ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 21, No. 9, pp. 596-603, 2015. 9
http://dx.doi.org/10.5626/KTCP.2015.21.9.596

위키피디아 기반의 3차원 텍스트 표현모델을 이용한 개념망 구축 기법

(Building Concept Networks using a Wikipedia-based 3-dimensional Text Representation Model)

홍 기 주 † 김 한 준 †† 이 승 연 †††
(Ki-Joo Hong) (Han-Joon Kim) (Seung-Yeon Lee)

요 약 개념망(Concept Network)은 시멘틱 검색, 개인화 검색, 추천, 텍스트마이닝 기법의 개선 등에 필수적인 지식베이스이다. 최근 효과적인 개념망 구축을 위해 온톨로지를 기반으로 하여 개념의 표현을 확장시키는 연구가 활발하다. 이에 본 논문은 World Knowledge로 평가받고 있는 위키피디아 데이터를 '개념' 집합의 원천으로 활용하여 3차원 텍스트 표현 모델 기반 개념망을 구축하는 기법을 제안한다. 사실상 개념들 간의 관계 정보는 시간의 흐름에 따라 변동하기 때문에, 텍스트 문서로부터 도출되는 '개념'은 Formal Concept Analysis 이론체계의 개념에 따르는 것이 바람직하다. 이를 위해 본 논문은 하나의 개념을 '단어'와 '문서' 간의 2차원 행렬로 표현하여 문서집합에 잠재된 개념간의 연관망을 보다 정확하게 생성하게 한다.

키워드: 개념망, 텍스트 마이닝, 위키피디아, 텍스트 표현 모델, 3차 텐서, 텍스트 큐보이드

Abstract A concept network is an essential knowledge base for semantic search engines, personalized search systems, recommendation systems, and text mining. Recently, studies of extending concept representation using external ontology have been frequently conducted. We thus propose a new way of building 3-dimensional text model-based concept networks using the world knowledge-level Wikipedia ontology. In fact, it is desirable that 'concepts' derived from text documents are defined according to the theoretical framework of formal concept analysis, since relationships among concepts generally change over time. In this paper, concept networks hidden in a given document collection are extracted more reasonably by representing a concept as a term-by-document matrix.

Keywords: concept network, text mining, Wikipedia, text representation model, 3-order tensor, text cuboid

· 이 논문은 2010년도 한국연구재단의 기초연구사업(과제번호: NRF-2010-0025212) 지원으로 이루어졌으며, 또한 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임. (No. NRF-2013R1A2A2A01017030)
· 이 논문은 제41회 동계학술발표회에서 '위키피디아 기반 3차원 텍스트 표현 모델을 이용한 개념망 구축 기법'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 서울시립대학교 전자전기컴퓨터공학부
lovingmoon1@paran.com
†† 종신회원 : 서울시립대학교 전자전기컴퓨터공학부 교수
(Univ. of Seoul)
khj@uos.ac.kr
(Corresponding author임)
††† 비 회 원 : 서울시립대학교 전자전기컴퓨터공학부
qltkdrn45@naver.com

논문접수 : 2015년 3월 18일
(Received 18 March 2015)
논문수정 : 2015년 6월 8일
(Revised 8 June 2015)
심사완료 : 2015년 6월 10일
(Accepted 10 June 2015)

정보과학회 컴퓨팅의 실제 논문지 제21권 제9호(2015. 9)

1. 서 론

개념망(Concept Network)은 텍스트마이닝, 시멘틱 검색, 개인화 검색 및 추천 기법 등에서 수행 성능(또는 정확도)을 높이기 위해 핵심적인 지식베이스로 활용된다[1]. 개념망은 효율적인 정보의 연계와 활용에 있어 그 효과가 크기 때문에, 빅데이터 시대를 맞아 고품질의 개념망 구축에 대한 요구가 매우 커지고 있다[2]. 최근 고품질의 개념망을 구축하는 방법론 중에서 개념간의 의미적 관계를 표현하기 위해 위키피디아(Wikipedia), Open Directory Project(ODP)와 같은 신뢰도 높은 온톨로지를 이용한 연구가 활발하다[3,4]. 영문 위키피디아의 경우 현재 460만여 개의 페이지가 구축되었고 앞으로도 다수의 협업을 통해 꾸준히 갱신되고 있기에, 위키피디아는 그 양과 질적인 측면에서 온톨로지로서의 가치가 매우 크다. 이런 맥락에서 본 논문은 위키피디아를 개념집합의 원천으로 간주하여 고품질의 개념망을 구축하는 새로운 기법을 제안하고자 한다.

