J. Korea Saf. Manag. Sci. Vol. 17 No. 4 December 2015
http://dx.doi.org/10.12812/ksms.2015.17.4.335

ISSN 1229-6783(Print)
ISSN 2288-1484(Online)

특허 문서로부터 키워드 추출을 위한 위한 텍스트 마이닝 기반 그래프 모델

이 순 근* · 임 영 문* · 엄 완 섭*
*강릉대학교 산업경영공학과

Text-mining Based Graph Model for Keyword Extraction from Patent Documents

Soon Geun Lee* · Young Moon Leem* · Wan Sup Um*
*Dept. of Industrial & Management Engineering, Gangneung-Wonju National University

Abstract

The increasing interests on patents have led many individuals and companies to apply for many patents in various areas. Applied patents are stored in the forms of electronic documents. The search and categorization for these documents are issues of major fields in data mining. Especially, the keyword extraction by which we retrieve the representative keywords is important. Most of techniques for it is based on vector space model. But this model is simply based on frequency of terms in documents, gives them weights based on their frequency and selects the keywords according to the order of weights. However, this model has the limit that it cannot reflect the relations between keywords. This paper proposes the advanced way to extract the more representative keywords by overcoming this limit. In this way, the proposed model firstly prepares the candidate set using the vector model, then makes the graph which represents the relation in the pair of candidate keywords in the set and selects the keywords based on this relationship graph.

Kewords : Relationship graph model, Patents, Keyword extraction, Text mining

1.서 론

특허권은 다양한 분야에서 특허기술을 가진 개인이나 기업과 같은 단체 및 기관만이 그 기술을 배타적으로 사용할 수 있는 권리이다. 현재와 같이 고도로 발단된 기술사회에서 그 필요성 및 중요성이 더욱 증대되고 있는 상황이다. 이에 세계적으로 많은 개인과 기업과 같은 단체 및 기관들이 특허를 출원하고 있으며, 그 양 또한 방대하다. 이렇게 출원된 특허문서는 그 편의성 및 효율성으로 인해 전자문서의 형식으로 저장되고 분류되며, 이에 대한 분석을 통해 해당 기술의 강점 및 약점을 평가하는데 사용되고 있다[4][10]. 이러한 평가를 통해 얻어진 특허정보는 인력, 기술적 능력 및 성능을 측정하기 위한 중요한 자원으로 여겨진다[1]. 이와 같이 특허문서에 대한 검색의 필요성이 큰 만큼 인터넷을 통해 국내 및 국제 특허를 검색할 수 있다. 우리나라에서는 키프리스 웹사이트를 통해 전세계의 특허를 검색할 수 있으며 미국의 경우 미국 특허상표청(USPTO)을 통해 검색할 수 있다. 특허문서에 대한 검색이 이루어지기 위해서는 일반적인 문서에서와 같이 저장된 특허문서로부터 키워드를 추출하여 이에 대한 색인(indexing)을 수행하여야 한다[6]. 따라서 특

†Corresponding Author: Young Moon Leem, Dept. of Industrial & Management Engineering, Kangnung-Wonju National University, E-mail: ymleem@gwnu.ac.kr
Received October 20, 2015; Revision Received December 02, 2015; Accepted December 04, 2015.

허에 대한 검색을 효율적으로 수행하기 위해서는 해당 특허문서로부터 그 문서를 대표할 수 있는 키워드들을 효율적으로 추출하는 것이 중요하다. 일반적인 문서로부터 키워드들을 추출하는 가장 대표적인 방법은 벡터 공간모델을 적용하는 것이다[11]. 그러나 이 모델은 단어들 간의 관계성을 충실히 표현하지 못한다는 단점이 있다. 이러한 한계점을 극복하기 위해 제안된 모델이 문서를 구성하는 요소들 간의 관계성을 반영한 그래프 모델이다. 하지만 그래프 모델은 처리속도와 복잡성에 있어 벡터공간 모델보다는 성능이 떨어지는 한계점을 갖고 있다[5]. 일반적인 문서에 적용되는 이러한 방법들을 특허문서에 적용하기 위해서는 일반문서와 다른 특허문서의 반구조적인 특성을 반영하여야 한다. 즉, 특허문서는 일반적인 문서와 달리 반구조적인 문서로서 <발명의 명칭>, <요약>, <청구항>, <발명의 상세한 설명> 등과 같은 여러 섹션으로 구성되며 모든 특허문서는 동일한 구조로 갖는다. 따라서, 특허문서로부터 효율적으로 키워드를 추출하기 위해서는 이와 같은 특허문서의 특성을 반영하여야 한다.

