한글 웹 문서 클러스터링 성능향상을 위한 자질선정 기법 비교 연구*

A Comparative Study of Feature Selection Methods for Korean Web Documents Clustering

김 영 기(Young-Gi Kim)**

목 차

초 록

이 연구는 한글 웹 문서를 클러스터링 하기 위한 자질 선정 방법에 대한 비교연구이다. 이 연구에는 두 개의 코퍼스가 사용되었다. 클러스터링을 위한 실험 문서는 Naver의 자연과학 범주에서, 자질 선정을 위한 학습문서는 Yahoo Korea의 같은 범주에서 수집하였다. 우선 실험 문서를 단어자질과 동시링크, 그리고 이 둘을 혼합한 방법으로 클러스터링 한 다음 그 성능을 비교하였다. 다음으로 학습문서에서 카이제곱 통계량(X^2), 정보획득량(IG), 그리고 상호정보량(MI)을 이용하여 용어자질을 선정한 다음, 이를 실험문서에 적용하여 클러스터링 성능을 비교하였다. 여기에 각 범주별로 최댓값을 갖는 용어들만을 해당 범주를 대표하는 자질로 선정하는 '최댓값 자질 선정기법'을 실험적으로 도입하여 적용해 보았다. 실험 결과 사용된 자질에 따른 한글 웹 문서 클러스터링 정확률은 자연어 72.3%, 동시링크 74.3%, 단어-링크 혼합 74.8%, X^2 79.6%, Max X^2 83.8%로 나타났다. 전통적 자질 선정 기법 중에서는 X^2가 약간 나은 성능을 보여 주었지만 큰 차이는 발견되지 않았다. 그러나 최댓값 자질 선정기법을 적용하였을 때 클러스터링 성능은 크게 향상되었다. 이 논문에서 제안된 최댓값 자질 선정 기법은 웹 문서의 자질 공간 축소와 한글 웹 문서의 클러스터링을 위한 간단하면서도 효과적인 수단이다.

ABSTRACT

This paper is a comparative study of feature selection methods for Korean web documents clustering. First, we focused on how the term feature and the co-link of web documents affect clustering performance. We clustered web documents by native term feature, co-link and both, and compared the output results with the originally allocated category. And we selected term features for each category using X^2, *Information Gain (IG)*, and *Mutual Information (MI)* from training documents, and applied these features to other experimental documents. In addition we suggested a new method named *Max Feature Selection*, which selects terms that have the maximum count for a category in each experimental document, and applied X^2 (or MI or IG) values to each term instead of term frequency of documents, and clustered them. In the results, X^2 shows a better performance than *IG* or *MI*, but the difference appears to be slight. But when we applied the *Max Feature Selection Method*, the clustering performance improved notably. *Max Feature Selection* is a simple but effective means of feature space reduction and shows powerful performance for Korean web document clustering.

키워드: 클러스터링, 자질선정 기법, 한글 웹 문서, 최댓값 자질선정 기법
Clustering, Feature Selection Methods, Korean Web Documents, Max Feature Selection

---

* 이 논문은 2004학년도 경성대학교 학술지원연구비에 의하여 연구되었음.
** 경성대학교 문과대학 문헌정보학과 교수(ykk@ks.ac.kr)
논문접수일자 2005년 1월 10일
게재확정일자 2005년 3월 12일

1. 서 론

1. 1 연구 내용

인터넷 정보자원의 급격한 증대로 효율적인 정보 관리 및 검색이 요구된다. 온라인상에서 얻을 수 있는 정보의 양이 급증함에 따라 이를 효과적으로 조직하여, 검색성능을 향상시키기 위한 다양한 연구와 실험이 수행되고 있다. 특히 유사한 문서를 자동으로 군집화 시키는 자동 문서 범주화 기법(범주화클러스터링)은 가장 기본적인 자료관리 행위이면서, 동시에 고도의 지식과 테크닉을 요구하는 일로서 정보자원의 효과적인 탐색과 이용을 위한 출발점이며, 검색 성능을 향상시킬 수 있는 매우 강력한 수단 중의 하나이다.