위키피디아를 이용한 기존의 개념망 구축 방법은 크게 하이퍼링크 기반 모델과 문서 기반 모델로 나눌 수 있다[5]. 첫째, 하이퍼링크 기반 모델에서 '개념'은 하이퍼링크가 가리키는 위키피디아 페이지(이하 위키페이지) 내에 존재하는 하이퍼링크와 카테고리의 집합으로 정의된다. 이 때 하이퍼링크가 가리키는 위키페이지의 타이틀은 개념의 명칭을 의미하며, 개념간의 관계는 하이퍼링크와 대응되는 위키페이지 내의 하이퍼링크 유사도나 카테고리 유사도 등으로 계산된다. 둘째, 문서 기반 모델은 위키페이지를 하나의 '개념'으로 정의하며, 위키페이지의 타이틀이 해당하는 개념의 명칭이 된다. 이러한 문서 기반 모델은 텍스트마이닝 분야에서 전통적으로 사용되는 'Bag-of-Words'(BOW)방식을 전제로 한다. BOW방식은 문서에 존재하는 단어와 그 출현 빈도를 벡터공간으로 매핑하여 단어 벡터를 생성하고 단어 벡터 간의 유사도를 통해 문서의 유사도를 측정하는 방식이다. 위키피디아의 경우 World Knowledge로 평가될 만큼 문서 자체의 완성도가 매우 높아서 문서를 기반으로 하는 개념망 구축 기법이 하이퍼링크 기반 모델에 비해 보다 유용할 수 있으며 이에 관련된 연구도 활발히 이루어지고 있다[6-8]. 이에 본 논문은 위키피디아 데이터를 활용하여 개념 집합을 창출하고 각 위키페이지를 하나의 개념으로 정의하는 전략을 취한다. 그런데 기존의 문서를 BOW방식으로 표현하는 경우, 하나의 위키페이지는 하나의 단어 벡터로 표현되며, 단어 벡터는 문서에 존재하는 단어와 그 출현 빈도만을 고려하므로 개념 간 관계를 도출할 만큼의 충분한 개념 정보를 포함하지 못한다. BOW방식은 단순히 문서에 출현하는 단어 리터럴(literal) 정보만을 기록하고, 그것의 의미적인 정보는 담고 있지 못하고 있어서 고품질의 개념망을 구축하는 근본적인 걸림돌이 된다.

그림 1은 기존의 BOW모델과 제안 모델간의 텍스트 표현 방식의 차이를 표현한 것이다. BOW모델에서의 텍스트 표현 방식의 한계를 극복하기 위해 본 논문은 Formal Concept Analysis (이하 FCA) 이론체계에서의 '개념' 정의를 활용하여, 텍스트 문서에서의 '개념'을 '단어'와 '문서'간의 관계를 표현한 2차원 형태의 행렬로 확장하고, 그 표현 방식을 기반으로 개념간의 연관망을 구축하고자 한다. 이를 통해 문서 집합에 잠재된 개념간의 관계 정보를 더욱 정확하게 도출하고 시간의 변화에 따

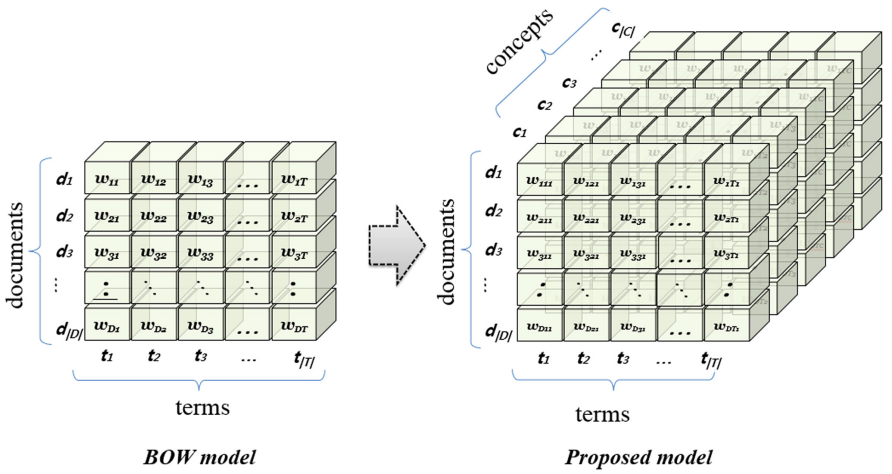

그림 1 BOW 포현모델과 제안 표현모델의 비교
Fig. 1 Comparison of the BOW text model and the proposed text model

른 관계 정보의 변화를 반영한 향상된 개념망을 생성함으로써 BOW기반 개념망 구성 방식의 단점을 보완한다.

2. 3차원 텍스트 표현 모델

2.1 개념의 구성

일반적으로 개념간의 관계는 시간에 따라 서서히 변화한다. 예를 들어 '스포츠'와 '경제'라는 개념은 본래 상호간 연관성이 거의 없었으며, 현재 해당 위키페이지에서도 서로 공유하고 있는 단어 역시 현저히 적다. 그러나 오늘날 스포츠 산업이 발전함에 따라 '스포츠'와 '경제'라는 개념의 관계는 가까워지고 있는데, 실제로 '스포츠 경영학', '스포츠 산업학' 등 관련 연구가 활발해지고 있음이 이를 입증한다. 이러한 점을 고려할 때, 기존의 BOW방식에 근거한 개념망 구축 기법은 개념 자체의 의미만을 반영하므로 위와 같은 개념 관계성의 변화를 반영하는데 한계가 있다. 이를 극복하고자 본 논문에서는 FCA 이론체계를 활용하여 텍스트 문서에서의 개념을 재정의 하고자 한다.