본 논문에서는 벡터공간모델을 이용하여 그래프모델의 복잡성을 감소시키고, 또한 반구조적인 특허문서의 특징과 특허문서 내에 존재하는 키워드들 간의 관계성을 반영하여 효율적으로 키워드를 추출하는 관계성 그래프 모델을 제시하고자 한다.

2 관련연구

2.1 텍스트 마이닝

텍스트 마이닝은 데이터 마이닝의 기술을 활용하여 자유롭거나 비구조화된 텍스트 형태의 전자문서들로부터 유용한 지식을 효율적으로 발견, 추출하기 위한 지식 집약적인 처리(knowledge- intensive process) 기술이라고 정의할 수 있다." [9].

텍스트마이닝은 데이터마이닝과 비슷한 개념이지만, 데이터마이닝이 관계형 데이터베이스, 트랜잭션 데이터베이스와 데이터웨어하우스의 데이터 , XML과 같은 구조화된 데이터를 대상으로 지식을 발견하고 추출한다. 하지만, 현실적으로 사용할 수 있는 지식은 텍스트 문서, 뉴스 기사, 연구 논문, e-메일과 같은 비구조(unstructured) 또는 반구조적(semi-structured)인 데이터로 저장되며, 이러한 데이터에 대해 기존의 데이터마이닝 기술을 적용하기 어렵다. 텍스트 마이닝은 이러한 데이터마이닝의 한계를 극복하기 위해 비구조적 또는 반구조적 텍스트들로부터 지식을 발견하고 추출하기 위한 새로운 기술이라고 볼 수 있다[2].

텍스트 마이닝은 비구조 또는 반구조의 데이터를 대상으로 검색, 문서분류, 문서요약 등을 포함한 기능들을 수행하기 위해서는 그 데이터들로부터 주요 특징(features)을 추출할 수 있어야 한다. 이를 위해서 비정형화된 문서를 정형화된 모델로 추상하는 작업이 중요하다. 텍스트를 정형화하기 위해 현재까지 많이 사용되고 있는 모델은 벡터공간모델(vector space model)이며, 최근에 그래프에 기반한 모델들이 제시되고 있다[5].

2.2 벡터 공간 모델

벡터공간모델은 문서를 다차원 유클리드언(euclidean)공간의 벡터로 표현하는 모델이다[3]. 각각의 문서 $d_j(j=1..N)$ 는 전체 문서 집합 D를 구성한다고 하자. 문서 d_j가 주어졌을 때, 벡터공간모델은 d_j을 다음과 같이 그 문서를 구성하는 단어들과 그 가중치의 집합으로 나타낼 수 있다.

$$d_j = \{t_{1,j} : w_{1,j}, ..., t_{n,j} : w_{n,j}\}$$

여기서 $t_{i,j}(i=1...n)$ 는 문서 d_j을 구성하는 단어를 나타내며, $w_{i,j}(i=1...n)$ 은 단어 $t_{i,j}$에 부여된 가중치를 의미한다. 가중치 $w_{i,j}$을 계산하는 방법으로는 단어 $t_{i,j}$가 그 문서에 포함되어 있는 경우에 $w_{i,j}$ 값을 1로 정하고 그렇지 않은 경우 0 값으로 단순히 정할 수 있다. 또한 단어의 빈도수를 이용하거나 모든 단어들의 총 발생 수에 대한 해당 단어의 상대빈도수로 할 수 있는 등 다양한 방법이 있다[7]. 이러한 방법들 중 가중치 $w_{i,j}$를 계산하기 위해 지금까지 가장 많이 사용되고 있는 방법은 문서집합으로부터 그 값을 계산하는 TF-IDF(Term Frequency-Inverse Document Frequecy)이다[12]. 이 이외에도 인터넷 검색을 위한 색인 생성에 사용되는 PageRank 알고리즘이 있으며[14], 또한 기계학습 분야에서 널리 활용되고 있는 Support Vector Machine 알고리즘[13] 이나 신경회로망[8] 알고리즘을 적용한 예도 있다. 그러나 이와 같은 방법들은 단일 문서가 아닌 여러 문서들로 구성되는 문서 집합으로부터 계산한다. 단일 문서로부터 이를 계산하는 방법은 단어의 빈도수에 기반하는 방법이 있으며[3], 다른 방법으로는 통계적 방법으로서 빈발단어(frequent term)를 먼저 추출하고 그 추출된 빈발 단어와 각 단어가 동일 문장 내에 동시에 나타나는 확률 분산(distribution)을 구해 이를 이용하는 방법이 있다[11].

