링크기반 클러스터링을 이용한 웹 문서 검색의 성능 향상에 관한 실험적 연구

An Experimental Study on Enhancing the Retrieval Performance for the Web Documents Using Link-Based Clustering Technique

김혜진, 문성빈, 연세대학교 문헌정보학과

Hea-Jin Kim, Sung-Been Moon
Dept. of Library and Information Science, Yonsei University

본 연구에서는 하이퍼텍스트나 웹 문서의 검색에서 링크로 연결된 문서들이 주제적으로 서로 관련되어 있다는 것을 기반으로 하여 링크정보를 참조한 웹 문서 클러스터링 기법을 제안하였고 이것을 이용하여 검색된 결과를 질의 근접 순위화함으로써 웹 문서 검색의 성능을 향상시키는 방안을 연구하였다. 본 연구에서 사용된 웹 문서 집단은 웹(WWW)을 통하여 직접 수집하였으며 웹 문서가 다른 웹 문서를 링크하고 있을 때를 OutLink, 다른 웹 문서로부터 링크를 받고 있을 때를 InLink로 구분하였다. 실험결과 OutLink를 참조하여 클러스터링을 수행하는 기법과 InLink를 참조하여 클러스터링을 수행하는 기법 모두 검색성능을 향상시켰다.

1 서 론

웹 문서는 일반 문서와는 달리 구조정보를 포함하고 있는데 이것은 HTML 태그와 하이퍼링크로 표현된다. 따라서 웹 문서의 검색과 관련된 연구들은 HTML 태그를 식별하여 추출된 정보와 링크정보를 식별하여 링크로 연결된 문서를 검색결과에 포함시키는 기법을 제시하고 있으며 이것은 링크로 연결된 이웃 문서들은 서로 주제적으로 관련되어 있다는 가설을 뒷받침한다.

본 연구에서는 하이퍼텍스트나 웹 문서의 검색에서 링크로 연결된 문서들이 주제적으로 서로 관련되어 있다는 것을 기반으로 하여 링크정보를 참조한 웹 문서 클러스터링 기법을 제안하였다. 본 연구의 목적은 검색된 웹 문서가 가지고 있는 구조정보 중 <A HREF> 태그로 표현되는 링크정보를 참조하여 링크로 연결된 문서와 공통으로 출현하는 용어에 가중치를 강화하고 클러스터링을 수행한 뒤 순위화하여

검색의 효율성을 향상시키는 것이다. 즉, 주어진 질의를 통하여 검색된 웹 문서와 InLink, OutLink로 연결된 문서에 공통으로 발견되는 용어를 웹 문서를 표현하는 보다 가치있는 정보라 판단하고 공통 용어 벡터의 가중치를 강화하여 보다 잘 된 클러스터를 형성시킨다. 그리고 그 결과를 질의 근접 순위화 함으로써 검색의 효율성을 향상시키는 기법을 제안하였다.

2 문헌 클러스터링

문헌 클러스터링은 자동분류의 한 종류로써 문헌들이 가지고 있는 고유한 자질들을 이용하여 문헌들을 집단화하는 기법이다. 자동분류가 외부 질의나 주제어, 이용자 프로파일 등에 대한 적합성 또는 유사도에 따라 문헌들을 분류하는 반면 클러스터링에 의한 분류는 문헌 집단 전체가 가지고 있는 내재적인 자질들을 이용하여 문헌들을 고유한 집단들로 분리한다 (Greengrass 2000).

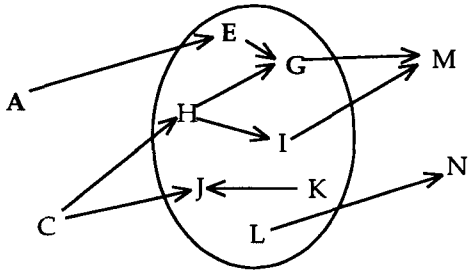
문헌 클러스터링은 문헌 집단 전체의 고유한 구조를 알게 해주며 계층적 클러스터링을 수행하였을 경우 문헌 집단은 포괄적인 범주를 가진 상위 수준의 군집들과 좀더 주제적으로 유사하고 응집력있는 하위 수준의 군집들로 구성된다. 따라서 문헌 클러스터링은 상위 수준의 군집들을 선택하여 문헌 집단이 가지고 있는 주제들을 점차 자세하게 탐색하여 정보요구에 적합한 하위 수준의 군집을 선택하거나 정보요구(질의)와 주제적으로 유사한 하위 수준의 군집을 선택함으로써 검색의 효율성을 높이는데 이용된다.

