J. Inf. Technol. Appl. Manag. 23(2): 143~152, June 2016
http://dx.doi.org/10.21219/jitam.2016.23.2.143

ISSN 1598-6284 (Print)
ISSN 2508-1209 (Online)

도메인 온톨로지에 의한 문서 군집화 기법

김우생* · 관향동**

Document Clustering Technique by Domain Ontology

Woosaeng Kim* · Xiang-Dong Guan**

Abstract

We can organize, manage, search, and process the documents efficiently by a document clustering. In general, the documents are clustered in a high dimensional feature space because the documents consist of many terms. In this paper, we propose a new method to cluster the documents efficiently in a low dimensional feature space by finding the core concepts from a domain ontology corresponding to the particular area documents. The experiment shows that our clustering method has a good performance.

Keywords : Document Clustering, Ontology

Received : 2016. 04. 08. Revised : 2016. 05. 24. Final Acceptance : 2016. 06. 24.

※ The work reported in this paper was conducted during the sabbatical year of Kwangwoon University in 2015.

* Coressponding Author, Professor, Department of Computer Science, Kwangwoon University, Computer Software Department, 20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea, Tel : +82-2-940-5217, e-mail : kwsrain@gmail.com
** Department of Computer Science, Kwangwoon University, e-mail : nicholas_gem@msn.com

1. 서 론

문서들에 대한 군집화는 유사한 문서들의 그룹을 만들어 구조화하고 검색과 관리와 처리를 용이하게 할 수 있다. 특히 웹과 SNS(Social Network Service) 등의 소셜 미디어 환경에서 문서들이 폭발적으로 증가함으로 문서 군집화에 대한 필요성이 증가하고 있다.

전통적인 문서 군집화는 문서를 용어들의 집합 즉, 용어 벡터(term vector)로 표현하고, 이들 간의 거리를 유사도 척도로 사용하여 군집화하는 방법을 주로 사용하고 있다. 이러한 방법은 문서 집합에 포함된 용어들의 의미적 관계를 고려하지 않고, 단지 문서에 포함된 용어들의 빈도만을 주로 이용하고 있다[Hu et al., 2009]. 또한 문서들은 많은 용어들을 포함하고 있기 때문에 높은 차원의 특징 공간(feature space)에서 군집화를 수행해야 된다.

온톨로지는 지식을 표현하기 위한 개념(concept) 들과 개념간의 관계를 통하여 개념에 대한 의미를 정의하고 구조화하기 위한 데이터 모델로, 최근의 문서 군집화 기법은 이러한 온톨로지를 이용하거나 문서 집합의 내부 구조를 나타내는 의미 특징을 많이 사용하는 추세이다.

온톨로지를 사용하면 용어 보다는 적은 숫자의 개념들로 문서를 표현할 수 있으나 여전히 많은 수의 개념들을 처리해야 하는 문제점이 있다. 따라서 본 연구에서는 특정 영역의 문서들에 대응하는 도메인 온톨로지(domain ontology)에서 중요한 개념들을 자동으로 추출해 이를 기반으로 낮은 차원의 특징 공간에서 군집화를 효율적으로 수행할 수 있는 방법을 제안한다. 또한 제안하는 기법은 핵심 개념들을 통한 각 군집의 주제를 쉽게 파악할 수 있으며 이를 통해 응용에 적절한 군집의 숫자도 결정할 수 있는 방법이다.

본 논문의 구성은 제 2장에서 관련 연구, 제 3장에서 도메인 온톨로지에 기반한 문서 군집화 기법을 설명하고, 제 4장에서는 기존 방법과의 성능을 비교하고, 제 5장에서 결론을 낸다.

2. 관련 연구

온톨로지란 특정 분야를 기술하는 데이터 모델로서 특정한 분야에 속하는 개념과, 개념 사이의 관계를 기술하는 정형 어휘의 집합으로 이루어진다. 먼저 개념은 여러 관념 속에서 공통된 요소를 추출하고 분석하여 얻은 하나의 공통 관념이라 정의 할 수 있고, 관계는 시소러스(thesaurus)에서 일반적으로 사용되는 상하, 동의, 유의, 부분 관계 등의 의미 관계 유형과 넓은 의미 관계인 구성원 관계, 위치 관계 등의 개념 관계 유형이 있다. 이러한 관계는 계층적 구조나 네트워크 구조를 형성하게 된다[Choi and Ok, 2004; Choi et al., 2006].

