제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

도메인 불용어 제거를 통한 효율적인 텍스트 마이닝 기법

송재선*, 주길홍, 이원석
*연세대학교 컴퓨터과학과
songjs, faholo, leewo@amadeus.yonsei.ac.kr

An Efficient Text Mining method based on Domain Stopword Elimination

Jae-Sun Song*, Kil-Hong Joo, Won-Suk Lee
*Dept of Computer Science, Yon-Sei University

요 약

정보 검색 분야에서 문서 클러스터링방법은 사용자에게 양질의 다양한 정보를 제공하기 위한 방법으로 이에 대한 많은 연구가 수행되었다. 그러나 기존의 문서클러스터링 방법들은 클러스터간의 포함관계를 나타내는 계층적 관계를 표현하지 않고 의미적으로만 비슷한 내용의 문서를 묶어 여러개의 클러스터로 나타내었다. 이에 본 논문에서는 각 문서가 속하는 도메인 별로 불용어와 키워드를 추출하여 문서클러스터링에 적용하는 알고리즘을 제안한다.

1. 서론

정보 검색(IR)[1]이란 대량의 자료 속에서 사용자의 요구에 따라 다양하고 유용한 정보를 효율적으로 찾는 방법이다. 정보 검색을 효율적으로 수행하기 위한 방법으로 문서 클러스터링이 있다. 문서 클러스터링 방법은 문서내의 단어를 문서의 의미를 표현할 수 있는 벡터로 변환하고, 이 벡터 값을 사용하여 각 문서들간의 유사도를 측정하고, 동일한 의미를 가진 문서들을 클러스터로 묶는 방법이다. 따라서 문서 클러스터링방법은 인간의 수작업에 의한 문서 분류를 대신하여 정보 검색의 효율성을 증가 시켰다.

문서 클러스터링방법의 정확성을 높이기 위하여 문서내에서의 불용어 판별, 단어의 가중치 계산, 가중치의 정규화, 유사도 계산 방법등에 관한 많은 연구가 시행되어져 왔으며, 다양한 방법[4]들이 제시되고 있다. 그러나 이러한 방법들은 문서클러스터링에 사용되는 데이터의 특성에 따라 성능이 크게 좌우되며, 카테고리의 계층적 관계를 표현 할 수 없는 문제점이 있다. 이에 본 논문에서는 문서집합을 계층적 관계를 갖는 여러 도메인으로 분류하고, 각 도메인에 따른 문서의 불용어를 판별하며, 불용어가 제거된 문서들로 응집도와 참여도를 고려한 문서 클러스터링 방법을 제안한다.

2. 관련 연구

일반적으로 문서 클러스터링에서 가장 널리 사용되고 있는 방법은 계층적 집적 클러스터링 방법(Hirarchical Agglomerative Clustering)이다. 계층적 집적 클러스터링방법이란 각 문서들이 하나의 클러스터가 되어 다른 모든 클러스터와의 유사도를 비교하여 가장 유사도가 높은 클러스터와 결합한다. 이러한 과정을 반복함으로써 의미있는 클러스터를 찾는다. 이러한 계층적 클러스터링 방법은 유사도의 비교 방법에 의해 다음과 같이 나눌 수 있다 [2].

(1) 단일 연결 방법(Single Linkage Method)
(2) 완전 연결 방법(Complete Linkage Method)
(3) 집단 연결 방법(Group Average Method)
(4) WARD 방법 (Ward Method)

일반적으로 불용어란 문장에서 의미 없이 쓰이는 조사, 혹은 정보 검색에서 인덱싱의 효율을 높이기 위해 정의해서 사용하는 부정어 사전(Negative Dictionary)에 포함된 단어를 의미한다. 일반적인 불용어 제거 방법으로는 단어의 출현 빈도를 이용한 통계적인 방법인 불용어 색인(index)를 구성하는 방법이 많이 이용된다. 또한, 단어의 가중치를 산정하는 방법으로는 단어의 빈도수에 기반한 $tf \times idf$ 공식이 많이 사용되고 있으며, 문서의 특성에 따라

다양한 성능을 나타낸다.