3 링크기반 클러스터링 실험

3.1 실험 설계

3.1.1 개념 정의

본 연구에서 사용된 'OutLink'라는 용어는 실험문서가 링크하고 있는 검색되지 않은 웹 문서를 뜻하며 'InLink'라는 용어는 실험문서를 링크하고 있는 검색되지 않은 웹 문서이다. <그림 1>은 Modha와 Spangler(2000)가 제시한 하이퍼텍스트의 개념지도인데 지도에서 E와 InLink로 연결된 문서는 A이고 OutLink로 연결된 문서는 G이다.



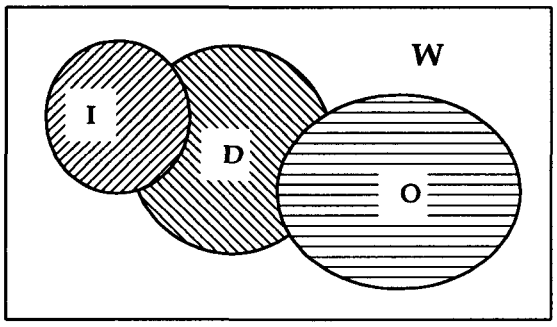
<그림 1> 하이퍼텍스트의 개념지도

3.1.2 웹 문서 집단

웹 문서 집단의 구축은 다음과 같은 2단계로 이루어졌다.

첫 번째 단계인 '실험문서 수집단계'에서는 MSN(www.msn.com) 디렉토리에서 제공하는 주제어들 중 17개를 추출하여 웹 문서를 검색하여 리스트를 작성하였다. 두 번째 단계인 '링크 문서 수집단계'에서는 첫 번째 단계에서 수집된 450개의 실험문서의 OutLink와 InLink를 파악하여 리스트를 작성하였다. 즉, OutLink는 링크를 나타내는 <A HREF> 태그의 정보를 이용하였고, InLink는 알타비스타(www.altavista.com)에서 제공하고 있는 링크검색(link search)을 이용하여 실험문서들을 링크하고 있는 웹 문서들의 주소를 파악하였다. 이렇게 작성된 실험문서, OutLink 문서, InLink 문서들을 웹집(Webzip)을 이용하여 수집하였다. <그림 2>는 본 연구에서 구축한 웹 문서 집단의 모형도

인데 D는 검색과 클러스터링 대상이 되는 '실험문서'이며 O, I는 실험문서 D와 OutLink, InLink로 연결된 링크 문서 집단이다.



<그림 2> 웹 문서 집단의 모형도

<표 1> 웹 문서 집단의 링크정보 통계치

구분		건 수
실험문서	MSN 수집 결과	450
실험문서에 대한 링크정보	InLink 건수	2,878
	OutLink 건수	7,710
	평균 InLink 건수	6
	평균 OutLink 건수	17
	InLink 최대 값	86
	OutLink 최대 값	163
총계	수집결과(중복문서 제거)	9,932

웹 문서 집단을 고려하여 10개의 새로운 질의를 작성하고 질의 당 상위 50위에서 절단하여 수작업으로 적합성 판정을 하였다. 질의 당 평균 적합문서 수는 15.2건이었다.

3.1.3 역문헌빈도

본 연구에서는 웹 환경에서 발생할 수 있는 상황을 고려하여 '전역적 역문헌빈도(Global_idf)'와 '지역적 역문헌빈도(Local_idf)'라는 개념을 새로이 정의하고 이것을 웹 문서의 용어 벡터를 조정하는 중요한 가중치의 하나로 제안하였다. 전자의 역문헌빈도는 일반적인 검색실험에서 사용되고 있는 실험문서집단 전체의 용어분포를 반영하고 있는 문헌빈도이고, 후자의 역문헌빈도는 검색된 결과의 문서 수내의 용어 분포를 반영하고 있는 문헌빈도이다. 이렇게 역문헌빈도를 전역적과 지역적으로 나눈 이유는 일반적인 검색 실험 환경에서와 달리 웹 환경에서는 검색된 웹 문서들이 가지고 있는 용어 정보로는 용어의 가중치로써 사실상 전역적 역문헌빈도를 적용하는 것이 불가능하기 때문이다.

