사도의 경우 상관계수보다는

코사인이, 단어 유사도의 경우는 상관계수보다는 역문헌빈도(idf)가 더 나은 클러스터링 성능을 보였다. 또한 사용된 평가함수 중에서도 대체적으로 I2의 경우에 성능이 높은 것으로 나타났다. 각 클러스터별 클러스터링 결과는 다음 표와 같다.

〈표 10〉 단어-링크 혼합 클러스터링 결과

20-way clustering: [I2=7.33e+002] [1312 of 1449], Entropy: 0.260, Purity: 0.748

cid	Size	ISim	ISdev	ESim	ESdev	Entpy	Purty	Aggr	Alt	Phys	Bio	Ecol	Math	Acou	Eart	Astr	Stat	Chem
0	312	+0.959	+0.077	+0.008	+0.002	0.478	0.388	121	4	97	90	0	0	0	0	0	0	0
1	15	+0.778	+0.249	+0.097	+0.027	0.164	0.867	13	0	0	2	0	0	0	0	0	0	0
2	34	+0.509	+0.186	+0.007	+0.004	0.110	0.941	0	0	0	1	0	0	0	1	32	0	0
3	39	+0.477	+0.167	+0.010	+0.008	0.099	0.949	0	0	0	1	0	0	0	1	37	0	0
4	51	+0.441	+0.203	+0.003	+0.002	0.040	0.980	0	0	0	0	0	1	0	0	0	50	0
5	22	+0.301	+0.192	+0.003	+0.002	0.546	0.545	0	0	0	12	0	3	0	2	3	0	2
6	72	+0.276	+0.114	+0.013	+0.008	0.061	0.972	0	0	0	1	0	0	0	1	70	0	0
7	82	+0.249	+0.160	+0.004	+0.003	0.157	0.927	0	0	0	2	0	76	1	1	0	1	1
8	78	+0.244	+0.109	+0.006	+0.004	0.029	0.987	0	0	0	1	0	0	0	0	0	0	77
9	37	+0.242	+0.079	+0.014	+0.009	0.052	0.973	0	0	0	1	0	0	0	0	36	0	0
10	30	+0.212	+0.105	+0.003	+0.002	0.298	0.700	0	0	0	8	1	0	0	21	0	0	0
11	34	+0.192	+0.064	+0.002	+0.002	0.000	1.000	0	0	0	0	0	0	0	0	34	0	0
12	101	+0.167	+0.076	+0.012	+0.006	0.370	0.762	0	0	0	77	9	1	1	8	4	0	1
13	44	+0.145	+0.067	+0.007	+0.004	0.000	1.000	0	0	0	44	0	0	0	0	0	0	0
14	55	+0.120	+0.049	+0.007	+0.004	0.038	0.982	0	0	0	0	1	0	0	54	0	0	0
15	54	+0.072	+0.048	+0.006	+0.005	0.419	0.611	0	0	0	33	0	1	2	1	1	0	16
16	65	+0.067	+0.025	+0.004	+0.002	0.238	0.862	0	0	0	3	0	1	0	4	56	0	1
17	72	+0.062	+0.025	+0.008	+0.006	0.172	0.903	0	0	0	65	1	0	0	4	0	0	2
18	62	+0.055	+0.023	+0.002	+0.001	0.523	0.565	0	0	0	6	0	11	1	35	1	0	8
19	53	+0.053	+0.025	+0.005	+0.004	0.286	0.736	0	0	0	39	12	0	0	2	0	0	0

전체적으로 보아 1,449개의 문서 중에서 1,312개의 문서가 클러스터링 되었으며, 클러스터로 묶이지 않은 137개의 문서에는 '단어개수와 동시링크 개수가 모두 0'인 문서 134개가 포함되어 있다. 클러스터별, 주제별 클러스터링 성능은 앞의 링크기반 클러스터링 성능과 거의 유사한 패턴을 보여주고 있다. 다만 링크기반 클러스터링보다 단어-링크 혼합 클러스터링의 성능이 약간 더 나아졌음을 알 수 있다. 그러나 <클러스터 0>으로 인해 전반적인 성능을 저하시키는 문제는 여전히 존재하는 것으로 나타나고 있다. 단어-링크 혼합 클러스터링이 링크기반 클러스터링과 비슷한 패턴을 보인 것은 전반적으로 단어빈도보다 동시링크 빈도 숫자가 크기 때문이다.

한편 이 문제를 보완하기 위해 문서에 포함된 단어 수가 5개 이하, 10개 이하, 20개 이하인 문서에만 각각 동시링크 빈도를 추가하여 복잡도와 순정도를 뽑아 보았지만 별다른 성능개선 효과는 보이지 않았다.

〈표 11〉 문서에 포함된 단어 개수에 따른 단어-링크 혼합 클러스터링 성능 비교

단어 개수	Entropy	Purity
5개 이하	0.343	0.724
10개 이하	0.359	0.686
20개 이하	0.339	0.711

Ⅳ. 결 론

웹 문서 클러스터링의 경우 단어기반 클러스터링 기법이 일반적이지만, 웹 문서에 텍스트 자체가 없거나 단어 수가 매우 적은 문서의 경우 클러스터링 자체가 불가능하게 된다. 이 연구에서는 네이버 디렉터리 중 '자연과학' 범주에 포함된 1,449개의 웹 문서를 대상으로 단어기반 클러스터링과 링크기반 클러스터링, 그리고 단어-링크 혼합 클러스트링 기법으로 클러스터링 해 보았으며, 그 결과를 네이버 디렉터리에 초기 할당된 범주와 비교해 보았다. 그 결과 단어빈도와 동시링크 빈도를 함께 이용한 방식의 클러스터링 성능이 가장 좋게 나타났다. 아울러 웹 문서에 포함된 단어 수와 단어기반 클러스터링 성능과의 관계를 조사해 본 결과 '단어개수 5' 이하에서 그 성능이 크게 떨어짐을 알 수 있었다.

