http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.21, No.4, November 2016, pp.1-14
http://dx.doi.org/10.7838/jsebs.2016.21.4.001

텐서공간모델 기반 시멘틱 검색 기법

A Tensor Space Model based Semantic Search Technique

홍기주(Kee-Joo Hong)*, 김한준(Han-Joon Kim)**,
장재영(Jae-Young Chang)***, 전종훈(Jong-Hoon Chun)****

초 록

시멘틱 검색은 검색 사용자의 인지적 노력을 최소화하면서 사용자 질의의 문맥을 이해하여 의미에 맞는 문서를 정확히 찾아주는 기술이다. 아직 시멘틱 검색 기술은 온톨로지 또는 시멘틱 메타데이터 구축의 난제를 갖고 있으며 상용화 사례도 매우 미흡한 실정이다. 본 논문은 기존 시멘틱 검색 엔진의 한계를 극복하기 위하여 이전 연구에서 고안한 위키피디아 기반의 시멘틱 텐서공간모델을 활용하여 새로운 시멘틱 검색 기법을 제안한다. 제안하는 시멘틱 기법은 문서 집합에 출현하는 '단어'가 텐서공간모델에서 '문서-개념'의 2차 텐서(행렬), '개념'은 '문서-단어'의 2차 텐서로 표현된다는 성질을 이용하여 시멘틱 검색을 위해 요구되는 온톨로지 구축의 필요성을 없앤다. 그럼에도 불구하고, OHSUMED, SCOPUS 데이터셋을 이용한 성능평가를 통해 제안 기법이 벡터공간모델에서의 기존 검색 기법보다 우수함을 보인다.

ABSTRACT

Semantic search is known as a series of activities and techniques to improve the search accuracy by clearly understanding users' search intent without big cognitive efforts. Usually, semantic search engines requires ontology and semantic metadata to analyze user queries. However, building a particular ontology and semantic metadata intended for large amounts of data is a very time-consuming and costly task. This is why commercialization practices of semantic search are insufficient. In order to resolve this problem, we propose a novel semantic search method which takes advantage of our previous semantic tensor space model. Since each term is represented as the 2nd-order 'document-by-concept' tensor (i.e., matrix), and each concept as the 2nd-order 'document-by-term' tensor in the model, our proposed semantic search method does not require to build ontology. Nevertheless, through extensive experiments using the OHSUMED document collection and SCOPUS journal abstract data, we show that our proposed method outperforms the vector space model-based search method.

키워드 : 시멘틱 검색, 텍스트 표현 모델, 위키피디어, 텐서공간모델, 벡터공간모델, 텍스트마이닝

Semantic Search, Text Representation Model, Wikipedia, Tensor Space Model, Vector Space Model, Text Mining

이 논문은 2015년도 한국연구재단의 개인기초연구사업(NRF-2015R1D1A1A09061299), 2015년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구(NRF-2015S1A5B6036362)이며, 또한 이 논문은 한성대학교 교내학술연구비 지원과제임.

* First Author, School of Electrical and Computer Engineering, University of Seoul(lovingmoon1@hanmail.net)
** 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)
**** Co-Author, Department of Data Technology, Myonji University(jonghoonchun@gmail.com)
Received: 2016-08-21, Review completed: 2016-10-07, Accepted: 2016-10-18

1. 서 론

전 세계적으로 스마트폰과 소셜미디어서비스로 인해 비/반정형 텍스트 데이터가 급격히 확산되면서 대용량 텍스트 데이터에서 사용자가 원하는 문서 또는 기사를 정확히 찾아주는 정보검색 기술의 중요성이 더욱 높아지고 있다[9]. 현재 대부분의 검색 엔진은 키워드(Keyword) 기반 알고리즘과 링크(Link) 기반 알고리즘으로 구현된다. 키워드 기반 알고리즘은 검색 질의어와 관련도가 큰 문서를 찾아서 정확도, 날짜순으로 정렬하는 방법이다. 키워드 기반 알고리즘의 대표적인 예는 벡터공간모델에서 TF-IDF(Term Frequency-Inverse Document Frequency) 가중치를 활용한 것이며, 이 가중치는 주어진 문서 집합에서 특정 단어가 해당 문서 내에서 얼마나 중요한지를 평가하는 통계적 수치이다[12]. 벡터공간모델에서 하나의 문서는 각 성분 값으로서 TF-IDF 가중치를 가지는 벡터로 표현되며 벡터간 코사인 유사도 함수를 이용하여 문서 검색이 이루어진다. 그러나 이를 웹 문서에 적용하는 경우에 문서간 연관 정보, 태그 정보 등을 활용하지 못하여 이를 해결하기 위해 링크 기반 알고리즘이 대두된다[7]. 이는 검색 질의와의 연관도뿐만 아니라 웹 페이지의 하이퍼링크 정보를 활용하여 질의와 무관하게 사용자가 선호할 가능성이 높은 문서를 찾는다. 여기서 웹 페이지의 인기도는 진입 링크의 수와 진출 링크의 수에 의해 결정된다.

이와 같은 노력에도 불구하고 현재 검색엔진은 사용자의 검색 의도를 정확히 파악하고 해당 의미에 적합한 검색 결과를 도출하기 위한 연구가 계속 진행되고 있다. 궁극적으로 검색엔진은 사용자의 인지적 노력을 최소화하면서 사용자 질의의 문맥을 이해하여 의미에 맞는 문서를 정확히 찾아주는 시멘틱 검색 시스템이 되어야 하기 때문이다[14]. 이러한 시멘틱 검색이 실현되기 위해서는 사용자가 제공한 검색 단어의 의미를 포착할 수 있는 온톨로지를 갖추고 있어야 한다[11]. 그런데 문제는 검색 대상이 되는 문서와 관련된 온톨로지를 구축하고 이를 유지 보수하는 과정이 많은 시간과 비용을 요구한다는 것이다.