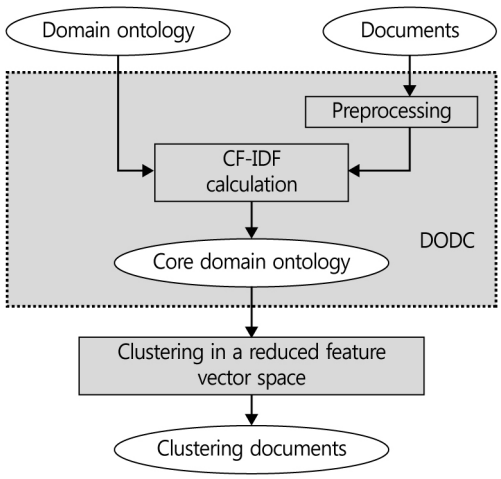
차세대 웹으로 불리는 시맨틱 웹(semantic web)은 온톨로지를 기반으로 하기 때문에, 시맨틱 웹의 연구와 더불어 다양한 분야의 도메인 온톨로지가 구축되고 있으며[Ra et al., 2012; Son et al., 2010; Jo and Kim, 2013; Hwang et al., 2012; Hwang et al., 2005] 또한 온톨로지를 자동 또는 반자동으로 구축하고자 하는 연구도 활발히 진행되고 있다[Kong et al., 2005; Min and Lee, 2008; Bae et al., 2007]. 시맨틱 웹과 온톨로지 관련 기술에 대한 연구 개발은 온톨로지와 유사한 구조를 가진 관련 결과물의 연구까지 확대되어, 기존의 시소러스, 어휘 의미망 등이 온톨로지로 활용되는 것은 온톨로지의 내부 구조 속에 포함되는 개념성, 관계성, 속성 등의 몇몇 특성이 시소러스나 어휘 의미망 등과 일치하는 부분들이 많기 때문이다. 그리하여 새로운 온톨로지를 생성 구축하는 연구 이상으로 기존의 WordNet, UMLS 등과 같은 연구 결과물을 기반으로 상위 수준의 온

톨로지나 특정 영역의 온톨로지 등으로 활용하는 연구도 국내외적으로 진행되고 있다[Mun and Woo, 2006; Kim and Choi, 2014; Jo and Lee, 2015].

근래에는 이러한 온톨로지의 다양한 방면의 연구와 더불어, 온톨로지의 개념이나 개념들 간의 관계를 활용하여 문서를 분류하거나 군집화하는 연구들도 진행되고 있다. 군집의 중요 용어와 위키피디아를 이용해 문서 군집을 향상 시키는 연구[Park et al., 2012], 의미 특징 기반의 용어 가중치 재 산정을 이용한 문서 군집의 성능 향상에 관한 연구[Park et al., 2013], 온톨로지를 이용해서 고객 서비스 관리 분야에서 고객과 요청을 군집화하는 연구[Smirnov et al., 2005], 유전자 온톨로지를 이용해서 군집화에 기반한 유전자 표현 분석에 대한 연구[Wang et al., 2005] 등이 있다. 반면에 COSA(concept selection and aggregation) 기법은 특정 문서들에 대한 도메인 온톨로지에서 중요하지 않은 개념들을 제거하여 특징 공간의 차원의 수를 줄인 후 군집화를 수행한다[Hotho et al., 2011]. COSA 기법이 기존의 SiVeR, TES보다 군집화 결과가 우수함을 보였지만, 이 방법은 한 개념의 중요도를 계산할 때 문서들에서 개념의 빈도만을 고려한 후 온톨로지의 각 레벨에서 우선순위가 낮은 개념들을 줄여 나가기 때문에 최적화된 해를 구하기 힘든 문제점을 갖고 있다.

3. 도메인 온톨로지 문서 군집화(Domain Ontology Document Clustering : DODC)

시스템 전체 구조는 <Figure 1>과 같다. 군집화하고자 하는 특정 영역의 문서들과 이와 연관된 도메인 온톨로지가 있을 때, 도메인 온톨로지로부터 핵심 도메인 온톨로지를 구성해, 이를 기반으로 축소된 특징 공간에서 문서 군집화를 수행한다.



〈Figure 1〉 System's Overall Structure

문서 전처리를 통해 먼저 문서들에서 도메인 온톨로지의 개념들을 찾는다. 이를 위해 문서들에 대해 불용어 제거,[1] 어간 추출,[2] 동의어 처리[3] 등을 수행 한다.