이전에는 수작업으로 문서마다 범주를 지정해 주는 방식이 사용되었으나 이 경우에는 사람의 노력, 시간, 비용 면에서 심각한 어려움을 초래할 수 있다. 이러한 작업을 자동 분류 시스템으로 교체하거나 보조 시스템으로 활용하면 비용과 노력을 크게 줄일 수 있을 것이다. 그러므로 클러스터링은 대량의 문서를 효율적으로 관리하고 검색할 수 있게 하는 동시에 방대한 양의 수작업을 감소시키는데 그 목적이 있다(Yang 1997, 고영중 2002).

정보검색 분야에서 클러스터링은 매우 다양한 방식으로 연구되고 있는데, 크게 단어 유사도에 기반을 둔 단어기반 클러스터링(term-based clustering)과 링크기반 클러스터링(link-based clustering), 그리고 이 둘을 혼합한 혼합형 클러스터링(hybrid clustering)으로 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 기반을 둔 자동색인 기법이 필수적이다. 최근에는 온라인 시소러스나 단어의 의미적 관련성을 활용하는 시도도 부분적으로 이루어지고 있다.

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 하이퍼링크와 같은 외부 정보원의 부가적 이용을 고려해 볼 수 있다(정성원 등 2002).

반대로 웹 문서에 나타나는 용어의 수가 너무 많은 경우에도 좋은 클러스터링 성능을 기대하기 어렵다. 이 경우 문서 클러스터링의 가장 큰 문제점은 자질 공간 차원(문서에 포함된 단어 수)이 너무 크다는 것이다. 자연어 자질 공간은 문서에 나타나는 단어나 구와 같은 용어로 구성되는데, 평균 수만에서 수십만에 이르기 때문에 문서에 나타나는 모든 용어가 자질로 선택된다면 그 시간이 매우 오래 걸리게 된다. 따라서 클러스터링 성능의 저하 없이 자질의 수를 줄이기 위해 문서에 나타나는 정보량을 계산하고, 정보량이 큰 단어만을 자질로 선택하려는 연구가 활발히 진행되어 왔다. 그러나 많은 자질 선택 방법이 개발되어 왔지만 대량의 문서, 특히 한글로 된 대량의 웹 문서를

대상으로 한 연구와 평가는 거의 이루어 지지 않았다.

이 연구에서는 우선 웹 문서에 포함된 단어 빈도와 역문헌 빈도, 동시 링크 빈도, 그리고 이 둘을 혼합한 방법으로 한글 웹 문서를 클러스터링 한 다음 그 성능을 비교해 보았다. 다음으로 별도의 학습 데이터를 통해, 범주를 대표할 수 있는 중요한 자질을 선별하는 기법으로 널리 알려진 카이제곱 통계량(X^2), 정보획득량(*Information Gain, IG*) 그리고 상호정보량(*Mutual Information, MI*)을 이용하여 주요 자질을 선정한 다음, 이를 다시 실험문서에 적용해 보았다. 여기에 덧붙여 기존의 자질 선택 방법에 각각의 용어를 최댓값을 갖는 범주에만 배타적으로 할당하는 최댓값 자질 선정 기법(*Max Feature*)을 제안, 적용하여 클러스터링 성능을 비교해 보았다.

1. 2 관련 연구

정보검색에 있어서 클러스터링은 다양한 방법으로 연구되어 왔다. 웹 정보검색의 경우 클러스터링은 주제별 커뮤니티의 확인(Kumar 1999, Mukherjea 2000)이나 웹 구조의 추출(Larson 1996, Pirolli 1996), 그리고 중복 문서의 발견(Broder 1997), 적합성 피드백을 통한 질의확장(Chang 1998) 등의 수단으로 연구되어 왔다. 단어기반 클러스터링은 각 문헌에 포함된 단어 유사도에 근거하여 전체 문헌 집단에 대한 조직으로 나아가고 있으며, 최근 인터넷 환경에서의 정보검색은 링크기반 클러스터링 알고리즘과 검색결과의 동적 클러스터링 양쪽을 포괄하는 연구경향을 보이고 있다.

2. 연구 방법

2. 1 클러스터링을 위한 문서 수집

이 연구에는 두 개의 코퍼스가 사용되었다. 클러스터링을 위한 실험 문서는 Naver의 자연과학 범주에서, 자질 선정을 위한 학습문서는 Yahoo Korea의 같은 범주에서 수집하였다. 우선 실험 문서는 검색 포탈 엔진인 네이버 디렉터리 중에서 '자연과학' 카테고리(http://dir.naver.com/Education__and__Science/Science) 내의 문서들을 선택하여 수집하였다. '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. 최종적으로 1,449개의 웹 문서가 실험대상으로 선정되었으며, 범주별 웹 문서 수는 〈표 1〉과 같다.

