의미정보의 효율적인 분류를 위한 계층적 중복 문서 클러스터링

강동혁[0] 주길홍 이원석
연세대학교 컴퓨터과학과
{wolfkang, faholo, leewo}@amadeus.yonsei.ac.kr

Hierarchical Overlapping Document Clustering for Efficient Categorization of Semantic Information

Dong-Hyeok Kang[0] Kil-Hong Joo Won-Suk Lee
Dept. of Computer Science, Yonsei University

요 약

기존의 문서 클러스터링 알고리즘은 모든 문서가 각각 하나의 클러스터에만 할당되도록 설계되어, 문서에 여러 개의 주제가 포함되어 있을지라도 문서는 유사도 비교에 의해 오직 하나의 클러스터에 포함된다는 단점이 있다. 본 연구에서는 이러한 문서 클러스터링 방법의 한계를 극복하기 위해 문서가 여러 개의 클러스터에 포함될 수 있는 계층적 중복 문서 클러스터링을 제안한다. 또한, 문서 클러스터링의 정확도를 높이기 위해서 불용어 제거 알고리즘을 이용해 불용어를 제거하여 클러스터링에 사용되는 키워드를 선별하고, 단어가중치 산출을 위한 TF*NIDF 공식을 제안한다.

1. 서 론

문서 클러스터링은 문서간의 유사도를 바탕으로 연관된 문서들을 군집화함으로써 문서들을 주제별로 통합하는 방법이다[1]. 문서 클러스터링은 대용량 문서들을 탐색하고, 검색하는 데 있어서, 효율성을 높이고, 검색의 정확성을 증대시키는 방법으로, 이에 대해 많은 연구가 이루어지고 있다. 문서 클러스터링은 대체로 명사 추출, 문서 클러스터링 알고리즘 적용의 순서로 수행된다. 명사 추출 단계에서는 형태소 분석을 통해 명사만 추출한다. 이 때 미리 작성한 불용어 목록에 의해 일부 단어들을 제거한다. 명사 추출 후, 문서의 유사도를 기반으로 문서 클러스터링 알고리즘을 적용한다.

명사 추출 단계에서 형태소 분석 시 발생하는 중의성으로 인해 제거되지 않은 불용어가 존재하거나, 불용어 목록이 불충분하여 불용어들이 남아있을 수 있다. 그리고, 기존의 문서 클러스터링 알고리즘들은 문서를 보는 관점에 따라 문서의 주제가 약간씩 다르거나 하나의 문서가 여러 주제를 포함할 수 있음에도 불구하고, 하나의 문서는 반드시 하나의 클러스터에만 포함되는 한계를 가지고 있다.

본 논문에서는 문서 클러스터링의 정확도를 높이기 위해 통계적인 기법으로 불용어를 찾아 제거하는 불용어 제거 알고리즘을 제안한다. 그리고, 기존의 문서 클러스터링 알고리즘이 갖는 한계를 극복하기 위해 하나의 문서가 유사한 주제를 갖는 여러 클러스터에 포함되는 계층적 중복 문서 클러스터링(HODC, hierarchical overlapping document clustering) 알고리즘을 제안한다. 또한, 문서 클러스터링에 필요한 단어가중치를 산출하기 위해 기존의 TF*IDF 공식을 수정한 TF*NIDF 단어가중치 산출 방법을 제안한다.

2. 관련 연구

클러스터링을 위한 많은 방법들[2,6]이 있는데, 문서 클러스터링과 관련하여 가장 널리 적용되고 있는 방법에는 계층적 집적 클러스터링(HAC, hierarchical agglomerative clustering) 방법과 반복 클러스터링(iterative clustering) 방법이 있다.

계층적 집적 클러스터링은 각 문서를 하나의 클러스터로 두고, 각 클러스터간의 유사도를 모든 클러스터 사이에 계산하여, 가장 가까운 클러스터를 새로운 클러스터로 결합하는 방법이다. 계층적 집적 클러스터링 방법에는 단일 연결(single link), 완전 연결(complete link), 그리고 집단 평균 연결(group average link) 방법이 있다.

반복 클러스터링 방법은 재배치(reallocation) 방법이라고 불리는 방법으로, 반복적으로 문서들을 가장 유사한 클러스터에 재할당함으로써, 클러스터링을 최적화하는 방법이다. 반복 클러스터링 방법은 수행 시간이 빠르지만, 계층적 집적 클러스터링 방법이 보다 정확한 클러스터링 결과를 유도하는 것으로 알려져 있다.