본 논문에서는 이러한 시멘틱 검색의 장애요소를 극복하기 위하여 이전 연구에서 제안한 위키피디아 기반 텐서공간모[2, 5]을 활용한 새로운 시멘틱 검색 기법을 제안한다. 이전 연구의 시멘틱 텐서공간모델에서는 FCA (Formal Concept Analysis) 이론체계에서의 '개념' 정의를 확장하여, 텍스트 문서를 '단어'와 '개념' 간의 관계를 표현한 2차 텐서로 표현하고, 결과적으로 하나의 문서 집합은 문서-단어-개념의 3차 텐서로 표현한다[5]. 제안하는 시멘틱 기법은 텐서공간모델에서 문서 집합의 '단어'가 텐서공간모델에서 '문서-개념'의 2차 텐서, '개념'은 '문서-단어'의 2차 텐서로 표현되는 성질을 이용하여, 시멘틱 검색을 위해 요구되는 온톨로지의 필요성을 없앤다. 또한 주어진 사용자 질의와 연관된 문서를 찾기 위해 텐서공간모델에서의 유사도 함수와 질의 단어와 의미적으로 연관된 개념을 융합함으로써 검색 정확도를 높인다.

본 논문의 구성은 다음과 같다. 제2장에서는 관련 연구를 서술하며, 제3장에서는 제안하는 시멘틱 검색 시스템 및 알고리즘을 포함한다. 제4장에서는 제안 기법의 성능을 검증하기 위해 실험 및 성능 평가 결과에 대해 기술한다. 마지막으로 제5장에서 결론을 맺는다.

2. 관련 연구

시멘틱 검색은 사용자의 검색의도를 능동적으로 파악하여 검색의 정확도를 향상시키는 일련의 활동 및 방법론을 통칭한다. 시멘틱 검색의 요소 기술은 '지식 획득', '지식 표현', 그리고 '지식 활용'으로 나눌 수 있다. 첫째, '지식 획득'은 주어진 콘텐츠로부터 시멘틱 메타데이터를 생성하는 과정이며, 주어진 콘텐츠에서 의미있는 시멘틱 메타데이터를 추출하기 위해 WSD(Word Sense Disambiguation)[8], NER (Named-Entity Recognition)[6] 등의 기술이 활용된다. 둘째, '지식 표현'은 개념과 단어 정보 간의 연결과 추론을 위한 온톨로지를 구축하는 것이다. 마지막으로, '지식 활용'은 검색 사용자 관점에서 시멘틱 검색 시스템과의 인터페이스 영역으로서 온톨로지 기반 질의 확장 등을 통해 시멘틱 검색을 처리하는 과정을 포함한다. 현재 시멘틱 검색의 상용화에 어려움을 겪는 이유를 요소 기술별로 기술하면 다음과 같다. 우선 '지식 획득' 측면에서 정확한 시멘틱 메타데이터 생성이 쉽지 않다. 텍스트 문서 내에 포함된 단어의 의미는 문장 속에서 다른 단어와의 관계를 통해서 그 의미를 판별해야 하며, 최근 이를 위해 복잡한 CRF(Conditional Random Field)와 같은 확률 이론이 사용되기도 한다[13]. '지식 표현'의 측면에서는 개념 또는 단어 간 관계성을 담고 있는 온톨로지를 지속적으로 유지 보수해야 하는 문제이다. 문서 집합의 특성 및 시간의 흐름에 따라 개념 간의 관계가 변동되기 때문에 정확한 시멘틱 검색을 성취하기 위해서 점진적으로 온톨로지를 갱신하는 작업이 중요하다. 현재 이와 관련한 많은 연구가 정적 온톨로지를 구성하는 것에 목표를 두고 있다[1]. 마지막으로 '지식 활용' 측면은 사용자의 검색의도를 인지하여 질의 단어의 의미를 정확히 판별하기 위한 문제와 관련되며 본 연구에서는 주어진 사용자 질의와 의미적으로 연결된 다수의 확장 질의를 자동 추천하는 방식을 취한다.

위와 같은 난제를 고려한 시멘틱 검색 시스템은 기존 정보검색 시스템이 가지는 다양한 기능들(예: 문서 수집, 가공, 인덱싱, 질의처리 등)을 온톨로지와 연관지어 보다 정교한 메커니즘으로 승화되어야 한다. 특히 이는 시멘틱 질의처리를 위해서 문서에 포함된 단어의 의미를 판별하고 검색어의 애매성을 제거하는 기능을 포괄해야 하며, 검색 정확도를 높이기 위해서 온톨로지 정보를 고려한 유사도 함수를 갖추어야 한다. 본 연구는 근본적으로 벡터 공간모델이 아닌 위키피디어 기반의 시멘틱 텐서공간모델을 가지고 텍스트 문서를 표현함으로써 온톨로지가 필요 없는 시멘틱 검색 시스템을 구성하여 시스템의 복잡도를 줄이고 검색 정확도를 높이고자 한다.