2. 2 자연어 기반 클러스터링

단어기반 클러스터링을 위해 실험 대상이 된 1,449개의 문서에서 부산대학교 자연어 처리 연구실의 색인 시스템을 이용하여 17,223개의 용어를 추출하여, 문서-단어빈도 행렬(document-term frequency matrix)을 작성하였다. 클러스터링 작업과 클러스터링 성능 분석을 위해서는 미네소타 대학에서 개발한 클러스터링 시스템(Clustering Toolkit)인 Cluto2.1.1(http://www-users.cs.umn.edu/~karypis/ cluto/)을 사용하였다. 그것은 Cluto가 지금까지 알려

〈표 1〉 범주별 실험 대상 문서

범 주	문서 수
농학(Aggr)	145
대체과학(Alt)	4
물리학(Phys)	102
생물학(Bio)	426
생태학(Ecol)	24
수학(Math)	102
음향학(Acou)	5
지구과학(Earth)	149
천문학(Astro)	323
통계학(Stat)	56
화학(Chem)	113
계	1,449

진 대부분의 클러스터링 성능분석기법을 다양하게 적용해 볼 수 있는 강력한 시스템이기 때문이다.

일반적으로 키워드를 통한 두 문헌간의 관계를 나타내기 위해서는 단어출현빈도(Term Frequency, TF)와 역문헌빈도(Inverse Document Frequency, IDF)를 이용한 단어벡터(word vector)와 문서 헤드의 거리비교(distance comparison)가 사용된다.

그리고 두 문서간의 유사도를 계산하여 각 문서를 특정 클러스터에 할당하기 위한 문서범주화 모델로는 베이지언 확률모델(Bayesian Probability Model), k-최근린법(k-Nearest Neighbor), 지지벡터(Support Vector Machine, SVM), 선형분류모델(Linear Classification Model) 등이 대표적이다. Cluto 시스템은 지금까지 알려진 대부분의 유사도 계산 모델을 옵션으로 제공하고 있는데, 여기서는 이러한 평가 함수 중 가장 좋은 결과를 보이는 다음의 함수를 사용하였다.

$$\max \sum_{i=1}^{k} \left(\sum_{v,u \in S_i} \sim(v,u) \right)$$

여기서 k는 전체 클러스터 수, S는 클러스터된 전체 문서 수, Si는 i번째 클러스터에 할당된 문서 집합, ni는 i번째 클러스터에 할당된 문서 수, v, u는 문서, $\sim(v, u)$는 두 문서간의 유사도를 각각 의미한다. 클러스터링 정확도를 얻기 위해 클러스터링 결과를 네이버 디렉터리에 초기 할당된 범주와 비교하였다.

2. 3 링크기반 클러스터링

웹 문서는 각종 정보원들이 링크를 통해 서로 연결되어 있는데, 링크 기반 클러스터링은 하이퍼링크에 포함된 정보를 이용하여 관련문서 군집화 시키는 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)로 나눌 수도 있다. 나가는 링크가 많은 사이트는 연결성(hubness)이 높으며, 들어오는 링크가 많은 사이트는 신뢰성(authority)이 높다고 말한다[Belew 2000].

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-link)의 경우 웹 문서의 파싱(parsing)을 통해 그 개수를 쉽게 구할 수 있지만, 들어오는 링크는 해당 분야의 전체 코퍼스(corpus)를 직접 갖고 있어야만 구할 수 있다.

이 연구에서는 실험대상이 된 웹 문서들을 대상으로 각 두 개의 문서를 동시에 링크하고 있는 사이트 수를 조사하기 위해 Alltheweb의 "Advanced Search"(http://www.alltheweb.com)를 이용하였다. 사용된 검색 식은 다음과 같다.

