웹 디렉토리 서비스를 위한 문서 클러스터링

이문기 ○ 권오욱 이종혁
포항공과대학교 컴퓨터공학과
{jjunior, ohwoog, jhlee}@kle.postech.ac.kr

Document Clustering for Web Directory Service

Moon-Ki Lee ○ Oh-Woog Kwon Jong-Hyeok Lee
Dept. of Computer Science & Engineering, POSTECH

요약

대부분의 검색 엔진에서는 사용자의 정보 검색 요구에서 나타나는 키워드 장벽의 문제점을 해결하고 사용자의 정보 검색 과정에 도움을 주기 위해 디렉토리 서비스를 제공한다. 하지만 디렉토리 서비스에서 새로운 웹 사이트를 지속적으로 인덱스하여 하나의 주제어에 너무 많은 수의 웹 사이트가 부여되어 있으면 사용자의 검색 편의를 위해서 재분류하여 세분류할 필요가 있다. 따라서 본 논문에서는 한 주제어에 과다하게 부여된 웹 사이트들을 세분류하기 위해 기존의 문서 클러스터링 기법을 사용하여 클러스터링을 할 때 생기는 문제점을 보완한 문서 클러스터링 시스템을 소개한다.

1. 서론

일반적으로 현재 사용되는 상용 웹검색 서비스는 사용자가 입력한 질의(query)에 적합한 문서를 인터넷에서 찾아서 정보를 필요로 하는 사용자에게 제공한다. 하지만 현재 상용화된 정보 검색 기술이 단지 키워드(keyword) 위주의 검색을 수행하기 때문에 키워드 장벽(keyword barrier)의 문제가 발생한다. 이러한 문제를 해결하기 위해서 많은 상용 웹검색 서비스 시스템들이 주제 범주 체계를 구축하고 이에 따라 웹 페이지들을 분류하여 제공하는 디렉토리 서비스를 하고 있다. 그러므로 사용자들은 자신이 찾고자 하는 정보를 주제 범주 체계에 따라 이동하면서 쉽게 찾을 수 있다. 하지만 이런 주제 범주 체계에서 지속적으로 추가되는 웹 사이트로 인해 하나의 주제어에 과다한 웹 사이트가 부여되어 있을 때는, 주제 범주 체계보다 더욱 세분화된 사용자의 정보 요구에 부합하지 못하여 디렉토리 서비스 제공의 초기 목적에서 벗어나 사용자들의 정보 검색 과정을 더욱 어렵게 한다. 따라서 이러한 경우에는 해당 주제 범주 체계로 분류되어 있는 웹 사이트들을 재분류할 필요가 있다. 이러한 재분류는 디렉토리 서비스 제공의 원래 목적에 부합할 수 있다. 하지만 디렉토리 서비스에 기존의 문서 클러스터링 기법을 사용해서는 만족할만한 분류 결과를 얻을 수 없다. 따라서 본 논문에서는 클러스터링 시스템의 성능 향상을 위하여 같은 단어가 많이 나타나는 문서는 그만큼 서로 유사하다는 가정에 바탕을 두고 같은 범주 내의 문서들을 차별화할 수 있는 문서유사도 식을 제안한다.

2. 기존 연구

문서 클러스터링은 정보 검색의 효율성(efficiency)과 유효성(effectiveness)을 증대시키기 위한 목적으로 사용한다[1]. 대표적인 문서 클러스터링의 방법론은 클러스터링의 결과로 생성되는 그룹의 구조에 따라서 계층적 클러스터링(hierarchical clustering method)과 비계층적 클러스터링(non-hierarchical clustering method)으로 나눌 수 있는데 개개의 방법론에 따라서 여러가지 구현 알고리즘이 있다[2,3,4].

비계층적 클러스터링은 입력되는 문서의 순서에 따라 클러스터링 결과가 달라지는 단일 처리 방법(single pass method)과 이의 단점을 보완한 재배치 방법(reallocation method)이 있다. 계층적 클러스터링은 문서간의 유사도 정보를 토대로 단계적으로 계층적인 클러스터를 형성하는 방법으로 응집 알고리즘(agglomerative method)과 분할 알고리즘(divisive method)이 있다. 계층적 응집 알고리즘(HACM : hierarchical agglomerative clustering method)에는 단일 링크 방법(single link method), 완전 링크 방법(complete link method), 그룹 평균 연결 방법(group average link method)등이 있다[2,4].

