http://www.calsec.or.kr/jsebs
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.19, No.3, Aug 2014, pp.107-123
http://dx.doi.org/10.7838/jsebs.2014.19.3.107

위키피디어 기반 개념 공간을 가지는 시멘틱 텍스트 모델

A Semantic Text Model with Wikipedia-based Concept Space

김한준(Han-Joon Kim)*, 장재영(Jae-Young Chang)**

초 록

텍스트마이닝 연구의 기본적인 난제는 기존 텍스트 표현모델이 자연어 문장으로 기술된 텍스트 데이터로부터 의미 또는 개념 정보를 표현하지 않는데 기인한다. 기존 텍스트 표현 모델인 벡터공간 모델(vector space model), 불리언 모델(Boolean model), 통계 모델(statistical model), 텐서공간 모델(tensor space model) 등은 'Bag-of-Words' 방식에 바탕을 두고 있다. 이러한 텍스트 모델들은 텍스트에 포함된 단어와 그것의 출현 횟수만으로 텍스트를 표현하므로, 단어의 함축 의미, 단어의 순서 및 텍스트의 구조를 전혀 표현하지 못한다. 대부분의 텍스트 마이닝 기술은 대상 문서를 'Bag-of-Words' 방식의 텍스트 모델로 표현함을 전제로 하여 발전하여 왔다. 하지만 오늘날 빅데이터 시대를 맞이하여 방대한 규모의 텍스트 데이터를 보다 정밀하게 분석할 수 있는 새로운 패러다임의 표현모델을 요구하고 있다. 본 논문에서 제안하는 텍스트 표현모델은 개념공간을 문서 및 단어와 동등한 매핑 공간으로 상정하여, 그 세 가지 공간에 대한 연관 관계를 모두 표현한다. 개념공간의 구성을 위해서 위키피디어 데이터를 활용하며, 하나의 개념은 하나의 위키피디어 페이지로부터 정의된다. 결과적으로 주어진 텍스트 문서집합을 의미적으로 해석이 가능한 3차 텐서(3-order tensor)로 표현하게 되며, 따라서 제안 모델을 텍스트 큐보이드 모델이라 명명한다. 20Newsgroup 문서집합을 사용하여 문서 및 개념 수준의 클러스터링 정확도를 평가함으로써, 제안 모델이 'Bag-of-Word' 방식의 대표적 모델인 벡터공간 모델에 비해 우수함을 보인다.

ABSTRACT

Current text mining techniques suffer from the problem that the conventional text representation models cannot express the semantic or conceptual information for the textual documents written with natural languages. The conventional text models represent the textual documents as bag of words, which include vector space model, Boolean model, statistical model, and tensor space model. These models express documents only with the term literals for indexing and the frequency-based weights for their corresponding terms; that is, they ignore semantical information, sequential order information, and structural information of terms. Most of the text mining techniques have been developed assuming that the given documents are represented as 'bag-of-words' based text models. However, currently, confronting the big data era, a new paradigm of text representation model is required which can analyse huge amounts of textual documents more precisely. Our text

이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(No. NRF-2013R1A2A2A01017030), 또한 본 연구는 한성대학교 교내학술연구비 지원과제임(장재영).

* Corresponding Author, School of Electrical and Computer Engineering, University of Seoul(khj@uos.ac.kr)
** Co-Author, Department of Computer Engineering, Hansung University(jychang@hansung.ac.kr)

2014년 07월 18일 접수, 2014년 08월 13일 심사완료 후 2014년 08월 19일 게재확정.

model regards the ‘concept’ as an independent space equated with the ‘term’ and ‘document’ spaces used in the vector space model, and it expresses the relatedness among the three spaces. To develop the concept space, we use Wikipedia data, each of which defines a single concept. Consequently, a document collection is represented as a 3-order tensor with semantic information, and then the proposed model is called text cuboid model in our paper. Through experiments using the popular 20NewsGroup document corpus, we prove the superiority of the proposed text model in terms of document clustering and concept clustering.

키워드 : 텍스트 표현모델, 텍스트마이닝, 텍스트 큐보이드, 위키피디어, 개념공간, 벡터공간, 텐서공간
Text Representation Model, Text Mining, Wikipedia, Text Cuboid, Concept Space, Vector Space, Tensor Space

1. 서 론

텍스트마이닝(text mining)은 데이터마이닝(data mining)의 한 분야로서, 비정형적(unstructured) 또는 반정형적(semi-structured) 데이터에 대하여 데이터마이닝, 기계학습, 자연어 처리, 정보검색, 웹공학 기술을 적용하여 유용한 지식 또는 패턴을 추출, 가공하는 기술이다. 최근 IDC 디지털 유니버스 연구보고서에 따르면 2011년 생성된 데이터의 양은 약 1.8Zettabytes로 추정, 향후 10년 동안 그 규모는 50배를 초과할 것이며, 그 중에서 비/반정형 데이터가 90%에 달할 것이라는 전망이다[8]. 의미 있는 대다수의 정보는 비정형적 텍스트의 형태로 존재하기 때문에 빅데이터(big data) 시대에 텍스트마이닝 기술은 매우 중요한 기술로 떠오르고 있다[8, 9, 11, 21, 22]. 특히 모바일 기기 확산에 따라 잠재적 가치가 큰 소셜 텍스트 데이터가 방대한 규모로 생산되고 있어, 텍스트 분석 기술의 중요도는 날로 커지고 있다[24].

텍스트마이닝은 세부적으로 자동문서분류(text classification), 문서클러스터링(document clustering), 연관관계분석(association mining), 자동문서요약(text summarization), 지능형 정보검색(intelligent information retrieval), 정보추천(information recommendation), 개념망(concept network) 등의 기술을 포함한다. 이러한 텍스트마이닝 기술들은 핵심적으로 요구되는 세부 작업이 유사하거나 동일한 부분이 많으며, 이는 텍스트 표현모델과 깊게 연관되어 있다. 다시 말해서 텍스트 문서를 텍스트마이닝 알고리즘의 입력으로 사용하기 위한 형태가 텍스트 표현모델에 의해 결정되는 것이다. 결국 텍스트마이닝 알고리즘의 성능은 텍스트 모델이 가지는 성격에 의존할 수밖에 없다. 이런 맥락에서 텍스트마이닝 연구의 주요 문제는 한가지로 요약할 수 있다. 이는 기존 텍스트 표현모델이 자연어 문장으로 기술된 텍스트 데이터로부터 의미 또는 개념 정보를 표현하지 못한다는 것이다.