그러나 도메인 온톨로지의 개념들은 여전히 많기 때문에 이들 중에서 중요한 개념들만을 추출할 필요가 있다. 따라서 문서들에서 찾은 특정 개념의 중요도를 나타내는 척도로 기존의 TF-IDF(Text Frequency-Inverse Document Frequency)를 확장한 CF-IDF(Concept Frequency-Inverse Document Frequency)를 사용한다[Snasel et al., 2005]. 개념 c가 문서 d_i에 대한 빈도 $cf_{i,c}$는 식 (1)과 같다. 여기서 c가 문서 d_i에 대한 빈도 $cf_{i,c}$는 식 (1)과 같다. 여기서 $n_{i,c}$는 개념 c가 문서 d_i 중에 나타나는 횟수이고 $\sum n_i$는 전체 문서의 총 개념수이다.

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

1) https://lucene.apache.org/core.
2) https://lsg3.nlm.nih.gov/LexSysGroup/Summary/lexicon.html.
3) http://lyle.smu.edu/~tspell/jaws.

개념 c의 전체 문서에 대한 역 문서 빈도 idf_i는 식 (2)와 같다. 여기서 N은 전체 문서의 수이고 $|\{d:c \in d\}|$는 개념 c를 포함하는 문서들의 수이다.

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

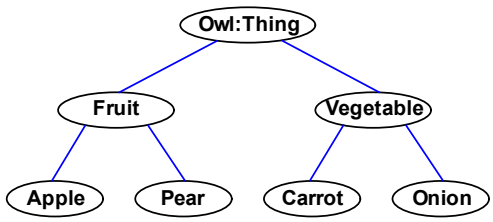
개념 c의 문서 d_i에 대한 CF-IDF 는 식 (3)과 같다. CF-IDF 값은 한 문서의 특정 개념에 대한 중요 정도를 나타내기 때문에, CF-IDF 값들의 합은 전체 문서의 특정 개념에 대한 중요 정도를 나타낸다.

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

도메인 온톨로지에서 한 개념이 특정 문서에서 나오면 그것의 상위 개념도 특정 문서에 포함된다고 볼 수 있다. 따라서 도메인 온톨로지의 리프 노드(개념)들 중에서 CF-IDF 합이 가장 작은 노드를 상위 노드와 합병함으로써 노드들의 즉, 개념들의 숫자를 줄일 수 있다. 이처럼 CD-IDF 합이 가장 작은 리프 노드를 상위 노드와 합병하는 과정을 반복하면 핵심 도메인 온톨로지를 구할 수 있으며, 이를 기반으로 축소된 특징 공간에서 전체 문서를 군집화 한다.

다음의 예를 통해 본 논문에서 제안하는 DODC 방법을 설명한다. 어떤 특정 영역의 문서들에 대한 도메인 온톨로지가 <Figure 2>와 같다고 가정할 때, 전체 문서는 총 6개의 개념을 포함하고 있으므로 대응하는 특징 공간은 6차원이 된다. DODC를 통하여 이를 3개 개념들만 포함된 핵심 도메인 온톨로지로 즉, 3차원 특징 공간으로 변환하고자 한다.

<Figure 2>에서 '사과'와 '배'의 상위개념은 '과일'이며, '당근'과 '양파'의 상위개념은 '야채'이다. 전체 문서가 6개(A, B, C, D, E, F)라고 가정할 때, 각 개념이 각 문서에 나온 횟수는 <Table 1>의 각 행의 위에 표시하고, 각 개념의 문서에 대한 CF-IDF 값은 각 행의 아래에 표시한다. <Table 1> 맨 아래에는 각 개념의 전체 문서에 대한 CF-IDF 값을 표시한다.