LINK:url_A AND LINK:url_B

동시링크 빈도를 구하는 작업은 매우 많은 시간과 네트워크의 트래픽 가중, 호스트의 부하를 요하는 작업이다. 이 실험에서는 14대의 인터넷에 연결된 PC로 1주일에 걸쳐 1,449개의 문서에 대해 모두 1,049,076{(1,449×1,449-1449)/2}회의 검색이 이루어 졌으며, 검색 결과에서 동시링크 빈도수를 추출하였다. 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. 이 실험에서는 TFIDF를 변형한 CCIDF (Common Citation × Inverse Document Frequency)를 이용하여 웹 문서들 간의 상대적인 유사도와 비유사도를 구했으며, 이를 단어기반 클러스터링 결과와 비교하였다.

2.4 단어-링크 혼합 클러스터링

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. 즉 웹 문서 색인을 통해 생산된 단어 빈도 벡터와 동시 링크 빈도 벡터를 결합하여 같은 방법으로 클러스터링 성능을 실험하였다.

2.5 자질 선정을 위한 문서 수집

범주별 자질 선정을 위한 학습 문서는 한국 야후의 자연과학 디렉터리(http://kr.dir.yahoo.

com/science/)에서 수집하였다. Naver와 Yahoo Korea의 자연과학 디렉터리 범주는 서로 유사하지만 포함된 문서는 충분히 다르기 때문에 학습문서로서 적당한 것으로 판단하였다. 〈표 2〉는 Yahoo Korea 자연과학 디렉터리의 각 범주 안에서 범주 이름, 수집된 문서 수, 문서 파싱 과정을 거쳐 생성된 전체 용어 수, 불용어 처리 과정을 거친 이후 채택된 단어 수를 요약한 것이다. 불용어 처리는 일반적인 불용어 처리 과정을 거친 후 다음에 해당하는 용어를 추가로 제거하였다.

첫째, 문헌빈도(DF)가 2 이하인 모든 용어

둘째, 철자 오류, 띄어쓰기 오류, 한자와 한글, 영어, 숫자 등이 혼합된 용어

셋째, 널리 알려지지 않은 사람 이름이나 ID와 같은 고유명사

셋째, 웹 문서 편집 도구에 의해 사용된 용어

넷째, 홈 페이지나 게시판, 로그 인 등과 같이 사이트에 따라 공통적으로 많이 포함된 용어

2.6 자질 선정 방법

단어기반 클러스터링의 가장 큰 문제점은 자질 공간 차원이 너무 크다는 것이다. 자연어 자질 공간은 문서에 나타나는 단어나 구와 같은 용어로 구성되는데, 평균 수만에서 수십만에 이르기 때문에 문서에 나타나는 모든 용어가 자질로 선택된다면 그 시간이 매우 오래 걸리게 된다. 따라서 클러스터링 성능의 저하 없이 자질의 수를 줄이기 위해 문서에 나타나는 정보량을 계산하고, 정보량이 큰 단어만을 자질로 선택하려는 연구가 활발히 진행되어 왔다(Yang 1997). 자질 선별이란 범주를 대표할 만한 중요한 용어를 얻는 방법이다. 많은 자질

〈표 2〉 자질선정을 위한 학습 문서

범주	문서 수	전체 색인어 수	선정된 색인어 수
건축학	73	199	25
기계공학	59	298	66
재료공학	10	632	98
전기·전자공학	60	727	69
토목공학	46	687	61
화학공학	30	371	44
환경공학	22	769	96
농학	40	983	145
컴퓨터공학	69	871	83
생명과학	38	1,856	245
물리학	65	341	85
생물학	368	5,785	1,240
지구과학	134	2,018	378
화학	61	307	45
수학	127	872	156
지리학	52	745	45
천문학	79	1,175	241
합계	1,333	18,636	3,122

선정 기법 중에서 일반적으로 카이제곱 통계량(X^2), 상호정보량(Mutual Information, MI) 그리고 정보획득량(Information Gain, IG) 등이 좋은 성능을 보인다. 이러한 기법들을 이용하여 벡터 공간을 줄일 수 있다(고영중 2002).

이 중에서 카이제곱 통계량은 중요 자질을 순위화하여 벡터 차원을 줄일 수 있으며, 용어 t와 범주 c와의 의존성을 측정하는데 사용된다. X^2을 계산하는 식은 다음과 같다.

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}$$

A : 범주 c에 속해 있는 문서 중 용어 t를 포함하고 있는 문서 수
B : 범주 c에 속하지 않은 문서 중 용어 t를 포함하고 있는 문서 수
C : 범주 c에 속해 있는 문서 중 용어 t를 포함하고 있지 않은 문서 수
D : 범주 c에 속하지 않은 문서 중 용어 t를 포함하고 있지 않은 문서 수
N : 학습에 사용된 전체 문서 수

카이제곱 통계량은 용어 t와 범주 c가 완전히 독립적이면 0의 값을 가진다.