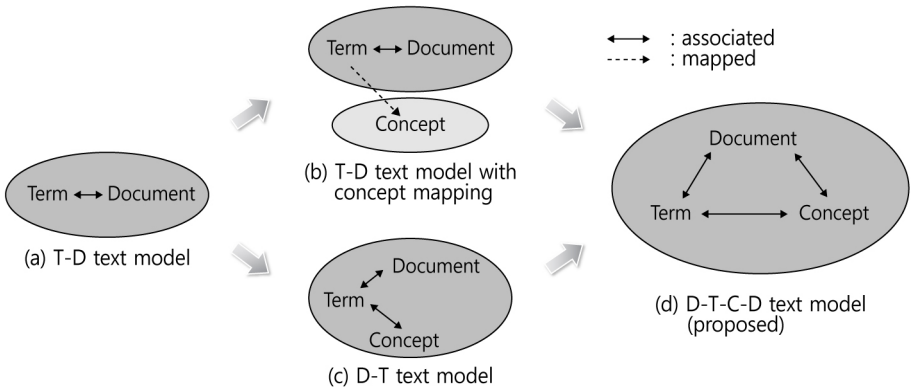
벡터공간 모델(vector space model), 불리언 모델(Boolean model), 확률모델(probability model), 텐서공간 모델(tensor space model) 등의 기존 텍스트 표현모델은 기본적으로 Bag-of-Words(BoW)방식에 바탕을 두고 있다. BoW 방식은 텍스트에 포함된 단어와 그것의 출현 횟수만으로 텍스트 문서를 표현한다. 따라서 BoW 기반 텍스트 표현모델은 단어의 함축 의미, 단어순서 및 텍스트의 구조 정보를 표현하지 못한다는 단점을 가진다. 실제 실험을 통해 BoW 기반 텍스트 모델에서 클러스터링 정확

도는 75% 미만임을 확인하였으며(제 4장 참조), 다른 관련 연구 문헌에서 제시한 정확도 역시 80% 수준을 넘어서지 못하였다. 본 논문은 BoW 기반 모델이 단어의 의미 정보를 담고 있지 않음에 주목하여, 문서에 출현한 단어가 해당 문맥에서 가지는 의미 정보를 표현모델에 담고자 한다. 제안 모델은 개념(concept) 정보를 문서(document) 및 단어(term)와 동등한 공간에서 세 가지 요소의 매핑 관계를 모두 표현한다. 개념공간의 구성을 위해서 위키피디어(Wikipedia) 데이터를 활용하며, 하나의 개념은 하나의 위키피디어 페이지로부터 정의된다. 결과적으로 주어진 텍스트 문서집합을 의미적으로 해석이 가능한 3차 텐서(3-order tensor)로 표현하게 되며, 이를 본 연구에서 '텍스트 큐보이드 모델'이라 명명한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 제안 모델과 관련된 연구와 배경 지식을 서술한다. 제 3장에서는 본 논문이 제안하는 텍스트 큐보이드 모델을 설명하며, 제 4장에서는 이를 검증하기 위한 실험 및 성능 평가 결과에 대해 기술한다. 마지막으로 제 5장에서 결론을 맺는다.

2. 관련 연구 및 배경 지식

기존 Bag-of-Words 방식에 근거한 대표적 텍스트 표현모델은 벡터공간 모델[17], 불리언 모델[12], 확률모델[12] 등이다. 특히 벡터공간 모델은 1975년 G. Salton이 정보 검색을 목적으로 창안한 이후, 현재까지 텍스트마이닝 연구의 기본 토대를 이루고 있다[17]. 이는 각 단어를 서로 독립적인 차원으로 간주하여, 텍스트 데이터를 n차원 벡터 $<t_1, t_2, \cdots, t_n>$형태로 표현한다. 이는 간단하면서도 성능의 안정성, 구현의 편의성으로 인해 꾸준히 활용되고 있다. 그리고 단어 의미의 불확정적 성질은 자연스럽게 확률모델의 정립에 이르렀으며, 이는 기본적으로 각 단어를 랜덤변수 t로 간주하여 문서 d가 확률값 $\Pr(d|t)$로 표현된다. 불리언 모델은 문서를 단어들의 집합으로 간주하여, 집합론에 기반하여 불리언 연산자로 표현된 검색 질의를 처리한다. 하지만 불리언 모델은 정보검색의 주요 이슈인 랭킹 문제를 해결하지 못하여 각광받지 못하였다. 이러한 초기 텍스트 모델은 <Figure 1>(a)와 같이 단어(term)와 문서(document)간의 연관관계를 모델링한 접근 방식으로 설명할 수 있다.



〈Figure 1〉 Evolution of Text Representation Models

최근까지 벡터공간 모델, 확률모델을 포함한 초기 텍스트 표현모델은 텍스트마이닝의 기반 모델로 활용되면서 여러 변형된 텍스트 모델이 제안되어 왔다. 이와 관련된 연구를 살펴보면, 기존 텍스트 모델의 단점을 극복하기 위해 인간의 지식과 연관된 의미(semantic) 또는 개념(concept) 요소를 포함시키려는 노력이 다수를 차지한다. 이에 대한 하나의 접근 방법은 <Figure 1>(b)와 같이 단어 또는 그것의 부분집합을 특정 개념으로 매핑(mapping)하여 텍스트 문서를 개념공간에서 처리하는 것이다. 또 하나의 접근 방법은 <Figure 1>(c)와 같이 텍스트 모델이 단어와 문서간의 연관 관계뿐만 아니라 단어와 개념간의 연관관계를 포함하는 것이다. 통계모델 기반의 추론망(inference network) 모델[7], 베이지안 신뢰망(Bayesian belief network) 모델[16]이 여기에 속한다. 그러나 이 모델들은 텍스트의 표현을 목적으로 한 것이라기보다 일종의 검색 모델이기 때문에, 사실상 텍스트 자체의 표현모델로서 보기는 어렵다. 요약하면 초기 텍스트 모델의 확장을 위해 개념 요소와의 연관성 측면의 연구 노력이 있었지만, 이는 모델의 구성인자로서 개념 정보를 직접 담고 있지 않아 정형성(formalism) 측면에서 한계가 있으며, 표현모델이 아닌 검색모델 차원에서 접근하여 다양한 텍스트 분석 응용에 활용되기 어렵다. 이와 관련하여 최근 텍스트 표현의 의미성을 포괄하기 위해 Wikipedia, Open Directory Project(ODP)와 같이 방대한 규모의 신뢰도가 높은 world knowledge 온톨로지를 활용하는 연구가 활발하다[5, 6]. World knowledge라 함은 현재 학계, 산업계에서 공인받은 신뢰도가 높은 온톨로지를 의미한다. <Figure 1>(b)에 해당하는 모델은 문서 또는 출현 단어를 world knowledge의 특정 개념에 매핑하는 방법을 포함한 것이다. 그리고 이는 아직 world knowledge가 표현모델 내부의 구성요소로 활용되지 않고, 기존 BoW 기반 모델의 개선에 있어 보조적 역할에 그치고 있다.

텍스트 표현모델로서 최근에 대두된 것으로 N-gram 모델[4], 행렬공간 모델(matrix space model)[1], 텐서공간 모델(tensor space model)[20, 25]을 들 수 있다. N-gram 모델은 텍스트 내부의 단어의 순서 정보를 담으려는 노력의 일환으로 제시된 것으로서, 특정 단어에 대한 확률값은 이전 (N-1)개의 단어에 대한 마코프체인(Markov chain) 확률값으로 계산된다. 이는 단어의 순서 정보를 활용하여 단어 비중값을 보다 정확히 산출하려 하지만 개념(의미) 수준에 접근하지 않으며, 또한 계산복잡도 문제를 극복하지 못하였다. 행렬공간모델은 텍스트 문서를 구조적으로 분할하여 문단/문장 수준의 정보를 정량화하여 이를 행렬로 표현한다. 더 나아가 텐서공간 모델은 텍스트의 단어 벡터를 2차원적으로 표현하여 단어간 상관도를 행렬 형태로 표현한 것이다. 하지만 이 모델 또한 단어의 의미적 개념공간을 고려하지 않을 뿐만 아니라, 차원(단어) 성분을 최적으로 분할해야 문제를 안고 있어 실용성이 낮다. 결과적으로 이 세가지 모델은 <Figure 1>(a)의 T-D 텍스트 모델의 범주에 포함시킬 수 있다.