3. 계층적 중복 문서 클러스터링

본 논문에서 제안하는 계층적 중복 문서 클러스터링의 수행 단계는 크게 세 단계로 나눌 수 있다. 첫 번째 단계는 명사를 추출하고, 불용어 제거 알고리즘을 통해 불용어를 제거하는 단계이다. 두 번째 단계는 단어가중치를 산출하고, 키워드를 선정하는 단계이다. 마지막으로, 세 번째 단계는 키워드와 키워드의 가중치를 토대로 HODC 알고리즘을 수행하는 단계이다. 그림1은 본 논문에서 제안하고 있는 계층적 중복 문서 클러스터링의 흐름도를 나타낸다.

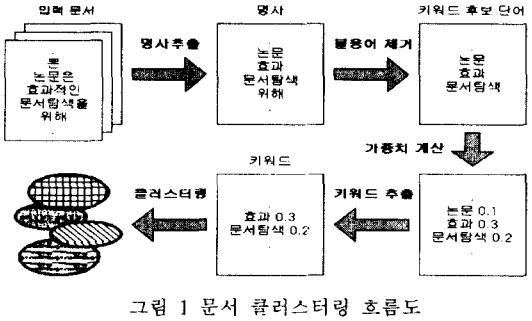


그림 1 문서 클러스터링 흐름도

3.1 명사 추출 및 불용어 제거

먼저, 클러스터링을 수행할 문서들을 대상으로 명사를 추출한다. 본 논문에서는 형태소 분석기인 HAM[3]을 사용하여 명사들을 추출하였다. 명사 추출 후, 해싱 기법을 이용한 불용어 제거 알고리즘을 통해 불용어를 판별하여 제거한다. 불용어 제거를 위해 문서들이 정치, 경제, 사회, 과학 등과 같은 대주제들로 분류되어 있다고 가정하고, 이렇게 분류되어 있는 문서의 집합을 사이트(site)라고 한다.

사이트 S_j에서 단어 t_k가 갖는 문서 지지도(document support) ds_{jk}는 식(1)과 같다.

$$ds_{jk} = \frac{(\text{사이트 } S_j \text{ 에서 단어} t_k \text{를 포함하는 문서 수})}{(\text{사이트 } S_j \text{ 에 포함된 전체 문서 수})} \quad \text{식(1)}$$

그리고, 해싱에 이용되는 해시 함수 h는 식(2)와 같다.

$$h(ds_{jk}) = \begin{cases} B-1, \text{ if } ds_{jk} = 1.0 \\ \lfloor ds_{jk} \times B \rfloor, otherwise \end{cases} \quad (\text{단, } B\text{는 버킷의 수}) \quad \text{식(2)}$$

버킷(bucket)에 포함되는 레코드는 (사이트 번호, 문서 지지도)로 구성된다. 단어 t_k가 불용어 인지 아닌지를 판별하기 위해, t_k의 각 사이트별 문서 지지도를 구하고 해시 함수를 통해 해시 테이블을 작성한다. 그후, 가장 많은 사이트를 포함하고 있는 버킷과 이 버킷에 포함된 사이트 문서 빈도수의 평균과의 차가 δ ($0 \le \delta \le 1$)이내인 버킷들을 하나로 합친다. 이렇게 합쳐진 버킷에 포함된 사이트 수가 *MinSite*(최소 사이트 수) 이상이고, 문서 지지도의 평균이 *MinSup*(최소 문서 지지도) 이상이면, 이 단어는 불용어 후보 단어로 간주한다. 그런데, 특정 사이트에서는 불용어 후보 단어라 하여도 아주 높은 문서 지지도를 갖는 경우가 있다. 이런 경우에는 단어가 특수한 의미로 사용되는 것으로 판단하여, 불용어 후보 단어의 문서 지지도가 *SpcSup*(특수 용어 판별을 위한 문서 지지도) 이상이 되는 사이트를 제외한 다른 사이트에서만 불용어 후보 단어를 삭제하도록 한다.