따라서 본 연구에서 웹 문서 집단에 적용할 '전역(global)'과 '지역(local)'을 나타내는 범위의 정의는 주제어를 가지고 웹 탐색엔진을 통하여 검색된 결과인 실험문서 450건을 '지역'으로, 실험문서로부터 발생된 InLink, OutLink 문서를 모두 합친 9,932건을 '전역'으로 정의하고 각각의 역문헌빈도를 산출하여 단어빈도와 함께 역문헌빈도를 용어 가중치로 주었다(TF · IDF).

실험에서 사용한 단순 단어빈도(TF)와 공식은 <표 2>과 같다.

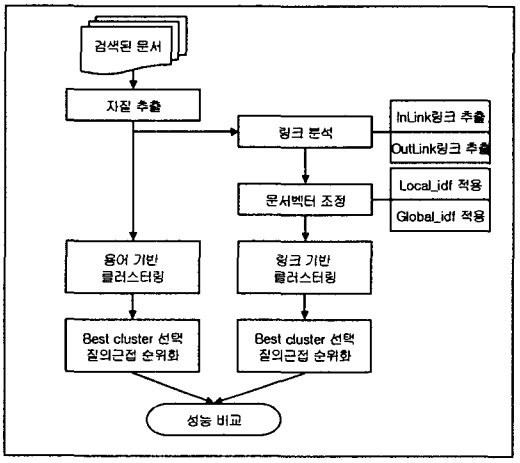
<표 2> 단어빈도 가중치 공식

이름	공식	기호
이진 TF	1,0	btf
단순 TF	tf	ttf
로그 TF	$1+\log(tf)$	ltf
루트 TF	$\sqrt{tf}$	rtf
보정 TF	$0.5+0.5\times\frac{tf}{\max tf}$	5tf

3.1.4 실험내용

실험은 검색된 웹 문서의 용어들만을 가지고 계층적 완전연결 클러스터링을 수행한 결과를 베이스라인으로 평가하고 웹 문서와 InLink, OutLink로 연결된 문서들을 참조하여 공통으로 출현하는 용어의 가중치를 강화한 후

클러스터링을 수행한 결과와 검색 성능을 비교하였다. <그림 3>은 링크기반 클러스터링 실험의 전체적인 설계도이다.



<그림 3> 링크기반 클러스터링 실험 설계도

3.2 성능평가 방법

웹 문서의 검색은 질의에 대한 적합문서의 총수를 알 수가 없으므로 재현율에 대한 평가는 불가능하다. 따라서 이용자들은 상위 n개의 문서의 정확률에 따라서 검색의 만족도가 달라질 것이다. 이것은 일반적인 정보검색 실험과는 매우 다른 환경을 의미한다. 일반적으로 대부분의 정보검색 실험에서는 효율성을 나타내는 지표로써 성능곡선을 사용하는데 이것은 각 순위에서의 재현율과 정확률의 값을 산출하여 그 쌍을 그래프로 나타낸 것이다. 보통 여러 개의 질의에 대해 재현율-정확률을 구한 후 같은 순위에서의 평균을 내어 곡선을 그린다(정영미 1993, p.302).

본 연구의 웹 문서 집단은 웹(WWW)의 실제 데이터의 모형이므로 평균 정확률만을 구하여 실험의 효율성을 평가하기로 결정하였다. 본 연구에서 사용한 방법은 재현율-정확률의 성능곡선을 응용한 방법으로 재현율-정확률의 쌍으로 성능곡선을 만드는 것이 아니라 검색순위-정확률로 성능곡선을 만드는 것이다. 본 연구에서 사용한 정확률 공식은 다음과 같다.

$$p = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{검색된 적합문헌수}}{\text{검색문헌 총수}}, \quad (\text{단}, i = \text{순위})$$

위의 공식을 적용하여 각 질의에 대한 순위-정확률을 구한 후 같은 순위에서의 평균을 내어 평균 정확률을 산출하였다.

실험의 평가는 클러스터 당 적합문서가 포함된 비율을 산출하고 두 개의 클러스터 중 적합문서의 비율이 높은 클러스터(best cluster)를 선택하여 질의 근접 순위화(close-ness to the query)를 하고 20위까지의 평균정확률을 산출하여 비교하였다. 즉, 실험에서 제안하는 기법이 얼마나 더 적합문서를 한 클러스터에 모으는지를 평가하였다.