앞서 설명한 바와 같이 기존 대부분의 텍스트마이닝 연구는 BoW 방식이어서 단어 간에 독립성을 가정하므로 이들 간의 연관성 및 의미적 정보를 텍스트 분석에 반영하지 않았다.

2000년 이후 이를 극복하기 위한 하나의 방법으로 텍스트에 출현하는 단어들의 상호 의미적 관계정보를 추출하여 이를 그래프(graph) 형태로 표현하는 연구가 활발히 진행되었다[10, 18, 23]. 이는 기존 그래프 이론을 활용하여 기계학습의 효과를 높일 수 있다는 장점이 있지만, 아직 텍스트 문서를 그래프로 추상화하는데 있어서 통일된 그래프 모델의 부재한 상황이다. 이는 그래프의 노드(node)와 간선(edge)의 정의가 텍스트 분석 목적에 따라 다양한 형태로 존재하기 때문에 자동분류, 클러스터링, 자동요약, 검색 등의 텍스트마이닝 응용분야에 따라 목적 의존적인 그래프 모델이 제시되는 제한성을 가지게 된다.

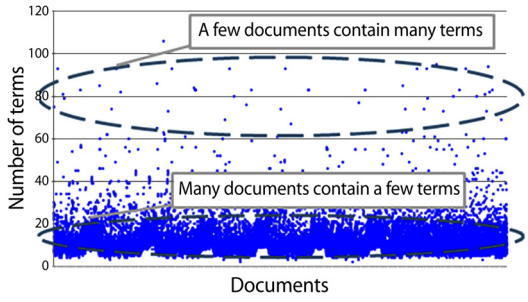
오늘날 빅데이터 시대를 맞이하여 엄청난 규모의 텍스트 데이터를 효과적으로 분석할 수 있는 새로운 패러다임의 표현모델을 요구하고 있다[3, 25]. 앞서 언급한 바와 같이, 정보검색 및 텍스트마이닝 기술의 기반이 되는 텍스트 모델의 발전 과정은 Bag-of-Words 모델, 시멘틱 연관 정보를 가지는 Bag-of-Words 모델로 요약할 수 있다(<Figure 1> 참조). 본 논문은 텍스트마이닝 기법의 한계가 기반이 되는 텍스트 표현모델이 개념 정보를 포함하고 있지 못함을 포착하여, 개념공간을 문서 및 단어공간과 동일한 요소로 상정한 텍스트 큐보이드 모델을 제안하고자 한다.

3. 3차원 큐보이드 텍스트 표현모델

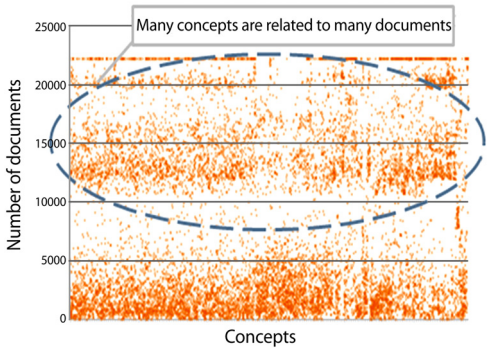
3.1 개념 공간의 중요성

본 연구가 추구하는 텍스트 표현모델은 '개념(의미)'을 '단어'와 동등한 수준에서 고찰해야 하는 요구에서 출발한다. <Figure 2>는 개념(concept) 공간이 텍스트 표현에 있어서 중요한 요소가 되어야 함을 단적으로 보여준다. <Figure 2>(a)는 단어 term(세로축)과 문서 document(가로축)간의 연관 관계를 표시한 것으로서, 대부분의 문서가 소수의 단어로 표현되고, 소수의 문서가 다수의 단어로 표현되는 저밀도(sparsity) 상황을 보여준다. 반면에 <Figure 2>(b), (c)는 개념 concept(세로축)와 문서 document(가로축)간의 연관 관계를 표시한 것으로서, 대부분의 문서가 다수의 개념에 의해 표현되고 있음을 보여준다. 여기서 <Figure 2>(b)는 개념공간으로서 ODP의 카테고리 집합을, <Figure 2>(c)는 위키피디어 페이지 집합을 사용하였으며, 위키피디어 페이지 집합을 개념공간으로 활용한 경우가 문서와 개념간의 연관 정도가 상대적으로 크게 나타나고 있음을 보여준다. 이는 문서를 표현하는 요소로서 '단어'보다는 '개념'의 표현력이 훨씬 크다는 것을 실증하는 것이며, 이러한 관찰 결과가 본 연구의 착안점이 되었다.

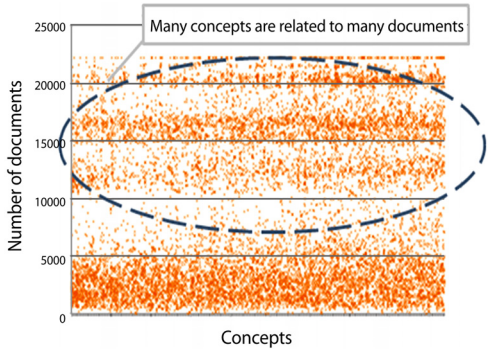
본 논문이 제안하는 텍스트 표현모델은 위키피디어 페이지 집합을 개념공간(concept space)으로 정의하고 이를 문서 및 단어공간과 동등한 관점에서 접근하여(<Figure 1>(c) 참조), 결국 하나의 문서를 2차 텐서(tensor)인 단어-개념 행렬(term-by-concept matrix)로 표현하고, 문서집합을 3차 텐서인 문서-단어-개념 텐서(term-term-concept tensor)로 정의한 것이다. 그래서 본 논문의 제안 모델을 '텍스트 큐보이드(text cuboid)' 모델이라 칭한다.



(a) *Term-document* Relationship



(b) *Concept*(ODP)-*document* relationship



(b) *Concept*(Wikipedia)-*document* relationship

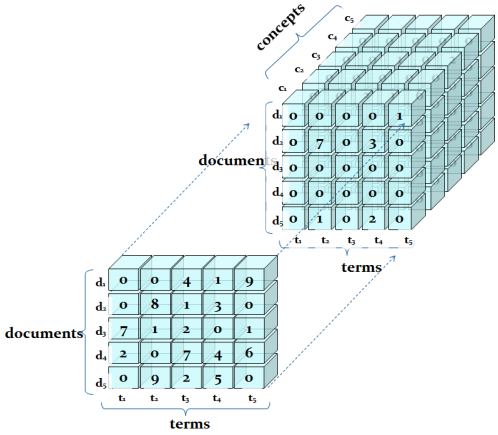
〈Figure 2〉 Observation of Significant Relationships Among Terms, Documents, and Concepts

3.2 개념 공간의 도입

<Figure 1>(c)에서 보는 바와 같이, 본 논문에서 제안하는 텍스트 표현모델은 개념(concept), 문서(document), 단어(term)간의 연관 관계를 모두 동등한 매핑 공간으로 간주하여, 결국 문서집합을 의미적으로 해석이 가능한 3차 텐서로 표현한다. <Figure 3>에서 보는 바와 같이, 기존 벡터공간 모델을 활용한 경우에는 문서집합이 단어 벡터(term vector)의 집합인 단어-문서의 연관 관계를 표현한 행렬로서 표현되는데 반해, 제안 모델에서는 문서집합이 단어와 개념간의 연관 관계를 갖는 행렬의 집합체인 3차원 큐보이드 형태를 취한다. 이를 실제 구축하기 위해서는 주어진 문서에 포함된 각 단어가 해당 문서 안에서 어떠한 의미를 가지는지 평가해야 하며, 그 값을 산출하기 위해 벡터공간 모델에서 사용하는 *tf-idf*와는 다른 형태의 가중치 기법을 고안해야 한다.