이와 같이 벡터공간모델은 정형화되지 않은 텍스트 문서를 정형화된 모델로 표현한다. 이 모델은 그 단순

ISSN 1229-6783(Print)
ISSN 2288-1484(Online)

함으로 기존의 데이터마이닝에서 사용되었던 기술들을 큰 수정 없이 그대로 사용할 수 있는 장점이 있는 반면, 단어들이 서로 독립적이므로 단어 간의 출현 순서나 관련성을 표현할 수 없는 단점이 있다[5].

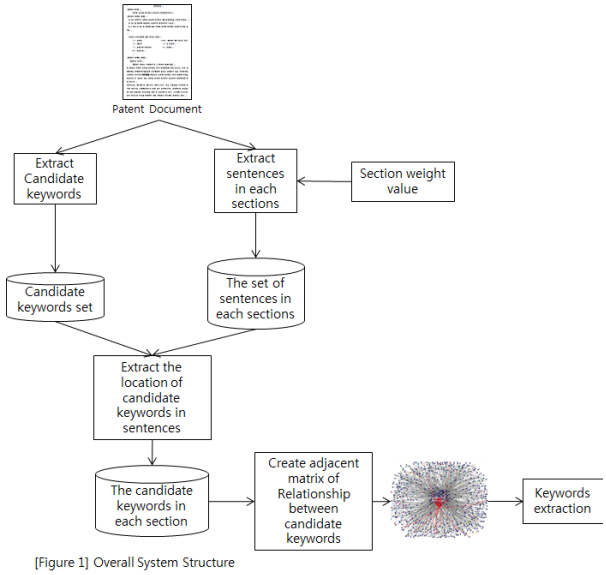
2.3 그래프 기반 모델

그래프 기반 모델은 벡터공간모델의 단점을 극복하기 위해 제안된 모델로서 텍스트 문서를 구성하는 요소들 즉, 단어, 문장, 문단, 문서 등을 그래프로 표현한다. 이 모델은 그래프를 구성하는 정점(vertex)과 간선(edge)에 이들 요소들을 어떻게 표현하는지에 따라 여러 종류로 나뉜다[5].

그래프를 통해 표현하고자 하는 내용에 따라 크게 3가지로 분류할 수 있다. 첫 번째로 문법적 연관성을 그래프로 표현하는 모델이다. 이 방법에서는 문서의 구성요소들을 자연어 처리 기법을 이용하여 품사별로 구분한다. 정점을 품사별로 구분하여 표현하고 이들 간의 관계를 간선의 레이블로 나타낸다. 다른 모델로는 의미적 연관성을 그래프로 표현하는 방법이 있다. 이에 해당하는 대표적인 방법으로 문서와 개념 간의 관계를 그래프로 표현하는 모델이다. 마지막으로 가장 일반적으로 사용되는 모델로 문서 구성 요소인 단어의 동시 발생 빈도수나 문장 간의 유사도 등을 그래프로 표현하는 모델이다. 이러한 그래프 모델은 대부분 문서 분류(classification)와 군집화(clustering), 그리고 문서요약에 이용되어 왔으며 키워드 추출에 이용된 예는 드물다[5].

3. 관계성 그래프 모델

본 논문에서 제시한 관계성 그래프 모델은 반구조적인 특허문서의 특징을 이용하기 위해 특허문서를 구성하는 각 섹션의 중요도에 따라 섹션별 가중치를 정한다. 다음으로 기존의 벡터공간 모델을 통해 후보 키워드군을 얻고 이 후보 키워드들 간의 관계성을 찾는다. 후보 키워드들 간의 관계성은 후보 키워드들이 동일한 문장 내에 나타나는지에 여부에 따라 결정된다. 후보 키워드들 간의 관계성의 정도는 후보 키워드들이 나타나는 섹션의 가중치와 그 빈도에 따라 계산된다. 이러한 정보를 바탕으로 후보 키워드들을 정점으로 하고 후보 키워드들 간의 관계성을 간선으로 하는 관계성 그래프를 구성한다. 이렇게 구성된 관계성 그래프에 대해 임계값 이상의 관계성을 갖는 키워드들을 선택함으로서 특허문서로부터 효율적으로 키워드를 추출한다. 전체적인 처리과정은 [Fig. 1]과 같다.