3.2 단어가중치 산출 및 키워드 선정

불용어를 제거한 후, 남아있는 단어들을 대상으로 단어가중치를 계산한다. 문서에서 단어가중치를 산출하는 방법은 TF*IDF(term frequency inversed document frequency)[4] 공식이 많이 사용된다. TF*IDF 공식에서는 문서수가 많으면 많을수록 IDF 값이 가중치를 결정하는 데 큰 비중을 차지하므로, 본 논문에서는 IDF 값이 일정 범위를 넘지 않도록 조종한 TF*NIDF(term frequency normalized inversed document frequency) 공식을 제안한다. 따라서, 문서 d_i에서 단어 t_j의 가중치 $tfnidf_{ij}$는 식(3)과 같다.

$$tfnidf_{ij} = tf_{ij} \times \left\{ \mu - \ln\left(\left(\frac{e^{\mu}-1}{N-1} \right)(df_j - 1) + 1 \right) \right\} \quad \text{식(3)}$$

tf_{ij}는 단어 빈도수로서 문서 d_i에서 단어 t_j가 나타난 횟수이고, df_j는 문서 빈도수로서 N개의 문서들 중에서 단어 t_j가 존재하는 문서 수이다. 그리고, μ (>0)는 IDF 값의 최대값이다.

다음으로, 단어가중치를 토대로 문서의 키워드를 선택한다. 단어가중치의 평균값을 구하여 단어가중치가 평균값 이상인 단어들을 키워드로 선정하고, 선정된 키워드들의 가중치를 다음의 코사인 정규화 식을 통해 정규화한다.

$$w(d_i,k_j) = \frac{tfnidf_{ij}}{\sqrt{\sum_{k=1}^{n} tfnidf_{ik}^2}} \quad \text{식(4)}$$

3.3 문서 클러스터링

본 논문에서 제안하는 계층적 중복 문서 클러스터링을 위해 다음의 정의를 이용한다.

문서들에 속해 있는 전체 키워드들의 집합을 K, 클러스터링을 수행할 문서들의 전체 집합을 D, 문서들의 집합인 클러스터를 C, 그리고 클러스터 C_i에 속하는 문서들에 있는 키워드들의 집합을 CK_i라 할 때,

$K = \{k_1, k_2, ..., k_n\}$
$D = \{d_1, d_2, ..., d_s\}$
$d_i \subset K$, $d_1 \cup d_2 \cup ... \cup d_s = K$
$C = \{d_1, d_2, ..., d_k\}$
$CK_i = \bigcup_{d_j \in C_i} d_j$
와 같이 정의한다.

상기의 정의들을 바탕으로 문서간의 유사도와 클러스터간의 참여도, 클러스터의 응집도를 정의한다. 두 문서 d_1, d_2의 유사도 $s(d_i,d_j)$는 식(5)와 같다.

$$s(d_i,d_j) = \frac{1}{2}\left(\frac{\sum_{k_l \in d_i \cap d_j} w(d_i,k_l)}{\sum_{k_l \in d_i} w(d_i,k_l)} + \frac{\sum_{k_l \in d_i \cap d_j} w(d_j,k_l)}{\sum_{k_l \in d_j} w(d_j,k_l)} \right) \quad \text{식(5)}$$

두 클러스터 C_m과 C_n가 있을 때, C_n에 대한 C_m의 참여도 $p(C_m|C_n)$는 식(6)과 같다.

$$p(C_m|C_n) = \frac{\sum_{d_i \in C_m} \left(\sum_{k_l \in (CK_m \cap CK_n)} w(d_i,k_l) \right)}{\sum_{d_i \in C_n} \left(\sum_{k_l \in CK_n} w(d_i,k_l) \right)} \quad \text{식(6)}$$

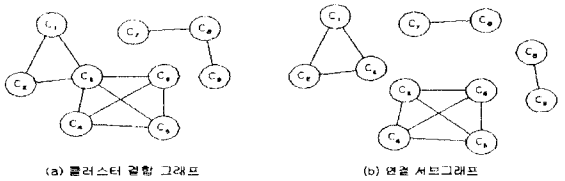
클러스터 C_u가 있을 때, C_u의 응집도는 문서간의 유사도를 바탕으로 계산되며 식(7)과 같다.

$$c(C_u) = \frac{\sum_{d_i \in C_u} \left(\sum_{d_j \in C_u - \{d_i\}} s(d_i,d_j) \right)}{|C_u| C_2} \quad \text{식(7)}$$

클러스터 간의 참여도와 클러스터의 응집도를 기반으로 *클러스터 결합 가능성*을 정의한다. *클러스터 결합 가능성*은, 두 클러스터 C_m과 C_n이 다음의 *클러스터 결합 조건*을 만족시킬 때, 두 클러스터는 *결합 가능하다*고 말한다.

