온톨로지 개념 합병 기반 문서 군집화 기법

관향동, 김우생
광운대학교 컴퓨터과학과

Text Clustering Algorithm Based on Ontology Concepts Combination

XiangDong Guan, Woosaeng Kim
Department of Computer Science, Kwangwoon University
nicholas_gem@msn.com, kwsrain@kw.ac.kr

요 약

문서 군집화를 통하여 문서를 효율적으로 조직, 관리, 검색 할 수 있다. 일반적으로 문서 군집화는 많은 단어와 개념들을 포함하고 있기 때문에 차원이 큰 벡터 공간 모델에서 군집화를 수행한다. 본 논문에서 문서 집합에 대응하는 온톨로지를 이용하여 문서 벡터 공간의 차원을 줄여 효율적으로 군집화하는 방법을 제안하고, 실험을 통하여 기존 방법보다 우수함을 보인다.

Ⅰ. 서 론

인터넷의 대중화로 많은 사용자가 온라인으로 문서를 검색한다. 문서 검색과 관련해 문서 분류나 문서 군집화는 정보 검색 시스템에서 방대한 양의 문서들을 구조화하는데 중요한 역할을 담당하고 있다. 근래에 특정 도메인의 개념들을 계층화하고 관계를 모델링 한 온톨로지를 사용해 문서를 분류하거나 군집화 하는 연구들이 진행되고 있다[1,2,3,4]

일반적으로 문서 군집화는 많은 단어와 개념들을 포함하고 있기 때문에 차원이 큰 벡터 공간에서 군집화를 하게 된다. 본 연구에서는 문서 집합에 대응하는 온톨로지에서 핵심 개념들을 추출해 벡터 공간의 차원을 줄여 군집화의 효율성을 높이는, 온톨로지 개념 합병 (Ontology Concepts Combination, OCC) 기반 문서 군집화 기법을 제안한다. 본 논문의 구성은 2장에서 온톨로지 개념 합병 기반 문서 군집화 기법을 설명하고, 3장에서 제안한 알고리즘을 실험을 통해 평가하고 4장에서 결론을 낸다.

Ⅱ. 온톨로지 개념 합병 알고리즘

논문에서는 문서 집합에서 핵심 개념들을 찾기 위해, 문서에서의 한 개념의 중요도를 나타내는 척도로 기존의 tf/idf를 확장한 CF-IDF (Concept Frequency - Inverse Document Frequency)를 사용한다[2]. 개념 c가 문서 d_i에 대한 빈도 $cf_{i,c}$는 식 (1)과 같다. 여기서 $n_{i,c}$는 개념 c가 문서 d_i중에 나타나는 횟수이고 $\sum n_i$ 는 문서 d_i의 총 개념수이다.

$$cf_{i,c} = \frac{n_{i,c}}{\sum n_i} \quad (1)$$

개념 c가 문서 세트 D에 대한 역문서 빈도 idf_c는 식 (2)와 같다. 여기서 N은 총 문서의 수이고 $|\{d : c \in d\}|$ 는 개념 c를 포함하는 문서들의 수이다.

$$idf_c = \log \frac{N}{|\{d : c \in d\}|} \quad (2)$$

따라서 개념 c가 문서 d_i에 대한 CF-IDF 값은 식 (3)과 같다.

$$cfidf_{i,c} = cf_{i,c} \times idf_c \quad (3)$$

CF-IDF 값은 한 문서의 특정 개념에 대한 중요 정도를 나타내므로, CF-IDF 값들의 합은 전체 문서의 특정 개념에 대한 중요 정도를 나타낸다. 따라서 각 개념의 CD-IDF 합을 비교하면 전체 문서에서 중요하거나 중요하지 않은 개념들을 찾을 수 있다. 또한 온톨로지 상에서 한 개념이 특정 문서에서 나오면 그것의 상위 개념도 특정 문서에

서 나오는 것을 의미한다. 따라서 온톨로지 상의 개념의 숫자를 줄이기 위해 온톨로지의 리프 노드들 중에서 CF-IDF 합이 가장 작은 개념을 상위 노드와 합병한다. 이처럼CD-IDF 합이 작은 리프 노드를 상위 노드와 합하는 과정을 반복하면 핵심 개념들로 구성된 핵심 온톨로지를 얻을 수 있다. 다음의 예를 통해 온톨로지 개념 합병을 알고리즘을 설명한다. 어떤 문서 집합에 대응하는 온톨로지가 그림 1과 같다고 가정할 때, 전체 문서는 총 6개 개념을 포함하고 있으므로 대응하는 벡터 공간은 6차원이 된다. OCC를 통하여 이를 3차원으로 축소하는 방법을 설명한다.