그림 2는 FCA 이론체계를 활용하여 개념을 재정의한 것을 표현한 것이다. FCA 이론체계에서는 하나의'개념(Concept)'이 '외연(Extent)'과 '내연(Intent)'의 쌍으로 정의된다. '외연'은 해당 개념에 포함되는 인스턴스(Instance)들의 집합이며, '내연'은 외연에 포함된 모든 인스턴스들의 공통된 속성들의 집합을 의미한다. 이를 반영하여 하나의 개념(위키페이지)을 나타내는 '외연'은 개념의 명칭(위키페이지 타이틀)을 질의어로 검색하여 얻어진 문서집합으로 구성되며, '내연'은 그 문서집합으로부터 추출한 주요 키워드의 집합으로 구성한다. 만약 '외연'이 뉴스나 사설 등 시간성이 존재하는 문서들의 집합으로 구성되었다면 이를 활용하여 시간에 따른 개념간의 관계를 파악하는 것이 가능해 진다.

본 연구에서는 FCA 이론에서의 개념에서 상응하는

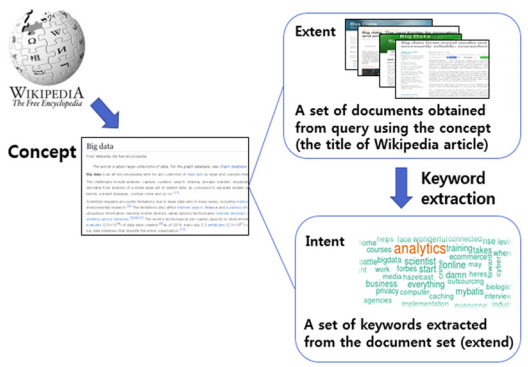

그림 2 개념의 정의

Fig. 2 Definition of the concept

'개념'을 생성하기 위해서 개념 - 문서(외연) - 단어(내연)로 표현되는 3차 텐서(Tensor) 형태의 텍스트 표현 모델을 구축하며, 이러한 표현 모델을 '3차원 텍스트 큐보이드(3-dimensional Text Cuboid)'라 명명한다. 3차원 텍스트 큐보이드를 활용하여 개념간의 유사도를 더욱 정확하게 측정함과 동시에, 시간의 흐름에 따른 유의미한 개념간의 관계 변화를 도출하고자 한다.

2.2 전처리 과정

3차원 텍스트 큐보이드 상에서 개념을 '외연'과 '내연'으로 표현하기 위해 우선적으로 문서 집합의 특징을 적절하게 선택하는 것이 중요하다. 이러한 양질의 특징 선택을 위한 전처리 방법으로 우선 품사 태깅(Part-of-speech tagging)을 통해 명사 단어를 추출하였으며, 아래에 제시된 과정을 통하여 문서에서 양질의 단어를 특징으로 선택하였다.

- 품사 태깅 도구를 이용하여 NN, NNS에 해당하는 명사 단어를 추출
- 문서의 의미를 결정하는데 기여하지 못하는 불용어(Stopword)를 제거
- 여러 변형을 가지는 동일 단어를 분별하기 위하여 어근 추출(Stemming)을 수행
- 문서에 출현하는 회수가 임계치 이하인 희소 단어를 제거

2.3 개념 벡터 생성

2.2절에서 기술한 문서 전처리 과정을 통해 얻은 단어 집합을 이용하여 각 문서에 존재하는 단어와 개념간의 비중치를 담은 개념 벡터(Concept vector)를 생성한다. 개념 벡터의 생성을 위해서 특정 단어에 대한 모든 개념들의 비중치를 부여하는 알고리즘이 요구되며 본 연구에서는 Lesk알고리즘(또는 gloss overlap)을 사용하였다. Lesk알고리즘은 문맥상 특정 단어가 일반적으로 주위의 단어들에 의존적인 성격을 고려한 알고리즘으로 Word Sense Disambiguation분야에서 많이 활용되어 왔다[9]. 결과적으로 문서 내의 모든 단어는 목표 단어를 중심으로 반경 r개의 주변 단어를 포함하는 '개념 윈도우'(Concept window)를 생성하여 개념과 비교함으로써 개념의 비중치를 결정하게 된다.

그림 3(a)는 단어 'java'와 반경 r개의 주변 단어를 포함하는 개념윈도우를 보여준다. 이러한 단어의 개념윈도우를 통해 주변 단어의 의미적 비중치를 'java'와 같은 목표단어(Target Word)에 더해줌으로써 목표단어의 의미를 보다 정확히 결정하는 것이 가능하다.

