웹 문서의 단어정보와 링크정보 결합을 이용한 클러스터링 기법*

부산대학교 전자계산학과,[1] 울산과학대학 컴퓨터정보학부[2]
이원희[1†] · 이교운[2] · 박 흠[1] · 김영기[1] · 권혁철[1]

Clustering Method Using the Union Information of Term Frequency and Link in Hypertext

Won-Hee Lee,[1] Kyo-Woon Lee,[2] Heum Park,[1] Young-Ki Kim,[1] Hyuck-Chul Kwon[1]
Department of Computer Science0,[1] Busan National University, Busan,
Department of Computer Informtion,[2] Ulsan College, Ulsan, Korea

요 약

최근의 웹 문서는 텍스트 위주의 구성이 아닌 이미지, 사운드, 동영상 등의 다양한 타입으로 구성되는 추세이다. 이에 따라 단순히 웹 문서 내의 단어 정보추출 만으로는 좋은 성능의 클러스터링을 기대하기 어렵다. 본 논문은 전통적인 문서 클러스터링 기법인 단어기반 클러스터링 기법의 취약점을 제시하고, 웹 문서 간의 링크구조정보 중 동시인용 정보를 이용하여 웹 문서 클러스터링 성능향상의 가능성을 보이고자 한다. 실험에서는 네이버 디렉토리 중 '자연과학' 범주에 포함된 문서를 대상으로 위의 두 가지 방식과 이 두 가지를 혼합한 단어-링크 혼합 클러스터링을 통해 기존의 방식보다 더 낳은 성능을 얻을 수 있었다.

서 론

인터넷 정보자원의 급격한 증대로 이를 효과적으로 조직하여, 검색성능을 향상시키기 위한 다양한 연구와 실험이 수행되고 있다. 특히 클러스터링은 정보자원의 효과적인 탐색과 이용, 그리고 검색 성능을 향상시킬 수 있는 수단 중의 하나이다.

정보검색 분야에서 클러스터링은 매우 다양한 방식으로 연구되고 있는데, 크게 단어 유사도에 기반한 단어기반 클러스터링(term-based clustering)과 웹 문서 간의 링크구조에 기반한 링크기반 클러스터링(link-based clustering), 그리고 이 둘을 혼합한 혼합형 클러스터링으로 나눌 수 있다.

이 중에서도 가장 활발히 연구되고 있는 클러스터링 기법은 단어기반 클러스터링이다. 이 방법은 웹 문서에 등장하는 단어를 추출하여 단어 벡터를 만들고, 이 벡터를 이용하여 관련문서를 군집화하는 기법이다. 최근에는 온라인 시소러스나 단어의 의미적 관련성을 활용하는 시도도 이루어지고 있다.

하지만, 단순히 문서 내의 단어정보만을 이용해서는 좋은 성능의 클러스터링을 기대하기 어렵다. 그 이유는 최근의 웹 문서가 텍스트 위주의 구성이 아닌 이미지, 사운드, 동영상 등의 다양한 타입으로 구성되는 추세이기 때문이다. 심지어 웹브라우저를 통해 보이는 텍스트가 이미지로 처리되는 경우도 많이 존재한다. 실제 웹 문서의 텍스트를 기계적인 파싱(parsing)을 통해 추출해 보면 텍스트가 문서 내에 아예 존재하지 않거나 매우 적은 수의 텍스트로 이루어진 문서도 상당히 많이 나타난다.

특히 웹 사이트의 대문페이지(door-way-page)의 경우 이러한 성향이 강하게 나타나며, 문서에서 일정량 이상의 단어 추출이 어려운 이러한 문서들은 단어기반 클러스터링 방법만으로는 좋은 성능의 문서분류가 어렵다. 따라서 하이퍼링크와 같은 외부 정보원의 부가적 이용을 고려해 볼 수 있다.

웹 문서의 링크정보는 문서의 "중요성"이나 "연관성"의 판단을 위한 중요한 정보이며, 웹 문서의 링크 정보를 이용하여 정보검색의 성능을 향상시키려는 많은 연구들이 진행되고 있다.

*이 논문은 과학 기술부(한국과학기술기획평가원)의 국가지정연구실 지원으로 이루어진 것임.
†E-mail : whlee@pusan.ac.kr

본 논문에서는 웹 문서에서 추출되는 단어의 수가 단어기반 클러스터링의 성능에 많은 영향을 미친다는 전제하에, 웹 문서에 포함된 단어 수와 클러스터링 성능과의 관계를 밝힌 다음, 이 부분을 웹 문서의 동시인용 빈도를 통해 보완할 수 있는 알고리즘을 제시한다.