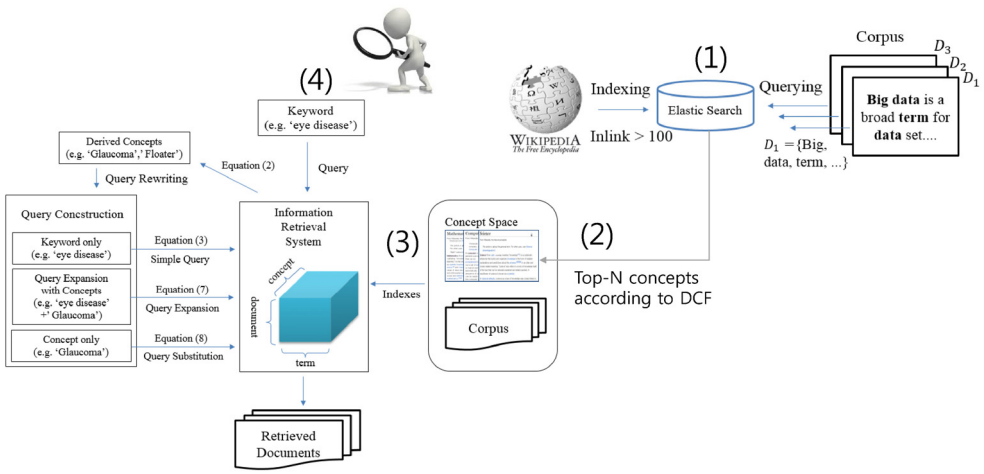
3. 시멘틱 검색

3.1 시멘틱 검색 시스템의 구조

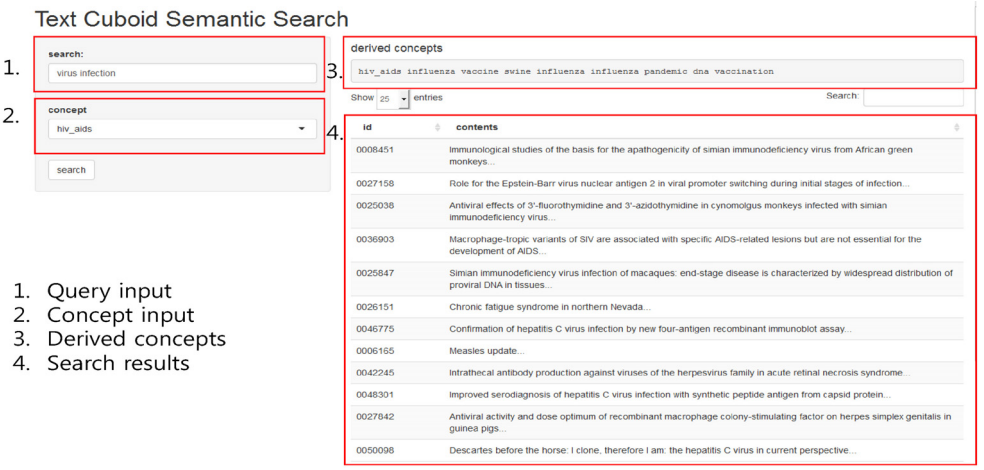
<Figure 1>은 본 논문이 제안하는 시멘틱 검색 시스템의 구조 절차를 보여준다. 우선 Berlanga et al.[2]에서 제안한 시멘틱 텐서공간의 구성을 위해 ① 주어진 문서 집합(Document Corpus)에 적합한 위키피디어 문서를 추출, ② 이에 기반하여 개념 공간을 자동 생성한다. 그리고 나서 ③ 각 문서의 출현 단어가 개념 공간상에

가지는 중요도를 평가함으로써 '개념-단어-문서'(concept-term-document)의 3차 텐서를 생성한다. 그 이후 ④ 사용자가 검색엔진에 질의를 전달하면, 주어진 질의 단어를 텐서의 '단어' 축과 비교하여 연관된 개념들을 도출하고, 3차 텐서공간에서의 유사도 계산 및 질의 확장 과정을 거쳐 개념적으로 연관된 문서들을 반환하게 된다.

<Figure 2>는 시멘틱 검색을 위한 사용자 인터페이스를 보여준다. 좌측 상단에 위치한 입력 영역은 (1) 질의 입력 search 부분과 (2) 개념 입력 concept 부분으로 구성된다. concept 부분은 제3.3.2절에서 소개하는 질의 확장 기법에 사용되는 정보로서, 사용자에 의해 직접 입력되거나 자동 입력될 수 있다. 우측 상단의 (3) derived concepts 부분에는 입력한 질의 및 개념과



〈Figure 1〉 The Proposed Semantic Search System Architecture



〈Figure 2〉 Semantic Search User Interface

관련된 연관 개념들의 집합이 출력된다. 또한 (4) 우측 하단의 (4) Search results 영역에는 제안하는 시멘틱 검색 알고리즘에 의해 도출된 문서 집합이 유사도 순서대로 출력된다.

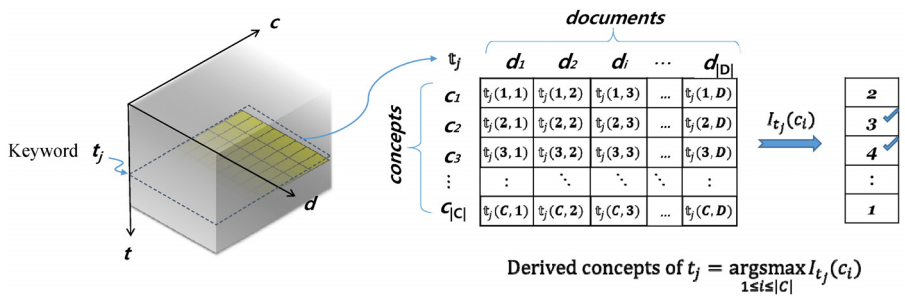
3.2 연관 개념의 추출

시멘틱 검색의 핵심은 사용자 질의의 애매성을 제거하여 질의 의도를 정확히 파악하는 것이다. 본 연구에서는 주어진 문서집합에 따라 자동 구성된 개념 공간으로부터 사용자 질의와 관련이 높은 개념들을 추출함으로써 보다 정확한 검색 결과를 얻게 된다. 예를 들어, 'eye disease(눈병)'라는 질의어와 연관된 'glaucoma(녹내장)', 'cataract(백내장)' 등의 개념을 도출하여 이를 활용하여 만족도가 높은 검색 결과를 얻을 수 있는 것이다. 이러한 연관 개념이 검색 정확도의 개선에 기여할 수 있는 것은 기 구축된 3차 텐서가 검색 대상이 되는 문서집합에 의해 자동 구성된 개념공간을 확보하고 있기 때문이다.