3. 웹 디렉토리 서비스를 위한 문서 클러스터링

웹 사이트 클러스터링의 경우 웹 사이트를 어떻게 표현 하느냐에 따라 다양한 접근 방법이 존재할 수 있다. 웹 사이트를 구성하고 있는 HTML 문서들에서 웹 사이트를 대표할 수 있는 정보를 뽑아내는 적당한 방법이 현

재까지 없기 때문에 웹 사이트의 특성을 대표할 수 있는 다른 정보가 필요하다. 이러한 정보는 웹 검색 서비스에서 전문가에 의해 만들어진 스크립트 파일에서 찾을 수 있다. 스크립트 파일에서 하나의 웹 사이트에 대해 기술한 정보를 하나의 문서로 보고 클러스터링 방법을 사용하여 웹 사이트들을 분류하였다.

3.1. 스크립트 파일의 문서 특성

디렉토리 서비스에서는 전문가에 의해 미리 정해진 주제 범주 체계에 따라 새로운 웹 사이트를 분류하고 해당 웹 사이트의 정보를 기술한다. 여타 다른 문서의 경우와 달리 스크립트 파일이 가지는 특성은 다음과 같다.

- 문서를 대표하는 단어의 수가 평균 20 개 미만
- 한 문서에서 단어의 출현 빈도는 아주 작다
- 같은 범주내의 문서에서 출현하는 단어는 유사

일반적으로 비계층적인 클러스터링 방법들이 지속적인 검색에서 유효성을 감소시킨다는 단점 때문에 계층적인 클러스터링 방법들이 주로 사용된다[3]. 하지만 스크립트 파일이 가지는 위와 같은 특성 때문에 같은 범주의 문서는 의미상 분포 형태가 밀집된 구형이라는 것을 짐작할 수 있다. 그러므로 분산되지 않고 모여 있는 자료와 구형(sphere)의 자료에 대해 좋은 성능을 보이는 재배치 방법의 결과가 좋을 것이다[2]. 따라서 본 논문에서는 웹 디렉토리 서비스를 위한 문서 클러스터링 방법으로 재배치 방법을 제안한다.

3.2. 재배치 방법

재배치 방법은 생성될 M 개의 그룹의 대표값(cluster centroid)을 임의로 선택하고, 모든 문서를 M 개의 그룹과 비교해서 유사도가 큰 그룹에 문서를 포함시키고, 그룹의 대표값을 다시 계산한다. 이러한 과정을 각 문서가 속한 그룹의 대표값이 변하지 않을 때까지 반복한다. 이 방법은 제곱 에러(squared error)를 최소화하는 방향으로 클러스터링을 하는 방법이다.

하지만 이미 같은 주제어에 속한 문서들이기 때문에 스크립트 파일이 가지는 특성 중 세번째 특성으로 인해 기존의 유사도를 사용하는 문서 클러스터링 방법으로는 문서들간의 유사도가 클 경우 잘 분류할 수 없다. 따라서 그 중에서도 서로 같은 단어가 많이 나타나는 문서들끼리의 유사도를 높일 수 있는 방법론이 제시되어야 한다.

3.3. 제안하는 유사도 계산식

기존의 코사인 유사도만으로는 앞의 문제점을 해결할 수 없기 때문에 이미 서로 유사하다고 묶여진 문서들 간의 차별성을 극대화시키기 위해서 비교가 되는 두개의 문서간에 같이 나타나는 단어의 수에 대한 가중치를 부여하는 방법을 제시한다. 그리고 이러한 가중치를 "매칭계수(matching factor)"라고 정의한다. 이것은 기존의 코사인 유사도식이 두개의 기저(basis)로 구성된 평면에 나타나는 문서의 벡터에 대해서는 유사도의 차이(dissimilarity)를 제대로 나타내지 못하기 때문이다. 매칭계수를 이용한 유사도 계산식은 다음과 같다.