그리고 개념공간을 생성하기 위한 개념들의 집합인 개념 도메인을 미리 정의해야 한다. 만약 의학 분야와 관련된 텍스트 문서를 인덱싱하고자 할 경우에는 의학 분야의 표준 온톨로지인 UMLS(Unified Medical Language System)를 개념 도메인으로 활용할 수 있다. 본 연구에서는 일반 문서를 다루기 때문에 world knowledge 수준의 온톨로지인

위키피디어 데이터의 일부를 개념 도메인으로 활용한다. 즉 하나의 위키피디어 페이지(이하 위키페이지)가 하나의 개념을 정의하게 되며, 해당 위키페이지의 제목을 개념의 명칭으로 한다.



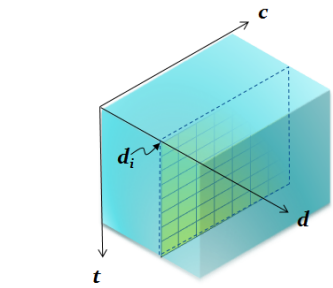
〈Figure 3〉 The Proposed Text Cuboid Model

3.3 텍스트 큐보이드 모델의 활용

<Figure 4>~<Figure 6>은 텍스트 큐보이드 모델의 구성 요소인 문서(document), 단어(term), 개념(concept)를 2차 텐서인 행렬로 표현한 것을 보여준다. 문서, 단어, 개념이 1차 텐서 벡터가 아닌 행렬로 표현됨으로써 텍스트 분석에 어떠한 장점이 있는지 고찰해 보자.

가. 문서의 표현(개념-단어 행렬)

<Figure 4>는 텍스트 큐보이드의 문서공간 축에 대한 단면인 문서 d_i의 '개념-단어 행렬(concept-by-term matrix)'을 보여준다. 이 행렬은 주어진 문서에 출현한 특정 단어가 어떤



terms

concepts	t_1	t_2	t_j	⋯	t_T
c_1	w_{11i}	w_{12i}	w_{1ji}	…	w_{1Ti}
c_2	w_{21i}	w_{22i}	w_{2ji}	…	w_{2Ti}
c_k	w_{k1i}	w_{k2i}	w_{kji}	…	w_{kTi}
⋮	:	⋱	⋱	⋱	:
c_C	w_{C1i}	w_{C2i}	w_{Cji}	…	w_{CTi}

〈Figure 4〉 Representing a Document : Concept-by-Term Matrix

개념을 가지고 있는지를 정량적으로 보여준다. 그 행렬의 행 또는 열 기준의 합을 구하면, 하나의 문서를 단어벡터 또는 개념벡터로 표현할 수도 있다. 이 맥락에서 제안 모델은 벡터공간 모델을 확장한 것이라 할 수 있으며, 또한 최근 연구의 일환인 문서 또는 단어를 개념공간으로 매핑하는 연구를 포괄한다 하겠다. 그리고 하나의 문서가 각 단어에 대한 개념 정보를 담는 행렬로 표현되기 때문에 문서 간 유사도 계산에 있어서 단어 리터럴(literal)과 개념 가중치를 동시에 고려하여 보다 정확한 계산이 가능하다. 두 행렬 X, Y간의 유사도는 다음과 같이 Frobenius 거리함수로 정의한다.

$$dist(X,Y) = \| X-Y \|_F \quad (1)$$

$$= \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |[X-Y]_{ij}|^2}$$

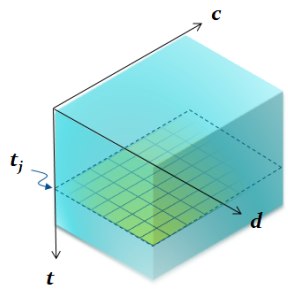
$$= \sqrt{trace((X-Y)^T \cdot (X-Y))}$$

여기서 $\|.\|_F$는 Frobenius 거리함수를, $[X]_{ij}$는 행렬 X의 i번째 행, j번째 열의 값을 의미한다. 식 (1)의 유사도 공식은 당연히 개념-단어 행렬뿐만 아니라, 개념-문서 행렬, 문서-단어 행렬에도 적용된다. 실제 실험으로 통해서 이 거리함수에 의해 수행한 문서 클러스터링의 정확도가 획기적으로 높아짐을 확인하였다(제 4.2절 참조).

또한 정보검색 관련하여 제안 모델상에서 자연스럽게 개인화 검색을 성취할 수 있다. 벡터공간 모델에서 검색을 실현하기 위해서는 주어진 검색 질의를 문서로 간주하여 이를 인덱싱된 문서들의 단어벡터와의 유사도를 계산한다. 이와 유사하게 제안 모델에서도 검색 질의를 하나의 문서로 간주하여 '개념-문서' 행렬로 확장한다면 개인화 검색을 쉽게 실현할 수 있다. 검색 이용자가 과거에 선택한 문서들 또한 '개념-단어' 행렬로 표현되며 이를 요약한 개념벡터 및 단어벡터를 사용자 프로파일로 정의할 수 있다. 그러한 개념 및 단어를 융합한 공간에서 표현된 사용자 프로파일을 활용함으로써 초기 검색 질의를 확장하여 개인화 검색을 실현할 수 있는 것이다.

나. 단어의 표현(개념-문서 행렬)

<Figure 5>는 텍스트 큐보이드의 단어공간 축에 대한 단면인 단어 t_j의 '개념-문서 행렬(concept-by-document matrix)'을 보여준다. 이 행렬은 주어진 문서와 관련된 특정 개념이 어떤 단어에 반영되고 있는지를 정량적으로 보여준다. 앞에서 제시한 식 (1)을 이용하여 유사 단어를 도출함으로써 단어 클러스터(term cluster) 또는 단어 네트워크(term network)를 쉽



documents

concepts	d_1	d_2	d_i	⋯	d_D
c_1	w_{1j1}	w_{1j2}	w_{1ji}	…	w_{1jD}
c_2	w_{2j1}	w_{2j2}	w_{2ji}	…	w_{2jD}
c_k	w_{kj1}	w_{kj2}	w_{kji}	…	w_{kjD}
⋮	:	⋱	⋱	⋱	:
c_C	w_{Cj1}	w_{Cj2}	w_{Cji}	…	w_{CjD}

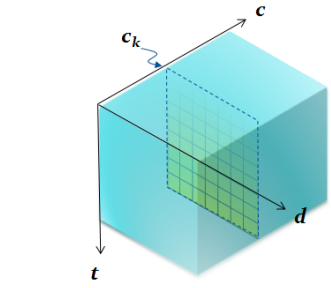
〈Figure 5〉 Representing a Term : Concept-by-Document Matrix

게 구성할 수 있다. 또한 2개 이상의 의미를 가지는 단어의 경우, 해당 '개념-문서' 행렬을 관찰함으로써 주어진 문서집합 안에서 어떠한 개념에 반영되고 있는지를 쉽게 파악할 수 있으며, 특정 단어가 상식적으로 알려져 있는 의미(개념)들보다 더 포괄적인 개념들을 추출할 수 있다. 이를 이용하여 자동문서분류(text classification)를 위한 특징 추출(feature extraction) 차원에서 풍부하고 신뢰도가 높은 특징 집합을 구성할 수 있다.