[Figure 1] Overall System Structure

3.1 후보 키워드군 추출

특허문서로부터 키워드 후보군을 얻는데 기존의 벡터 기반 모델을 이용하여 키워드 후보군을 추출한다. 기존의 알고리즘들은 문서로부터 불용어들을 제거한 후 남은 단어(term) 혹은 구(phrase) 전체에 대해 통계적인 방법이나 빈도수에 기반하여 그 중요도를 측정하고 이를 순서화한다. 본 논문에서는 이들 순서화된 목록으로부터 중요도가 높은 일정 수의 키워드들을 후보키워드 군으로 설정한다. 일정 수의 키워드를 후보군으로 설정하는 이유는 중요도가 낮은 키워드들은 관계성 그래프 모델에 의해 제거될 것이기 때문이다.

3.2 섹션별 후보 키워드군의 문장 내 위치 정보 추출

한 문장에 나타나는 키워드들은 직관적으로 관계성이 있다고 생각할 수 있다. 후보 키워드들이 한 문장에 나타나는지를 확인하기 위해서 해당 문서를 섹션에 따라 문장 단위로 구별하여 저장한다. 한 문장은 마침표(.)로 끝나는 것이 일반적이지만 마침표만으로 문장단위를 구별해 내기는 어렵다. 그 이유로는 특허문서의 경우 다음과 같은 단어들을 포함하는 것이 일반적이기 때문이다.

U.S.
No.
e.g.
i.e.
jan.
Pat.
fig.

이러한 경우에는 심볼 "." 는 마침표가 아니라 도트를 의미함으로 마침표와 구별해야 한다.

본 논문에서는 일반적으로 마침표 다음에는 빈 공간이 나타나는 것을 이용하여 마침표를 구분한다. 하지만 위에 나열한 단어들 다음에도 공백이 나타나는 것이 일반적이기 때문에 이것만으로는 위에 나열한 도트와 구별할 수 없다. 따라서 벡터 모델 방식에서와 같이 특허문서에서 나타날 수 있는 위와 같은 단어구들을 저장해 놓고 문서 내에 나타나는 경우에 이들을 예외적으로 처리함으로써 마침표와 구별하였다. 분리된 문장들은 [Fig. 2]와 같이 섹션별로 저장된다.

Number of section	Number of sentence	Weight value of each section	sentence

[Figure 2] Data structure for Statement

위 그림의 각 섹션의 가중치는 특허문서를 구성하는 각 섹션의 중요도에 따라 섹션별 가중치를 정한다. 특허문서의 경우 <발명의 명칭>과 <요약>이 다른 섹션에 비해 특허의 중요 내용을 포함하고 있을 것이기 때문에 다른 섹션에 비해 높은 가중치를 주는 것이 합리적이다. [Fig. 2]의 자료구조에 저장된 문장단위 데이터와 후보 키워드군에 포함된 후보 키워드들을 서로 비교하여 각 후보 키워드가 포함된 문장번호, 섹션번호를 추출하여 [Fig. 3]과 같은 자료구조에 저장한다.

Candidate Keyword	Number of section	Number of sentence	The weight value of section

[Figure 3] Data structure for Candidate Keyword

3.3 관계성 기반 인접행렬

두 개의 키워드들이 동일한 문장 내에 나타나는 경우 그 키워드들은 의미적으로 관계가 있다고 생각할 수 있다. 또한 후보 키워드들 중에 동일한 문장에 나타나는 횟수가 많은 경우 그 후보키워드들 간에는 관련성이 다른 후보 키워드들 간보다 높다고 생각할 수 있다. 더불어, 동일한 키워드들을 포함한 문장이 중요 섹션에 나타나는 경우 그렇지 않은 경우보다 해당 문서를 대표하는데 더 적절한 후보 키워드들이라고 생각할 수 있다. 이러한 휴리스틱에 기반하여 본 논문에서는 이전에 처리했던 정보들을 활용하여 각 후보 키워드간의 관계성을 그래프로 표현한다. 후보 키워드간의 관계성 그래프는 [Fig. 4]와 같은 인접행렬에 저장한다.