$$sim(D_i, D_j) = \left(1.2 + \frac{mf}{T(D_i)} + \frac{mf}{T(D_j)}\right)^{mf-1} \times \frac{\sum_{k=1}^{L}(weight_{ik} \times weight_{jk})}{\sqrt{\sum_{k=1}^{L} weight_{ik}^2 \times \sum_{k=1}^{L} weight_{jk}^2}}$$

where

mf : D_x와 D_y에서 같이 발생하는 용어의 수

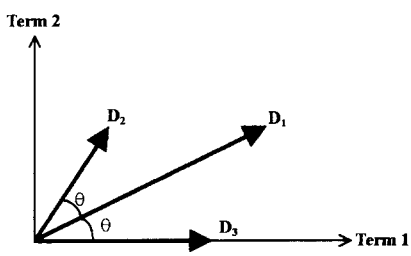
$T(D_x)$: D_x에서의 전체 용어의 수

L : 용어 벡터(term vector) 의크기

매칭계수에 대한 4 차원 공간상의 예는 아래와 같다.

문서	용어벡터	매칭계수	$T(D_x)$
1	(1, 3, 0, 0)	1	2
2	(0, 2, 1, 5)		3

위의 유사도 계산식을 2 차원 공간상에서 예로 나타내면 다음과 같다.



위의 그림과 같은 상황에서 코사인 유사도만으로는

$$sim(D_1, D_3) = sim(D_1, D_2)$$

이다. 하지만, 매칭계수를 사용하면 다음과 같다.

$$(1.2 + \frac{2}{2} + \frac{2}{2})^1 sim(D_1, D_2) > (1.2 + \frac{1}{2} + \frac{1}{1})^0 sim(D_1, D_3)$$

$$3.2 \times sim(D_1, D_2) > 1 \times sim(D_1, D_3)$$

4. 실험

실험은 한국통신의 한미르 검색 엔진에서 사용하는 32,442 개의 웹 사이트에 대해 기술된 스크립트 파일에서 불용어를 제거하고 용어(term)를 추출했다. 용어 추출시에는 기대 상호 정보 척도(EMIM : expected mutual information measure)와 상호 정보 척도(MI : mutual information measure)를 이용하여 스크립트 파일에서 나타나는 용어의 60%만 선택하여 역색인 파일을 작성하였다.

4.1. 실험 방법

본 논문의 실험을 위해서, 2 개의 주제 범주들에 속하는 문서들의 합집합을 과다한 하나의 주제 범주로 보고, 최대신장트리(MST : maximum spanning tree)를 이용한 단일

링크 방법과 제안하는 재배치 방법으로 클러스터링하여 결과를 서로 비교하였다. 두가지 클러스터링 방법에 의해 원래의 주제 범주들로 각각 클러스터링이 되는지를 평가하였다. 따라서 결과로 생성되는 클러스터의 수는 2개로 고정된다. 재배치 방법에서는 초기 클러스터의 대표값을 모든 문서간 유사도 중에서 가장 작은 2개의 문서로 각각 선택했다.

4.2. 실험 집합

실험을 위한 실험 문서의 집합은 다음과 같다.

- Set A : 서로 상이한 범주의 문서 집합
- Set B : 서로 비슷한 범주의 문서 집합(중복된 범주에 속하는 문서가 존재)

[표 1] 실험 문서 집합 A

실험 데이터 이름		실험 데이터의 범주(웹 사이트 수)
A-1	A-1-1	여행,레저스포츠|스포츠|종목별|야구(10)
	A-1-2	여행,레저스포츠|스포츠|종목별|축구(28)
A-2	A-2-1	건강,의학|병원|과별|성형외과(57)
	A-2-2	과학,기술|회사|화학(65)
A-3	A-3-1	컴퓨터,인터넷|회사|소프트웨어(66)
	A-3-2	건강,의학|간호|대학,학과|학과,연구실(11)

[표 2] 실험 문서 집합 B

실험 데이터 이름		실험 데이터의 범주(웹 사이트 수)
B-1	B-1-1	기업,회사|업종별|통신|한국통신|서비스(5)
	B-1-2	기업,회사|업종별|통신|한국통신|전화국(9)
B-2	B-2-1	여행,레저스포츠|스포츠|종목별|축구|구단(15)
	B-2-2	여행,레저스포츠|스포츠|종목별|축구|기관,단체(6)
B-3	B-3-1	건강,의학|대체의학|명상,기공,단학|수련원(19)
	B-3-2	건강,의학|대체의학|명상,기공,단학(20)