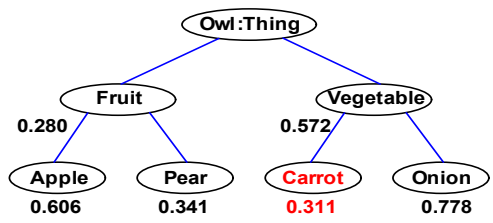


〈Figure 2〉 Example of Domain Ontology

〈Table 1〉 CF-IDF Value

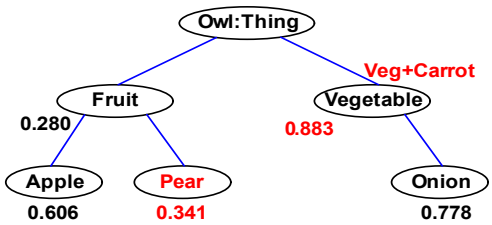
	Fruit	Apple	Pear	Veg	Carrot	Onion
A		2				
		0.447				
B	1	1	1			
	0.100	0.159	0.100			
C	2		3			
	0.120		0.181			
D	1		1	3		
	0.060		0.06	0.286		
E				3	2	
				0.286	0.311	
F						4
						0.778
T	0.280	0.606	0.341	0.572	0.311	0.778

각 개념의 전체 문서에 대한 CF-IDF 값을 도메인 온톨로지의 각 노드에 표시하면 <Figure 3>과 같다.



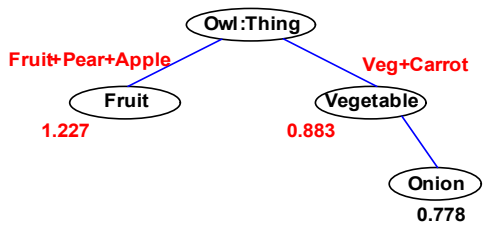
〈Figure 3〉 Ontology Marked with CF-IDF Sum

<Figure 3>의 리프 노드 중에서 전체 문서에 대한 CF-IDF 값이 가장 작은 것은 '당근'이니까, '당근'을 그의 상위 개념 '야채'로 합병시킨 도메인 온톨로지는 <Figure 4>와 같다.



〈Figure 4〉 Process of Merger between Concepts

다음으로 리프 노드 중에서 전체 문서에 대한 CF-IDF 값이 가장 작은 '배'를 상위 개념인 과일과 합병시킨다. 이처럼 3개의 개념만 남을 때까지 노드 간의 합병을 반복하면 결과는 <Figure 5>와 같다. <Figure 5>는 3개의 개념만을 포함하는 핵심 도메인 온톨로지로서 3차원 특징 공간에 대응 된다.



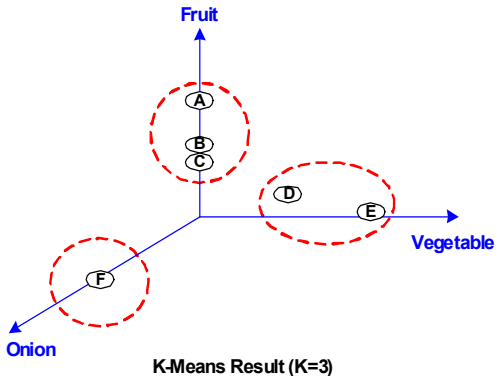
〈Figure 5〉 Core Domain Ontology

이제 전체 문서를 세 개의 군집으로 분류한다고 가정할 때, 3개의 핵심 개념들로 각 문서에 대한 CF-IDF 값을 다시 계산한 결과는 <Table 2>와 같다. <Table 2>를 참조하여 3차원 특징 공간에서 각 문서에 대응하는 특징 벡터를 구하면, 문서 A는(0.447, 0, 0)이고, 문서 B는(0.359, 0, 0)이며, 나머지 문서들도 같은 방법으로 대응하는 특징 벡터를 구할 수 있다.

〈Table 2〉 CF-IDF Value of the Core Concepts

	Fruit	Vegetable	Onion
Fomula	Fruit+Pear+Apple	Veg+Carrot	Onion
A	0.447		
B	0.359		
C	0.301		
D	0.120	0.286	
E		0.597	
F			0.778