문서 클러스터링을 측정하기 위하여 자주 사용되는 방법으로 문서의 쌍에 기반한 RAND방법과 CSIM방법[5]이 있다. 다음과 같은 두 가지의 클러스터링방법 A와 B를 비교할 경우 짝지어진 쌍의 데이터가 A와 B에서 동일한 클러스터에 속하는 가에 따라 다음과 같이 평가한다.

		B	
		1	0
A	1	a	b
	0	c	d

1 : 짝지어진 쌍이 하나의 클러스터에 속한 경우
0 : 짝지어진 쌍이 다른 클러스터에 속하는 경우

Rand 척도에서 두 클러스터링 결과간의 유사도는 식(1)과 같다.

$$Rand(C_A, C_B) = \frac{a + d}{a + b + c + d} \quad (1)$$

또한, 분할표에 대해 다이스 계수 공식을 적용한 다음과 같은 CSIM은 식(2)와 같다.

$$CSIM(C_A, C_B) = \frac{2a}{2a + b + c} \quad (2)$$

수작업으로 분류한 클러스터를 A라하고, 문서 클러스터링에 의해 만들어진 클러스터를 B라 하면, CSIM 척도에 의해 문서 클러스터링의 정확도를 측정할 수 있다.

3. 도메인 불용어 제거를 통한 문서 정규화 방법

정보검색에서 일반적으로 사용되는 불용어 제거 방법은 단어가 출현하는 문서수에 기반한 통계적 방법에 의해 부정어 색인(Negative Index)를 구성하여 사용하는 방법이 주로 사용된다. 따라서, 본 논문에서는 다음과 같은 방법으로 불용어를 판별하여 제거한다.

3. 1 도메인 불용어 제거 방법

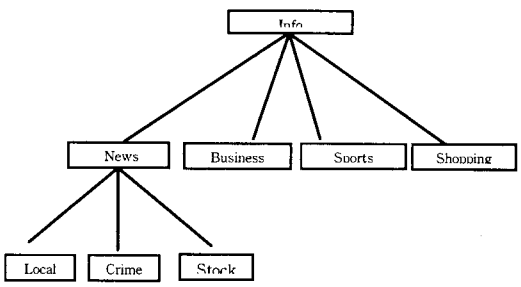

그림 1

그림 1 의 전체 도메인집합 $D=\{ d_1, d_2, \cdots, d_n \}$ 의 카테고리에서 도메인 지지도와 문서지지도는 다음과 같이 정의한다.

정의 1. 도메인 지지도 (Domain Support)

$$S_{d_i}(w) = \frac{d_i\text{의 하위도메인중 단어}w\text{의 출현 도메인수}}{\text{도메인 } d_i\text{의 하위도메인수}}$$

정의 2. 문서 지지도 (Document Support)

$$T_{d_i}(w) = \frac{d_i\text{ 에서 단어}w\text{를 포함하는 문서 수}}{\text{도메인 } d_i\text{ 에 포함된 전체 문서 수}}$$

여러 도메인과 많은 문서에 공통적으로 출현하는 단어일수록 불용어로 판별될 가능성이 높다. 즉 도메인 지지도 값과 문서지지도 값이 높을 수록, 또한 문서 지지도 값의 편차가 적을 수록 불용어로 판별될 가능성이 크다. 따라서 다음과 같은 단계에 의해 불용어를 판별하여 제거한다. 이때 도메인 지지도값을 측정하기 위하여 제일 상위의 카테고리인 News, Business, Sports는 미리 정의하여 알고리즘을 수행한다.

[단계 1] 도메인에 포함되는 모든 단어들에 대해 도메인 지지도와 문서 지지도 값을 계산한다.
[단계 2] 도메인 지지도 값이 최소 도메인 지지도 이상이고문서 지지도 값이 최소 문서지지도 값 이상이며, 문서지지도의 표준 편차가 범위 이내인 값이면 불용어 리스트에 추가 한다.
[단계3] 불용어 리스트를 각 문서에 적용하여 각 문서에서 클러스터링에 사용되지 않을 단어들을 추출한다.

