Journal of Korean Institute of Intelligent Systems
Vol. 25, No. 5, October 2015, pp. 470-476
http://dx.doi.org/10.5391/JKIIS.2015.25.5.470

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

JKIIS

통계적 텍스트 마이닝을 이용한 빅 데이터 전처리

A Big Data Preprocessing using Statistical Text Mining

전성해
Sunghae Jun

청주대학교 통계학과
Department of Statistics, Cheongju University

요 약

빅 데이터는 여러 분야에서 다양하게 사용되고 있다. 예를 들어, 컴퓨터학과 사회학에서 빅 데이터에 대한 서로 간의 접근방법에 대한 차이는 있겠지만 빅 데이터의 분석을 통한 활용 측면에서는 공통적인 부분을 갖는다. 따라서 대부분의 분야에서 빅 데이터에 대한 의미 있는 분석과 활용은 필요하게 된다. 통계학과 기계학습은 빅 데이터의 분석을 위한 다양한 방법론을 제공한다. 본 논문에서는 빅 데이터분석 과정에 대하여 알아보고 수집된 빅 데이터의 원천에서부터 분석을 거쳐 최종적으로 분석결과를 활용하는 전체 과정을 위한 효율적인 빅 데이터 분석방법에 대하여 연구한다. 특히, 빅 데이터의 특성을 갖는 여러 데이터 중 하나인 특허문서 데이터에 대하여 빅 데이터분석을 적용하여 효과적인 특허분석을 수행하고 이 결과를 연구개발 기획에 적용하는 방법론에 대하여 제안한다. 제안방법에 대한 실제적용을 위하여 전 세계 특허데이터베이스로부터 실제 기업의 전체 출원, 등록 특허문서를 수집, 분석하고 연구개발 업무에 활용하는 전 과정에 대한 사례연구를 수행하였다.

키워드 : 빅 데이터분석, 통계학, 자연어처리, 텍스트 마이닝, 특허분석, 선형모형

Abstract

Big data has been used in diverse areas. For example, in computer science and sociology, there is a difference in their issues to approach big data, but they have same usage to analyze big data and imply the analysis result. So the meaningful analysis and implication of big data are needed in most areas. Statistics and machine learning provide various methods for big data analysis. In this paper, we study a process for big data analysis, and propose an efficient methodology of entire process from collecting big data to implying the result of big data analysis. In addition, patent documents have the characteristics of big data, we propose an approach to apply big data analysis to patent data, and imply the result of patent big data to build R&D strategy. To illustrate how to use our proposed methodology for real problem, we perform a case study using applied and registered patent documents retrieved from the patent databases in the world.

Key Words : Big Data Analysis, Statistics, Natural Language Processing, Text Mining, Patent Analysis, Linear Model.

Received: Aug. 28, 2015
Revised : Sep. 17, 2015
Accepted: Sep. 19, 2015
† Corresponding author
shjun@cju.ac.kr

이 논문은 2014학년도에 청주대학교 산업과학연구소가 지원한 학술연구조성비(특별연구과제)에 의해 연구되었음

1. 서 론

인터넷 환경의 보편화와 데이터 저장 시스템의 획기적인 발전에 의해 분석대상이 되는 데이터의 크기는 지속적으로 증가하여 빅 데이터의 시대에 진입하였다 [1],[2],[3]. 이에 따라 이질적이고 거대한 데이터 속에 숨겨진 패턴을 찾기 위한 빅 데이터분석에 대한 연구도 다양한 분야에서 활발히 진행되고 있다 [4],[5],[6],[7]. 특히 기술경영(management of engineering; MOT) 분야에서 대표적인 빅 데이터인 특허문서의 분석을 통하여 기업의 R&D 계획을 위한 기술예측 및 혁신에 대한 연구가 이루어지고 있다 [8],[9],[10],[11],[12]. 전 세계 특허청에 출원, 등록된 특허문서는 매우 방대하고 각 특허에는 출원날짜, 출원인, 특허명칭, 발명의 요약, 특허분류번호, 기술상세도면, 청구항, 등 다양하고 서로 이질적인 데이터 형태로 구성되어 있다 [13],[14]. 따라서 특허문서는 빅 데이터 구조를 갖는다 [15]. 통계학 및 기계학습(machine learning) 알고리즘 등 전통적인 데이터분석 방법은 정형화된