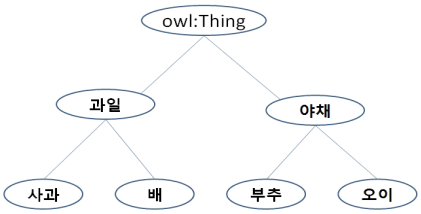

그림 1. 한 온톨로지의 예

그림 1에서 '사과'와 '배'의 상위개념은 '과일'이며, '부추'와 '오이'의 상위개념은 '야채'이다. 전체 문서에 총 6 개(A, B, C, D, E, F)의 문서가 있다고 가정할 때, 각 개념이 각 문서에 나온 횟수는 표 1의 각행 위에 표시하고, CF-IDF 값은 각 행 밑에 표시하였다. 표 맨 밑에는 각 개념의 CF-IDF 합을 기록하였다. 각 개념의 CF-IDF 합을 온톨로지에 표시하면 그림 2와 같다.

표 1. CF-IDF 값 표

	과일	사과	배	야채	부추	오이
A		2				
		0.447				
B	1	1	1			
	0.100	0.159	0.100			
C	2		3			
	0.120		0.181			
D	1		1	3		
	0.060		0.060	0.286		
E				3	2	
				0.286	0.311	
F						4
						0.778
합	0.280	0.606	0.341	0.572	0.311	0.778

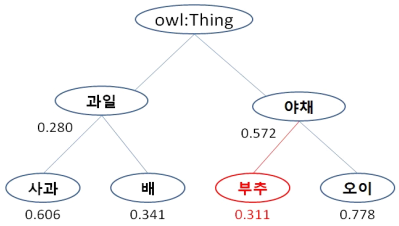

그림 2. CF-IDF 합이 있는 온톨로지

그림 2의 리프 노드 중에서 CF-IDF 값이 가장 작은 것은 '부추'이니까 '부추'를 그의 상위개념 '야채'로 합병시킨 결과는 그림 3과 같다.

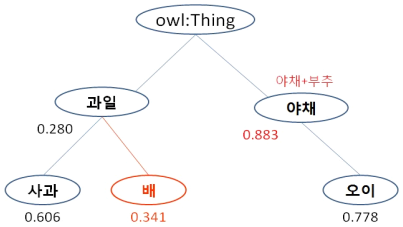

그림 3. 개념 합병 과정

다음으로 리프 노드 중에서 CF-IDF 값이 가장 작은 '배'를 상위 개념인 과일과 합병시킨다. 3 개의 개념만 남을 때까지 반복하면 결과는 그림 4와 같다. 그림 4는 핵심 개념을 3개 포함하는 핵심 온톨로지이며 3차원 벡터 공간에 대응 된다. 3개의 핵심 개념들을 갖고 CF-IDF 값을 다시 계산한 결과는 표 2와 같다.

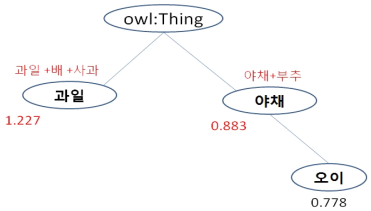

그림 4. 개념 합병 결과

표 2. 핵심개념들의 CF-IDF 값

	과일	야채	오이
공식	과일 +배 +사과	야채+부추	오이
A	0.447		
B	0.359		
C	0.301		
D	0.120	0.286	
E		0.597	
F			0.778

이제 전체 문서를 k-means 알고리즘을 이용하여 세 개 군집으로 분류할 수 있다. 표 2를 참조하여 각 문서에 대응하는 벡터의 좌표를 구한다. 예를 들면 문서A의 벡터 좌표는 (0.447, 0, 0)이고, 문서B는 (0.359, 0, 0)이다. 그림 5는 각 벡터에 대응하는 점의 공간 분포와 k-means로 군집화한 결과를 보여 준다.