4 실험결과

4.1 OutLink 기반 클러스터링

<표 3>은 OutLink 기반 클러스터링 결과를 각 가중치별로 비교한 것이다. 즉, 링크 문서만을 참조하여 클러스터링을 수행하였을 때와(TF), 링크 문서를 참조하고 지역적 역문헌빈도를 적용하였을 때(TF · Local_idf), 링크 문서를 참조하고 전역적 역문헌빈도를 적용하였을 때(TF · Global_idf)의 성능을 비교하였다.

이진 TF의 경우 세 가지의 경우 모두 베이스라인보다 좋은 성능을 나타내었으며(4.34%, 4.99%, 3.75% 향상), 단순 TF의 경우 Local_idf를 함께 사용하였을 경우에만 약간의 성능향상이 있었다(0.85%). 로그 TF의 경우도 모두 베이

<표 3> OutLink 기반 클러스터링 결과

	btf	향상률	ttf	향상률	ltf	향상률	rtf	향상률	5tf	향상률
baseline	0.6807		0.6823		**0.7159**		0.6618		**0.7007**	
TF	**0.7102**	4.34%	0.6813	-0.15%	0.6856	-4.22%	0.6834	3.26%	0.6895	-1.61%
TF · Local_idf	**0.7146**	4.99%	0.6881	0.85%	**0.7086**	-1.02%	0.6977	5.43%	**0.7122**	1.63%
TF · Global_idf	**0.7062**	3.75%	0.6740	-1.23%	**0.7143**	-0.22%	0.6990	5.63%	**0.7254**	3.53%

<표 4> InLink 기반 클러스터링 결과

	btf	향상률	ttf	향상률	ltf	향상률	rtf	향상률	5tf	향상률
baseline	0.6807		0.6823		**0.7159**		0.6618		**0.7007**	
TF	**0.7224**	6.13%	0.6778	-0.66%	**0.7046**	-1.57%	0.6896	4.20%	**0.7009**	0.02%
TF · Local_idf	0.6931	1.82%	0.6867	0.64%	**0.7178**	0.27%	**0.7249**	9.54%	0.6935	-1.03%
TF · Global_idf	**0.7206**	5.87%	0.6725	-1.44%	**0.7101**	-0.80%	**0.7033**	6.27%	**0.7182**	2.50%

스라인보다 성능이 저하되었다(-4.22%, -1.02%, -0.22%).

그러나 루트 TF는 베이스라인은 비록 저조한 성능을 나타내었지만 각 가중치 영역에서 모두 좋은 성능을 나타내었다(3.26%, 5.43%, 5.63% 향상). 마지막으로 보정 TF 역시 Local_idf, Global_idf를 함께 가중치를 주었을 경우 각각 1.63%, 3.53%의 성능향상이 있었다.

4.2 InLink 기반 클러스터링

<표 4>는 InLink 기반 클러스터링 결과를 각 가중치별로 비교한 것이다.

이진 TF의 경우 OutLink기반과 마찬가지로 TF, TF · Local_idf, TF · Global_idf 적용 모두 베이스라인보다 좋은 성능을 나타내었으며(6.13%, 1.82%, 5.87% 향상) 단순 TF도 역시 OutLink기반과 같은 결과인 TF · Local_idf만이 약간의 성능향상이 있었다(0.64%). 그러나 로그 TF의 경우에서는 Out Link를 참조한 실험결과가 모두 베이스라인보다 성능이 저하되었던 반면 TF · Local_idf에서 약간의 성능 향상을 가져왔다(0.27%). 루트 TF도 역시 베이스라인보다 InLink를 참조한 경우 실험에서 모두 좋은 성능을 나타내었다(4.20%, 9.54%, 6.27% 향상). 마지막으로 보정 TF는 OutLink기반과 달리 InLink만을 적용하였을 때가 0.02% 향상되었고 TF · Local_idf가 -1.03%로 성능저하를 나타내었지만 TF · Global_idf에서 2.50% 향상된 0.7182의 평균 정확률을 기록하였다.