다음 단계로서, 앞서 소개한 개념윈도우에 속한 단어들이 각 개념(위키페이지)에 존재하는지 검사한다. 결과적으로, 해당 단어가 특정 위키페이지에 존재한다면 '1'로 표시하고, 아니면 '0'으로 표시한다. 특정 문서에서 목표단어가 가지는 개념에 대한 비중치는 개념윈도우의

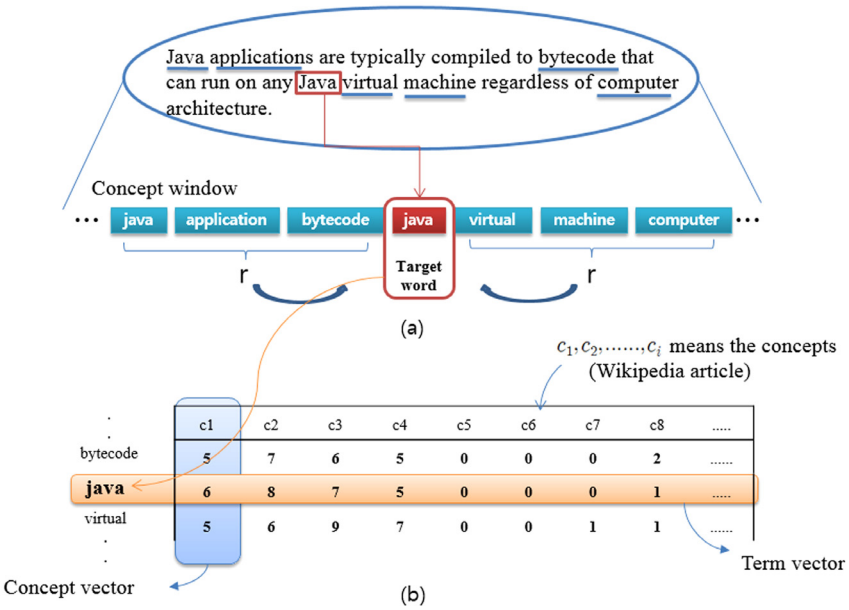

그림 3 (a) (개념 윈도우의 생성) (b) (개념 윈도우를 이용한 개념 벡터의 생성)

Fig. 3 (a) Creating the concept window (b) Creating a concept vector through the concept window

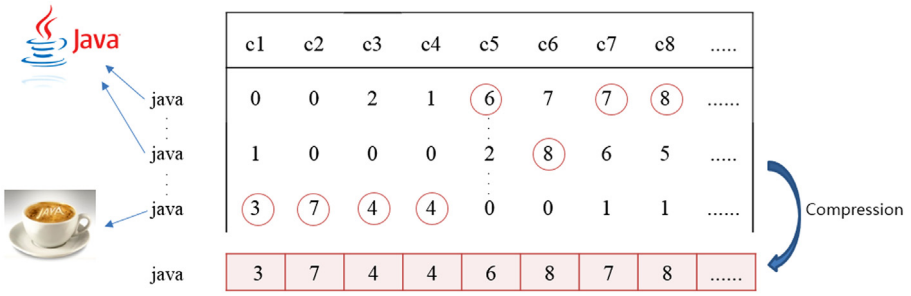

그림 4 중복 단어에 대한 개념 벡터의 생성

Fig. 4 Creating a concept vector for duplicate words

총합에 윈도우의 사이즈를 나눈 값이 된다. 여기서 주목할 점은 문서에서 같은 단어가 여러 번 출현하는 경우, 해당 수만큼의 단어 벡터가 생성되는데 문서를 단어-개념의 행렬로 표현하기 위해 행렬 상의 모든 단어들은 유일한 값이어야 하므로 이를 하나로 압축하는 방안이 필요하다는 것이다. 본 연구에서는 각 개념 성분에 해당하는 비중치의 최댓값을 취함으로써 다수의 단어 벡터를 하나의 벡터로 압축한다.

그림 4는 하나의 문서에서 'java'라는 단어가 여러 번 출현하는 경우 개념 벡터의 각 성분에 대한 비중치를 계산하는 방법을 보여준다. 예를 들어 'c1'개념에 대하여 첫 번째로 출현한 'java'에 개념윈도우의 총합은 0, 두 번째로 출현한 'java'에 대한 개념윈도우의 총합은 1, 세 번째로 출현한 'java'에 대한 개념윈도우의 총합은 3일 때 해당 단어의 개념에 대한 비중치는 최댓값인 3으로 취하여 계산하며 이러한 방식으로 압축된 하나의 단어 벡터가 최종적으로 해당 문서에서'java'라는 단어를 표현하는 단어벡터로 사용된다.

그림 5는 'Computer' 관련 카테고리에 속하는 문서를 이용하여 2가지 개념인 'Computer', 'Coffee'에 대한 단어의 비중치를 계산한 결과를 개념벡터의 형태로 표현한 것이다. 개념으로 사용된 'Computer' 관련 개념($c_1 \sim c_4$)과 'Coffee' 관련 개념($c_5 \sim c_8$) 중 'Computer' 관련 개념들의 비중치가 확연히 높은 것을 확인할 수 있다.