관련연구

정보검색에 있어서 클러스터링은 다양한 방법으로 연구되어 왔다. 클러스터링은 주제별 커뮤니티의 확인(Kumar SR et al. 2000)이나 웹 구조의 추출(Larson RR, 1996), 그리고 중복 문서의 발견(Broder AZ et al. 1997), 적합성 피드백을 통한 질의확장(Chang CH & Hsu CC, 1998) 등의 수단으로 연구되어 왔다.

전통적인 문서 클러스터링 기법이라 할 수 있는 단어기반 클러스터링은 각 문서에 포함된 단어의 유사도에 근거하여 전체 문서를 분류하는 기법이다.

링크기반 클러스터링은 특정 문서를 링크하고 있는 두 문서는 연관성이 존재한다는 아이디어를 다양한 링크 구조에 적용하였다. 특히 링크 구조 중 동시인용(Co-citation) (Below RK, 2000)과 서지결합(coupling)은 두 개의 문서 간 유사도 판단에 가장 많이 이용되는 기법이다. 동시인용은 두 문서를 동시에 인용(out-link)하는 문서 수를 클러스터링에 적용하는 방법이고, 서지결합은 두 문서를 동시에 연결하는(in-link) 문서 수를 클러스터링에 적용하는 방법이다.

실험데이터 선정 및 실험방법

1. 실험데이터 선정

실험 대상으로 검색 포탈 사이트인 네이버 디렉토리(http://dir.naver.com/)의 상위 14개 범주 중 '학문과학' 분야의 하위 영역인 '자연과학'을 선택하였다. 네이버 디렉토리의 문서를 선정한 이유는 적절한 분류기준에 따라 수작업으로 분류된 문서와 실험에 의해 분류된 문서를 비교함으로써 문서분류 성능을 쉽게 구할 수 있기 때문이다. 그리고 '자연과학' 분야의 경우 웹 문서가 풍부하고 범주 내의 하위 범주 구분이 명확하기 때문에 실험 대상으로 적합하다고 판단하였다. '자연과학'은 다시 16개의 하위 범주로 나누어지는데 이 중에 다른 범주와 겹치지 않는 11개의 범주가 최종 실험대상이 되었다. 그 이유는 "자연과학/지구과학/해양학/해양법"에 속하는 문서가 "사회,문화/법/해양법"의 문서에도 분류되어 있기 때문에 이러한 문서는 실험대상에서는 제외하기로 하였다.

최종 분석대상으로 선정된 범주와 대상 문서 수는 Table 1과 같다.

2. 단어기반 클러스터링(Term-based clustering)

실험 대상이 되는 1,449개의 문서는 부산대학교 한국어정보처리연구실의 색인기를 이용하여 전체 문서의 단어 벡터를 작성하였다. 그리고 클러스터링 작업과 클러스터링 성능 분석은 미네소타 대학 컴퓨터 과학과(University of Minnesota, Department of Computer Science)에서 개발한 클러스터링 툴킷(Clustering Toolkit)인 Cluto2.1을 사용하였다.

Table 1. 범주별 분석대상 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

Table 2. 유사도 평가함수

Criterion Function	Optimization Function
I1	Maximize $\sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{v,u\in s_i} sim(v,u)\right)$
I2	Maximize $\sum_{i=1}^{k}\left(\sum_{v,u\in s_i} sim(v,u)\right)$
E1	Minimize $\sum_{i=1}^{k}\frac{\sum_{v\in Si,u\in S} sim(v,u)}{\sum_{v,u\in Si} sim(v,u)}$
G1	Minimize $\sum_{i=1}^{k} n_i^2\frac{\sum_{v\in Si,u\in S} sim(v,u)}{\sum_{v,u\in Si} sim(v,u)}$
G1P	Minimize $\sum_{i=1}^{k} n_i\frac{\sum_{v\in Si,u\in S} sim(v,u)}{\sqrt{\sum_{v,u\in Si} sim(v,u)}}$
H1	Maximize $\frac{I1}{E1}$
H2	Maximize $\frac{I2}{E1}$

문서 간 유사도의 평가는 다음 Table 2의 일곱 가지 함수를 사용하였다. Table 2에서 k는 전체 클러스터의 수이며, S는 클러스터 된 전체 문서 수, Si는 i번째 클러스터에 할당된 문서의 집합이다. ni는 i번째 클러스터에 할당된 문서수, v,u는 문서, $sim(v,u)$는 두 문서 간의 유사도 함수이다.