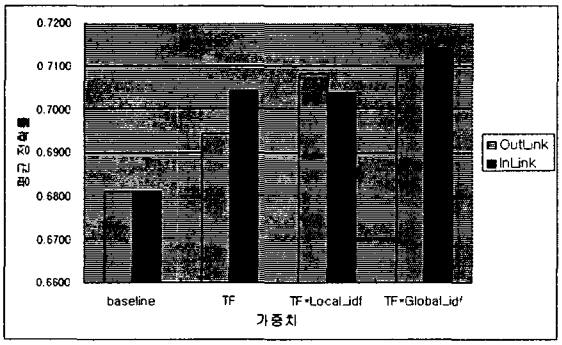
4.3 OutLink, InLink 성능 비교

<표 5>와 <그림 4>는 링크기반 클러스터링의 성능을 비교한 표와 그림이다. 전체적으로 부정적인 결과를 보인 단순 TF(ttf)와 로그 TF(ltf)를 제외한 세 가지의 단어빈도(btf, rtf, 5tf)의 각 링크 확장에 따른 평균 정확률을 구하여 OutLink와 InLink의 성능을 비교하였다. <표 5>에 나타난 링크기반 클러스터링의 성능을 비교하자면 다음과 같다.

첫째, OutLink, InLink의 적용 모두 베이스라인보다 성능 향상을 보였으며(1.95%, 3.41%) 특히 단어빈도만을 가중치로 사용하였을 경우 OutLink보다 InLink의 참조가 더 우수한 성능을 나타내었다.

<표 5> 링크기반 클러스터링 성능 비교표

	OutLink	향상률	InLink	향상률
baseline	0.6811		0.6811	
TF	0.6944	1.95%	0.7043	3.41%
TF · Local_idf	0.7082	3.98%	0.7038	3.34%
TF · Global_idf	0.7102	4.28%	0.7140	4.84%



<그림 4> 링크기반 클러스터링 성능 비교

둘째, Local_idf를 TF와 함께 가중치로 주어 클러스터링을 수행한 결과에서는 OutLink가 InLink보다 더 좋은 결과를 나타내었다(3.98%, 3.34%향상).

셋째, Global_idf를 TF와 함께 가중치로 주어 클러스터링을 수행한 결과에서는 OuLink보다 InLink가 더 좋은 성능을 나타내었다(4.28%, 4.84% 향상).

넷째, TF · Local_idf의 적용보다 TF · Global_idf를 링크 문서의 정보와 함께 사용하는 것이 더 향상된 실험 결과를 나타내었다.

마지막으로 InLink 문서를 참조하고 TF · Global_idf를 가중치로 적용한 결과가 다른 어느 것보다도 가장 좋은 검색 성능을 나타내었다.

실험결과 웹 문서를 검색하는데 있어서 링크정보가 가지는 유용성이 잘 증명되었는데 특히 InLink의 정보가 OutLink보다 더 효율적인 까닭은 InLink로 연결되어있는 문서들이 Out Link보다 주제적으로 더 다양하기 때문이라 판단된다. 즉, 웹 문서의 저자가 알 수 없는 불특정 사이트로부터 링크를 받는 것이 InLink이고 저자의 시야에 속한 특정 사이트로 링크를 하는 것이 OutLink이다. 따라서 InLink의 문서가 주제나 용어의 측면에서 더 다양할 수밖에 없다. 이러한 InLink의 특징이 웹 문서에 반영될 때 해당 웹 문서를 주제적으로 더 잘 표현할 수 있는 것이다.

4 결 론

본 연구에서 제안한 웹 문서의 링크정보를 이용하여 클러스터링을 수행한 후 이 결과를 검색결과의 순위화에 적용하는 기법이 검색 성능을 향상시키는데 효율적인 방법임을 실험을 통해 증명하였다. 특히 InLink의 정보가 Out Link의 정보보다 더 유용하였으며, InLink를 참조하여 공통으로 출현하는 용어의 가중치(TF)를 강화한 후 Global_idf를 역문헌빈도로 사용한 클러스터링 기법이 가장 성능이 우수하였다.

참고문헌

정영미. 1993. 정보검색론. 구미무역(주). pp176-202.

Greengrass, Ed. 2000. Information retrieval: A survey. UMBC Center for Architecture for Data-Driven Information Processing. p111-129.

Modha, Dharmendra S. and W. Scott Spangler. 2000. Clustering hypertext with applications to web searching. *Proceedings of the eleventh ACM on Hypertext and hypermedia* p143 - 152.