<Figure 3>은 본 논문에서 제안하는 3차 텐서 모델상의 단어 공간 축에 대한 한 단면인 목표 단어 t_j의 '개념-문서'(concept-by-document)의 2차 텐서(즉, 행렬)를 표현한 것으로서, 이는 목표 단어 t_j에 대하여 주어진 문서와 관련된 개념과의 관계를 정량적으로 보여준다. 이 행렬의 행 또는 열 기준의 합을 구하면, 하나의 단어를 개념벡터 또는 문서벡터로 표현할 수 있다. 특히 개념벡터를 통해 각 개념 성분에 해당하는 비중값을 확인함으로써 다수의 의미를 가지는 다의어들의 개념 분포를 산출할 수 있다. 제안 시스템은 주어진 문서 집합에서 질의어와 관련이 높은 개념 집합을 추출해야 하는데, 이를 위해 질의어에 포함된 단어를 목표 단어로 간주하여 이에 대한 '개념-문서' 2차 텐서에서 문서 공간 축의 모든 컬럼(column) 값들을 합산하여 수치가 높은 개념들을 선별한다. 이를 정형적으로 표현하기 위해, '개념-문서' 2차 텐서를 이용하여 목표 단어 t_j에 대한 개념 c_i의 중요도(importance)를 식 (1)의 $I_j(c_i)$로 정의한다.

$$I_{t_j}(c_i) = \sum_{k=1}^{|D|} t_j(c_i, d_k) \qquad (1)$$

여기서 $t_j(c_i, d_k)$은 단어 t_j가 문서 d_k내에서 개념 c_i의 의미를 가지는 비중값을 의미하며, $|D|$와



〈Figure 3〉 Extracting Associated Concepts Using the 'Concept-by-Document' Tensor(Matrix)

$|C|$는 각각 문서집합 내 문서의 총 개수, 개념 공간 내 개념의 총 개수를 의미한다. <Figure 3>은 목표 단어 t_j에 대한 연관 개념을 도출하는 과정을 도식화 한 것이다.

개념 공간에 존재하는 모든 개념에 대하여 $I_{t_j}(c_i)$ 값을 평가하여 임계값 이상이 되는 연관 개념을 판별하는 것이다. 이를 식 (2)와 같이 표현할 수 있다.

$$Concepts\ derived\ with\ t_j = \mathrm{argsmax}_{1 \le i \le |C|} I_{t_j}(c_i) \quad (2)$$

여기서 $\mathrm{argsmax}_{x \in X} F(x)$는 함수값 $F(x)$을 최대로 하는 한 개 이상의 인자 x들로 구성된 집합 X를 의미한다. 식 (2)에 주어지는 임계값은 문서집합, 질의어에 따라서 변동될 수 있다. 만약 질의어에 2개 이상의 단어가 포함된 경우에는 각 단어에 해당하는 '개념-문서' 2차 텐서들을 합산하여 식 (2)에 따라 중요도가 높은 개념들을 선정하면 된다. 예를 들어, 질의어가 'big data'와 같이 복합 명사일 경우 'big'과 'data'를 분할하여, 단어 'big'에 대한 '개념-문서' 행렬과 단어 'data'에 대한 '개념-문서' 행렬을 더하고, 개념 공간 내 모든 개념에 대한 $I_{t_j}(c_i)$ 값을 산출하며, 'big data'와 연관된 개념들을 얻을 수 있다.

3.3 시멘틱 검색 알고리즘

3.3.1 유사도 기반 시멘틱 검색

기존 벡터공간모델에서는 하나의 문서를 N-차원 단어에 대한 벡터로 표현하며, 벡터의 각 성분은 출현단어의 TF-IDF 가중치를 사용한다. 일반적으로 사용자 질의에 적합한 문서를 검색하기 위해서 벡터로 표현된 문서와 질의 간의 코사인(cosine) 유사도를 계산한다[3].

$$\mathrm{sim}(\vec{d}, \vec{q}) = \frac{\langle \vec{d}, \vec{q} \rangle}{\| \vec{d} \| \cdot \| \vec{q} \|} \quad (3)$$

이와 유사하게, 3차 텐서공간모델에서는 문서가 '단어-개념'의 2차 텐서인 행렬로 표현되므로, 행렬간의 유사도를 산출하여 문서와 질의간의 유사성을 평가할 수 있다. 사실 행렬간의 유사도는 아직 수학적으로 정의되어 있지 않지만, 식 (3)의 코사인 유사도를 확장하여 텐서공간모델에서의 문서 $\mathbb{D}$와 질의 $\mathbb{Q}$ 간의 유사도 함수를 식 (4)와 같이 정의할 수 있다.[2)]

$$\mathrm{sim}(\mathbb{D}, \mathbb{Q}) = \frac{\langle \mathbb{D}, \mathbb{Q} \rangle_F}{\| \mathbb{D} \|_F \cdot \| \mathbb{Q} \|_F} \quad (4)$$

여기서 $\langle \mathbb{D}, \mathbb{Q} \rangle_F$는 행렬 $\mathbb{D}$와 $\mathbb{Q}$에 대한 프로베니우스 곱(Frobenius product)을 의미한다. 또한 $\| \mathbb{D} \|_F$는 행렬 $\mathbb{D}$에 대한 프로베니우스 노옴(Frobenius norm)을 의미하며, 이는 $\sqrt{\langle \mathbb{D}, \mathbb{D} \rangle_F}$로 정의된다[10]. $\langle \mathbb{D}, \mathbb{Q} \rangle_F$은 식 (5)에서 보는 바와 같이, 행렬 $\mathbb{D}$와 $\mathbb{Q}$의 동일한 행과 열 위치에 존재하는 성분값의 곱을 모두 더한 것이다.

$$\langle \mathbb{D}, \mathbb{Q} \rangle_F = \sum_{i=1}^{n} \sum_{j=1}^{m} (\mathbb{D}_{ij} \cdot \mathbb{Q}_{ij}) \quad (5)$$

여기서, $\mathbb{D}_{ij}$는 행렬 $\mathbb{D}$에서 i-번째 행, j-번째 열에 위치한 성분값을 의미하며, n과 m은 각각 총 단어의 개수, 총 개념의 개수를 의미한다. 결론적으로 식 (4)를 기반으로 주어진 질의와

2) 본 논문에서 문서 d에 대한 벡터 표현 심벌은 $\vec{d}$, 2차 텐서인 행렬 표현 심벌은 $\mathbb{D}$로 표기한다.