실험 대상문서의 벡터를 이용하여 클러스터링을 한 다음, 그 결과를 네이버 디렉터리에 초기 할당된 범주와 비교하였으며, 그 일치 여부를 클러스터링 성능으로 간주하였다.

3. 링크기반 클러스터링(Link-based clustering)

웹 문서는 링크를 통해 서로 연결되어 있는데, 링크 기반 클러스터링은 문서 간의 링크구조를 이용하여 관련문서를 군집화시키는 것으로, 하이퍼링크가 두 문서 간에 의미적 연관성이 있을 것이라는 것을 전제로 하고 있다. 이러한 연관성은 패스(path)의 길이에 반비례하며, 링크 수에 비례한다고 볼 수 있다.

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 hubness가 높으며, 들어오는 링크가 많은 사이트는 authority가 높다고 한다(오효정 등, 1999).

해당문서의 들어오는 링크(in-link) 수는 AltaVista와 Google, Alltheweb 같은 검색 시스템의 "link : url"검색옵션을 통해, 나가는 링크(out-link)의 경우 웹 문서의 링크정보 파싱을 통해 그 수를 구할 수 있다.

본 논문의 실험에서는 링크기반 클러스터링에 적용할 수 있는 동시인용과 서지결합 중 동시인용을 적용해 보았다. 특정 문서에 들어오는 링크(in-link)는 많은 문서의 링크정보를 반영할 때 좋은 클러스터링 성능을 기대할 수 있어, 이에 따라 충분히 많은 문서를 보유한 검색시스템에서 동시링크 정보를 이용하기로 하였다.

실험 대상인 1,449개의 웹 문서를 대상으로 동시에 링크하고 있는 문서의 수를 조사하기 위해 Alltheweb(http://www.alltheweb.co.kr)의 "AdvancedSearch"옵션을 이용하였다. 사용된 검색 식은 다음과 같다.

LINK : url_A AND LINK : url_B

이러한 방법으로 웹 문서의 쌍을 동시에 링크하고 있는 웹 문서들의 수를 구하여 동시인용빈도 행렬을 작성하였다. 이 행렬은 1,449×1,449 크기의 동시인용빈도로 나타난다. 시간의 절약을 위해 하삼각 행렬의 빈도를 먼저 구하고 상삼각의 행렬위치에 빈도를 채워 넣었다. 자동으로 검색 식을 생성하고, 생성된 검색 식을 이용하여 동시인용빈도를 구하기 위해 부산대학교 한국어 정보처리 연구실의 웹 로봇을 이용하였다.

웹 문서들 간의 관련성의 정도, 즉 상대적인 유사성과 비유사성을 나타내기 위해 동시인용 빈도는 새로운 형태로 변형될 필요가 있다. 일반적으로 단어 출현빈도를 통한 두 문헌 간의 관계를 나타내기 위해서는 TFIDF(TF×IDF)를 이용한 단어벡터(word vector)와 문서의 거리비교(distance comparison)가 사용된다. 또한, TFIDF를 변형하여 CCIDF (Common Citation×Inverse Document Frequency) 알고리즘을 고려해 볼 수 있다. 이 실험에서는 CCIDF를 이용하여 웹 문서들 간의 상대적인 유사도와 비유사도를 구했으며, 실험 환경은 단어기반 클러스터링과 마찬가지로 Cluto 시스템을 사용하였다.

4. 혼합 클러스터링(Hybrid clustering)

본 논문의 실험에서는 단어기반 클러스터링 기법으로 클러스터링이 잘 되지 않는 문서들의 공통된 특성을 분석한 다음, 이런 문서들에 한해서 링크기반 클러스터링을 하는 방법이 있을 것이다. 단어기반 클러스터링에서 각 문서의 색인어를 자질(feature)로 한 것에 문서 간 동시인용 빈도를 또 하나의 자질로 추가하는 방법을 사용하였다.