다. 개념의 표현(문서-단어 행렬)

<Figure 6>은 텍스트 큐보이드의 개념공간 축에 대한 단면인 개념 c_k의 '문서-단어 행렬(document-by-term matrix)'을 보여준다. 이 행렬은 하나의 개념이 문서공간과 단어공간의 조합으로서 표현됨을 보여주는데, 이는 Formal Concept Analysis(FCA) 이론[19]에서 제

시한 개념(concept)의 정의와 일맥상통한다. FCA 이론에 따르면, 하나의 개념은 '외연(extent)'과 '내연(intent)'으로 구성되며, '외연'은 해당 개념에 포함되는 인스턴트(문서)들의 집합이며, '내연'은 외연에 포함된 모든 인스턴스들의 공통된 속성(단어)들의 집합으로 정의된다. FCA이론의 개념과 정확히 일치하는 개념을 생성하기 위해서는 문서-단어 행렬에 대한 별도의 작업이 필요하다. 즉 문서-단어 행렬에서 모든 단어에 대한 비중값의 합이 임계점을 초과하는 문서들을 가려내어 이를 '외연'으로 정하고, 그 문서들이 공통적으로 가지는 단어들을 '내연'으로 정하면 된다. 그러한 개념에 대한 2차 텐서 행렬간의 유사도 및 포함관계를 계산함으로써, 주어진 텍스트 큐보이드에 속한 개념(위키페이지)들의 네트워크 또는 계층트리 형태를 쉽게 구성할 수 있게 된다.



terms	d_1	d_2	d_i	⋯	d_D
t_1	w_{k11}	w_{k12}	w_{k1i}	…	w_{k1D}
t_2	w_{k21}	w_{k22}	w_{k2i}	…	w_{k2D}
t_j	w_{kj1}	w_{kj2}	w_{kji}	…	w_{kjD}
⋮	:	⋱	⋱	⋱	:
t_T	w_{kT1}	w_{kT2}	w_{kTi}	…	w_{kTD}

documents

〈Figure 6〉 Representing a Concept : Term-by-Document Matrix

3.4 텍스트 큐보이드의 구축

텍스트 큐보이드의 구축을 위해 주요 관건이 되는 것은 개념 차원을 생성하기 위한 개념 도메인을 설정하는 것과 이에 대한 매핑을 수행하는 기법이다. 또한 문서에 존재하는 각 단어에 대하여 개념공간에서의 비중치를 담은 개념벡터를 생성해야 하는데, 이는 개념기반 역인덱싱(inverted indexing)을 통해서 가능하다.

가. 개념 도메인의 정의

본 연구에서 채택한 위키피디어 온톨로지 데이터는 현재 450만 개 이상의 매우 많은 위키페이지를 포함하고 있다. 텍스트 큐보이드의 개념공간을 창출하기 위해 사용되는 위키페이지는 상위 수준의 일반성(generality)을 가져야 하고, 해당 위키페이지의 콘텐츠 품질이 적정한 수준 이상이어야 한다. 이러한 요구 조건을 만족하는 위키페이지들을 선별하기 위해, 다각적인 분석 실험을 통해 다음과 같은 필터링 휴리스틱을 도출하였다.

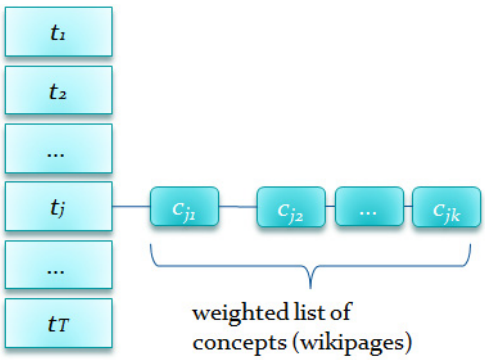
① 페이지 분량이 5000바이트 미만의 위키페이지
② 백링크(backlink) 개수가 20개 미만의 위키페이지
③ 고유명사 수준의 타이틀을 가지는 위키페이지
④ 타이틀에 특수문자가 포함된 위키페이지
⑤ 타이틀이 숫자로 시작하는 위키페이지
⑥ 타이틀이 불용어(stopword)에 속하는 위키페이지

1, 2번 휴리스틱은 저품질의 위키페이지를 제거하기 위한 것이다. 현재 영문 위키피디어 데이터에 대한 실험을 통해, 고품질의 위키페이지는 작성된 콘텐츠의 분량이 대략 5000바이트 이상이고, 다른 위키페이지에서 하이퍼링크로 참조하는 백링크의 개수가 20을 초과함을 관찰하였다. 물론 이 값은 실험적으로 결정한 임계값이다. 일반적으로 텍스트마이닝 시스템은 분석 이전에 특징선택(feature selection)과정을 수행하기 때문에, 개념 도메인의 범위를 지나치게 좁게 잡으려고 노력할 필요는 없다. 즉 1, 2번 휴리스틱에서 제시한 임계치를 낮게 수정해도 무방하다. 나머지 3, 4, 5번 휴리스틱은 일반성을 가지는 위키페이지를 선별하기 위한 것이다. 인명, 지명, 기관명 등의 고유명사에 해당하는 위키페이지는 낮은 수준의 개념성을 가지기 때문에 제외해야 한다. 그리고, 타이틀이 특수문자(예 : (,), & 등)를 포함하거나 숫자로 시작하는 위키페이지들 또한 일반성이 높지 않은 것으로 관찰하여 개념 도메인에서 제외시킨다. 추가적으로 불용어에 해당하는 위키페이지도 개념 도메인에 넣기에는 적합하지 않으므로 필터링 휴리스틱에 포함시킨다. 2014년 6월 현재 약 450만 개인 영문 위키페이지는 위에서 제시한 필터링 휴리스틱에 따라 약 21만 개의 위키페이지 집합으로 축소되며, 이는 개념 공간을 생성하기에 적당한 크기로 평가한다.

나. 개념벡터의 생성

텍스트 큐보이드의 품질은 출현 단어에 대한 개념벡터의 정확성에 의해 결정된다. Navigli [15]에 따르면, 유사한 문맥에서 출현한 단어들은 비슷한 의미를 가질 가능성이 크다. 이는 특정 단어의 의미가 그것의 인접한 주변 단어에 의해서 결정될 수 있다는 것을 암시한다. 그리고 실제 단어의 의미는 그것의 속한 문장, 문법적 구조 등을 고려하여 개념벡터로 표현될 수 있다. 본 논문에서는 단어의 의미를 결정하는 문맥 범위를 그것이 속한 문장으로 한정한다. 그리고 어떤 단어가 특정 개념의 위키페이지에 존재한다면 그 단어는 해당 개념과 관련이 있다고 가정한다. 그러한 가설을 바탕으로 문서 d_i 내 특정 단어 t_j에 대한 개념벡터 $cv_{d_i}(t_j)$를 생성하고자 한다.