문서간의 유사도를 계산하여 유사도 순으로 문서 집합을 반환할 수 있다. 그런데 여기서 질의 행렬 $\mathbb{Q}$는 텐서공간모델에서의 문서와 동일한 구조의 행렬로 표현됨을 가정하였다. 초기에 주어진 사용자 질의 자체는 단순한 단어의 나열이므로 이를 행렬 구조로 변환하는 작업이 필요하다. 이를 위해 제3.2절에서 소개한 연관 개념 도출 기법을 활용하여 쉽게 질의 행렬을 구성할 수 있다. 구체적으로, 사용자 질의는 1개 이상의 단어를 포함하는데, 그 단어에 대하여 개념공간에서의 연관 정보를 구하여 문서와 동일한 구조의 단어-개념(term-by-concept) 행렬을 구성할 수 있는 것이다. 여기서 단어에 대한 개념 벡터의 성분값은 간단히 연관성 여부를 나타내는 값으로 정의한다. 즉 특정 단어에 대해 연관된 개념 성분은 1, 그렇지 않은 개념 성분은 0으로 표현한다. <Figure 4>는 2개의 단어 t_{q1}, t_{q_2}를 포함하는 질의에 대한 행렬을 구성하는 예를 보여준다. 단어 t_{q1}를 표현하는 개념-문서(concept-by-document) 행렬로부터 $I_{t_{q1}}(c_i)$ 값이 임계값 이상인 연관 개념 c_2, c_3를 도출하였고, 단어 t_{q1}의 행렬로부터 연관 개념 c_1, c_3를 도출하였다. 이를 기반으로 t_{q1}, t_{q_2}의 개념 벡터 구성하여 <Figure 4>의 우측 하단의 단어-개념(term-by-concept) 행렬을 생성하게 된다.

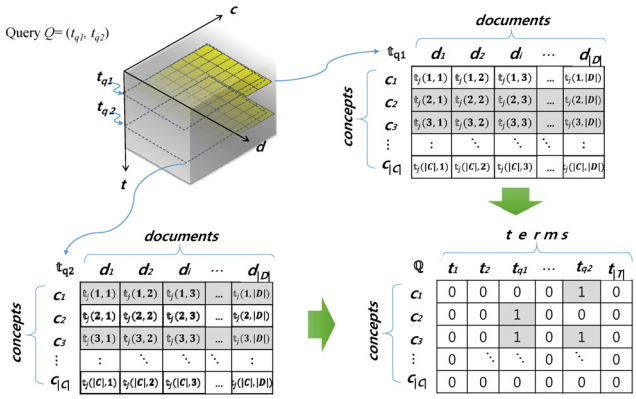
3.3.2 개념 기반의 질의 확장

제3.3.1절에서 제안한 기법과 더불어 본 절에서는 사용자가 단어와 개념을 함께 제시하는 질의 기법을 소개한다. 텐서공간모델에서 사용자 질의와 개념을 효과적으로 검색에 활용함으로써 검색 정확도를 더욱 높일 수 있다. 식 (6)은 초기 질의 q에 개념 c_i를 포함한 질의 q'를 표현한 것이다.

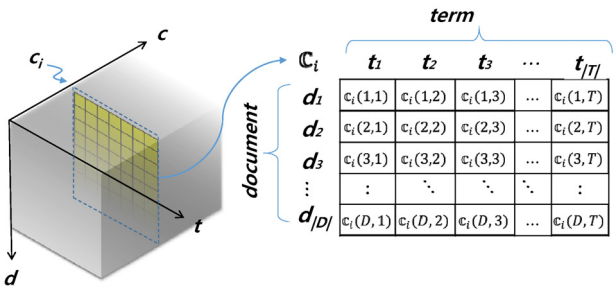
$$q' = q + c_i \tag{6}$$

<Figure 5>는 텐서공간에서 개념공간 축의 한 단면으로서 개념 c_i의 '문서-단어'(document-by-term) 행렬을 보여준다. 본 연구에서 단어 수준의 질의와 개념 수준의 정보를 동시에 적용할 수 있는 유사도 함수를 다음의 식 (7)과 같이 제안한다.

$$\text{sim}\ (d, q') = \beta \cdot \text{sim}\ (\mathbb{D}, \mathbb{Q}) + (1-\beta) \cdot I_{c_i}(d) \tag{7}$$



〈Figure 4〉 Composing the Concept-by-term Tensor(Matrix) for the User Query



〈Figure 5〉 Term-by-Document Matrix for the Concept c_i in the Concept Space

$$I_{c_k}(d) = \frac{\sum_{j=1}^{|T|} c_i(d, t_j)}{|d|} \quad (8)$$

식 (7)에서 $\mathrm{sim}(\mathbb{D}, \mathbb{Q})$는 문서 d와 초기 질의 q의 유사도(식 (4) 참조)를 의미하며, $I_{c_i}(d)$는 연관 개념 c_i에 대한 문서 d의 중요도를 나타낸다. 식 (8)에서 $|T|$는 단어공간에 포함된 단어의 총 수, $|d|$은 문서 d에 출현한 단어의 수를 의미한다. 여기서 $|d|$는 단어 수가 많은 문서의 중요도가 부적절하게 높아지는 것을 방지하기 위해 정규화 차원에서 사용되었다. β는 $I_{c_i}(d)$의 가중치를 제어하기 위해 사용되며, 만약 질의어 없이 오로지 개념만으로 질의를 수행하고자 한다면 β값을 0으로 설정한다. 본 연구에서 실험적으로 얻은 최적의 β값은 0.7이다.