다시 말해서 웹 문서 색인 과정을 통해 만들어진 단어 벡터와 웹 문서 간 링크구조를 이용하여 만들어진 동시링크 빈도 행렬을 하나의 행렬로 합침으로써 서로 다른 2가지 형태의 문서 자질을 하나의 평가함수에 쉽게 적용될 수 있다는 장점이 있다.

Fig. 1은 단어-링크 혼합 클러스터링에 사용되는 벡터구조이다.

실험결과

1. 단어기반 클러스터링

분석대상 문서는 모두 1,449개이며, 문서 당 평균 포함된 단어 수는 202.6개, 파싱을 통한 전체 색인어 수는 17,223

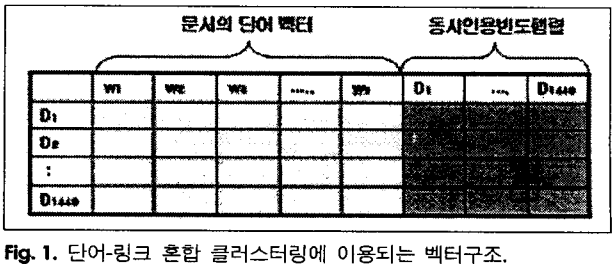

Fig. 1. 단어-링크 혼합 클러스터링에 이용되는 벡터구조.

개이다. 이 실험에서 사용된 벡터의 크기는 1,449×17,223 이다.

크게 상향식 클러스터링과 하향식 클러스터링 기법에 앞에서 제시한 일곱 가지 평가함수(criterion function)를 적용하였으며, 유사도 측정 함수는 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 엔트로피(entropy)와 순정도(purity)로 비교해 보면 다음 Table과 같다.

클러스터링은 엔트로피가 낮을수록, 순정도가 높을수록 그 성능이 높다고 볼 수 있다. 따라서 이러한 결과들을 종합해 보면 본 연구의 실험 대상 문서의 단어기반 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 대체로 I2, H2 등의 경우에 성능이 높은 것으로 나타났다. 가장 클러스터링 성능이 좋은 것은 문서 유사도에 코사인, 단어 유사도에 상관계수를 사용한 하향식 클러스터링으로 그 중에서도 I2의 경우에 가장 높은 성능을 보였다.

따라서 다음의 결과는 대상 웹 문서에 클러스터링 방식으로 하향식을, 문서 유사도에는 코사인을, 단어 유사도에는 상관계수를, 평가함수로는 I2를 사용하여 클러스터링하였다.

Table 3. 단어기반 하향식 클러스터링 성능(문서 유사도 : 코사인, 컬럼(단어) 유사도 : idf)

평가함수	Entropy	Purity
I1	0.501	0.586
I2	0.363	0.708
E1	0.356	0.699
G1	0.426	0.675
G1P	0.367	0.705
H1	0.394	0.676
H2	0.382	0.684

Table 4. 단어기반 하향식 클러스터링 성능(문서유사도 : 코사인, 컬럼(단어) 유사도 : 상관계수)

평가함수	Entropy	Purity
I1	0.457	0.604
I2	0.352	0.723
E1	0.374	0.712
G1	0.374	0.705
G1P	0.398	0.652
H1	0.427	0.619
H2	0.397	0.666

그 결과는 1,449개의 문서 중에서 856개의 문서가 네이버의 분류기준과 일치하였다.

Fig. 2는 클러스터 6의 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 군집화되는 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. 출현 단어의 수와 단어기반 클러스터링 성능

웹 문서에 포함된 단어 수가 매우 적거나 이미지만으로 구성되어 있는 경우는 단어기반 클러스터링 자체가 불가능하다. 이번 실험의 경우에도 실험 대상문서 1,449개 중에서 71개(약 5%) 문서가 단어개수가 0으로 나타났으며, 단어 수 다섯 개 이하의 문서가 355개로 전체 실험 대상의 25%