일단 해당 단어가 출현하는 위키페이지(개념)를 신속하게 산출하기 위해 <Figure 7>의 개념 위키페이지들에 대한 역인덱스를 구축한다. 이때 역인덱스에 사용되는 단어는 선정된 위키페이지 집합에서 주요 단어로 결정된 것들이다. 그림에서 보는 바와 같이 단어 t_j에 연결된 포스팅 리스트에는 그 단어가 출현한 개념들인 c_{j1}, c_{j2}, …, c_{jk}가 존재한다. 이때 정확한 개념벡터를 구성하기 위해 각 개념 노드마다 해당 단어에 대한 비중값을 설정해야 한다. 만약 단어 t_j가 포스팅 리스트의 개념들에 골고루 분포되어 있다면 그것의 비중값은 작아져야 하며, 그렇지 않다면 커지도록 설정하는 것이 바람직하다. 기존 *tf-idf* 가중치 기법을 사용할 수 있지만, 인덱싱 대상이 텍스트 문서가 아닌 온톨로지 수준의 개념에 대한 메타적 기술(description) 텍스트 정보이기 때문에 엔트로피 기반 가중치 기법이 적합하다. 이를 반영하여 단어 t_j의 개념 c_k에 대한 가중치를 식 (2)와 같이 정의한다.



〈Figure 7〉 Weighted Inverted Index for Building Concept Vectors

$$weight(t_j, c_k) = 1 + \frac{1}{\log_2|C|} \cdot \Pr(t_j, c_k) \cdot \log_2 \Pr(t_j, c_k) \quad (2)$$

여기서, $\Pr(t_j, c_k)$는 단어 t_j가 개념 c_k에 출현할 확률값을 의미하고, $|C|$는 개념 도메인에 포함된 위키페이지의 개수를 의미한다. 확률값 $\Pr(t_j, c_k)$은 0부터 1사이의 값을 가지고, 엔트로피 $\Pr(t_j, c_k) \cdot \log_2 \Pr(t_j, c_k)$의 값은 최소 -0.5이므로 가중치를 양수화하기 위해 1을 보태준다. <Figure 7>의 역인덱스와 식(2)의 가중치 공식을 이용하여 다음 4단계의 절차를 거쳐 개념벡터 $cv_{d_i}(t_j)$를 생성한다. 여기서 t_j는 개념벡터에 대한 목표단어로서 그것의 앞뒤 주변단어가 어떤 위키페이지와 관련을 맺고 있는지를 조사함으로써 목표 단어의 개념벡터를 결정한다. 이는 특정 단어의 의미가 해당 문맥 안에 존재하는 주변 단어에 의해 결정된다는 WSD 기법 또는 언어학적인 고찰에 근거한다[15].

① 단계 : <Figure 7>의 역인덱스를 활용하여 목표 단어의 문장에 속한 단어들에 대하여 해당 단어가 출현한 위키페이지 개념을 알아낸다.

② 단계 : 식 (2)를 활용하여 기정의된 개념공간(위키페이지 집합)에 대하여 각 단어들의 비중값을 계산한다.

③ 단계 : 주변 단어에 대한 각 개념 차원 비중값의 평균을 계산하여 목표단어의 개념벡터를 산출한다.

④ 단계 : 한 문서에서 동일한 목표 단어가 2개 이상인 경우에 해당 수만큼의 단어벡터가 생성되므로, 각 개념 차원의 비중값들의 최대값만을 취함으로써 하나의 개념벡터를 생성한다.

위 단계를 거쳐 우리는 결과적으로 문서 d_i에 출현한 단어 t_j에 대하여 개념벡터 $cv_{d_i}(t_j) = < w_{ij1}, w_{ij2}, \cdots, w_{ijk}, \cdots, w_{ijC} >$를 얻게 된다. 여기서 w_{ijk}는 문서 d_i에 대한 식 (2)의 $weight(t_j, c_k)$값을 의미한다. 결과적으로 주어진 문서집합에 대한 3차 텐서인 텍스트 큐보이드를 얻게 된다.

단어의 의미를 결정하는 범위에 따라 개념벡터를 생성하는 알고리즘이 수정되어야 한다. 사실 이 문제는 2개 이상의 의미(sense)를 가지는 단어가 해당 문맥에서 하나의 의미를 결정하는 기법인 단어중의성해소기법(word sense disambiguation, WSD)과 고유명사 인식기술(named entity recognition, NER)과 관련된 것이다. 최근 WSD 및 NER과 관련된 연구 결과[14]를 활용하여 목표 단어가 속한 문장의 범위를 넘어서 주변 단어를 더 포괄적으로 고려할 수 있으며, 또한 conditional random field(CRF) 이론을 활용하여 단어의 의미를 확률적으로 해석하여 보다 정

확한 개념벡터 생성에 기여할 여지를 가지고 있다[15]. 이는 향후 연구 내용으로 남긴다.

4. 성능 평가

4.1 실험 방법

가. 실험 데이터 셋업

본 논문에서는 제안 텍스트 모델의 효능을 검증하기 위해서 20NewsGroup 문서집합(http://qwone.com/~jason/20Newsgroups)을 이용한다. 20Newsgroup 문서집합은 USENET의 20개 뉴스그룹(rec.autos, soc.religion.christian, sci.electronics, comp.graphics, rec.sport.hockey, sci.space 등)에 기고된 기사들로 구성된 것이며, 자동문서분류 및 문서클러스터링 기법들의 평가를 위해 자주 사용된다[2]. 그리고 개념공간을 생성하기 위해서, 제 3.4절에서 제시한 필터링 휴리스틱에 따라 정제된 위키페이지 집합에서 20NewsGroup 문서집합과 관련된 50개의 위키페이지를 선택하였다.

나. 평가 방법 및 평가 척도

본 실험에서 제안한 텍스트 큐보이드 모델이 기존 벡터공간 모델과 비교하여 2가지 측면에서 우수함을 보일 것이다. 하나는 문서 수준에서 클러스터링 정확도를 평가하는 것이고, 다른 하나는 개념 수준에서 클러스터링 정확도를 평가하는 것이다.

이미 하나의 뉴스그룹(카테고리)으로 분류된 문서 그룹은 클러스터링 관점에서 하나의 클러스터로 간주할 수 있다. 본 실험에서는 20Newsgroup의 데이터에서 10개의 카테고리를 선정하여, 각 카테고리는 100개의 문서를 담게 하였다. 그래서 수작업에 의해 분류된 문서 그룹이 클러스터링 알고리즘에 의해 클러스터로 생성되는 정도를 정량적으로 평가하면 될 것이다. 그 정량적 수치는 F1-측정치로 정의할 수 있으며 그 식은 다음과 같다.

$$F_{i,t} = \frac{2 \cdot p_{i,t} \cdot r_{i,t}}{p_{i,t} + r_{i,t}},$$

$$p_{i,t} = \frac{\textit{Number of documents on category } t \text{ in cluster i}}{\textit{Number of documents} \text{ in cluster i}} \quad (3)$$

$$r_{i,t} = \frac{\textit{Number of documents on category } t \text{ in cluster i}}{\textit{Number of documents on category } t \text{ in the corpus}}$$

$p_{i,t}$는 클러스터 i에서 카테고리 t의 문서들이 인식된 정확도, $r_{i,t}$는 카테고리 t의 문서들에 대해서 클러스터 i에서 인식된 재현율을 의미한다. $F_{i,t}$는 $p_{i,t}$와 $r_{i,t}$의 조화평균으로서 0에서 1까지의 값을 가진다. 전체 클러스터에 대하여 $F_{i,t}$를 계산한 후 이들의 평균을 계산하면 이 값을 통해 10개의 카테고리에 속한 문서에 대하여 클러스터링을 수행하여 본래의 뉴스그룹 카테고리로 복원되는 정도를 알 수 있게 된다.