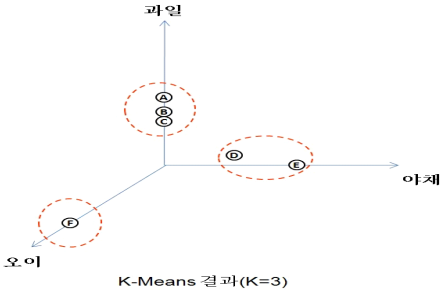

그림 5. 군집화 결과

III. 실험 및 성능 평가

본 논문에서 사용한 온톨로지는 총 2636개 개념을 포함한 생물 의학 연구 온톨로지(The Ontology for Biomedical Investigations, OBI)이다[5]. 그림 6은 OCC알고리즘을 통해 얻은 15 개 핵심 개념들로만 구성 된 핵심 온톨로지이다.

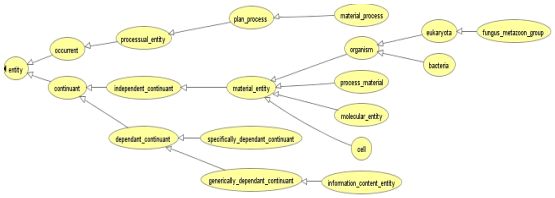

그림 6. 15-차원 공간의 핵심 온톨로지

성능 비교를 위하여 OCC와 COSA[1] 알고리즘으로 생물 의학 연구 온톨로지를 15차원으로 축소해 군집화한 결과를 실루엣 계수와 평균 제곱 오차의 척도로 비교하였다. 실루엣 계수는 군집의 질을 측정하는 방법으로, 계수의 값이 0.7 이상이면 군집화가 제대로 되었음을 보여 준다[6]. 그림 7은 k를 3으로 할 때 실루엣 계수가 차원을 따라 변화하는 함수를 보인다. 모든 차원에서 OCC가 COSA보다 좋음을 알 수 있다. 그리고 실루엣 계수가 0.85까지 가는 것을 보면 문서들이 제대로 군집화 된 것을 알 수 있다.

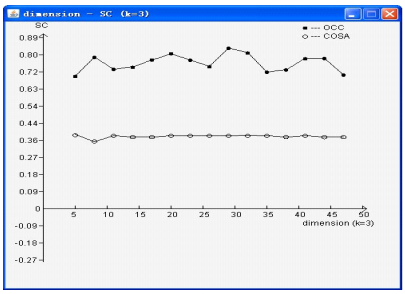

그림 7. k는 3일 때 dimension-SC 함수도

그림 8은 k를 3으로 고정할 때 차원을 따라 변화하는 평균 제곱 오차를 보여 준다. OCC는 전체적으로 COSA 보다 MSE가 적음을 보여 준다. 그것은 OCC가 좀 더 좋은 핵심 개념들을 사용해서 군집화를 수행하기 때문이다.

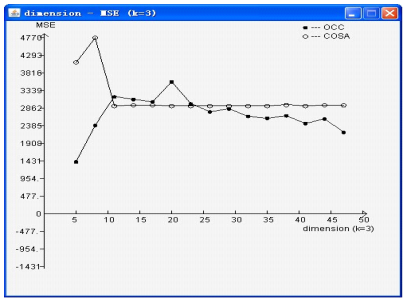

그림 8. k는 3일 때 dimension-MSE 함수도

IV. 결 론

온톨로지를 사용하여 문서를 군집화하는 방법들이 연구되고 있다. OCC 알고리즘은 문서 집합의 핵심 개념을 온톨로지를 이용해서 핵심 개념들을 찾아 낸 후 축소된 벡터 공간에서 군집화를 시도한다. 실험을 통하여 OCC알고리즘의 성능을 우수함을 보였다.

참고문헌

[1] A.Hotho, A.Maedche, and S.Staab. "Ontology -Based Text Clustering," Proceedings of the IJCAI-001 Workshop "Text Learning: Beyond Supervision," 2001.8.

[2] M.H.Pham, D.Bernhard, G.Diallo, et al. SOM-based Clustering of Multilingual Documents Using an Ontology, Data Mining with Ontologies: Implementations, Findings and Frameworks , 2007

[3] 문헌정, 우용태. "지식문서에서 도메인 온톨로지를 이용한 개념 추출 기법", 한국정보처리학회 논문지, 2006.

[4] 박은진, 김재훈, 옥철영, "온톨로지를 이용한 단어 군집화 성능 개선", 정보처리학회 논문지 B, 2006.

[5] http://obi-ontology.org/page/Main_Page

[6] L.Kaufman and P.J.Rousseeuw, "Finding Groups in Data: An Introduction to CLuster Analusis", Wiley, New York, 1990.