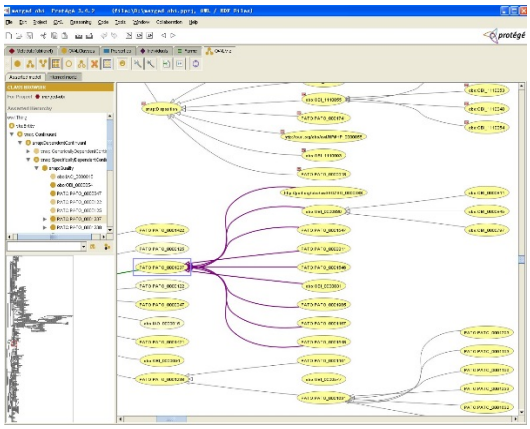
<Figure 6>은 3차원 특징 공간에서 전체 문서들의 특징 벡터 분포와, 이들을 k-means 군집화 방식에서 k = 3으로 군집화 한 결과를 보여 준다. 특징 공간이 6차원에서 3차원으로 줄어, 개념들 간의 관계가 좀 더 명확해 졌음을 알 수 있고, 또한 각 군집의 주제를 핵심 개념들을 통해 쉽게 파악할 수 있다. <Figure 6>에서 '과일' '양파'와 '야채'는 각 군집의 주제가 된다. k-means 등 기존 군집화 기법의 문제점은 적절한 군집의 숫자를 구하는 방법이 없다는 점이다. 그러나 DODC 기법은 초기의 k값에 의한 핵심 개념들을 알 수 있기 때문에, 응용에 필요한 개념들이 결과에 포함 되어 있는지 또는 없는지에 따라 k값을 조절하여 가장 적절한 군집의 숫자도 결정할 수 있는 방법이다.



〈Figure 6〉 Clustering Result in a 3 Dimensional Feature Vector Space

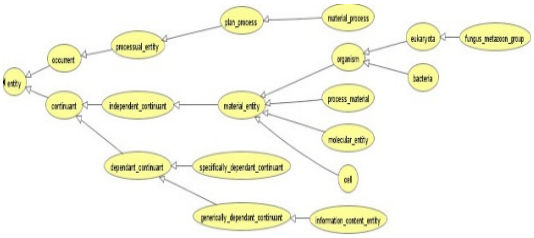
3. 실험 및 성능 평가

본 논문에서 군집화를 위해 사용한 특정 영역 문서 들은 BioMed Central 사이트[4]의 생물 의학 문서들이며, 이와 연관된 도메인 온톨로지는 총 2,636개의 개념을 포함한 생물 의학 연구 온톨로지(The Ontology for Biomedical Investigations : OBI[5])이다. Protege API[6]를 사용해 온톨로지를 처리하였고, 이를 통한 OBI 온톨로지의 일부분은 <Figure 7>과 같다.



〈Figure 7〉 Part of OBI Ontology

<Figure 8>은 DODC를 통해 19개의 개념들로 구성된 핵심 생물 의학 연구 온톨로지이다.



〈Figure 8〉 Core Ontology

4) http://www.biomedcentral.com.
5) http://obi-ontology.org/page.
6) http://protegewiki.stanford.edu/wiki/ProtegeOWL_API_Programmers_Guide.

DODC의 성능을 평가하기 위해 생물 의학 연구 온톨로지를 대상으로, DODC와 COSA의 실루엣 계수와 평균 제곱 오차를 구하여 비교하였다[Kaufman and Rousseeuw, 1990]. 우선 실루엣 계수(Silhouette Coefficient)는 군집의 수와는 무관하게 군집화 질을 측정하는 척도로 계수의 값이 커지면 군집들이 잘 분리된 것을 나타내며, 반대로 계수의 값이 작아지면 군집들이 제대로 분리 되지 못한 것을 나타낸다. 실험치에 의하면 실루엣 계수 값이 0.7에서 1.0사이에는 군집들이 잘 분리되며, 0.25 이하는 의미 있는 군집들이 이루어지지 않았다. 실루엣 계수는 다음과 같이 정의된다.

정의 1. 실루엣 계수
최대 가능한 군집화 결과를 $D_M = \{\overline{D_1}, \cdots, \overline{D_k}\}$로 가정할 때, 한 문서 $d \in D$부터 군집 $\overline{D_l} \in D_M$까지 거리는 식 (4)와 같다.

$$dist(d, \overline{D_l}) = \frac{\sum_{p \in \overline{D_l}} dist(d, p)}{|\overline{D_l}|} \quad (4)$$

문서 $d \in D$에서 그의 군집까지의 거리를 $a(d, D_M)$, 문서 d에서 가장 가까운 이웃 군집까지의 거리를 $b(d, D_M)$라고 가정할 때, 문서 d의 실루엣 $s(d, D_M)$은 식 (5)와 같고,

$$s(d, D_M) = \frac{b(d, D_M) - a(d, D_M)}{\max\{a(d, D_M), b(d, D_M)\}} \quad (5)$$

실루엣 계수 $SC(D_M)$는 식 (6)으로 표현된다.

$$SC(D_M) = \frac{\sum_{p \in D} s(p, D_M)}{|D|} \quad (6)$$

평균 제곱 오차(Mean Squared Error : MSE)는 군집화 결과를 간단하게 분석하는 한 방법으로, 평균 제곱 오차가 클수록 오차가 더 크다는 의미이며, 평균 제곱 오차 정의는 다음과 같다.