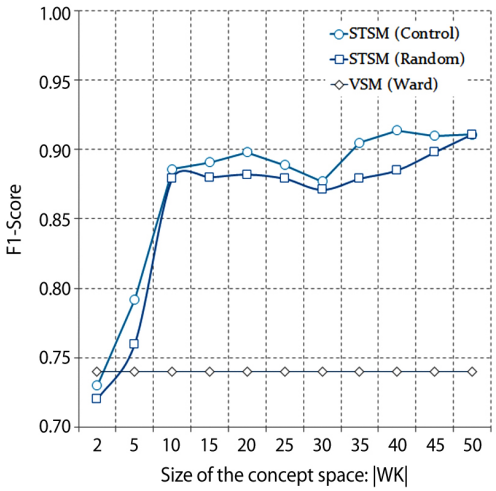
제안 모델에서는 하나의 문서가 단어-개념 행렬로 표현되며, 그러한 행렬로 표현된 문서들은 Frobenius 거리함수에 따라 서로간의 거리값을 가진 정방행렬을 도출할 수 있다. 사실 행렬간의 거리 측정은 수학적으로는 의미가 없다. 하지만 개념적으로 행렬간 유사도는 정량화가 가능한 것이어서, 이를 위한 함수를 Frobenius 거리함수의 역수로 정의하였다. 다

른 형태의 거리함수는 행렬의 크기(norm)으로부터 도출될 수 있다. 일반적으로 행렬 A의 크기는 $L_{p,q} = \left[\sum_{j=1}^{n}\left\{\sum_{i=1}^{m}|a_{ij}|^p\right\}^{q/p}\right]^{1/q}$ 으로 정의하며, Frobenius norm은 $L_{p,q}$가 p = 2, q = 2인 경우에 해당한다. 그런데 실험을 통해 p, q값의 변화에 따른 거리 (또는 유사도) 함수가 성능면에서 거의 차이가 없는 것으로 파악되어 본 실험에서는 Frobenius norm 기반 유사도 함수만을 사용한다. 그 유사도값을 입력으로 받아 동일한 클러스터링 기법을 수행하여 기존 벡터공간 모델과 제안 큐보이드 모델에서 클러스터링 정확도를 비교한다. 본 실험에서 채택한 클러스터링 알고리즘은 문서 데이터에 가장 적합하다고 평가받는 HAC(Hierarchical Agglomerative Clustering) Ward 알고리즘이다.

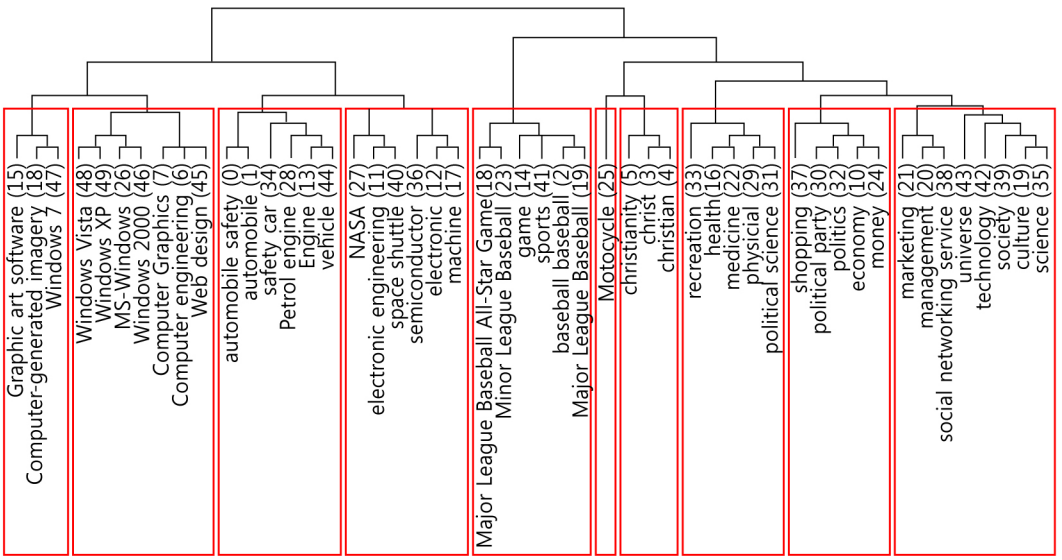
또 하나의 평가 방법은 개념 수준의 클러스터링 결과를 관찰하는 것이다. 하나의 개념은 '단어-문서' 행렬로 표현되기 때문에, 문서 수준의 클러스터링과 똑같이 개념들간의 거리값을 계산함으로써 클러스터링을 수행할 수 있다. 다만 개념공간의 생성에 위키페이지들간의 유사성에 대한 공인된 결과가 존재하지 않기 때문에 본 실험에서는 정성적인 기술을 하였다.

4.2 실험 결과

<Figure 8>은 제안 모델(STSM)과 벡터공간 모델(VSM)에서 HAC Ward 클러스터링 알고리즘을 수행한 결과를 비교한 것이다. 가로축은 개념공간의 크기를, 세로축은 F1-측정치를 의미한다. 벡터공간 모델에서 Ward 알고리즘의 정확도는 74% 수준이며, 제안 모델에서는 평균적으로 90% 수준에 이르러, 대략 22%의 개선 효과를 보였다. 그림에서 보는 바와 같이 개념공간의 크기(위키페이지의 개수)는 클러스터링 정확도 즉 텍스트 큐보이드 모델에 기반한 문서 표현의 정확도를 결정하는 중요한 요소가 됨을 알 수 있다. 개념공간의 크기가 10을 초과해야 적정 수준의 정확도를 보이고 있으며, 그 이후부터는 증가 추세이기는 하나 큰 변화를 보이지는 않는다. 텍스트 큐보이드의 품질을 결정하는 또 하나의 요소는 개념공간의 이질성이다. 개념공간을 구성하는 위키페이지가 다양할수록 문서들이 보다 정확하게 표현된다는 것이다. 그림에서 STSM(Random)은 개념공간을 구성하기 위해 위키페이지를 랜덤하게 결정한 경우에 해당하고, STSM(Control)은 다양한 위키페이지가 포함되도록 수작업에 의해 개념공간을 구성한 경우에 해당한다. 예상한 바대로, STSM(Control)의 경우가 STSM(Random)에 비해 F1-측정치가 최대 4%의 차이를 보인다.



〈Figure 8〉 Changes of F1-score from Varying the Size |WK| of Concept Space



〈Figure 9〉 The Cluster Dendrogram of Concepts(Wikipages)

<Figure 9>는 개념공간을 구성하는 50개의 위키페이지가 '단어-문서' 행렬로 표현되어 이에 대한 클러스터링을 수행한 결과인 클러스터 덴드로그램을 보여준다. 전반적으로 유사한 의미의 위키페이지가 동일한 클러스터에 묶이고 있음을 확인할 수 있다. 컴퓨터, 자동차, 스포츠, 종교, 건강 관련 위키페이지는 동일한 클러스터에 포함되었다. 다만 맨 우측 클러스터에는 경제 및 과학/문화 관련 개념이 섞여 있는데, 다시 이를 2개의 클러스터로 분할한다면 경제 관련 위키페이지인 marketing, management 개념과 과학/문화 관련 universe, technology, society, culture, science 개념들로 나눠짐을 알 수 있다.