그림 6은 목표 단어에 대한 개념의 비중치를 부여하는 방법을 보여준다. 앞서 기술한 방법은 개념 윈도우

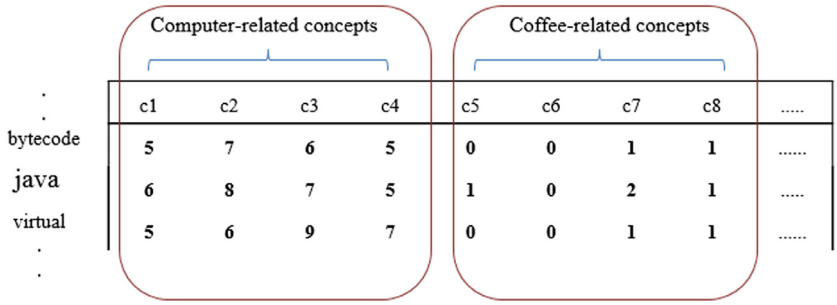

그림 5 개념 벡터에서 개념 성분의 가중치 비교

Fig. 5 Comparing the weights of conceptual dimensions in a concept vector

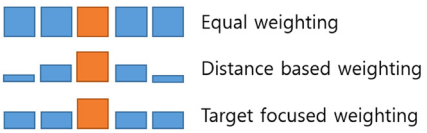

그림 6 개념의 비중치를 부여하는 다양한 방법들

Fig. 6 The variety of ways to weight the concept

내의 모든 단어에 동일한 가중치를 부여하는 Equal weighting이며 추가적으로 개념 윈도우 내의 단어의 위치에 따라 서로 다른 가중치를 부여하는 방법도 고려할 수 있다. Distance based weighting는 목표 단어와 거리가 멀어질수록 더 낮은 가중치를 부여하여 목표 단어와 거리가 먼 단어들은 목표 단어에 더 적은 영향을 미치도록 조정하는 방법이고, Target focused weighting는 목표 단어에 가중치 β를 부여하고 나머지 단어들에 나머지인 1-β를 공평하게 부여하여 목표 단어에 가장 큰 가중치를 부여하는 방법이다.

2.4 3차원 텍스트 큐보이드를 통한 개념의 표현

2.3절에서 주어진 문서에 대한 개념 벡터를 도출하는 방안을 제시하였다. 도출된 개념 벡터를 이용하여 하나의 문서집합은 '단어'와 '개념'간의 연관 관계를 갖는 3차 텐서(Tensor)의 형태로 표현하는 것이 가능하다.

그림 7은 3차원 텍스트 큐보이드 상에서 2개의 개념 성분에 대한 단면을 단어-문서(term-by-document) 행렬로 표현한 것이다. 이 그림에서 보는 바와 같이, 단어-문서 행렬로 표현된 개념은 앞서 기술한 FCA의 개념 정의와 상통할 뿐만 아니라 개념 간 관계 정도를 보다 정확하게 측정할 수 있는 형태가 된 것이다.

2.5 유사도 측정

그림 8은 기존의 BOW방식과 3차원 텍스트 큐보이드에서 텍스트 표현 방식의 차이를 보여준다. 기존의 BOW 방식에서는 하나의 문서가 각 단어에 대한 TF-IDF 가

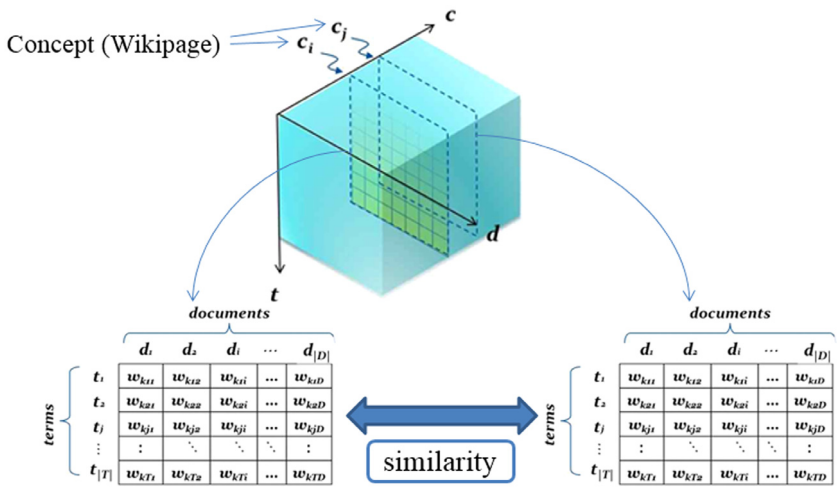

그림 7 텍스트 큐보이드에서 2개의 개념성분에 대한 단면

Fig. 7 Two slices of concepts in the text cuboid

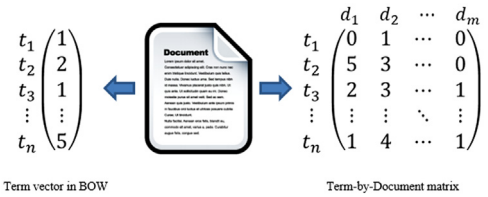

그림 8 (텍스트 표현 방식의 차이
Fig. 8 Difference between text representations

중치의 계산을 통해 단어 벡터의 형태로 표현되며, 그 벡터들 간의 코사인(Cosine) 유사도를 계산함으로써 문서 간 유사도를 측정한다. 이에 반해, 3차원 텍스트 큐보이드 상에서 하나의 개념은 단어-문서의 행렬로 표현되므로 행렬간의 유사도를 산출하여 개념간의 유사도를 구할 수 있다. 식 (1)은 그림 7에서 표기한 개념 c_i와 c_j에 대한 유사도를 측정하기 식으로서, 두 벡터간의 코사인 유사도를 행렬에 대한 유사도로 확장한 것이다.