지금까지의 실험결과를 종합하여 전체 실험대상 문서 1,449개에 대한 각 클러스터링 방식별 성능을 표로 제시하면 다음과 같다.

〈표 12〉 클러스터링 방식별 성능 비교

	단어기반	링크기반	단어-링크 혼합
클러스터링된 문헌수	1068	1154	1312
클러스터링되지않은 문헌수	381	295	137
단어/링크/혼합 '0'인 문헌수	350	293	134
Entropy	0.352	0.263	0.260
Purity	0.723	0.743	0.748

이 연구에서 제시된 방법 중 클러스터링 성능을 가장 크게 향상시킬 수 있는 것은 단어-링크 혼합 클러스터링으로서, 그 중에서도 동시링크 빈도 전체를 단어출현빈도에 혼합한 것이 가장 좋은 결과를 보였다.

그러나 실험대상 문서 수가 웹 전체 문서에 비해 너무 작다는 것과, 특정 검색엔진의 하위 디렉터리 하나만을 대상으로 실험이 이루어졌다는 점 등은 이 실험의 제한점으로 볼 수 있다. 앞으로 보다 실험범위를 넓혀 웹 문서에 포함된 단어 개수와 클러스터링 성능과의 관계, 그리고 단어와 문서, 동시링크, 클러스터 등의 유사도 계산에 관련된 보다 정밀한 알고리즘의 개발과 실제 웹 문서에 대한 적용 등이 후속 과제로 남아있다.

참고문헌

국민상, 정영미. "인용문헌을 이용한 검색 성능 향상에 관한 실험적 연구", 제19회 한국정보관리학회 학술대회 논문집(2002. 8). pp.235-240.

김영기, 이원희, 권혁철, "동시링크를 이용한 웹 문서 클러스터링 실험", 한국도서관·정보학회지, 제34권 2호(2003. 6). pp.233~253.

오효정, 임정묵, 이만호, 맹성현. "점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 모델", 제11회 한글 및 한국어 정보처리 학술대회(1999). pp.89~96.

정상화, 이종혁, "문서구조 정보에 기반한 웹 페이지 범주화 모델", 제10회 한글 및 한국어 정보처리 학술대회(1998). pp.91~96.

정성원, 이원희, 김영기, 권혁철. "웹 문서 중 의미 있는 표의 추출", 한글 및 한국어 정보처리, 제14집(2002. 10). pp.332~339.

최정태, 양재한, 도태현. 문헌분류의 이론과 실제. 부산대학교 출판부, 1998.

Aptè, Chidanand and Damerau, Fred and Weis, Sholom M. "Towards language independent automated learning of text categorization models", *Proc. of the 17th Annual International ACM-SIGIR* (1994). pp.233~251.

Baker, L. Douglas and Maccallum, Andrew K. "Distributional clustering of words for text classification", *Proc. of the 21th Annual International ACM-SIGIR*, 1998.

Belew, R. K. *Finding Out About: A Cognitive perspective on search engine technology and the WWW.* Cambridge University Press, 2000.

Chakrabarti, Soumen and Dom, Byron and Piotr Indyk, "Enhanced hypertext categorization using hyperlinks", *Proc. of International Conference on SIGMOD '98* (1998). pp.307~318.

Joachims, Thorsten. "Text categorization with support vector machines", *Proc. of European Conference on Machine Learning, ECML '98*, 1998. pp.137~142.

Kessler, M. M. "Bibliographic coupling between scientific papers", *American Documentation.* vol.14 no.1(1963). pp.10~25.

Larkey, Leah S. "Automatic essay grading using text categorization techniques", *Proc. of the 21th Annual International ACM-SIGIR* (1998). pp.90~95.

Lewis, David L. and Ringuette, Marc. " A comparison of two learning algorithms for text categorization", *Proc. of the 3rd Annual Symposium on Document Analysis and Information Retrieval* (1998). pp.96~103.

Lewis, David L. and Schapire, Robert E. and Callan, James P. and Papka. Ron "Training algorithms for linear text classifier", *Proc. of the 19th Annual International ACM-SIGIR* (1996). pp.298~315.

Small, H. "Co-citation in the scientific literature: A new measure of the relationship between two documents", *Journal of American society for Information Science.* vol.24(1973). pp.265~269.

Zhao, Ying and Karypis, George, "Criterion functions for document clustering - experiment and analysis", *Technical Report* TR #01-40, Department of Computer Science, University of Minnesota, 2001.

Zhao, Ying and Karypis, George, "Evaluation of hierarchical clustering algorithms for document datasets", *Technical Report* TR #02-22, Department of Computer Science, University of Minnesota, 2002.

Karypis, George, "CLUTO: A Clustering Toolkit", *Technical Report* TR #02-017, Department of Computer Science, University of Minnesota, 2002.