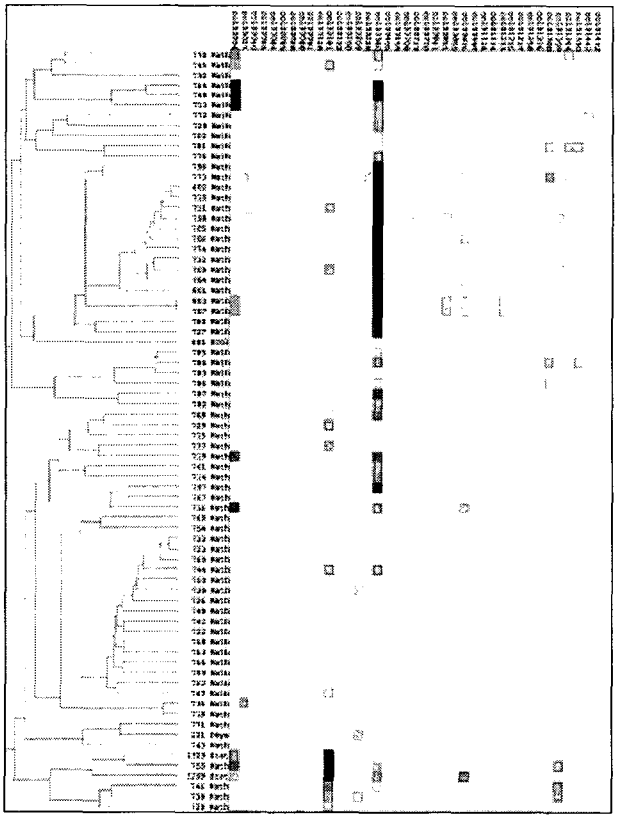

Fig. 2. 단어기반 클러스터링 결과 중 클러스터 6.

에 달하였고, 이 문서들의 평균 클러스터링 성능은 24.2%를 나타내었다. 링크 정보를 이용하여 이러한 문서들의 클러스터링 성능을 높인다면 전체 클러스터링 성능의 향상을 기대할 수 있다.

Table 5는 실험을 통한 웹 문서에 등장하는 단어 수와 클러스터링 성능과의 관계를 보여주고 있다.

위의 Table에 따르면 단어 수에 대한 클러스터링은 단어 수가 특정한 개수까지는 성능이 향상되나 어느 순간부터는 단어 수에 관계없이 비슷한 성능을 보여주고 있다. 특히 전체 실험대상 문서 1,499개의 약 1/4에 달하는 355개의 문서가 다섯 개 이하의 단어를 포함하고 있으며, 이들 중 86개 문서만이 클러스터링 전의 범주와 일치하게 클러스터링 되어 약 24.2%의 클러스터링 성공률을 보여줌으로써, 전체 클러스터링 성능에 결정적인 영향을 미치고 있다. 따라서 '단어개수 5' 이하인 문서에 대해서는 별도의 클러스터링 알고리즘의 개발이 요구되며, 이 연구에서는 두 문서를 동시에 링크하는 웹 문서의 수를 두 문서에 동시에 등장하는 단어 수와 같은 비중의 자질로 간주하는 방법을 고려하게 된 것이다.

Table 5. 웹 문서 내 단어 수와 클러스터링 성능과의 관계

단어 개수	문서 수	성공	실패	성공률(%)
0	71	0	71	0.0
1	39	3	36	7.7
2	77	23	54	29.9
3	51	17	34	33.3
4	75	33	42	44.0
5	42	10	32	23.8
6	73	48	25	65.8
7	8	4	4	50.0
8	59	29	30	49.2
9	55	20	35	36.4
10	58	32	26	55.2
11	5	3	2	60.0
12	27	17	10	63.0
13- 14	15	9	6	60.0
15- 16	32	20	12	62.5
17- 18	20	14	6	70.0
19- 20	15	8	7	53.3
21- 22	14	12	3	85.7
23- 24	11	10	1	90.9
25- 26	21	16	5	76.2
27- 28	16	11	5	68.8
29- 30	13	13	0	100.0
31- 32	10	4	6	40.0
33- 34	7	5	2	71.4
35- 36	15	10	5	66.7
37- 38	9	7	2	77.8
39- 40	16	7	9	43.8
41- 42	9	5	4	55.8
43- 44	12	9	3	75.0
45- 46	11	6	5	54.5
47- 49	9	4	5	44.4
50- 99	143	116	27	81.1
100-149	125	99	26	79.2
150-199	82	67	15	81.7
200-299	78	68	10	87.2
300-399	37	82	5	86.5
400-499	22	17	5	77.3
500-599	22	16	6	72.7
600 이상	45	32	13	71.1
계	1,449	856	593	59.1

3. 링크기반 클러스터링

분류대상이 되는 1,449개의 웹 문서를 클러스터링하기 위해서는 먼저 1,449×1,449 크기의 동시링크빈도 행렬을 작성하였다. 이를 단어기반 클러스터링과 같은 여러 가지 방법으로 클러스터링해 본 결과는 아래의 Table과 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.