그리고 정보 획득량(Information Gain)은 기계학습 분야에서 자주 사용되는 기법으로, 용어의 출현 빈도뿐만 아니라 출현하지 않은 빈도까지 고려하여 각 범주에서의 용어의 정보량을 계산한다. 용어 t의 정보 획득량은 다음과 같이 정의된다.

$$G(t) = -\sum_{i=1}^{m} \Pr(c_i) \log \Pr(c_i) + \Pr(t) \sum_{i=1}^{m} \Pr(c_i|t) \log \Pr(c_i|t) + \Pr(\bar{t}) \sum_{i=1}^{m} \Pr(c_i|\bar{t}) \log \Pr(c_i|\bar{t})$$

상호정보량은 통계적 언어모델에서 널리 사용되는 기법으로서 용어 t의 범주 c에서의 상호정보량은 다음과 같이 정의된다.

$$I(t,c) = \log \frac{\Pr(t \wedge c)}{\Pr(t) \times \Pr(c)}$$

상호정보량의 근사값은 다음의 식으로 계산된다.

$$I(t,c) \approx \log \frac{A \times N}{(A+C) \times (A+B)}$$

각 용어에 대한 범주별 정보량은 다음의 과정을 통해 구해진다.

$$(X^2/MI/IG)_{\max}(t) = \max_{i=1}^{m} \{X^2/MI/IG(t,c_i)\}$$

여기에 덧붙여 기존의 자질 선택 방법에 각각의 용어를 최댓값을 갖는 범주에만 배타적으로 할당하는 최댓값 자질 선정 기법(Max Feature Selection Method)을 제안, 적용하여 클러스터링 성능을 비교해 보았다. 최댓값 자질을 선정하는 과정은 다음과 같다.

우선 각 문서는 다음과 같이 표현할 수 있다.

Doc(t_1, t_2, t_3, ... , t_k) 또는

Doc(t_1^{c1}, t_2^{c1},...,t_i^{c2}, t_{i+1}^{c2},..., t_j^{c3}, t_{j+1}^{c3},..., t_k^{cm}),

즉 하나의 문서는 t^1부터 t_k까지의 단어를 포함하고 있으며, 각각의 단어는 서로 다른 범주 C_1에서 Cm 중의 하나에 속하게 된다. 이 때 문서 내의 각 단어들 중에서 가장 많은 값을 갖는 범주에 속하는 단어들만 남겨 해당 범주에 배타적으로 할당하고 나머지 단어들은 버린다. 즉,

Doc(t_1^{c1}, t_2^{c1},..., t_i^{c2}, t_i+1^{c2},..., ... ,t_j^{Cf}, t_{j+1}^{Cf}, ... t_k^{cm})는
Doc(t_j^{Cf}, t_{j+1}^{Cf},t_{j+2}^{Cf}, ... , t_k^{Cf})로 최적화 하게 된다.

이런 과정을 거쳐 학습문서에서 각 용어와 범주 사이의 X^2, MI, IG, Max X^2, *Max MI* 그리고 *Max IG*를 각각 구한 다음 이를 실험문서의 각 범주를 대표하는 자질로 적용하였다. 이 값들을 실험문서에 적용한 다음 그 클러스터링 성능을 비교하였다.

3. 실험결과 및 평가

3. 1 자연어기반 클러스터링

실험 문서를 대상으로 각 문서별 단어출현 빈도 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 기법으로 나누어 진행되었으며, 앞에서 제시한 평가함수(criterion function)를 적용하였다. 유사도 측정으로는 코사인(cosine)과 상관계수(correlation coefficient), 그리고 역문헌 빈도(idf)를 각각 적용하였다. 그 결과를 Naver 디렉터리에 초기 할당된 것과 비교하였다. 결과적으로 1,449개의 문서 중에서 1,068개의 문서가 클러스터링 되었으며, 클러스터링 정확도는 72.3%로 나타났다.

본 연구의 실험 대상 문서의 경우 클러스터링 방식은 상향식(agglomerative clustering)보다는 하향식(partitional clustering)이, 문서 유사도의 경우 상관계수보다는 코사인(cosine)이, 단어 유사도의 경우는 역문헌빈도(idf)보다는 상관계수가 더 나은 클러스터링 성능을 보였다.