	Candidate keyword 1	Candidate keyword 2	...	Candidate keyword i	...	Candidate keyword n
Candidate keyword 1						
Candidate keyword 2						
...						
Candidate keyword i						
...						
Candidate keyword n						

[Figure 4] Adjacent Matrix of Relationship between candidate keywords

위 인접행렬의 각 행렬요소의 값은 그래프의 간선(edge)을 나타내는 것으로서 두 후보 키워드가 동일 문장에 나타나는 경우 뒤에 설명할 관계성 척도 값이 저장된다. 이는 그래프에서 두 후보 키워드를 나타내는 정점들 간에 간선이 있음을 의미하며 저장된 관계성 척도 값은 그 간선의 레이블로 표시된다. 만일, 두 후보 키워드가 동일 문장에 나타난 적이 없는 경우 두 후보 키워드는 관계성이 없는 것으로 보고 0 값이 저장되며, 두 정점사이에는 간선이 없음을 의미한다. 두 키워드 후보 간의 관련성의 척도는 다음과 같이 가중치 확률로서 나타낸다.

$$e_i = \frac{\sum_{j=1}^{m} n_{i_j} \cdot w_{s_j}}{N^2} \cdots\cdots\cdots\cdots\cdots\cdots ①$$

여기서 e_i는 후보키워드 쌍 i에 대한 관련성 척도이다. N은 후보 키워드군의 크기이며 m은 해당문서를 구성하는 섹션의 개수이다. n_{i_j}는 j 섹션에 나타난 후보키워드 쌍 i의 빈도수를 의미하며, w_{s_j}는 섹션 j에 부여된 가중치 값이다.

3.4 간선 제거에 의한 키워드 추출

후보 키워드군에 속한 키워드들의 쌍이 동일문장에 나타나는 사건은 서로 독립적이라고 할 수 있다. 따라서 후보 키워드 쌍이 한 문장 내에 나타날 가중치 확률의 기댓값은 다음과 같다.

$$E(i) = \frac{\sum_{j=1}^{m} n_{i_j}}{N^2} \cdot \frac{\sum_{j=1}^{m} w_{s_j}}{m} \cdots\cdots\cdots\cdots ②$$

만약 후보키워드 쌍의 관련성을 나타내는 간선의 값, 즉 관계성 척도의 값이 위의 기댓값보다 더 적은 경우에는 후보 키워드쌍 i는 기대했던 것보다 더 적은 빈도수로 동일 문장 내에 나타났다는 것을 의미한다. 따라서 이와 같은 키워드 쌍간에 존재하는 간선들은 관

ISSN 1229-6783(Print)
ISSN 2288-1484(Online)

계성 그래프로부터 제거한다. 이와 같은 과정을 통해 부속 간선이 가장 많은 후보 키워드가 문서내의 다른 후보 키워드들과 관계성이 높고 따라서 그 문서를 대표하는 키워드가 된다. 간선 제거 후의 그래프로부터 후보 키워드 간에 간선이 남아 있다는 것은 위에서 서술한 것과 같이 평균이상의 관계성이 존재함을 의미한다. 또한 부속된 간선이 개수가 많을수록 더 많은 후보 키워드들과 관계성을 갖는다는 것을 의미한다. 따라서, 부속된 간선이 개수가 많을수록 특허문서를 대표할 수 있는 키워드라고 생각할 수 있다. 그러므로 간선 제거 후의 그래프에서 부속 간선이 많은 순으로 정렬함으로서 그 특허문서를 가장 잘 대표할 수 있는 키워드들을 추출할 수 있다.

4. 실험 및 평가

본 논문에서는 단일의 특허 문서로부터 그 문서를 대표할 수 있는 키워드들을 추출하기 위해 관계성 그래프 모델을 제시하였다. 이를 평가하기 위해 미국 특허상표청(USPTO)으로부터 임의로 한 개의 특허문서를 선택하였다. 선택된 특허문서의 발명의 명칭은 “은은한 조명의 임계값을 조절하기 위한 메소드 및 장치” 관한 것이다. 이 문서로부터 후보 키워드군을 선택하기 위해 사용한 기존의 벡터 모델은 단일 문서로부터 단어와 구(phrase)의 출현 빈도수만을 사용하여 키워드들을 구하는 TF(Term-Frequency) 방법과 통계적 방법을 이용한 키워드 추출 방법을 이용하였다. 빈도수에 기반한 방법으로 얻어진 키워드는 총 354개 였으며 통계적 방법으로 얻어진 키워드는 1663개 이다. <Table 1>은 단순 빈도수에 의해 추출된 키워드의 일부이며 <Table 2>는 통계적 방법을 통해 구해진 키워드일부이다.

<Table 1> Cadidate Keyword Extraction by Frequency

Candidate Keywords	Frequency
ambient light	169
level	148
load	129
controller	96
control	80
device	65
...	...
application	1
essentially	1

후보 키워드군을 선택하기 위해 빈도수에 의해 추출된 키워드로부터 그 빈도수가 30개 이상인 키워드를 후보키워드 군으로 선정하였으며 후보 키워드군의 크기는 21개이다. 또한 통계적 방법에 의해 추출된 키워드에 대해서는 χ^2의 값이 20,000 이상인 값을 가진 키워드들을 후보 키워드군으로 선정하였으며 그 크기는 37개이다. 특허문서를 구성하는 각 섹션에 대한 가중치 값은 <Table 3>과 같다.