$$\mathrm{sim}(\boldsymbol{X},\boldsymbol{Y})=\frac{\langle \boldsymbol{X},\boldsymbol{Y}\rangle_F}{\|\boldsymbol{X}\|_2\cdot\|\boldsymbol{Y}\|_2} \quad (1)$$

여기서 $<X,Y>_F$는 행렬 X, Y에 대한 *Frobenius* 곱을 의미하며, $\|X\|_2$는 행렬 X에 대한 L_2-노옴(norm)을 의미한다. 식 (2), (3)을 통해 행렬 X, Y에 대한 *Frobenius* 곱과 행렬 X에 대한 L_2-노옴(norm)을 구할 수 있다.

$$<X,Y>_F=\sum_{i=1}^{n}\sum_{j=1}^{m}X_{ij}\cdot Y_{ij} \quad (2)$$

$$\|X\|_2=\sqrt{<X,X>_F} \quad (3)$$

위의 방법으로 주어진 개념들 간의 유사도를 계산하여 보다 정확한 개념망을 구성하는 것이 가능하다. 이는 개념을 단어-문서의 행렬로 표현할 때 개념에 존재하는 단어와 더불어 주어진 문서에 공통적으로 존재하는 단어를 동시에 고려하기 때문이다. 즉 서로 직접적인 연관성이 적은 두 개의 개념일지라도 각 개념 내에 존재하는 단어가 주어진 문서에서 함께 출현하는 빈도가 높다면 해당 개념간의 유사도가 커지는 특징을 가진다.

3차원 텍스트 큐보이드에서 각각의 문서는 단어-개념의 행렬로 표현되기 때문에 식 (1)을 통해 개념 간 유사도를 측정하는 방식과 동일한 방법으로 유사도를 측정할 수 있다. 개념들 간의 유사도는 주관성에 따라 다르게 해석되는 반면, 문서의 경우 클러스터링을 통한 정량적인 평가가 가능하다. 또한 개념과 문서는 서로 동등한 차원 공간에 위치하므로 문서 클러스터링의 성능 향상은 개념의 표현력 개선과 간접적으로 영향이 있다고 할 수 있다.

그림 9는 기존의 BOW방식과 제안 기법간의 문서 클러스터링 결과를 비교한 것이다. 클러스터링을 위하여

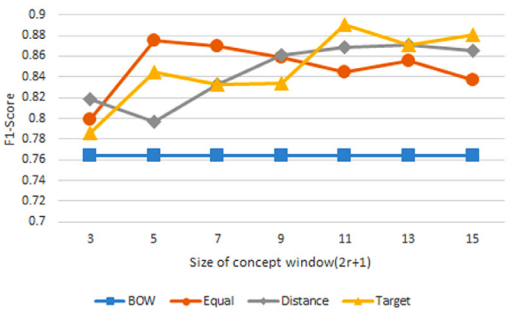

그림 9 BOW방식과 제안 기법 간 클러스터링 성능 비교
Fig. 9 Comparison of the clustering performance of BOW text model and the proposed text model

'20Newsgroups' 데이터셋의 8개의 카테고리(autos, baseball, chiristian, electrocins, graphic, med, mideast, space)에서 각각 50개의 문서를 선택하고, Hierarchical clustering알고리즘으로 원래의 카테고리를 복원하는 정도를 측정하였다. 실험 결과 제안 기법이 기존 BOW기법에 비해 최대 12%포인트 가량 향상된 클러스터링 성능을 보이는 것을 확인할 수 있으며, 이는 3차원 텍스트 큐보이드에서 기존의 방식보다 문서를 더욱 잘 표현한 결과로 해석할 수 있다. 단, Equal weighting의 경우 개념 윈도우의 사이즈가 커질수록 F1-score는 오히려 하락하는 특징을 보이는데 이는 주변의 단어들이 목표 단어에 지나치게 많이 고려되어 Word Sense Disambiguation의 성능이 오히려 감소한 결과로 해석된다. 이를 개선하기 위해 제안한 Distance based weighting과 Target focused weighting의 경우 개념 윈도우의 사이즈가 커질수록 Equal weighting비해 조금 더 높은 성능을 보였다.

3. 실험 및 평가

3.1 실험 데이터

개념 - 문서(외연) - 단어(내연)의 3차원 텍스트 큐보이드 생성을 위한 문서 집합을 구성하기 위해 2014년 9월부터 10월 기간 동안 Google News 사이트에서 'Big Data'를 주제로 하는 뉴스 기사 400개를 수집하였다. 또한 개념 집합을 구성하기 위해서 'Big Data'와 관련된 60개의 위키페이지를 선택하였다. 수집된 뉴스 기사 문서로부터 주요 단어를 선별하기 위해 2.2절에서 소개한 전처리 과정을 활용한다. 비교하여 위키페이지들로 구성된 개념집합의 경우 각 개념을 표현하는 적절한 단어들을 획득하기 위해 TF-IDF 가중치 기법을 이용하여 임계치를 초과하는 단어만을 선택하였다.