이전 실험의 단어기반 클러스터링과 링크기반 클러스터링 성능을 단순 비교해 보면, 각각 엔트로피와 순정도가 0.352,

Table 6. 링크기반 하향식 클러스터링 성능(문서유사도 : 코사인, 칼럼(링크 수) 유사도 : idf)

평가함수	Entropy	Purity
I1	0.319	0.695
I2	0.263	0.743
E1	0.278	0.725
G1	0.265	0.755
G1P	0.297	0.707
H1	0.280	0.721
H2	0.269	0.741

Table 7. 링크기반 하향식 클러스터링 성능(문서유사도 : 코사인, 칼럼(링크 수) 유사도 : 상관계수)

평가함수	Entropy	Purity
I1	0.338	0.667
I2	0.266	0.744
E1	0.296	0.711
G1	0.290	0.738
G1P	0.313	0.692
H1	0.301	0.713
H2	0.297	0.711

0.723과 0.263, 0.743으로, 링크기반 클러스터링의 성능이 더 나은 것으로 나타났다. 그 결정적인 요인이 된 것은 단어기반 클러스터링으로는 좋은 성능을 기대하기 어려운 단어 개수 5개 미만' 인 문서도 링크기반 클러스터링에서는 단어 수와 관계없이 클러스터링이 가능했기 때문이다.

그러나 링크기반 클러스터링을 통해서도 다른 문서와의 '동시링크 개수가 0' 인 문서 93개를 비롯한 동시링크 개수가 매우 적은 문서에 대한 클러스터링 누락과 실패에 대한 문제가 여전히 남게 된다. 이러한 문제는 링크기반 클러스터링에서도 성능저하의 원인이 된다.

따라서 웹 문서에서 추출할 수 있는 단어 수가 일정한 개수 이하인 문서에만 동시링크 빈도를 적용하거나, 단어출현 빈도에 동시링크 빈도를 단어출현 빈도와 같은 자질로 추가한다면 더 좋은 성능을 기대할 수가 있다.

Fig. 3은 단어기반 클러스터링과 링크기반의 클러스터링 성능곡선을 비교한 것이다.

단어기반 클러스터링은 이전의 Table 5의 성능에 따라서 나타내었고, 링크 기반 클러스터링은 문서 내부의 단어 수와 관계없이 전체 평균 성능으로 나타내었다.

문서 내의 출현 단어 빈도가 10개 미만인 문서에 링크기반의 클러스터링을 적용할 경우 성능 향상을 기대할 수 있을 것이다.

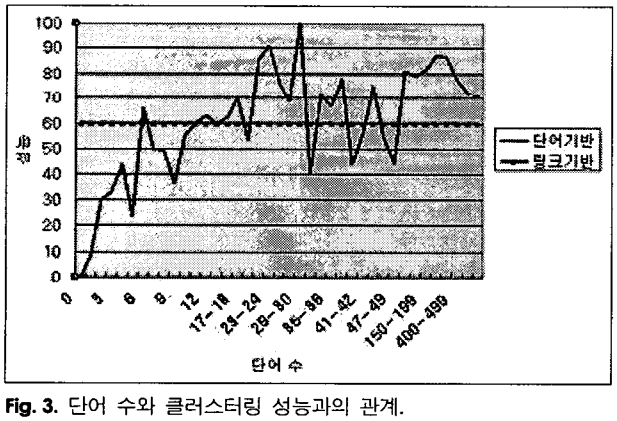

Fig. 3. 단어 수와 클러스터링 성능과의 관계.

Table 8. 단어-링크 혼합 클러스터링 성능(문서 유사도 : 코사인, 칼럼 유사도 : idf)

평가함수	Entropy	Purity
I1	0.400	0.599
I2	0.260	0.748
E1	0.293	0.702
G1	0.277	0.727
G1P	0.301	0.707
H1	0.298	0.709
H2	0.293	0.704

4. 단어-링크 혼합 클러스터링

전체문서의 단어 벡터에 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 클러스터링해 본 결과는 다음과 같다.

단어-링크 혼합 클러스터링의 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. 단어 링크 혼합 클러스터링 또한 사용된 평가함수 중에서도 대체로 I2의 경우에 성능이 높은 것으로 나타났다.