4. 실험 및 평가

4.1 실험데이터

본 연구에서 제안한 시멘틱 기법의 성능을 평가하기 위해 OHSUMED, SCOPUS 데이터셋을 활용하여 개념-문서-단어(concept-document-term)의 3차 텐서공간을 생성하였다.

4.1.1 OHSUMED 데이터셋

OHSUMED 데이터셋은 1987년부터 1991년까지 5년간 온라인 의학 정보 데이터베이스인 MEDLINE에서 발췌한 270개 의학 저널의 초록 348,566개를 모은 데이터셋으로 텍스트 마이닝 알고리즘의 평가를 위해 자주 사용된다[4]. 실제 성능평가를 위해 6개 카테고리(virus diseases, musculoskeletal diseases, eye diseases, female genital diseases, cardiovascular diseases)에서 총 3,000개의 문서를 선정하였다.

4.1.2 SCOPUS 데이터셋

SCOPUS 데이터셋은 SCOPUS 인덱스에 등재된 학술 저널 기사에 대한 초록 및 인용을 제공하는 데이터베이스이다. 본 연구에서 시간의 흐름에 따른 연관 개념의 변화를 측정하기 위하여 2009년부터 2014년까지 총 1,000개의 'Big data' 관련 논문 초록을 선정하였다.

4.2 성능평가

본 연구는 시멘틱 검색 기법의 성능 평가 척도로서 사용자의 검색 만족도를 정량적으로 평

가하는 $nDCG$(Normalized Discounted Cumulative Gain)를 사용하였다. p개의 검색 결과에 대한 $nDCG_p$의 계산을 위해 먼저 식 (9)의 DCG_p를 계산한다.

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i} \quad (9)$$

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (10)$$

여기서 rel_i은 실제 검색된 문서들 중에서 i번째 문서의 관련도를 나타낸다. 이 때 관련도는 실험 데이터 관련 전문가 3인에 의해 결정되었다. 제출한 결정상위에 랭크된 문서들은 하위에 랭크된 문서보다 중요하다는 가정하에 식 (9)는 순위 정보에 따라 패널티를 부여한 형태가 된다. 식 (10)의 $nDCG_p$는 DCG_p를 정규화한 척도이며, 구체적으로 DCG_p값을 분자로 하고 관련도를 내림차순으로 정렬하여 DCG를 계산한 $IDCG$를 분모로 하여 계산한다. 따라서 $nDCG$ 척도는 0과 1 사이의 값을 가지며 1에 가까울수록 우수한 검색 결과임을 의미한다.

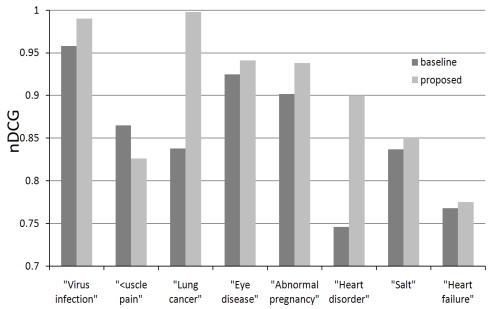
우선 제안하는 시멘틱 검색 기법의 효능을 단적으로 확인하기 위해 제3.2절에서 소개한 연관 개념 도출 기법에 따라 도출된 연관 개념들을 살펴보자. <Table 1>은 OHSUMED 데이터셋에서 'virus infection', 'muscle pain', 'lung cancer', 'eye disease', 'abnormal pregnancy', 'heart disorder', 'salt', 'heart failure'등의 질의어를 통해 도출된 개념들을 중요도에 따라 순서대로 나열한 것이다.

예를 들어, 질의어 'eye disease'에 대하여 'Glaucoma', 'Cataract surgery', 'Floater' 등의 연관 개념이 도출되었다. 이 연관 개념들은 제3.3절에서 질의 단어의 행렬화와 질의 확장 기법에 활용되어 사용자 질의 의미를 명확히 하는데 기여한다. <Figure 6>은 제3.3.1절에서 제안한 시멘틱 검색 기법에 따라 OHSUMED 데이터셋에서 'virus infection', 'muscle pain', 'lung cancer', 'eye disease', 'abnormal pregnancy', 'heart disorder', 'salt', 'heart failure' 등의 질의어를 통해 반환된 검색결과 집합의 만족도를 $nDCG$ 척도를 가지고 평가한 것이다. 그림에서 보는 바와 같이 'muscle pain'을 제외한 질의에 대해서 제안 기법이 2~15% 가량 검색 만족도를 개선하였다. 질의 단어 'muscle'가 주어진 경우에,

〈Table 1〉 Examples of the Concepts Associated with User Queries in the OHSUMED Dataset

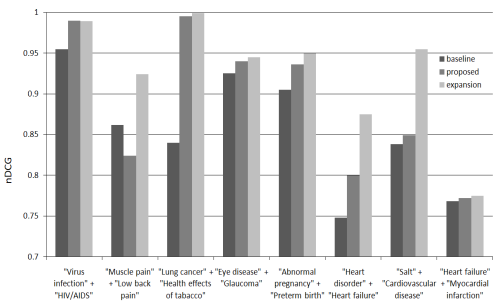
Queries	Top-3 derived concepts
'virus infection'	HIV/AIDS, DNA vaccination, Swine influenza
'muscle pain'	Low back pain, Anterior cruciate ligament injury, Back pain
'lung cancer'	Lung cancer, Health effects of tobacco, Acute respiratory distress syndrome
'eye disease'	Glaucoma, Cataract surgery, Floater
'abnormal pregnancy'	Preterm birth, Abortion, Pre-eclampsia
'heart disorder'	Heart failure, Health effects of tobacco, Hypertension
'salt'	Glaucoma, Hypertension, Cardiovascular disease
'heart failure'	Hypertrophic cardiomyopathy, Hypertension, Myocardial infarction