(structured) 데이터구조를 대상으로 한다 [16]. 그러나 크기(volume), 속도(velocity), 다양성(variety) 등의 특징으로 정의되는 빅 데이터는 통계학이나 기계학습에 의해 바로 분석이 수행되기에는 어려움이 존재한다. 왜냐하면 빅 데이터 그 자체는 아직 정형화되지 않았기 때문이다. 이와 같은 문제점을 해결하기 위하여 본 연구에서는 빅 데이터의 전처리를 통하여 정형화된 데이터 구조를 만들고 이를 이용한 빅 데이터 분석이 가능한 방법을 제안한다. 즉 하나의 과정 속에서 빅 데이터의 전처리와 분석이 가능할 수 있는 방법론을 제시한다. 또한 기술경영 분야에서 적용 가능한 사례분석을 통하여 제안하는 방법론의 실제적용에 대하여 알아본다.

2. 빅 데이터 전처리

일반적으로 빅 데이터는 크기, 속도, 그리고 다양성의 3가지 특징을 통하여 설명 된다 [2]. 먼저 빅 데이터는 단어의 의미 그대로 데이터의 크기에 이전과 비교할 수 없을 정도로 커졌다. 컴퓨터 저장 공간 및 네트워크 속도의 획기적인 발전에 의해 실시간으로 쌓이는 데이터의 크기는 대단히 빠른 속도로 증가하고 있다. 분석 관점에서 볼 때 과거에 비해 제공되는 데이터의 양은 훨씬 많아 졌지만 의미 있는 지식을 추출하는 과정은 더 복잡해지고 어려워 졌다. 빅 데이터가 갖는 또 다른 특성은 데이터가 갖는 다양성이다. 빅 데이터는 이질적이고 복잡한 데이터 구조를 갖는다. 주로 관계형 데이터베이스로 구성된 기존의 데이터와는 달리 빅 데이터 환경에서는 문자, 숫자, 그림 등 다양한 형태의 데이터가 서로 연관되어 한꺼번에 포함되어 있다. 이들 간의 숨겨진 패턴을 찾아내는 것은 빅 데이터에서 매우 중요한 작업 중 하나이다. 이와 같은 빅 데이터의 특성 때문에 통계학과 기계학습을 이용한 빅 데이터의 분석에는 어려움이 따른다. 기본적으로 통계학과 기계학습 기반의 대부분 분석 기법들은 다음과 같은 테이블 구조에 의한 정형화된 데이터 구조를 요구한다 [8],[15].

Table = Structured data		Variable = Column = Field			
		Var.1	**Var.2**	⋯	**Var.p**
Observation = Row = Record	**Obs.1**	N.11	N.12	⋯	N.1p
	Obs.2	N.21	N.22	⋯	N.2p
	⋮	⋮	⋮	⋱	⋮
	Obs.n	N.n1	N.n2	⋯	N.np

그림 1. 정형화된 데이터 구조

Fig. 1. Structured data

그림 1에서 각 열(column)은 변수(variable) 또는 필드(field)를 나타내고 각 행(row)은 관측치(observation) 또는 레코드(record)를 나타낸다. 그림 1은 총 p개의 변수(Var.1, Var.2, …, Var.p)와 n개의 관측치(Obs.1, Obs.2, …, Obs.n)를 갖는 테이블(table) 형태를 나타낸다. 각 행과 열이 만나는 곳의 원소는 구체적인 값을 나타낸다. 예를 들어 N.21은 첫 번째 변수에 대한 두 번째 관측결과를 나타낸다. 이와 같은 정형화된 데이터 구조를 구축하면 본격적인 통계분석이 가능하게 된다. 본 논문에서는 통계학, 텍스트 마이닝, 자연어 처리 기법 등을 이용하여 빅 데이터의 효율적 전처리 방법과 통계분석에 대한 방법론을 제안한다.

3. 빅 데이터 전처리를 위한 통계적 텍스트 마이닝

통계학을 이용한 빅 데이터 분석을 위하여 우선적으로 필요한 것은 분석이 가능한 형태로 수집된 빅 데이터를 변형시키는 작업이 필요하다. 본 논문에서는 빅 데이터의 효율적인 전처리와 통계분석을 함께 수행하는 통계적 텍스트 마이닝 (statistical text mining; STM) 방법에 대하여 연구한다. 특히 특허 빅 데이터의 STM에 대하여 연구한다.