5. 결 론

본 논문은 기존의 Bag-of-Words 방식의 한계점을 극복하기 위해 개념공간을 문서 및 단어공간과 동등한 수준으로 고려한 3차 텐서공간 모델을 제안하였다. 하나의 문서, 단어, 개념이 벡터 형태가 아닌 2차 텐서인 행렬 형태로 표현되어 보다 풍부한 정보를 담고 있기 때문에 텍스트마이닝 알고리즘의 성능을 개선하는데 크게 기여할 것으로 전망한다. 개념공간을 구성하기 위해 외부 지식베이스로서 위키피디어 백과사전 데이터를 이용하였으며, 이에 기반하여 출현 단어에 대한 개념벡터를 구성하는 방안을 제시하였다. 제안 모델에 의해 표현된 텍스트 큐보이드는 직관적이면서 고신뢰도의 텍스트마이닝 작업을 가능하게 한다. 예를 들어, 문서/단어/개념 클러스터링, 개념트리 및 연관망 구성, 단어 연관망 구성은 행렬 행태의 데이터간의 유사도 연산을 통해 수월하게 성취할 수 있다. 자동문서분류를 위해서는 분류모델의 구성 요소인 특징(feature)을 단어뿐만 아니라 단어와 개념이 융합된 형

태의 특징을 생성하면 된다. 또한 제안 모델은 텍스트 큐보이드 형태로 저장된 사용자 검색 프로파일을 활용하여 개인화 정보검색을 성취할 수 있다. 정보검색과 관련한 연구는 단어 리터럴을 개념 공간으로 매핑하여 보다 정확한 검색 결과를 찾는 문제와 검색 사용자별로 차별화된 검색 결과를 리턴시키기 위한 문제에 집중하고 있다. 제안 모델은 이 2가지 문제를 동시에 해결할 수 있는 단초가 될 것으로 판단한다. 향후 텍스트 큐보이드에 기반한 개인화 정보검색 모델을 구체화할 것이며, 대용량 텍스트 데이터에 대한 큐보이드의 저장, 분석을 수행하기 위하여 큐보이드 생성을 위한 맵리듀스(MapReduce) 알고리즘의 개발을 진행할 것이다. 또한 세부적으로 보다 정확한 개념벡터를 생성하기 위해 Conditional Random Field(CRF) 확률이론을 활용하고자 한다.

References

[1] Antonellis, I. and Gallopoulos, E., Exploring term-document matrices from matrix models in text mining, SIAM Text Mining Workshop, SIAM Conference on Data Mining, 2006.

[2] Berry, M. W., Survey of text mining : Clustering, Classification, and Retrieval, Springer-Verlag, 2003.

[3] Cai, D., He, X., Wen, J. R., Han, J., and Ma, W. Y., Support Tensor Machines for Text Categorization, Technical Report UIUCDCS-R-2006-2714, 2006.

[4] Cavnar, W. B. and Trenkle, J. M., N-Gram-Based Text Categorization, Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval, pp. 161-175, 1994.

[5] Faulkner, A., Automated Classification of Stance in Student Essays : An Approach Using Stance Target Information and the Wikipedia Link-Based Measure, Science, Vol. 376, No. 12, p. 86, 2014.

[6] Gabrilovich, E. and Markovitch, S., Feature generation for text categorization using world knowledge, Proceedings of International Joint Conferences on Artificial Intelligence, pp. 1048-1053, 2005.

[7] Howard, T. and Croft, W. B., Inference networks for document retrieval, Proceedings of International ACM SIGIR, pp. 1-24, 1989.

[8] http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

[9] http://www.statsoft.com/textbook/text-mining/.

[10] Jiang, C., Coenen, F., Sanderson, R., and Zito, M., Text Classification Using Graph Mining-Based Feature Extraction, Knowledge-Based Systems, Vol. 23, No. 4, pp. 302-308, 2009.

[11] Kimbrough, S., Executive Briefing : Text Mining for Business Intelligence, INSEAD-UNILEVER workshop, 2006.

[12] Lancaster, F. W. and Fayen, E. G., Information Retrieval On-Line, Melville Publi-

shing Co., 1973.

[13] Maron, M. and Kuhns, J., On relevance, probabilistic indexing and information retrieval, Journal of the Association for Computing Machinery, Vol. 7, pp. 216-244, 1960.

[14] Martinez, D. and Baldwin, T., Word sense disambiguation for event trigger word detection, Proceedings of the ACM fourth international workshop on Data and text mining in biomedical informatics, pp. 41-48, 2010.

[15] Navigli, R., Word sense disambiguation : A survey, ACM Computing Surveys, Vol. 41, No. 2, pp. 1-69, 2009.

[16] Ribeiro, B. and Muntz, R. A., Belief Network Model for IR, Proceedings of International ACM SIGIR, pp. 253-260, 1996.

[17] Salton, G., Wong, A., and Yang, C. S., A Vector Space Model for Automatic Indexing, Communications of the ACM, Vol. 18, No. 11, pp. 613-620, 1975.

[18] Schenker, A., Last, M., Bunke, H., and Kandel, A., Classification of Web Documents Using a Graph Model, Proceedings of 7th International Conference on Document Analysis and Recognition, pp. 240-244, 2003.

[19] Sui, Z., Zhao, Q., and Liu, Y., Inducting Concept Hierarchies from Text based on FCA, Proceedings of Fourth International Conference on Innovative Computing, Information and Control, pp. 1080-1083, 2009.

[20] Tamara, G. K. and Bader, B., Tensor Decompositions and Applications, SIAM Review, Vol. 51, No. 3, pp. 455-500, 2009.

[21] The Value and Benefits of Text Mining, JISC Digital Infrastructure, 2012.

[22] Witten, I. H., Text Mining, http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf.

[23] Wu, J., Xuan, Z., and Pan, D., Enhancing Text Representation for Classification Tasks with Semantic Graph Structures, International Journal of Innovative Computing, Information Control, Vol. 7, No. 5(B), pp. 2689-2698, 2011.

[24] Yeon, J., Shim, J., and Lee, S. G., Outlier Detection Techniques for Biased Opinion Discovery, Journal of Society for e-Business Studies, Vol. 18, No. 4, pp. 315-326, 2013.

[25] Zhang, B., Yan, J., Chen, Z., Liu, W., Bai, F., and Chien, L., Text representation : from vector to tensor, Fifth IEEE International Conference on Data Mining, pp. 725-728, 2005.

저 자 소 개

김한준 (E-mail : khj@uos.ac.kr)
1994년 서울대학교 계산통계학과 (공학사)
1996년 서울대학교 전산과학과 (이학석사)
2002년 서울대학교 컴퓨터공학부 (공학박사)
2002년~현재 서울시립대학교 전자전기컴퓨터공학부 교수
관심분야 텍스트마이닝, 데이터베이스, 기계학습, 정보검색, e-비즈니스 기술

장재영 (E-mail : jychang@hansung.ac.kr)
1992년 서울대학교 계산통계학과 (이학사)
1994년 서울대학교 계산통계학과 (이학석사)
1999년 서울대학교 계산통계학과 (이학박사)
2000년~현재 한성대학교 컴퓨터공학과 교수
관심분야 데이터베이스, 정보검색, 데이터마이닝