전체적으로 1,449개의 문서 중에서 1,137개의 문서가 제대로 클러스터링 되었으며, 제대로 클러스터링 되지 않은 312개의 문서에는 '단어개수와 동시링크 개수가 모두 0' 인 문서 42개가 포함되어 있다. 클러스터별, 주제별 클러스터링 성능은 앞의 링크기반 클러스터링 성능과 거의 유사한 패턴을 보여주고 있다. 다만, 링크기반 클러스터링보다 단어-링크 혼합 클러스터링의 성능이 좀 더 향상되었음을 알 수 있다. 문서의 링크정보와 단어정보를 혼합적용해도 클러스터링 성능을 저하시키는 문제는 여전히 존재하는 것을 확인할 수 있었다.

결 론

웹 문서 클러스터링의 경우 단어기반 클러스터링 기법이 일반적이지만, 웹 문서에 텍스트가 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 클러스터링 기법으로 클러스터링해 보았으며, 그 결과를 네이버 디렉터리

Table 9. 단어-링크 혼합 클러스터링 성능(문서 유사도 : 코사인, 칼럼 유사도 : 상관계수)

평가함수	Entropy	Purity
I1	0.392	0.616
I2	0.271	0.745
E1	0.309	0.687
G1	0.283	0.747
G1P	0.318	0.693
H1	0.318	0.682
H2	0.299	0.701

Table 10. 클러스터링 방식별 성능비교

	단어 기반	링크 기반	단어-링크 혼합
평균 정확률	59.1%	60.2%	73.6%
단어/링크/혼합 0인 문헌 수	71	93	42
Entropy	0.352	0.263	0.254
Purity	0.723	0.743	0.748

에 초기 할당된 범주와 비교해 보았다. 그 결과 단어빈도와 동시링크 빈도를 함께 이용한 방식의 클러스터링이 가장 높은 성능을 보였다.

지금까지의 실험결과를 종합하여 전체 실험대상 문서 1,449개에 대한 각 클러스터링 방식별 성능을 비교하였다.

Table 10에서 보는 바와 같이 제시된 방법 중 클러스터링 성능을 가장 크게 향상시킬 수 있는 것은 단어-링크 정보를 혼합한 것이 가장 좋은 결과를 보였다.

하지만 실험대상 문서가 특정 범주의 하위 범주 하나만을 대상으로 실험이 이루어졌다는 점은 이 실험의 제한점으로 볼 수 있다. 그리고 단순히 문서 간의 동시인용 빈도만을 적용하여 성능향상을 보였지만 또 다른 링크 구조인 서지결합 정보도 이용을 하였다면 좀더 좋은 성능을 기대할 수도 있을 것이다.

향후 단어-링크 혼합 클러스터링의 성능 향상을 위해서는 단어와 문서, 동시링크, 클러스터 등의 유사도 계산에 관련된 더욱 정밀한 알고리즘의 개발과 실제 웹 문서에 대한 적용 등이 필요할 것이다.

REFERENCES

Kumar SR, Raghavan P, Rajagopalan S, Tomkins A (2000) : *"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, pp133-141. Mukherjea S, "WTMS : a system for collecting and analyzing topic-specific Web information", Proceedings of the 9th International World Wide Web Conference, pp457-471*

Larson RR (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, 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, pp213-220*

Broder AZ, Glassman SC, Manasse MS, Zweig G (1997) : *"Syntactic clustering of the Web", Proceedings of the 6th International WWW Conference, pp391-404*

Chang CH, Hsu CC (1998) : *"Integrating query expansion and conceptual relevance feedback for personalized Web information retrieval", Proceedings of the 7th International WWW Conference*

Belew RK (2000) : *Finding Out About : A Cognitive perspective on search engine technology and the WWW. Cambridge University Press, p196*

오효정, 임정묵, 이만호, 맹성현(1999) : "점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 모델", 제11회 한글 및 한국어 정보처리 학술대회, *pp89-96*

정상화, 이종혁(1998) : "문서구조 정보에 기반한 웹 페이지 범주화 모델", 제10회 한글 및 한국어 정보처리 학술대회, *pp91-96*

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, pp265-269*

Ying Z, George K (2001) : *"Criterion functions for document clustering-experiment and analysis", Technical Report TR #01-40, Department of Computer Science, University of Minnesota*

Ying Z, George K (2002) : *"Evaluation of hierarchical clustering algorithms for document datasets", Technical Report TR #02-22, Department of Computer Science, University of Minnesota*

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