3. 2 문서 정규화 방법

불용어를 제거한 후에 단어의 가중치를 식(3)과 같이 계산한다. 이때 tf_{ij}는 문서내에서 단어의 출현 빈도를 의미하고, idf_i는 전체 문서에서 단어가 출현한 문서수의 역수를 의미한다.

$$w_{ij} = tf_{ij} \times idf_i \quad (3)$$

각 문서는 포함되는 단어의 수와 문서의 길이가 상이하기 때문에 식(4)와 같이 코사인 정규화요소로 나누어 정규화 하는 과정을 수행한다.

$$\sqrt{w_1^2 + w_2^2 + \ldots + w_t^2} \quad (4)$$

4. 계층적 집적 문서 클러스터링 방법

각 문서들에 대해 불용어 제거와 가중치 산출및 문서 정규화 과정을 수행한 후 문서간의 유사도를 비교하기 위하여 문서의 특성을 벡터 값으로 표현 한다. 본 논문에서는 계층적 중복 문서 클러스터링[6]에서 사용된 유사도와 응집도, 참여도의 개념을 사용한다.

정의 3. 문서간의 유사도

문서간의 유사도는 두 문서에서 공통 단어의 가중치합을 문서 전체의 가중치합으로 나누어준 값으로

다음과 같이 정의한다.

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

문서간의 유사도를 바탕으로 클러스터에 포함된 모든 문서쌍의 유사도를 평균낸 값을 응집도라고 하며 다음과 같이 정의한다.

정의 4. 응집도

응집도는 클러스터에 속하는 문서들이 단단하게 결합 되는지의 여부를 측정하는 척도로 다음과 같다.

$$\alpha(C_u) = \frac{\sum_{d_i \in C_u} \left(\sum_{d_i \in C_u - \{d_i\}} s(d_i, d_j) \right)}{|C_u| C_2}$$

정의 5.참여도

클러스터의 참여도는 두 클러스터 간의 공통 단어의 웨이트 합을 클러스터의 키워드 웨이트 합으로 나눠 준 값이 된다. 참여도는 결합되는 두 클러스터 간의 유사도를 나타내는 척도가 된다.

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

정의4와 정의5의 응집도와 참여도를 바탕으로 문서 클러스터링을 수행하는 알고리즘은 다음과 같다.

[단계 1] 도메인의 각 문서를 하나의 클러스터로 취급하여 모든 문서쌍간의 응집도를 비교한다.

[단계 2] 응집도 값이 최대인 쌍부터 시작하여 응집도 값이 최소 응집도 이상이며, 두 클러스터 간의 참여도를 계산하여 최소 참여도 이상의 값을 가지면 클러스터로 결합한다.

[단계 3] 결합된 클러스터와 다른 클러스터간의 응집도를 계산한다.

[단계 4] 단계 2와 3을 더이상 결합되는 클러스터가 없을 때까지 반복 수행한다.

[단계 5] 생성된 클러스터와 결합되지 않은 문서들간의 참여도를 계산한다.

[단계 6] 계산된 참여도의 값이 가장 큰 클러스터와 문서간의 응집도를 계산하여 최소 응집도 이상일 경우 클러스터로 결합하는 작업을 반복한다.

[단계 7] 생성된 클러스터들을 하나의 도메인으로 매핑하여, 생성된 각 도메인에 대해 도메인 불용어 제거와 키워드 가중치 산정, 계층적 집적 문서클러스터링 과정을 반복한다.

[단계 1~4]의 과정은 문서 클러스터링을 위한 기본 클러스터를 만들어 내는 단계로 작고 밀집된 클러스터를 만들어 내지만, 결합되지 않은 문서가 많이 존재한다. 따라서 생성된 기본 클러스터에 결합되지 않은 문서를 결합하기 위하여 [단계 5~6]을 수행하는 혼합적인 클러스터링 방법을 사용한다.

5. 실험

본 논문에서는 실험의 정확성을 위해 가장 널리 알려진 YAHOO![7]에서 제공하는 카테고리 서비스를 대상으로 실험을 수행하였으며, 실험 데이터의 특성은 표 1과 같다.