정의 2. 평균 제곱 오차
한 군집 집합 $D_M = \{\overline{D_1}, \cdots, \overline{D_k}\}$에 대해 전체의 평균 제곱 오차 MSE가 식 (7)일 때,

$$MSE(D_M) = \Sigma_{i=1}^{k} MSE(\overline{D_l}) \quad (7)$$

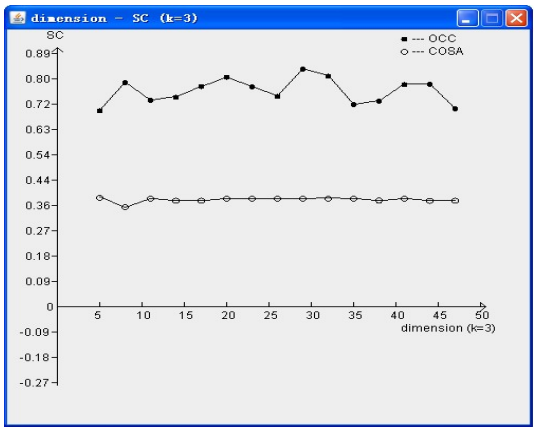
한 군집에 대한 평균 제곱 오차 $MSE(\overline{D_l})$는 식 (8)로 정의한다. 단, $\mu\overline{D_l}$는 군집 $\overline{D_l}$의 중심이다.

$$MSE(\overline{D_l}) = \Sigma_{p \in \overline{D_l}} dist(p, \mu\overline{D_l})^2 \quad (8)$$

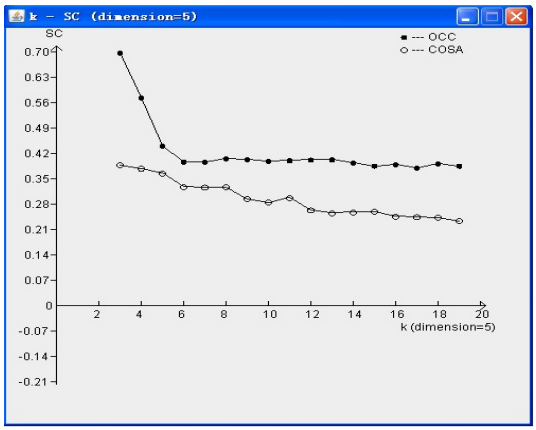
<Figure 9>는 k = 3으로 군집화 할 때 서로 다른 특징 공간 차원에서의 실루엣 계수를 보여 준다. 모든 특징 공간 차원에서 DODC가 COSA 보다 우수함을 알 수 있다. 그리고 DODC 경우 실루엣 계수가 0.85에 접근하는 것을 보면 문서들이 제대로 군집화 된다는 것을 알 수 있다. 이는 DODC 방식은 한 개념의 문서 집합에 대한 중요도를 계산할 때 개념 빈도뿐 아니라 역문서 빈도도 고려하여 온톨로지 전체에서 가장 우선순위가 낮은 개념들을 줄여 나가기 때문이다.

<Figure 10>은 특징 공간 차원이 5일 때 서로 다른 k값에서의 실루엣 계수를 보여 준다. <Figure 10>을 보면 군집 숫자 k가 클 때보다 작을 때 문서 군집들이 더 잘 분리되는 것을 알 수 있으며, 전체 구간에서 DODC 기법이 COSA 기법보다 우수한 것을 알 수 있다.