이는 'Extraocular muscles', 'Cardiac muscle' 등 다양한 신체 부위의 'muscle'에 대한 애매성이 존재하여 *nDCG*값이 낮게 평가된 것으로 분석된다. 이와는 반대로 질의 단어 'heart disorder'가 주어진 경우에 제안 기법의 성능이 상대적으로 매우 높은 것은 단어 'disorder'가 '질환'이란 의미를 가지는 용어로 흔히 사용하는 의학 용어일 뿐만 아니라 '무질서'의 의미를 가지는 일상용어이기 때문이다. 이러한 이유로 기존 기법은 'heart disorder'에 의한 검색 결과로서 '심장 질환'이 아닌 다른 의미의 질환 관련 문서 및 '무질서'의 의미를 가지는 문서를 다수 검색하는데 반해서, 제안 기법은 의미적으로 명확한 '심장 질환'과 관련된 문서를 검색하게 된다.



〈Figure 6〉 Comparison of *nDCG* for Various Queries

위와 같은 질의 애매성을 적극적으로 해소하기 위해 제3.4.2절에서 제안한 질의 확장 기법을 활용할 수 있다. 질의어 'muscle pain'에 대하여 연관 개념 'Low back pain'을 가지고 질의 확장을 시도하고 검색의도를 더욱 명확하게 하였다. 여기서 질의 확장을 위해 선정된 연관 개념은 최상위 가중치를 가지는 개념에 해당하며, 질의 단어에 연관된 개념이 다수 개 존재하는 경우에 사용자가 선택할 수 있는 과정이 포함될 수 있다. 식 (7)을 활용하여 질의 처리를 수행한 결과, <Figure 7>과 같이 벡터공간모델 기반 검색 기법에 비해 평균 6% 가량 *nDCG* 값이 개선된 것을 확인할 수 있다. 이 그림의 가로축에는 각 질의 단어에 연관된 개념이 제시되어 있다. 여기서 상대적으로 높은 검색 성능을 보이는 질의어가 'Salt+cardiovascular disease'인데, 이는 <Figure 6>의 질의어 'heart disorder'에 대한 성능 차이에 대한 동일한 이유로 설명할 수 있다. 즉 단어 'disease'는 '질환'이라는 의미를 가지며 이는 다른 단어에 비해 질병 명칭에 사용되는 사례가 많기에 이를 포함한 질의에 대해 잘못된 검색 결과를 넘겨줄 수도 있는 것이다. 이는 본 논문의 제안 기법이 다중 의미를 가지는 또는 빈번히 활용되는 특정 단어가 다른 단어가 결합하여 검색이 될 때 그 질의 단어의 명확한 의미를 포착하는 효과가 우수함을 증명하는 것이다.

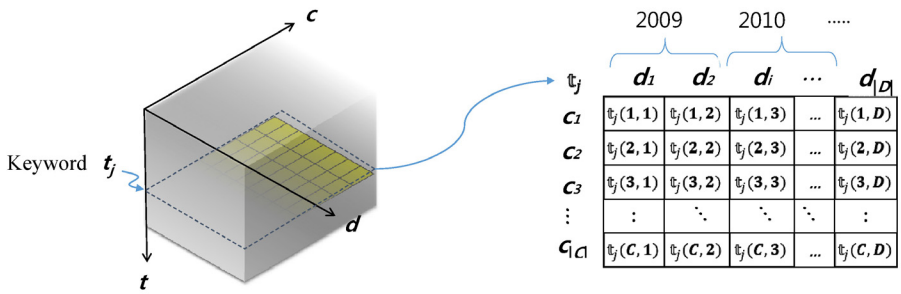


〈Figure 7〉 Improving the Search Algorithm Through Query Expansion Using Associated Concepts

특정 질의와 관련된 개념들은 시간의 흐름에 따라 유동적이므로 해당 시기에 적합한 연관 개념이 도출되어야 할 것이다.

<Figure 8>과 같이 문서 집합을 연도별로

분할한 후 각 분할별로 제3.2절에서 제안한 연관 개념 도출 알고리즘을 활용하여 연관 개념을 도출하였다. <Table 2>는 SCOPUS 데이터 셋에서 연도별로 'Big data'라는 질의를 통해 도출된 개념의 순위를 나열한 것이다. 실험 결과, 2009~2012년에는 'Cloud Computing', 'Predictive analytics'이 상대적으로 빅데이터와 관련성이 높은 개념이었고, 비교하여 2013~2014년은 IoT(Internet of Things)가 빅데이터와 가장 관련이 높은 개념으로 평가되었다. 그리고 2011 ~ 2012년에는 'Apache Hadoop'이 새롭게 빅데이터 관련 개념으로 도출되었다. 이렇듯 시간에 흐름에 따라 변화하는 연관 개념들을 반영한 결과라 할 수 있다.



〈Figure 8〉 Partitioning the Document Corpus for Extracting Associated Concepts Over Time

〈Table 2〉 Extracting the Concepts Associated with 'Big Data' Over Time

Rank	2009		2010		2011		2012		2013		2014	
	Concept	Weight	Concept	Weight	Concept	Weight	Concept	Weight	Concept	Weight	Concept	Weight
1	Big data	0.119	Big data	0.112	Big data	0.532	Big data	0.644	Big data	0.583	Big data	0.512
2	Cloud computing	0.1	Predictive analytics	0.098	Cloud computing	0.471	Predictive analytics	0.523	Internet of Things	0.511	Internet of Things	0.469
3	Predictive analytics	0.1	Clould computing	0.093	Predictive analytics	0.437	Cloud computing	0.49	Predictive analytics	0.468	Predictive analytics	0.432
4	Distributed file system for cloud	0.093	Chemical biology	0.087	Distributed file system for cloud	0.375	Analytics	0.41	Cloud computing	0.437	Cloud computing	0.387
5	Chemical biology	0.09	Distributed file system for cloud	0.079	Chemical biology	0.335	Social networking service	0.405	Social networking service	0.358	Social networking service	0.341
6	Digital preservation	0.078	Social networking service	0.079	Analytics	0.328	Distributed file system for cloud	0.402	Analytics	0.358	Chemical biology	0.335
7	Social networking service	0.078	Health informatics	0.078	Social networking service	0.325	Chemical biology	0.38	Chemical biology	0.343	Data mining	0.322
8	Health informatics	0.075	Digital preservation	0.076	Data mining	0.318	Data mining	0.375	Distributed file system for cloud	0.337	Analytics	0.32
9	Data mining	0.074	Data mining	0.07	Health informatics	0.296	Health informatics	0.327	Data mining	0.326	Health informatics	0.315
10	Metagenomics	0.072	Artificial neural network	0.07	Apache Hadoop	0.281	Apache Hadoop	0.327	Health informatics	0.311	Distributed file system for cloud	0.297