표 1. 실험 데이터 특성

도메인의 수	93
도메인 별 평균 문서 수	21.6
전체 단어수	974996
unique 한 전체 단어수	38663
도메인의 최대 깊이	3

알고리즘을 수행하기 위해 도메인의 깊이가 1인 세 개의 카테고리(News, Business, Sports)를 미리 정의하여 불용어 제거와 문서클러스터링 실험을 수행하며, 카테고리를 기준으로 생성된 클러스터들의 유사성을 각 레벨 별로 비교한다.

실험에 사용된 척도는 CSIM척도 외에 정보 검색 분야에서 널리 사용되는 정확도(Precision)과 재현율(Recall)[8]을 사용하여 효율성을 평가하는 방법을 사용한다. 정확도는 분류된 문서중에서 적합한 문서의 수를 나타낸 값으로, 부적합한 문서를 찾지 않는 능력을 나타내며, 재현율은 전체 적합한 문서중에서 올바르게 분류된 문서수의 값으로서 적합한 문서를 검색해 내는 능력을 측정한다. 따라서 정확성과 재현율값의 상호 관계는 다음과 같다.

표 2 정확성-재현율 간의 상호 관계

정확성	재현율	생성되는 클러스터의 특징
Low	Low	정확성이 낮은 클러스터가 생성된다.
High	Low	클러스터가 크기가 커지고, 불필요한 문서가 다량으로 포함된 경우.
Low	High	클러스터의 크기가 작고, 묶여야 할 문서를 찾지 못한 경우
High	High	의미적으로 정확한 클러스터가 생성된 경우.

두 값은 상호 보완적인 관계를 가지므로 식(5)와 같은 두 값을 조합한 정확성과 재현율 분기점(Precision-Recall Break Even Point)을 사용하여 문서 클러스터링의 정확성을 측정한다. 이하에서정확성과 재현율 분기점을 BEP로 표현한다.

$$BEP = \frac{2 \times \text{정확율} \times \text{재현율}}{\text{정확율} + \text{재현율}} \quad (5)$$

각각의 레벨에 대해 문서 클러스터링의 정확성을 측정한 실험은 다음과 같다. 불용어를 제거하는 최소 도메인지지도와 최소 문서지지도, 표준 편차의 값은 각각 0.8, 0.5, 0.1로 정하였으며, 이 값은 레벨1에서 불용어를 판별한 값중에서 정확성이 가장 높을 경우의 값이다. 그림 2는 도메인 레벨 2에서의 BEP 값을 나타낸 것으로 최소 참여도 값이 0.6 최소 응집도 값이 0.1일 경우 가장 높은 정확성을 나타내었다. 응집도는 클러스터의 결합 정도를 나타내므로 레벨 2에서는 클러스터를 느슨하게 구성 할 수록 더 높은 정확성을 보인다. 응집도가 낮을 수록 클러스

터들은 분리되지 않고, 결합될 가능성이 높게 되며, 레벨의 깊이가 낮을 수록 도메인은 많은 수의 문서를 포함하게 되므로, 낮은 최소 응집도를 가질 수록 정확한 클러스터를 만들게 된다.

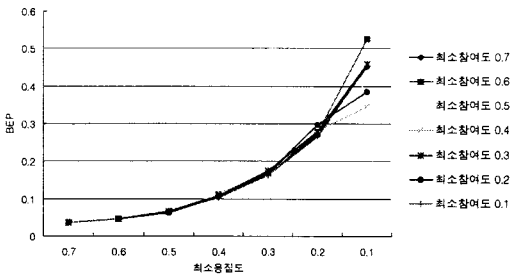

그림 2 도메인 레벨 2에서의 BEP

다음의 그림3은 도메인 레벨 3에서의 BEP값으로서 최소 참여도 값이 0.7, 최소 응집도 값이 0.2일때 가장 높은 정확성을 보여 주었다. 최소 응집도 값이 0.1일 경우 정확성이 크게 떨어지며, 이는 분리 되어야 할 여러 클러스터들이 결합되어 정확성이 낮아지게된다. 참여도의 값이 높을 수록 더 높은 정확성을 보이며, 이는 기본 클러스터의 구성 단계에서 정확한 클러스터가 생성될수록 높은 성능을 보임을 나타낸다. 도메인의 레벨이 깊어질수록 정확성이 높은 최소응집도의 값이 낮아 지며, 이는 의미적으로 명확하게 클러스터가 분류되었음을 나타낸다.