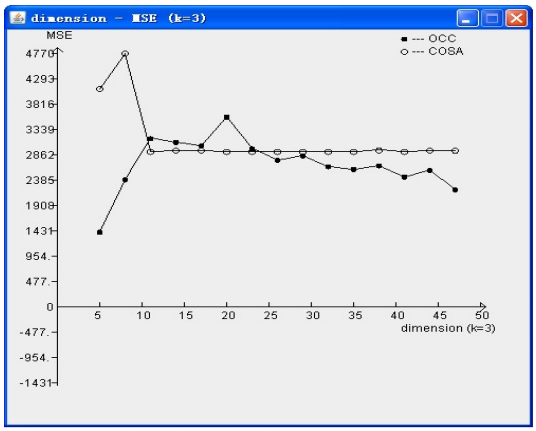
<Figure 11>은 k = 3으로 군집화 할 때 서로 다른 특징 공간 차원에서의 평균 제곱 오차를 보여 준다. DODC 기법이 전체적으로 COSA 기법보다 MSE가 작음을 보여 준다. 그것은 DODC 기법이 문서들에서 더 의미 있는 핵심 개념들을 추출해서 군집화를 수행하기 때문이다.



<Figure 9> Dimension-SC Function with k = 3

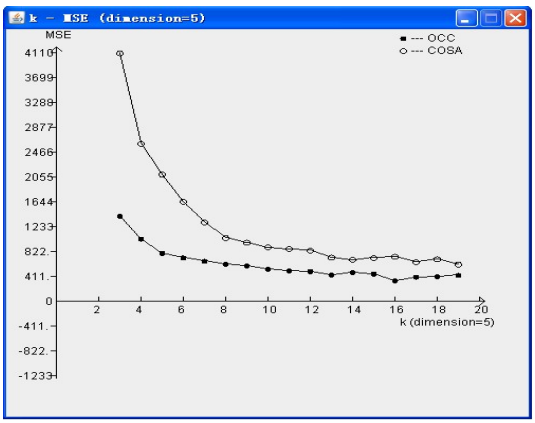


<Figure 10> K-SC Function in a 5 Dimensional Feature Vector Space



<Figure 11> Dimension-MSE Function with k = 3

<Figure 12>는 특징 공간 차원이 5일 때 서로 다른 k값에서의 평균 제곱 오차를 보여 준다. k가 커지면 한 개 군집의 평균 제곱 오차가 줄어드는 속도가 k가 커지는 속도보다 빠르기 때문에 k가 커질수록 전체 오차는 줄어드는 모습을 보이나, 전체적으로 DODC 기법이 COSA 기법보다 평균 제곱 오차가 작음을 알 수 있다.



〈Figure 12〉 k-MSE Function in a 5 Dimensional Feature Vector Space

4. 결 론

차세대 웹으로 불리는 시맨틱 웹은 온톨로지를 기반으로 하기 때문에, 시맨틱 웹의 연구와 더불어 다양한 분야의 도메인 온톨로지가 구축되고 있으며 또한 온톨로지를 자동 또는 반자동으로 구축하고자 하는 연구도 활발히 진행되고 있다. 따라서 최근의 문서 군집화 기법은 이러한 온톨로지를 이용하거나 문서 집합의 내부 구조를 나타내는 의미 특징을 많이 사용하는 추세이다. 본 논문에서 제안하는 DODC 기법은 특정 문서들에 대응하는 도메인 온톨로지로부터 핵심 개념들을 찾아 내, 이를 기반으로 낮은 차원의 특징 공간상에서 군집화를 효율적으로 수행한다. 또한 사용자는 핵심 개념들을 통해 각 군집의 주제를 쉽게 파악할 수 있고 이를 통해 응용에 적절한 군집의 숫자도 결정할 수 있게 된다. 추후 과제로는 특정 영역의 문서들을 군집화 할 때 시스템이 응용에 가장 적합한 군집의 숫자를 자동으로 결정하는 방법에 대한 연구가 필요하다.

References

[1] Bae, Y., Kim, J., Ok, D., and Choi, H. S., "Development of Concept and Instance Classification System for Automatic Construction of Ontology", Proceedings of The 34th KIISE Spring Conference, 2007.

[2] Choi, H., Lim, J., Bae, Y., Choi, S., and Ok, C. Y., "Ontology Construction Method and Example", Communications of the Korean Institute of Information Scientists and Engineers, 2006.

[3] Choi, H. S. and Ok, C. Y., "Information Retrieval and Ontology", *Communications of the Korean Institute of Information Scientists and Engineers*, Vol. 22, No. 4, 2004, pp. 62-71.

[4] Hotho, A., Maedche, A., and Staab, S., "Ontology-Based Text Clustering", Proceedings of the IJCAI-001 Workshop, "Text Learning : Beyond Supervision", 2011.

[5] Hu, X., Zhang, X., Lu, C., Park, E. K., and Zhou, X., "Exploiting Wikipedia as External Knowledge for Document Clustering", Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009.