3.2 실험 및 결과

본 실험의 목적은 2.5절에서 소개한 개념간의 유사도

측정 방법을 이용하여 개념망을 구축하는 것이다. 제안 기법의 경우 3차원 텍스트 큐보이드 상에서 단어-문서 행렬 형태로 표현된 개념들 간의 유사도를 측정하며 여기에 식 (1)이 사용된다. 기존 BOW 표현 기반의 개념망을 구성하는데 있어서 하나의 개념은 해당 위키페이지에 출현하는 명사들의 TF-IDF값으로 이루어진 단어 벡터로 표현되며, 개념간의 유사도는 단어 벡터간의 코사인 유사도 함수로 측정된다. 개념간의 유사도 값이 임계치를 초과하면 '1', 그렇지 않으면 '0'을 표시하여 개념간의 관계를 인접행렬(adjacency matrix)로 표현하고 이를 활용하여 개념망을 구축할 수 있다. 다만 위키페이지인 개념들 간의 유사성은 공인된 결과가 존재하지 않으므로 수집된 문서의 관찰을 통해 기대하는 개념 관계가 도출되는지 정성적인 관점에서 평가한다. 현재 수집한 문서집합은 'Big data'를 주제로 한 최신 기사들로 구성되어 있기 때문에, 정통 이론적 관점에서 유사한 개념(예: Data mining, Database등) 뿐만 아니라 개념의 확장 및 발전에 따라 같이 자주 언급되는 개념(예: Cloud computing, Apache Hadoop)과 연관되는 것을 기대할 수 있다.

표 1은 BOW 방식과 제안 기법에서 'Big data'와 가까운 개념과 유사도를 기술한 것이다. BOW방식에서 상위 개념이었던 'Artificial intelligence', 'Machine learning'등의 개념을 대신하여 제안 기법에서는 'Apache Hadoop', 'Cloud computing', 'Facebook' 등의 개념이 상위 개념으로 나타난 것을 확인할 수 있다. 제안 기법에 의해 새롭게 도출된 개념들을 고려해 볼 때 현재 'Big data'는 인프라 구축이나 데이터 활용 측면에서 빈번하게 언급되는 것을 유추해 볼 수 있다.

그림 10은 기존의 BOW방식을 통해 생성된 'Big data' 관련 개념망이다. 위 그림에서 'Big data'라는 개념은 'Data mining', 'Data mining', 'Data warehouse' 등의

표 1 개념 'Big data'와 가까운 상위 10개의 개념
Table 1 Top 10 concepts closest to the concept 'Big data'

BOW		TextCuboid	
Concept	Similarity	Concept	Similarity
BI	0.205	SAS software	0.98
Data analysis	0.195	Apache Hadoop	0.95
Data mining	0.174	BI	0.94
SAS software	0.166	Data analysis	0.94
Data warehouse	0.163	Statistics	0.93
Database	0.153	Data warehouse	0.93
R programming	0.143	Data mining	0.89
AI	0.133	Database	0.87
Computer science	0.128	Facebook	0.86
Machine learning	0.113	Cloud computing	0.84

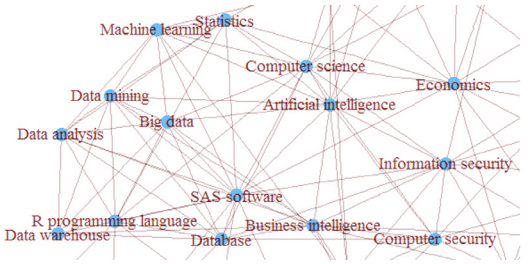

그림 10 BOW방식을 활용한 개념망
Fig. 10 Part of the Concept network utilizing the BOW model

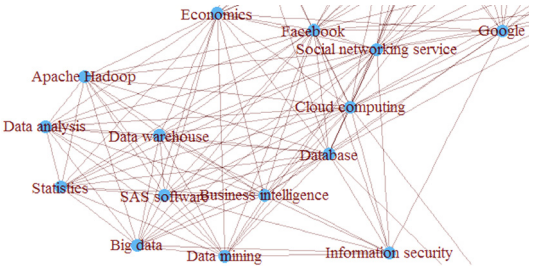

그림 11 제안 기법을 활용한 개념망
Fig. 11 Part of the Concept network utilizing the proposed model

개념들과 함께 개념망을 이루는 것을 확인할 수 있다. 개념망에 존재하는 이 그림에서 상호 연결된 개념들은 관련 위키페이지에 서로 공유하는 단어가 많고 그 출현 횟수가 유사한 것이다. 다시 말해서 개념적으로 유사한 위키페이지가 서로 공유하는 단어가 없다면 개념망에서 연결되지 않는다. 비교하여 그림 11은 제안 기법을 통해 생성된 'Big data'관련 개념망이다. 여기서 'Big data'라는 개념은 'Apache Hadoop', 'Cloud Computing', 'Facebook', 'Social networking service' 등의 개념들과 함께 연관되는 것을 확인할 수 있다. 이러한 개념들은 제안 기법에 의해 새롭게 도출된 관계들로 최근 'Big data'라는 개념이 이 같은 주제들과 함께 주로 언급되는 현재의 트렌드를 반영한 결과라 할 수 있다.