<table 2> Cadidate Keyword Extraction by Stochastics

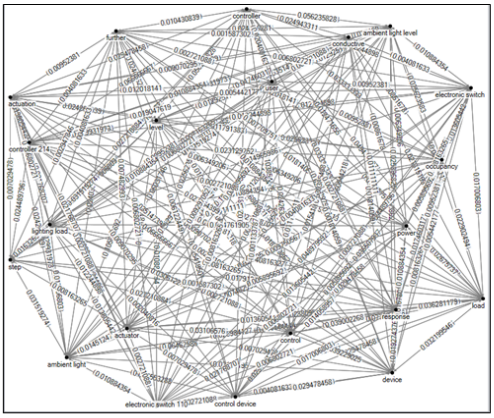
Candidate Keywords	chi-square
light	44679.7
load	42184.4
ambient light	40051.9
ambient	40051.9
controller	39976.2
threshold	39199.0
level	36370.1
control	36274.2
ambient light level	36090.6
light level	36090.6
level threshold	35213.7
light level threshold	35213.7
...	...
independently	366.6
issued nov 9	317.5
nov	317.5
issued nov	317.5
claimed	52.9

<발명의 명칭>, <청구항>, <요약> 등이 <발명의 상세한 설명> 섹션보다 해당 특허에 중요한 키워드를 포함하고 있을 것이기 때문에 더 높은 가중치을 부여하였다.

<Table 3> Weight for Section

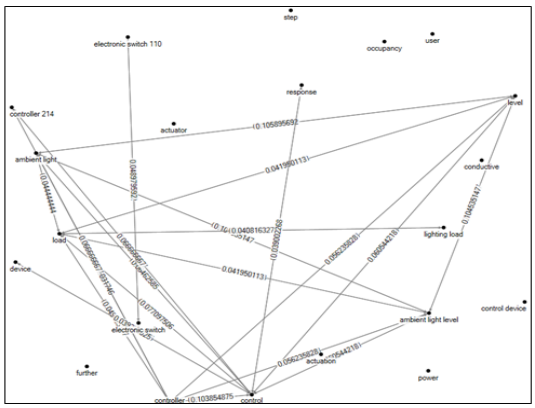
Section	Weight
TITLE	0.8
ABSTRACT	0.5
CLAIMS	0.7
DESCRIPTION	0.1
CROSS REFERENCES TO RELATED APPLICATION	0.1
BACKGROUND OF THE INVENTION	0.2
SUMMARY OF THE INVENTION	0.7
BRIEF DESCRIPTION OF THE DRAWINGS	0.3
DETAILED DESCRIPTION OF THE INVENTION	0.6

빈도수에 의해 추출된 키워드 후보군에 대한 관계성 그래프는 [Fig. 5]와 같다. 그래프의 정점들은 후보 키워드들을 나타내며 해당 후보 키워드를 그 정점의 레이블로 표시하였다.



[Figure 5] Relationship Graph between Keywords before Edge Cut-Off

간선은 인접한 두 후보 키워드들이 한 번 이상 동일 문장 내에 나타났음을 의미한다. 간선의 레이블은 두 후보 키워드 간의 관계성 척도 값을 나타내는 가중치 확률값이다. [Fig. 6]은 평균값 이하인 관계성을 갖는 간선을 제거 한 후의 그래프이다. 두 그림을 비교해보면 많은 간선들이 제거 되었음을 알 수 있다. 이것은 동일 문장 내에 나타난 키워드 쌍이 많다고 하더라도 기대값 이하의 관계성을 갖고 있었음을 의미한다. 또한 [Fig. 6]에서 볼 수 있듯이 간선이 존재하지 않는 정점들이 다 수 있음을 알 수 있다. 이들 정점은 다른 후보 키워드들과 기대값 이하의 관계성 밖에 없음을 의미한다. 따라서 이러한 후보 키워드들은 본 논문에서 만들어낼 키워드 리스트에서 삭제된다. <Table 4>는 기대값 이하 간선을 제거 한 후 부속 간선의 수가 1 이상인 정점들을 정렬한 키워드 리스트이다. 기존의 빈도수 기반 벡터 모델의 키워드 리스트인 <Table 3>과 비교해봤을 때 “ambient light”와 “level”은 동일하게 높은 순위를 갖지만 “ambient light level” 키워드는 본 논문에서 제시한 관계성 그래프 방법으로 키워드를 추출하였을 때 높은 순위를 가짐을 알 수 있다.