[6] Hwang, C., Lee, M., and Jung, G., "Design of Merchandise Retrieval System based on Ontology on EC", 2015 KSII Spring Conference, 2005.

[7] Hwang, M., Jeong, D. H., Cho, M., Jung, H.,

Kim, P., Yoon, S., and Han, K., "On Construction of National History Ontology", On Construction of National History Ontology", 2012 KSII Fall Conference, 2012.

[8] Jo, D. and Kim, D., "Study on Legal Ontology Construction and RDF Inference Method", The 39th KIPS Fall Conference 2013, 2013.

[9] Jo, D. H. and Lee, K. S., "Query Expansion based on UMLS and Wikipedia Knowledge Information for Clinical Decision Support", Proceedings of The 42nd KIISE Spring Conference, 2015.

[10] Kaufman, L. and Rousseeuw, P. J., Finding Groups in Data : An Introduction to Cluster Analysis, Wiley, New York, 1990.

[11] Kim, J. and Choi, K. S., "Automatic Construction of Korean WordNet based on CoreNet and Dictionaries", Proceedings of The 41st KIISE Fall Conference, 2014.

[12] Kong, H., Hwang, M., Kim, W., and Kim, P., "The Study on the Autometic Ontology Building Methodology about the Specific Domain Knowledge", Proceedings of The 32nd KIISE Fall Conference, 2005.

[13] Min, Y. and Lee, B., "Predicate Ontology for Automatic Ontology Building", The 34th KIPS Spring Conference 2008, 2008.

[14] Mun, H. J. and Woo, Y. T., "Concept Extraction Technique from Documents Using Domain Ontology", *The KIPS Transactions : Part D*, Vol. 13-D, No. 3, 2006, pp. 309-316.

[15] Park, S., Kim, K. J., Kim, K. H., and Lee, S., "Enhancing Document Clustering Using Term Re-weighting Based on Semantic Features", Journal of KIICE, 2013.

[16] Park, S., Lee, Y., Jung, M. A., and Lee, S., "Enhancing Document Clustering using Important Term of Cluster and Wikipedia", Journal of IEIE, 2012.

[17] Ra, M., Yoo, D., No, S., Shin, J., and Han, C., "National Defense Domain Ontology Development Using Mixed Ontology Building Methodology", The 38th KIPS Spring Conference 2012, 2012.

[18] Smirnov, A., Pashkin, M., Chilov, N., Levashova, T., Krizhanovsky, A., and Kashevnik, A., Ontology-Based Users and Requests Clustering in Customer Service Management System, Springer-Verlag GmbH, *Lecture Notes in Computer Science*, Vol. 3505, 2005, pp. 231-246.

[19] Snasel, V., Moravec, P., and Pokorny, J., "WordNet Ontology based Model for Web Retrieval", International Workshop on Challenges in Web Information Retrieval and Integration, 2005.

[20] Son, J., Kim, D., and Jung, I., "Representation of drug information and their relations using ontology", Proceedings of The 37th KIISE Spring Conference, 2010.

[21] Wang, H., Azuaje, F., and Bodenreider, O., "An ontologydriven clustering method for supporting gene expression analysis", In Proc. of the 18th IEEE International Symposium on Computer-Based Medical Systems, in press, 2005.

[22] http://lyle.smu.edu/~tspell/jaws/.

[23] http://obi-ontology.org/page.

[24] http://protegewiki.stanford.edu/wiki/Protege OWL_API_Programmers_Guide.

[25] http://www.biomedcentral.com.

[26] https://lsg3.nlm.nih.gov/LexSysGroup/Summary/lexicon.html.

[27] https://lucene.apache.org/core.

저자소개

Woosaeng Kim

Woosaeng Kim is currently a professor of Computer Software department of Kwangwoon University. He received the bachelor's degree in the department of Computer Science from University of Texas at Austin. He received the MS and Ph. D. degree in the department of Computer Science from University of Minnesota. He had worked as a system engineer at Hyundai Electronics Co. His current research interests include database, multimedia, web application, etc.

Xiang-Dong Guan

Xiang-Dong Guan received the bachelor's degree in the department of Electronics and Communications Engineering from University of Science and Technology Beijing. He received the MS degree in the department of Computer Science from Kwangwoon University. He works as a software engineer at Sinosoft Co. Ltd in China. His research interests include database, semantic web, etc.