(클러스터 결합 조건) $p(C_m|C_n) \ge MinPart$ *and* $p(C_n|C_m) \ge MinPart$ *and* $c(C_u) \ge MinCoh$ (단, C_u는 C_m과 C_n이 결합한 클러스터, *MinPart*는 최소 클러스터 참여도, *MinCoh*는 최소 클러스터 응집도)

그리고, *클러스터 결합 그래프*란 클러스터를 노드로 하고, 결합 가능한 클러스터들을 연결한 그래프를 말한다. 그림2(a)는 9개의 클러스터가 있을 때 클러스터간의 결합 가능성을 조사하여 작성한 *클러스터 결합 그래프*의 한 예이고, 그림2(b)는 그림2(a)의 연결 서브그래프이다.



(a) 클러스터 결합 그래프 (b) 연결 서브그래프

그림 2 클러스터 결합 그래프의 예

HODC 알고리즘을 수행하는 방법은 다음과 같다.

단계1) 하나의 문서를 포함하는 클러스터를 각각 생성한다.

단계2) 클러스터간의 결합 가능성을 모두 조사하여 클러스터 결합 그래프를 작성한다.

단계3) 클러스터 결합 그래프 중 다른 연결 서브그래프(connected subgraph)에 포함되지 않는 연결 서브그래프를 찾아서, 연결 서브그래프에 포함되어 있는 클러스터들을 모두 결합한다.

단계4) 더 이상 클러스터링이 이루어지지 않을 때까지, 다음의

과정을 되풀이한다.

- *클러스터 결합 조건*을 만족하는 두 클러스터들 중에서 응집도가 최대가 되는 두 클러스터를 결합한다.

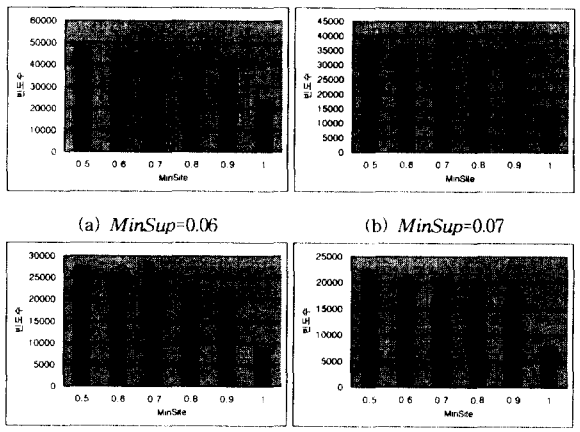
4. 실험

본 논문에서 제안하고 있는 불용어 제거 알고리즘, 계층적 중복 문서 클러스터링 알고리즘을 실제 문서들을 대상으로 실험을 수행하였다. 실험을 위해 야후!코리아 뉴스[5]에서 제공하고 있는 신문기사 중에서 경제, IT, 정치, 사회 등 10개 분야(사이트)의 기사를 추출하여 사용하였고, HAM을 이용하여 명사를 추출하였다. 표1은 실험에 사용된 데이터의 특성을 나타낸 것이다.

표 1 데이터 특성

사이트 별 평균 문서 수	1026
전체 명사 수	819,835
명사 수	112,860

추출된 명사를 대상으로 불용어 제거 알고리즘을 이용하여, 불용어 제거 실험을 수행하였다. 이 실험에 사용된 대부분의 단어들이 문서 지지도가 0.3 이하를 기록하여, 특수 용어 판별을 위한 문서 지지도인 *SpcSup*을 0.3으로 설정하였다. 그림3은 *MinSite*와 *MinSup*에 따라서 제거된 불용어들의 빈도수를 나타낸 것이다.



(a) *MinSup*=0.06 (b) *MinSup*=0.07

(c) *MinSup*=0.08 (d) *MinSup*=0.09

그림 3 불용어 빈도수

그림3에서 보듯이, MinSup과 MinSite가 높을수록 불용어들의 비율이 낮아지며, MinSup과 MinSite를 통해 제거되는 불용어들의 빈도수를 조절할 수 있다.