3. 2 링크기반 클러스터링

분석대상이 된 문서 전체를 대상으로 각 문서별 1,449×1,449의 동시링크빈도 행렬을 작성하였다. 이를 앞의 단어기반 클러스터링과 같은

여러 가지 방법으로 클러스터링 해 보았다.

전체적으로 보아 1,449개의 문서 중에서 1,154개의 문서가 클러스터링 되었으며, 클러스터로 묶이지 않은 295개의 문서에는 '동시링크 빈도 0'인 문서 293개가 포함되어 있다. 단어기반 클러스터링과 링크기반 클러스터링 성능을 단순 비교해 보면, 링크기반 클러스터링의 정확률이 74.3%로 단어기반 클러스터링 보다 성능이 더 나은 것으로 나타났다. 그 요인 중의 하나는 단어기반 클러스터링으로는 불가능했던 '단어 개수 0'인 문서도 링크기반 클러스터링에서는 클러스터링이 가능했다는 점을 들 수 있다. 그러나 링크기반 클러스터링을 통해서도 '동시링크 개수 0'인 문서를 비롯해 동시링크 개수가 매우 적은 문서에 대한 클러스터링 누락과 실패에 대한 문제가 여전히 남게 된다.

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

3. 3 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 앞의 실험과 같은 방법으로 클러스터링 해 본 결과 클러스터링 정확률은 74.8%로 나타났다.

전체 1,449개의 문서 중에서 1,312개의 문서가 클러스터링 되었으며, 클러스터로 묶이지 않은 137개의 문서에는 '단어개수와 동시링크 개수가 모두 0'인 문서 134개가 포함되어 있다. 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. 한편 이 문제를 보완하기 위해 문서에 포함된 단어 수가 5개 이하, 10개 이하, 20개 이하인 문서에만 각각 동시링크 빈도를 추가하여 성능을 비교해 보았지만 별다른 성능개선 효과는 보이지 않았다.

3. 4 자질 선정기법을 통한 클러스터링

Yahoo Korea의 자연과학 디렉터리에서 수집된 1,333개의 문서에서 3,122개의 단어를 추출하여 각 용어들의 카이제곱 통계량, 상호정보량, 정보획득량을 계산한 다음, 이를 통해 각 범주를 대표하는 용어들을 선정하였다. 예를 들어 X^2를 통해 선정된 범주별 자질과 X^2값의 일부를 보이면 〈표 3〉과 같다.

Yahoo Korea 디렉터리에서 학습된 범주별 색인어의 카이제곱 통계량, 상호정보량, 정보획득량을 Naver의 같은 디렉터리를 대표하는 색인어로 간주하여 문서-정보량 행렬을 작성하였다. 〈표 4〉는 이러한 전통적인 자질선정 기법과 최대값 자질 선정 기법을 통한 클러스터링 실험결과를 요약한 것이다.

〈표 4〉에서 보는 것처럼 클러스터링 성능은 학습을 하지 않았을 때와 비교하여 크게 향상되었지만, 자질 선정 방법에 따른 차이는 그다지 크게 나타나지 않았다. 그러나 여기에 Max

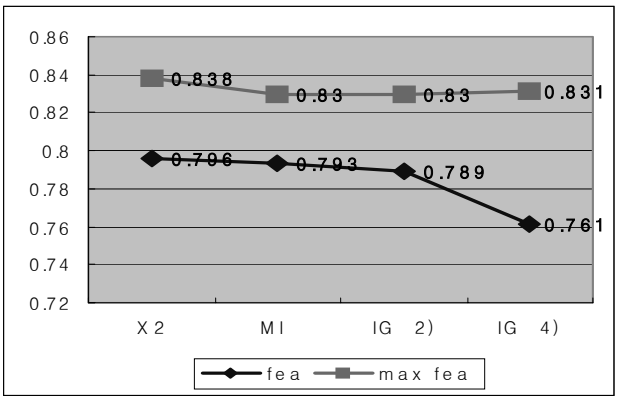
〈표 3〉 범주별 자질 및 X^2(일부)