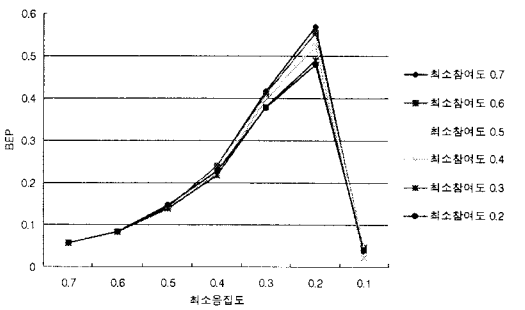

그림 3 도메인 레벨 3에서의 BEP

그림 4는 일반적인 클러스터링 방법과의 실험을 비교한 결과이다. 실험결과 단일연결 방법(SL)과 Ward의 방법(WARD)은 비슷한 성능을 나타내었으며, 본 논문에서 제안한 알고리즘(DHODC)와 집단 연결 방법(GAL)의 정확성이 높게 나왔다. 이는 논문에서 제안된 응집도의 개념과 집단 평균 연결방법의 유사도 개념이 유사하기 때문으로 보이며, 참여도를 고려한 본 알고리즘의 정확성이 높다는 것을 알 수 있다.

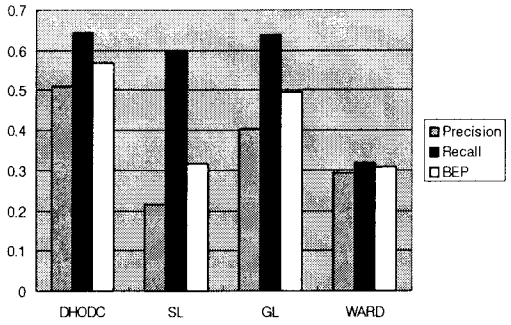

그림 4 일반적일 클러스터링 방법과의 비교

5. 결론 및 향후 연구

본 논문에서는 문서 클러스터링의 정확도를 높이기 위한 불용어 제거 방법과 응집도와 참여도를 고려한 문서 클러스터링 방법을 제안하였다. 제안된 도메인별 불용어 제거 알고리즘의 성능을 시험하기 위해 도메인 지지도와 문서지지도의 변화에 따른 정확도 변화 실험을 수행하였으며, 응집도와 참여도를 고려한 클러스터링 방법의 실험을 위해 일반적으로 쓰이는 문서 클러스터링 알고리즘과 비교 실험을 수행하였고, 도메인 레벨의 깊이가 깊을 수록 다른 알고리즘에 비해 높은 정확성을 보여 주었다. 본 논문에서 제안하는 알고리즘은 문서의 집합이 크게 변경될 경우 문서 클러스터링을 재수행해야 하기때문에 문서의 수가 증가함에 따라 효율적으로 대처할 수 있는 점진적 문서 클러스터링방법 에 관한 연구가 진행되어야 할 것이다.

참고문헌

[1] C.J. van Rijsbergen "Information Retrieval"

[2] Salton, G., Automatic Text Processing, Addison-Welsley Publishing Company, 1989.

[3] I.Aalbersberg, " A Document Retrieval Model based on Term Frequency Ranks", 17th Intern-ational ACM SIGIR Conference on Research and Development in Information Retrieval,163-172, 1994

[4] Cutting, D.R., Karger, D.R., Perderson, J.O., Tukey, J.W., "Scatter/Gather: a cluster-based approach to browsing large document collections," SIGIR'92, 318-329, 1992.

[5] 한승희, 이재윤. 1999. 문헌 클러스터링을 위한 유사계수간의 연관성 측정. 제 6회 정보 관리 학회 학술대회 논문집, 25-28

[6] 강동혁, 주길홍, 이원석 대용량 문서 데이터베이스를 위한 효율적인 점진적 문서 클러스터링 기법 정보처리학회 논문지 제10-D권 제1호 2003.2

[7] Yahoo! http://www.yahoo.com

[8] Aixin Sun, Ee-Peng Lim "Hierarchical Text Classification and Evaluation" ICDM2001 521-528 2001