특허문서 데이터는 특허제목, 발명자, 출원날짜, 요약, 특허분류코드, 도면, 등 다양한 형태의 데이터를 포함한다 [14]. 기술 분야에 따라 수집된 특허문서는 보통 수천에서 수백만 건 이상이 되기 때문에 저장된 특허데이터의 용량은 매우 크다. 또한 매일 수많은 특허가 출원, 등록되고 있기 때문에 특허문서는 빅 데이터의 여러 특성을 갖는다. 그림 2는 본 논문에서 제안하는 특허 빅 데이터의 STM에 대한 전체적인 구조를 나타내고 있다.

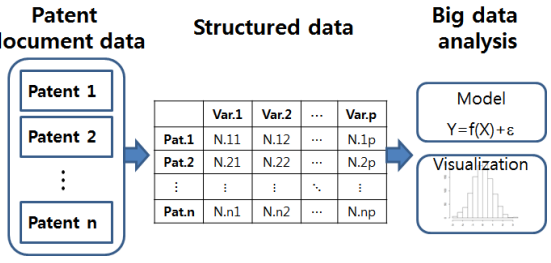

그림 2. 통계적 텍스트 마이닝 절차

Fig. 2. Statistical text mining process

먼저 키워드 검색식에 의해 분석의 대상이 되는 목표기술과 관련된 특허문서를 전 세계의 특허 데이터베이스로(patent DB)부터 수집한다. 수집된 특허 빅 데이터는 자연어처리와 텍스트 마이닝의 전처리 과정을 거쳐 정형화된 데이터(structured data)로 변환된다. 정형화된 데이터를 이용하여 모형화(modeling)와 시각화(visualization) 등의 빅 데이터 분석이 이루어진다. 그러므로 본 논문의 STM은 그림 3과 같이 통계학, 데이터 마이닝, 그리고 자연어처리의 학제적(interdisciplinary) 연계에 의해 구성된다.

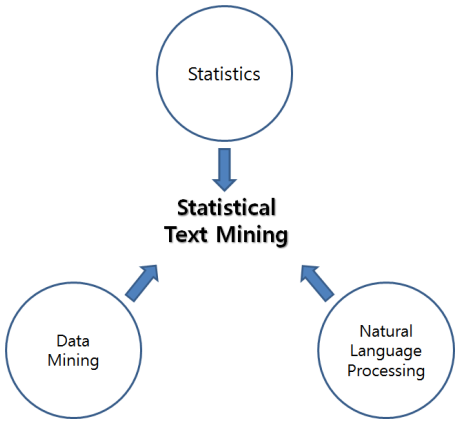

그림 3. 통계적 텍스트 마이닝 구조
Fig. 3. Structure of statistical text mining

즉 자연어 처리에 의한 텍스트 문서 데이터의 언어적 처리, 텍스트를 포함한 데이터 마이닝에 의한 데이터 통합과 정제, 그리고 통계학을 이용한 분석과 시각화를 통하여 STM의 전 과정이 수행된다. 그림 4는 STM의 각 절차에 대하여 세부적으로 설명하고 있다.