범주	범주별 자질, X^2값(〉=100)
물리학	물리 223, 물리학 201, 물리학과 196, 핵물리 176, 핵물리학 176, 가속기 137, 상대론 137
생물학	암컷 151, 곤충 137, 애견 127, 먹이 126, 번식 126, 강아지 124, 동물 123, 수컷 121, 야생화 121, 야생 118, 다리 116, 벌레 116, 수염 110, 열매 106, 습성 106
지구과학	기상 303, 저기압 273, 지표면 261, 북태평양 235, 고기압 235, 피해 224, 태평양 224, 강수 216, 황사 210, 가뭄 208, 예보 203, 강수량 196, 기상청 194, 기온 192, 해양 184, (중략), 안개 103, 기압 102, 경보 101
화학	화학과 565, 고분자 165, 자연과학대학 117, 화학 115, 유기 114, 분자 103, 생화 102
수학	수학과 369, 수학 277, 도형 188, 미분 169, 수열 163, 해석학 154, 기하 147, 수학자 146, 피타고라스 142, 무리수 134, 대수 133, 방정식 125, 대수학 124, 선분 124, 정수 114, 미적분 112, 복소수 105, 정수론 104
천문학	망원경 637, 태양계 585, 행성 512, 목성 438, 천체 437, 혜성 421, 자전 420, 토성 401, 금성 389, 공전 384, 궤도 372, 명왕성 370, 천왕성 338, 수성 337, 탐사선 337, 해왕성 322, 천문 296, (중략), 초속 103, 중력 102

〈표 4〉 자질선정 기법에 따른 클러스터링 정확률

자질선정 기법	정확률	최댓값 자질선정 기법	정확률
X^2	*0.796*	*Max X^2*	*0.838*
MI	*0.793*	*Max MI*	*0.830*
IG(threshold 2)	*0.789*	*Max IG(threshold 2)*	*0.830*
IG(threshold 4)	*0.761*	*Max IG(threshold 4)*	*0.831*

feature 기법을 적용해 본 결과 그 성능이 눈에 띄게 향상되었다. 특히 X^2를 Max X^2로 했을 때 경우 클러스터링 정확률이 79.6%에서 83.8%로 크게 향상되었다. 그 차이를 보다 명확하게 보이기 위해 그래프로 나타내면 〈그림 1〉과 같다.



〈그림 1〉 전통적 자질 선정 기법과 최댓값 자질선정 기법의 클러스터링 성능 비교

4. 결 론

이 연구는 한글 웹 문서 클러스터링을 위해 자질 선정 방법에 따른 성능을 비교하고, 여기에 덧붙여 최댓값 자질선정 기법을 제안하여 실험적으로 적용해 보았다.

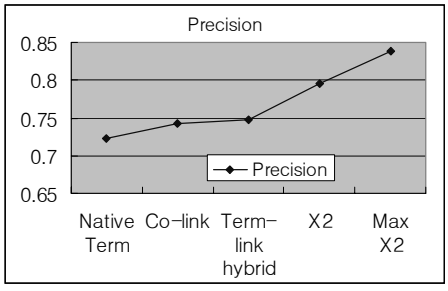
이 연구에는 두 개의 코퍼스가 사용되었는데, 클러스터링을 위한 실험 문서는 Naver의 자연과학 디렉터리에서 1,449개의 문서를, 범주별 자질 선정을 위한 문서는 Yahoo Korea의 같은 디렉터리에서 1,333개의 문서를 각각 수집하였다.

먼저 Naver 문서를 대상으로 단어기반 클러스터링과 링크기반 클러스터링, 그리고 단어-링크 혼합 클러스터링 기법으로 클러스터링해 보았으며, 그 결과를 네이버 디렉터리에 초기 할당된 범주와 비교해 보았다. 그 결과 단어 빈도와 동시링크 빈도를 함께 이용한 방식의 클러스터링 성능이 가장 좋게 나타났다.