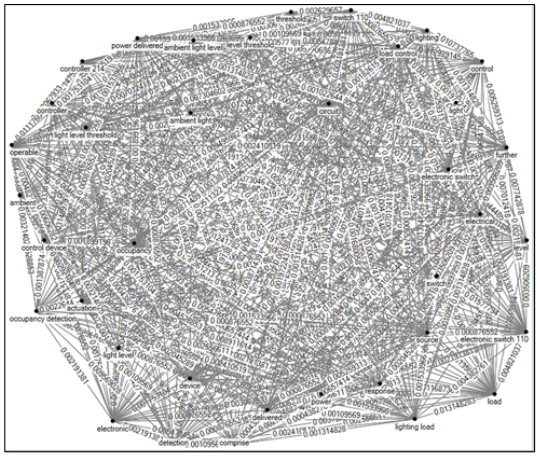
[Figure 6] Relationship Graph between Keywords after Edge Cut-Off

<Table 4> The Number of Edge Attached to Keyword after Edge Cut-Off

Keyword	Number of Edge
ambient light	5
level	4
ambient light level	3
load	3
control	3
controller	2
electronic switch	1

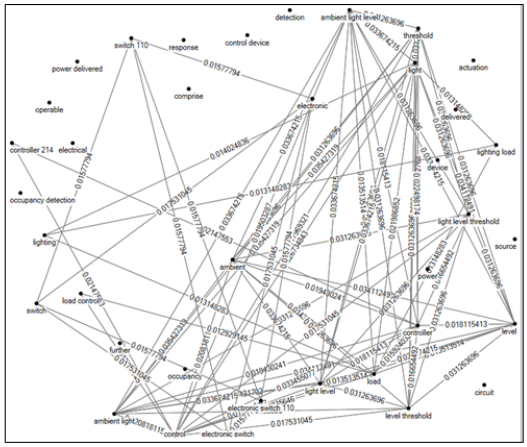
또한, 후보 키워드 군의 크기가 21개에서 7개로 줄었으며, <발명의 명칭>과 좀 더 관계성이 있는 키워드들이 포함되어 있음을 알 수 있다.

통계적 방법을 이용한 벡터 모델에 의해 구해진 후보 키워드군에 대해서 간선 제거 전과 후의 그래프는 [Fig. 7]과 [Fig. 8] 이다. [Fig. 7]에서 볼 수 있듯이 관계성 간선 제거 전의 그래프는 후보 키워드들 사이에 많은 간선들이 존재한다. [Fig. 5]에서와 같이 빈도수에 의해 구해진 관계성 그래프에 비해 간선이 수가 훨씬 많음을 알 수 있는데 이는 후보 키워드군의 크기의 차이에서 비롯된다.



[Figure 7] Relationship Graph between Keywords before Edge Cut-Off

ISSN 1229-6783(Print)
ISSN 2288-1484(Online)



[Figure 8] Relationship Graph between Keywords after Edge Cut-Off

이 후보 키워드군도 이전의 빈도수에 의해 구해진 후보 키워드에서와 같이 많은 간선이 줄었음을 알 수 있고, 부속된 간선이 없는 정점들도 있음을 알 수 있다. <Table 5>는 간선 제거 후 키워드들에 부속된 간선의 수가 1 이상인 것들만을 내림차순으로 정렬한 리스트이다. <Table 2>와 <Table 5>를 비교하였을 때 통계적 방법을 이용한 벡터 모델에 대해 후보 키워드 군의 경우 상위 6개의 키워드에 대해서는 차이가 없었으나 벡터 모델에서는 낮은 순위였던 "switch", "electronic switch" 키워드가 높은 순위를 가짐을 알 수 있다. 특허 주제와의 관련성에서는 통계적 방법을 이용한 벡터 모델에 의해 추출된 키워드들이나 본 논문에서 제시한 관계성 그래프 모델에 의해 추출된 키워드들이나 비슷한 것을 알 수 있었다. 하지만 키워드 군의 크기에 있어서는 통계적 방법을 이용한 벡터 모델의 경우 37개 이지만 본 논문에서 제시한 관계성 그래프 모델의 경우 15개로 큰 폭으로 줄었음을 알 수 있다.