표 2 최근접 개념 간 차이
Table 2 Difference between the most closest concepts

Method	Concepts	Similarity
BOW	Apache Hadoop - Cloudera	0.75
	SAS software - R programming	0.288
	Database - Data warehouse	0.212
TextCuboid	Apache Hadoop - Data analysis	0.97
	SAS software - Big data	0.98
	Database - Cloud computing	0.89

표 2는 BOW방식과 제안 기법에서 최근접 개념 간 차이를 기술한 것이다. 우선적으로 BOW방식에서 최근접 개념으로는 'Apache Hadoop'과 'Cloudera', 'SAS software'와 'R programming' 등 단지 개념간 유사도가 높은 개념들이 서로 연결되었다. 이에 반해 제안 기법의 경우 'Apache Hadoop'과 'Data anaysis', 'SAS software'와 'Big data' 등 개념간의 유사도를 고려할 뿐만 아니라 입력되는 문서를 통해 시간의 흐름에 따른 유의미한 개념간의 관계 변화를 도출하는 것 또한 가능하다.

4. 결론 및 향후 연구 방안

본 논문은 위키피디아로 정의된 개념을 FCA 이론체계에 입각하여 2차원 '단어-문서'간 행렬로 표현한다. 이를 위해서 문서집합을 '개념-단어-문서' 공간을 가지는 3차원 텍스트 큐보이드로 표현한다. 제안 기법을 이용하여 기존 BOW방식의 문제점을 개선하고, 주어진 문서집합을 반영하여 개념 간 유사도를 보다 정확하게 도출할 수 있다. 이를 통해 기존의 정적인 개념망 구축 기법에서 벗어나 시간의 변화에 따라 동적으로 변화하는 개념망 구축이 가능하다.

향후 대용량 텍스트 데이터에서 3차원 텍스트 큐보이드를 생성하고 분석을 수행하기 위한 맵리듀스(Map-Reduce) 알고리즘을 고안하여 대용량 문서 데이터에 대하여 실시간으로 개념망을 구축할 수 있는 기법으로 발전시키고자 한다. 또한 단어-문서 행렬로 표현된 개념의 특성을 십분 활용하여 개념망을 검색 및 추천 기능에 연계시키는 방안을 고안할 것이다.

References

[1] H. Yune, J. Noh, H. Kim, B. Lee, S. Kang, and J. Chang, "Concept Network-based Personalized Web Search Systems," *Journal of Korean Society for Internet Information*, Vol. 12, No. 2, pp. 63-73, 2011. (in Korean)

[2] V. Nastase, and M. Strube, "Transforming Wikipedia into a large scale multilingual concept network," *Artificial Intelligence*, Vol. 194, pp. 62-85, 2013.

[3] M. Daoud, L. Tamine, and M. Boughanem, "A personalized graph-based document ranking model using a semantic user profile," *Proc. of the 18th international conference on User Modeling, Adaptation, and Personalization*, pp. 171-182, 2010.

[4] D. Milne, and IH. Witten, "An open-source toolkit for mining Wikipedia," *Artificial Intelligence*, Vol. 194, pp. 222-239, 2013.

[5] IH. Witten, and D. Milne, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," *Proc. of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pp. 25-30, 2008.

[6] A. Moro, and R. Navigli, "WiSeNet: Building a Wikipedia-based semantic network with ontologized relations," *Proc. of the 21st ACM international conference on Information and knowledge management*, pp. 1672-1676, 2012.

[7] E. Gabrilovich, and S. Markovitch, "Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge," *Proc. of AAAI'06*, pp. 1301-1306, 2006.

[8] A. Huang, D. Milne, E. Frank, and IH. Witten, "Clustering documents using a Wikipedia-based concept representation," *Proc. of 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 628-636, 2009.

[9] S. Banerjee, and T. Pedersen, "An adapted Lesk algorithm for word sense disambiguation using WordNet," *Proc. of International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 136-145, 2002.

홍 기 주

2009년 한동대학교 경영경제학과(경영학사). 2014년~현재 서울시립대학교 전자전기컴퓨터공학부 석사과정. 관심분야는 텍스트마이닝, 온톨로지, 기계학습, 데이터베이스, 빅데이터 분석

김 한 준

1994년 서울대학교 계산통계학과(이학사) 1996년 서울대학교 전산과학과(이학석사) 2002년 서울대학교 컴퓨터공학부(공학박사). 2002년~현재 서울시립대학교 전자전기컴퓨터공학부 교수. 관심분야는 텍스트마이닝, 데이터베이스, 기계학습, 정보검색, 빅데이터 분석

이 승 연

2014년 안동대학교 정보과학교육과(공학사). 2015년~현재 서울시립대학교 전자전기컴퓨터공학부 석사과정. 관심분야는 텍스트마이닝, 온톨로지, 기계학습, 데이터베이스, 빅데이터 분석