5. 결론 및 향후 연구 방안

차세대 검색 엔진은 사용자가 원하는 정보를 좀 더 쉽게 찾을 수 있도록 시멘틱 정보를 활용하고 개인화 검색과 융합하면서 발전하게 될 것이다. 본 논문은 시멘틱 검색 시스템을 구축할 때 가장 난제로 꼽히는 온톨로지 구축 문제를 해결하기 위하여, FCA 이론체계에 입각하여 문서집합을 '개념-단어-문서'의 3차 텐서로 표현하는 텐서공간모델에서의 새로운 시멘틱 검색 기법을 제안하였다. 이 텍스트 모델에서 '단어-개념' 행렬로 표현된 문서 행렬은 개념 수준의 위키피디어 문서와 견주어 출현 단어의 시멘틱 정보를 함유하고 있기 때문에, 온톨로지 없이 자연스럽게 사용자 질의에 대한 시멘틱 검색이 가능하다. 이는 기본적으로 텐서공간모델에서 질의 단어와 연관된 개념의 추출과 행렬로 표현된 질의와 문서간의 유사도 비교를 통해 이루어진다. 또한 사용자 질의로부터 추출한 개념을 활용하여 관련 문서를 정확히 추출할 수 있다. 향후 제안한 시멘틱 검색 기법을 개인화 검색이 가능하도록 확장할 예정이며, 이것의 기본 아이디어는 사용자 검색 기록 정보를 담는 프로파일 정보를 '개념-단어'(concept-by-term) 2차원 텐서로 표현하고, 이를 기반으로 초기 사용자 질의에 포함된 의미 정보와 검색 의도를 정확히 반영한 질의로 확장하는 것이다.

References

[1] Baeza-Yates, R. and Ribeiro-Neto, B., Modern information retrieval: The Concepts and Technology behind Search, New York: ACM Press, Chapter 3, 2011.

[2] Berlanga, R., Nebot, V., and Pérez, M., "Tailored semantic annotation for semantic search," Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 30, pp. 69-81, 2015.

[3] Gantz, J. and David R., "The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east," IDC iView: IDC Analyze the Future 2007, pp. 1-16, 2012.

[4] Heck, L. P., Hakkani-Tür, D., and Tür, G., "Leveraging knowledge graphs for web-scale unsupervised semantic parsing," INTERSPEECH, pp. 1594-1598, 2013.

[5] Kim, H. J. and Chang, J. Y., "A Semantic Text Model with Wikipedia-based Concept Space," The Journal of Society for e-Business Studies, Vol. 19, No. 3, pp. 107-123, 2014.

[6] Kim, H. J., Hong, K. J., and Chang, J. Y., "Semantically enriching text representation model for document clustering," Proceedings of the 30th Annual ACM Symposium on Applied Computing, pp. 922-925, 2015.

[7] Nadeau, D. and Sekine, S., "A survey of named entity recognition and classification," Lingvisticae Investigationes, Vol. 30, No. 1, pp. 3-26, 2007.

[8] Navigli, R., "Word sense disambiguation: A survey," ACM Computing Surveys

(CSUR), Vol. 41, No. 2, pp. 1-69, 2009.

[9] Page, L., Brin, S., and Motwani, R., Winograd, T., "The PageRank citation ranking: bringing order to the Web," 1999.

[10] Rossi, R. G., Marcacini, R. M., and Rezende, S. O., "Benchmarking text collections for classification and clustering tasks," Institute of Mathematics and Computer Sciences, University of Sao Paulo, 2013.

[11] 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.

[12] Sudeepthi, G., Anuradha, G., and Babu, M. S. P., "A survey on semantic web search engine," IJCSI International Journal of Computer Science Issues, Vol. 9, No. 2, pp. 241-245, 2012.

[13] Tablan, V., Bontcheva, K., and Roberts, I., Cunningham, H., "Mímir: An open-source semantic search framework for interactive information seeking and discovery," Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 30, pp. 52-68, 2015.

[14] Yang, K. and Shahabi, C., "A PCA-based similarity measure for multivariate time series," Proceedings of the 2nd ACM international workshop on multimedia databases, pp. 65-74, 2004.

저 자 소 개

홍기주 (E-mail: lovingmoon1@hanmail.net)
2009년 한동대학교 경영경제학과 (경영학사)
2015년 서울시립대학교 전기컴퓨터공학부 석사과정
2015년~현재 이스트소프트 연구원
관심분야 텍스트마이닝, 온톨로지, 기계학습, 데이터베이스, 빅데이터 분석

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

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

전종훈 (E-mail: jonghoonchun@gmail.com)
1986년 Computer Science, University of Denver (학사)
1988년 Computer Science, Northwestern University (공학석사)
1992년 Computer Science, Northwestern University (공학박사)
1992년~1995년 University of Central Oklahoma Department of Computing Science 조교수
1995년~현재 명지대학교 융합소프트웨어학부 정교수
관심분야 빅데이터, 데이터베이스 시스템, 정보검색, 스마트사이니지, e-비지니스 기술