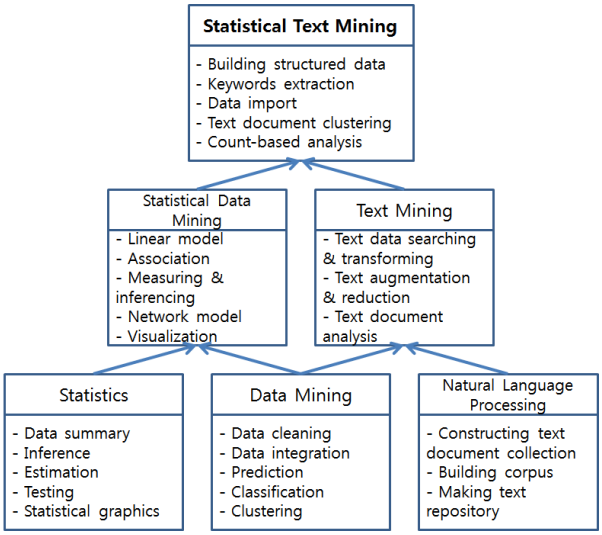

그림 4. 통계적 텍스트 마이닝 세부작업
Fig. 4. Tasks of statistical text mining

데이터의 요약과 추정, 검정의 추론(inference)을 이용하여 통계 모형화가 이루어지고, 통계 그래픽스(statistical graphics)를 이용하여 빅 데이터의 시각화 결과를 얻는다. 데이터 마이닝 단계에서는 데이터의 정제(cleaning)와 통합(integration)을 통하여 예측, 분류, 그리고 군집화가 이루어진다. 자연어 처리 기법을 이용하여 수집된 텍스트 문서 데이터로부터 코퍼스(corpus)를 만들고, 텍스트 마이닝의 전처리 전 단계로서 텍스트 저장소(text repository)를 구축한다. 통계학과 데이터 마이닝의 여러 기법들을 이용하여 통계적 데이터 마이닝 단계가 구축되며 이 단계에는 선형모형(linear model), 연관성(association) 분석, 측도와 추론 및 네트워크 모형, 그리고 정교한 시각화 작업이 포함된다. 데이터 마이닝과 자연어처리가 함께 적용되어 텍스트 마이닝 절차가 나타나며 이 단계에서는 텍스트 데이터의 검색과 변형(transforming), 텍스트 데이터의 확장(augmentation)과 축소(reduction), 그리고 텍스트 문서의 분석이 이루어진다. 마지막으로 통계적 데이터 마이닝과 텍스트 마이닝의 절차가 합쳐지면서 STM의 최종 결과를 얻게 된다. 마지막 단계에서 정형화된(structured) 데이터가 구축되고 이로부터 키워드를 추출하고 이를 이용하여 빅 데이터분석을 위한 여러 가지 기법들이 적용될 수 있게 된다.

본 논문에서는 제안방법의 구체적 적용을 위하여 대표적인 데이터 언어(data language)인 R을 이용한다. R은 소스가 공개된 무료 소프트웨어이다 [17]. R은 데이터의 조정(manipulation), 계산(calculation), 그리고 그래픽(graphical display) 기능을 모두 가지고 있는 통합된 데이터 분석 환경을 제공한다 [18]. 처음 R을 설치하면 기본적인 통계분석과 시각화 기능을 포함한 R 기본(R base)이 생성된다. R 기본을 이용하여 데이터의 입출력, 변형, 통계적 추론, 선형회귀분석, 산점도, 히스토그램, 상자그림(box plot) 등 다양한 데이터의 조정, 분석, 시각화를 수행할 수 있게 된다. 이와 같이 R 기본과 추가적으로 제공되는 R 패키지를 이용하여 빅 데이터를 위한 데이터 처리 및 분석 환경을 구축할 수 있다. 본 연구의 STM을 위한 R 구조는 그림 5와 같다.

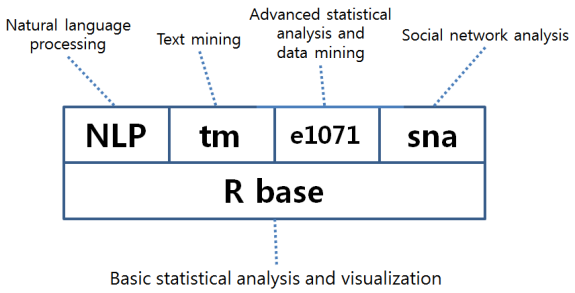

그림 5. 통계적 텍스트 마이닝을 위한 R 구조
Fig. 5. R structure for statistical text mining

기본적인 통계분석과 시각화는 R 기본을 통하여 이루어진다. 자연어처리를 위한 "NLP" 패키지와 텍스트 마이닝을 위한 "tm" 패키지가 사용된다 [18],[19]. R 기본에서 제공하지 않는 고급 통계분석과 데이터 마이닝 기법을 사용하기 위하여 R 패키지인 "e1071"을 사용한다 [20]. 또한 특허 문서 안에 포함되어 있는 기술 간 연관성을 파악하기 위하여 사회네트워크 분석(social network analysis)인 "sna" 패키지를 이용한다 [21]. 그림 6은 본 논문에서 제안하는 특허 빅 데이터의 STM 과정을 단계별로 설명하고 있다.