<Table 5> The Number of Edge Attached to Keyword after Edge Cut-Off

Keyword	Number of Edge
light	13
load	10
ambient light	9
ambient	8
controller	8
threshold	6
control	6
level	5
switch	4
ambient light level	3
electronic switch	3
light level	2
electronic	2
level threshold	1
lighting	1

5. 결론

본 논문에서는 단일의 반구조적인 특허 문서로부터 그 문서를 대표할 수 있는 키워드들을 추출하기 위해 관계성 그래프 모델을 제시하였다. 실험의 결과처럼 본 논문에서 제시한 관계성 그래프 모델은 그래프를 구성하는 노드의 수를 줄임으로서 기존 그래프 모델의 복잡성을 줄일 수 있었다. 또한 특허의 내용과의 관련성 측면에서 단일 문서로부터 키워드를 추출하기 위한 기존의 벡터 모델보다 더 좋거나 비슷한 정도의 관련성이 있는 키워드들을 추출하였다. 더욱이, 특허 문서를 대표하는 키워드 리스트의 크기에 있어서 기존의 벡터 모델보다 본 논문에서 제시한 관계성 그래프 모델이 그 크기를 많이 줄였음을 보였다. 결론적으로 본 논문에서 제시한 관계성 그래프 모델이 반구조적 특성을 갖는 단일의 특허 문서로부터 키워드들을 추출하는데 기존의 벡터 모델과 그래프모델에 의한 키워드 추출보다 더 효율적임을 알 수 있다.

6. References

[1] Coombs, J. E. & Bierly, P. E.(2006), "Measuring technological capability and performance" R&D Management, 36(4):421-438

[2] Feldman. R., and J. Sanger(2007), "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data" New York, NY Cambridge University Press.

[3] G. Salton, A. Wong and C. S. Yang(1975), "A vector space model for automatic indexing" Communications of the ACM, 18:613-620

[4] I.V. Wartburg, T. Teichert, K. Rost(2005), "Inventive progress measured by multistage patent citation analysis " Research Policy 34 (10), 1591-1607.

[5] Jae Young, Chang(2013), "A study on research trends of graph-based text representations for text mining" The Journal of The Institute of Internet, Broadcasting and Communication 13: No. 5

[6] Jens-Erik Mai(2005), "Analysis in indexing: document and domain centered approaches" Information Processing and Management

41:599-611
[7] Jiawei Han, Micheline Kamber(2011), "Data mining concepts and techniques" 2nd-edition Morgan Kaufmann press, 614-628
[8] Jo, Taeho, Lee, Malrey, and Gatton, T. M.(2006), "Keyword extraction from documents using a neural network model," ICHIT' 06, 2:194-197.
[9] Kao. A. and S. R. Poteet.(2007), "Natural Language Processing and Text Mining" London Springer-Verlag, 1-7
[10] Li, Y.R, Wang, L.H., & Hong, C. F.(2009), "Extracting the significant-rare keywords for patent analysis" Expert System with Applications, 36(6):5200-5204
[11] Matsuo, Y., and Ishizuka, M.(2004), "Keyword extraction from a single document using word co-occurrence statistical information," International Journal on Artificial Intelligence Tools, 13:157-169.
[12] Roberston, S.(2004), "Understanding inverse document frequency: On theoretical argument for IDF" Journal of Documentation, 60(5):503-520.
[13] Yu, J. X., Kitsuregawa, M., and Leong, H. V.(2006), "Keyword Extraction using Support Vector Machine," Lecture notes in computer science, 4016:85-96.
[14] Wang, J., Liu, J., Wang, and Cong(2007), "Keyword extraction based on PageRank," Lecture notes in computer science, 857-864.

저 자 소 개

이 순 근

인하대학교에서 학사 및 공학석사를 취득하였다. 현재 국립강릉원주대학교 산업경영공학과에 재학중이며, 주요관심사는 텍스트 마이닝이다.

임 영 문

연세대학교에서 학사 및 이학석사를 취득하였으며, 텍사스주립대학교 산업공학과에서 공학박사를 취득 하였다. 현재 국립강릉원주대학교 산업경영공학과 정교수를 재직 중이며 관심분야로는 인간공학, 정보시스테, 정보이론 응용 이다.

엄 완 섭

서울대학교에서 학사 및 공학석사를 취득하였으며 동 대학 산업공학과에서 공학박사를 취득하였다. 현재 국립강릉원주대학교 산업경영공학과 정교수를 재직 중이며 관심분야로는 전사적자원관리(ERP), 공급망관리(SCM), 생산운영관리이다.