4.3. 평가 방법

전통적인 정보 검색 시스템에서 가장 기본이 되는 평가 척도인 정확률(precision)과 재현율(recall)에 대해 평가하였다. 또한 기존의 클러스터링 연구에서 제시된 클러스터링 결과에 대한 평가방법[3]이 본 실험의 평가방법으로는 적합하지 않기 때문에 병합 정확률(union precision)을 정의한다.

$$\text{병합 정확률} = \frac{\text{(각 클러스터에 정확히 분류된 문서 수의 합)}}{\text{(전체 클러스터링 대상 문서의 수)}}$$

4.4. 결과 및 분석

[표 3] 실험 결과

		Single Link Method with MST			Reallocation Method		
		P	R	U.P	P	R	U.P
A-1	A-1-1	100(100)	100(100)	100 (100)	100(100)	100(100)	100 (100)
	A-1-2	100(100)	100(100)		100(100)	100(100)	
A-2	A-2-1	100(100)	100(100)	100 (100)	100(100)	100(100)	100 (100)
	A-2-2	100(100)	100(100)		100(100)	100(100)	
A-3	A-3-1	100(100)	100(100)	100 (100)	100(100)	100(100)	100 (100)
	A-3-2	100(100)	100(100)		100(100)	100(100)	
B-1	B-1-2	100(35.7)	20(100)	71.4 (42.9)	100(35.7)	100(100)	100 (42.9)
	B-1-2	69(100)	100(11.1)		100(100)	100(11.1)	
B-2	B-2-1	77.7(83.3)	100(100)	80.9 (85)	100(100)	92.8(26.7)	95.2 (50)
	B-2-2	100(100)	40(33.3)		85.7(37.5)	100(100)	
B-3	B-3-1	51.3(100)	100(5.3)	53.8 (56.8)	76.2(90.9)	84.2(52.6)	87.2 (75.7)
	B-3-2	100(55.6)	5(100)		81.3(69.2)	65(85.7)	

[표 3]에서 볼 수 있듯이 최대 신장 트리를 이용한 단일 링크 방법이 재배치 방법보다 낮은 성능을 보임을 알 수 있다. 따라서 디렉토리 서비스를 위한 클러스터링에는 재배치 방법이 효과적임을 실험을 통해 알 수 있다. [표 3]에서 괄호안의 수치는 매칭 계수를 적용하지 않고 원래의 코사인 유사도를 이용한 결과이다. 결과를 통해 매칭 계수를 사용함으로써 성능이 향상된 것을 알 수 있다.

5. 결론 및 향후 과제

본 논문에서는 재배치 방법과 최대 신장 트리를 이용한 단일 링크 방법을 사용해서 디렉토리 서비스의 문서를 클러스터링 하여 비교하였고, 재배치 방법이 이러한 경우에 적합하고 문제 영역의 특성상 일반적인 코사인 유사도식보다 매칭계수를 사용한 코사인 유사도식이 더 적합함을 알 수 있다.

재분류된 각각의 클러스터의 이름을 사람의 손을 거치지 않고 정하기 위한 적절한 방법론이 없기 때문에 이에 대한 연구가 후행되어야 한다. 또한 재배치 방법은 non-linear optimization algorithm 에서 복잡도를 줄이기 위한 EM(expectation maximization) 알고리즘[5]의 특수한 형태로 볼 수 있기 때문에 초기 파라미터의 설정에 따라 부분 극대(local maxima)에 빠질 수 있다. 실제로 재배치 방법에서 가장 유사도가 작은 문서쌍을 각각 클러스터의 초기 대표값으로 정했을 때의 결과가 최적은 아니었다. 따라서 부분 극대를 벗어나 전체 극대(global maxima)에 도달할 수 있는 방법론에 대한 연구가 후행되어야 한다.

참고문헌

[1] Van Rijsbergen, C. J. "Information Retrieval", London: Butterworths, 1979

[2] William B. Frakes, Ricardo Baeza-Yates, "Information Retrieval: Data Structures & Algorithms", Prentice Hall, 1992

[3] Peter Willett, "Recent Trends in Hierarchical Document Clustering: A Critical Review", Information Processing & Management Vol. 24, No. 5, 1988

[4] Anil K. Jain, Richard C. Dubes, "Algorithms for clustering data", Prentice Hall, 1988

[5] Christopher M. Bishop, "Neural Networks for Pattern Recognition", Clarendon Press OXFORD, 1995