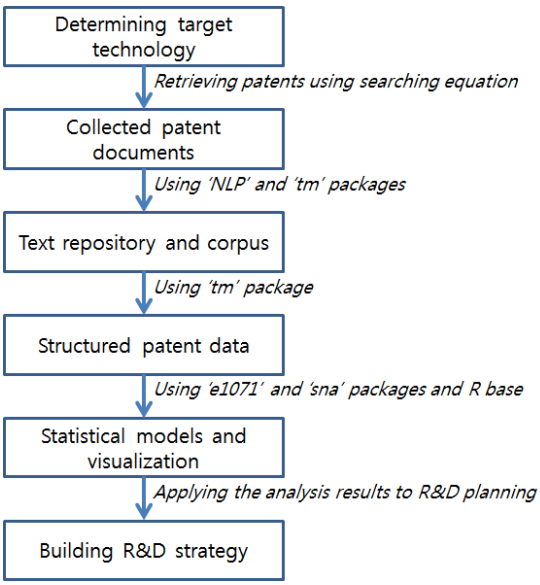

그림 6. 특허 빅 데이터의 통계적 텍스트 마이닝 절차

Fig. 6. Statistical text mining process of patent big data

가장 먼저 특허 빅 데이터의 분석의 주제에 해당하는 목표기술을 결정한다. 목표기술이 결정되면 키워드 검색식에 의해 목표기술에 해당하는 특허문서를 수집한다. 수집된 문서 데이터에 대하여 R의 'NLP'와 'tm' 패키지를 이용하여 문서 코퍼스(corpus)와 데이터 저장소를 구축하고, 이를 바탕으로 정형화된 특허 데이터를 만든다. R의 'e1071'와 'sna' 패키지, 그리고 R 기본 모듈을 이용하여 분석을 수행하여 최종적으로 통계적 모형과 시각화 결과를 얻는다. 이와 같은 분석 결과는 기업의 R&D 전략수립 및 신상품 개발을 위하여 사용된다. 본 논문에서 제안하는 방법론을 실제 R&D 전략구축에 적용하는 과정을 보이기 위하여 다음 절에서는 실제 사례분석을 수행한다.

4. 사례분석

제안하는 STM 기반 빅 데이터 전처리의 실제 사례분석을 위하여 본 논문에서는 대표적인 컴퓨터기업인 휴렛팩커드(Hewlett Packard; HP)의 출원, 등록 특허를 수집하여 분석하고 HP의 R&D 전략수립을 위한 결과를 얻었다. 먼저 특허문서의 수집을 위하여 미국특허청인 USPTO(United States Patent and Trademark Office)[22]와 특허검색 전문회사인 WIPSON(WIPS Corporation)[23]을 이용하였다. HP가 2013년까지 출원, 등록한 전체 특허 수는 29,043 건이었다. 본 논문의 STM을 수행하기 위하여 수집된 특허데이터로부터 코퍼스를 만들고 각 특허의 제목(title)과 요약(abstract) 정보를 추출하여 하나의 텍스트저장소를 구축하였다. 자연어처리 과정과 텍스트 마이닝의 과정을 통하여 정형화된 데이터구조를 만들었다. 본 사례분석에서 만들어진 정형화된 데이터의 각 행은 특허를 나타내고 각 열은 전체 특허문서에 나타난 단어(term)를 나타내었다. 그림 7은 HP 특허문서 데이터로부터 구축된 정형화된 데이터 구조를 나타내고 있다.

	a	abort	about	⋯	zone	zoo	zoom
US4991175	*Frequency of occurred term in each documents*						
US5657443							
US4985900							
⋮							
US8275818							
US8331131							
US8144555							

그림 7. 정형화된 HP의 특허 데이터

Fig. 7. Structured patent data of HP

위 데이터 구조는 각 행이 관측치(특허)가 되고 각 열이 변수(단어)로 이루어진 테이블 구조를 이루고 있기 때문에 통계학 및 기계학습 알고리즘 기반의 데이터 분석이 가능하게 된다. 먼저 정형화된 특허 데이터로부터 키워드를 추출하기 위하여 본 논문에서는 HP의 기술과 경영에 관한 여러 가지 자료를 이용하였다 [4],[24],[25],[26],[27]. 왜냐하면 제안하는 STM 방법론을 이용한 사례분석의 목적이 HP의 효율적인 R&D 전략기획이기 때문이다. 이를 통하여 선정된 키워드 리스트는 다음과 같다. 'analysis', 'application', 'business', 'cloud', 'computer', 'customer', 'data', 'enterprise', 'hardware', 'information', 'internet', 'management', 'medicine', 'mobile', 'monitor', 'network', 'printer', 'scanner', 'server', 'service', 'smart', 'software', 'storage', 'tablet', 'three-dimensional', 'web'.