문서 클러스터링 실험에서는 경제 분야의 신문기사 1000건에 대하여 기존의 HAC 방법과 본 논문에서 제안하고 있는 HODC 방법을 비교하는 실험을 수행하였다. 표2는 *MinPart*를 0.3으로 설정하였을 때, *MinCoh*에 따라서 HAC 알고리즘과 HODC 알고리즘에 의해 생성되는 클러스터의 개수와 두 문서 이상 포함하고 있는 클러스터들의 평균 문서 수, 클러스터의 평균 응집도, 그리고 중복되어 클러스터에 포함된 문서 수를 조사한 것이다.

표 2 문서 클러스터링 실험 결과

알고리즘	HAC					HODC				
MinCoh	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
클러스터 개수	84	70	50	25	12	143	108	57	26	13
클러스터의 평균 응집도	0.70	0.74	0.80	0.88	0.94	0.60	0.69	0.77	0.88	0.94
중복된 문서 수	0	0	0	0	0	26	14	4	1	1

HODC 알고리즘을 적용한 경우 HAC 알고리즘을 적용했을 때보다 중복된 문서들에 의해 클러스터의 수가 증가하였고, 중복을 허용함으로써 다양한 주제별로 클러스터들이 생성될 수 있음을 알 수 있다. 여기서, HODC 알고리즘에 의해 만들어진 클러스터들의 평균 응집도가 HAC 알고리즘에 의해 만들어진 클러스터들의 평균 응집도보다 낮은 것은 계층적 클러스터링 알고리즘이 응집도가 높게 형성되는 클러스터들을 먼저 결합하므로, 문서수가 많을수록 응집도가 낮아지기 때문이다.

5. 결론 및 향후 연구

본 논문에서는 문서 클러스터링의 정확도를 높이고자, 불용어를 판별하여 제거하는 불용어 제거 알고리즘을 제안하였다. 그리고, TF*IDF 공식이 갖는 약점을 극복하기 위해, TF*NIDF 공식을 제안하여 단어가중치를 산출하였다. 또한, 문서가 연관된 주제의 여러 클러스터에 중복적으로 포함되는 계층적 중복 문서 클러스터링 알고리즘을 제안하였다.

불용어 제거 실험에서는 불용어 제거 알고리즘을 적용하여 불용어를 판별하였다. 문서 클러스터링 실험에서는 기존의 HAC 알고리즘과 본 논문에서 제안하는 HODC 알고리즘으로 경제 분야 신문기사들의 클러스터링을 수행하여 그 결과를 비교하여, 문서가 중복적으로 클러스터에 포함되는 것을 허용함으로써 HODC 알고리즘이 다양한 주제별로 의미정보들을 효율적으로 분류할 수 있음을 확인하였다.

본 논문에서 제안하는 불용어 제거 알고리즘을 적용하기 위해서는 문서들을 미리 대주제로 분류해야 하는 단점을 가지고 있다. 이를 해결하기 위해서 클러스터링을 수행한 후, 클러스터링된 문서들을 토대로 불용어 제거 알고리즘을 수행하여 불용어를 판별하는 방법이 이용될 수 있을 것이다. 그리고, 불용어 제거 알고리즘과 계층적 중복 문서 클러스터링 알고리즘은 모두 주어진 문서 집합을 토대로 수행되므로, 새로운 문서들이 추가되었을 때, 처음부터 다시 각 알고리즘을 수행해야 하는 문제점을 안고 있다. 따라서, 새로운 문서가 추가될 때, 이전에 수행된 결과를 토대로 불용어 제거와 문서 클러스터링이 가능한 점진적인 방법에 관한 연구가 후행되어야 한다.

참고문헌

[1] Cutting, D. R., Karger, D. R., Pedersen, J. O., Tukey, J. W., "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections", 15th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.318-329, 1992

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

[3] 강승식, "HAM: 한국어 분석 모듈", http://nlp.kookmin.ac.kr

[4] Salton, G., Buckley, C., "Term-weighting approaches in automatic text retrieval", Information Processing and Management Vol. 24 No. 5 pp. 513-523, 1988

[5] 야후!코리아 뉴스, http://kr.dailynews.yahoo.com/

[6] Jain, A. K., Dubes, R. C., "Algorithms for Clustering Data", Prentice Hall, 1988