다음으로 Yahoo 문서를 대상으로 카이제곱 통계량, 정보획득량, 상호정보량을 통해 각 범주를 대표하는 용어를 선정하였으며, 이를 Naver 문서에 적용해 보았다. 여기에 추가로 각 범주별로 최곳값을 갖는 용어만을 각 범주에 할당하는 최댓값 자질선정 기법을 적용한 실험도 병행하였다. 클러스터링 성능은 학습을 하지 않았을 때와 비교하여 크게 향상되었지만, 자질 선정 방법에 따른 차이는 그다지 크게 나타나지 않았다. 그러나 여기에 최댓값 자질 선정 기법을 적용해 본 결과 그 성능이 눈에 띄게 향상되었다. 클러스터링에 사용된 자질에 따른 성능을 종합적으로 비교해 보면 〈표 5〉, 〈그림 2〉와 같다. 사용된 자질에 따른 한글 웹 문서 클러스터링 정확률은 자연어 72.3%, 동시링크 74.3%, 단어-링크 혼합 74.8%, X^2 79.6%, Max X^2 83.8%로 나타났다.

〈표 5〉 자질별 클러스터링 성능 비교

사용된 자질	정확률
자연어	0.723
동시링크	0.743
단어-링크 혼합	0.748
X2	0.796
Max X2	0.838



〈그림 2〉 자질별 클러스터링 성능 비교

〈표 5〉, 〈그림 2〉에서 보는 것처럼 Max X^2를 자질로 선택했을 때 클러스터링 성능이 가장 높았다. 이상의 연구 결과는 최댓값 자질 선정 기법이 한글 웹 문서를 클러스터링 하기 위한 간단하면서도 효과적인 자질 공간 축소 방법이며, 한글 웹 문서의 클러스터링을 위한 간단하면서도 효과적인 수단임을 보여주고 있다.

참 고 문 헌

고영중, 서정연. 2002. 문서관리를 위한 자동문서범주화에 대한 이론 및 기법. 『정보관리 연구』, 33(2): 19-32.

김영기, 이원희, 권혁철. 2003. 동시링크를 이용한 웹 문서 클러스터링 실험. 『한국도서관·정보학회지』, 34(2): 233-253.

정성원, 이원희, 김영기, 권혁철. 2002. 웹 문서 중 의미 있는 표의 추출. 『한글 및 한국어 정보처리』, 14: 332-339.

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

Barfourosh, A. Abdollahzadeh, M. L. Anderson, H. R. Motahary and D. Perlis, 2003. "Information Retrieval on the World Wide Web and Active Logic : A Survey and Problem Definition". 〈http://citeseer.ist.psu.edu/barfourosh02information.html.〉

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

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

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

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

He, Xiaofeng, Hongyuan Zha, Chris H. Q. Ding and Horst D. Simon, 2002. "Web document clustering using hyperlink structures," *Computational Statistics & Data Analysis*, vol.41: 19-45.

Jansen, M. B., A. Spink, J. Bateman, and T. Saracevic, 1998. "Real Life

Information Retrieval : A Study of User Queries On The Web", *ACM SIGIR Forum Archive* vol 32.

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

Kumar, S. R., P. Raghavan, S. Rajagopalan and A. Tomkins, 1999. "Trawling the Web for emerging cybercommunities", *Proceedings of the 8th WWW Conference*.

Larson, R. R. 1996. "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*.

Lee Kyo-Woon, Kim Young-Gi and Kwon Hyuk-Chul, 2004. "Clustering of web documents with the use of term frequency and co-link in hypertext", *The 3rd Asia Pacific International symposium on Information Technology*, Jan., 2004: 122-127.

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

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

Mukherjea, S, 2000, "Organizing topic-specific Web information", *Proceedings of the 11th ACM Conference on Hypertext:* 133-141.

Mukherjea, S, 2000. "WTMS: a system for collecting and analyzing topic-specific Web information", *Proceedings of the 9th International World Wide Web Conference:* 457-471.

Pirolli, P., P. Schank, M. Hearst and C. Diehl, 1996. "Scatter/ Gather browsing communicates the topic structure of a very large text collection", *Proceedings of the Conference on Human Factors in Computing Systems:* 213-220.

Salton, G. and M. J. McGill, 1983. "Introduction to Modern Information Retrieval", McGrawHill.

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

Smith, Kate A. and Alan Ng, 2003. "Web page clustering using a self-organizing

map of user navigation patterns" *Decision Support systems*, vol.35: 245-256.

Wang, Yitong and Masaru Kitsuregawa, 2001. "Line Based Clustering of Web Search Results", *Second International Conference on Advances in Web - Age Information management (WAIM)*.

Yang, Yiming and Jan O. Pederson, 1997. "A comparative study on feature selection in text categorization", *Proceeding of ICML-97, 14th International Conference on Machine Learning*.

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