즉 HP 기업이 보유한 다양한 기술은 위의 26개 키워드를 바탕으로 이루어진다고 할 때, HP의 키워드 간의 연관성을 파악하면 HP의 보유 기술 간의 상호 의존성을 확인할 수 있다. 또한 HP를 대표할 수 있는 중심 기술이 무엇인지를 찾기 위하여 SNA를 이용한 시각화 작업을 수행하였다. 그림 8은 HP 특허 데이터의 시각화 결과를 나타내고 있다.

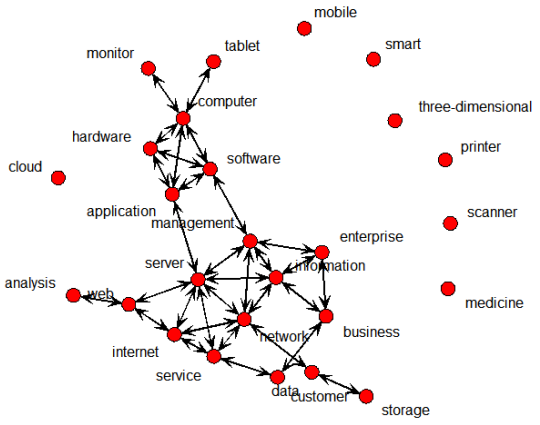

그림 8. 전체키워드를 이용한 SNA 그래프

Fig. 8. SNA graph by total keywords

그림 8의 SNA 그래프를 얻기 위하여 먼저 정형화된 특허

데이터를 이용하여 HP 특허 데이터의 키워드 간 상관분석을 실시하였고, 이 결과를 이용하여 SNA 그래프 작성에 필요한 인접행렬을 구하였다. 최종적으로 인접행렬을 이용한 SNA 그래프를 통하여 HP의 보유기술을 대표할 수 있는 중심 단어를 찾았다. 즉 그림 8의 SNA 그래프를 통하여 표 1과 같이 각 키워드 노드에 대한 차수(degree)를 계산하였다.

표 1. SNA 그래프를 이용한 각 키워드의 차수
Table 1. Degree of each keyword using SNA graph

Keyword	Deg.	Keyword	Deg.
server	7	customer	2
network	6	data	2
computer	5	analysis	1
information	5	monitor	1
management	5	storage	1
application	4	tablet	1
internet	4	cloud	0
service	4	medicine	0
software	4	mobile	0
business	3	printer	0
enterprise	3	scanner	0
hardware	3	smart	0
web	3	three-dimensional	0

키워드 노드의 차수는 해당 노드와 연결된 다른 노드의 개수를 의미한다. 따라서 해당 노드의 차수가 클수록 전체 키워드에서 중심적인 역할을 수행하게 된다. 키워드 server가 가장 높은 차수인 7의 값을 갖고 있다. 즉 7개의 다른 노드가 server 노드와 연결되어 있음을 알 수 있다. 다음으로 network이 6의 차수를 가지고 있고 computer, information, 그리고 management가 5의 차수를 나타내고 있다. 따라서 SNA 그래프 결과를 통하여 HP의 기술을 나타내는 중심 키워드는 server와 network이 우선적으로 고려될 수 있음을 알 수 있다. 이들 키워드에 연결된 다른 노드를 간의 관계를 좀 더 자세하게 파악하기 위하여 HP의 중심 키워드와 이에 연결된 키워드들로 이루어진 SNA 그래프는 그림 9와 같다.

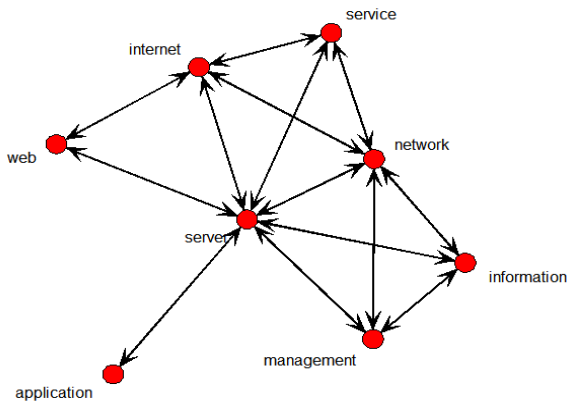

그림 9. 주요키워드를 이용한 SNA 그래프
Fig. 9. SNA graph by major keywords

위 그림을 통하여 server와 network에 영향을 미치는 키워드에 대한 다음과 같은 모형을 만들 수 있다. 즉 server와 network를 종속변수(dependent variable)로 하고, 이들에 연결된 키워드들을 독립변수(independent variable)로 하였다.

표 2. Server와 network에 영향을 미치는 키워드
Table 2. Keyword influencing server and network

dependent	independent
server	application, web, internet, service, network, information, management
network	information, management, server, internet, service

본 연구에서는 표준화된 선형회귀모형을 이용하여 키워드 간 기술 연관모형을 구축하였다. 그림 10은 Server에 영향을 미치는 키워드에 대한 기술모형을 나타내고 있다.

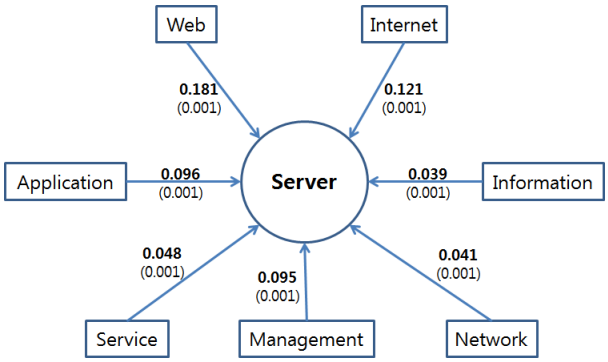

그림 10. 키워드 기반 기술모형: Server
Fig. 10. Technology model based on keywords: Server

위 그림에서 각 키워드를 연결하는 가중치는 회귀계수를 나타내고 괄호 안의 값은 각 회귀계수에 대한 유의확률(probability value)를 나타낸다. 일반적으로 95% 신뢰수준에서 유의확률이 0.05보다 작게 되면 해당 회귀계수는 통계적으로 유의함을 나타낸다 [28]. 키워드에 대한 유의확률이 0.05보다 작기 때문에 각 연결은 모두 통계적으로 의미가 있다. Server에 가장 큰 영향을 미치는 키워드는 Web 임을 알 수 있다. 다음으로 Internet, Application, Management, Service, Network, Information의 순서로 Server에 영향을 미치고 있음을 알 수 있다. 즉 server 기반 기술에 가장 영향을 미치는 것은 web 관련 기술이고 다음으로 Internet, Application 기반 기술들임을 알 수 있다. 그림 11은 HP에서 또 하나의 중심 키워드인 Network에 대한 기술모형 결과를 나타낸다.

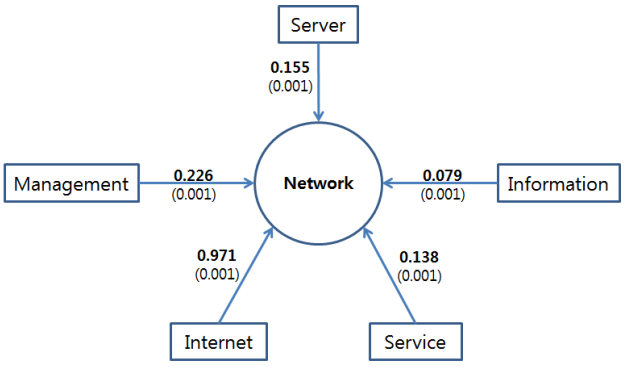

그림 11. 키워드 기반 기술모형: Network
Fig. 11. Technology model based on keywords: Network

Network에 영향을 미치는 5개의 키워드 중에서 Internet이 가장 큰 영향을 미치는 것으로 나타났다. 다음으로 Management, Server, Service, Information의 순서였다. Server의 경우와 마찬가지로 Network에 영향을 미치는 키워드들도 모두 통계적으로 유의함을 확인 할 수 있다. 그림 12는 그림 10과 11의 결과를 종합하여 HP의 R&D 전략을 위한 STM 결과를 나타내고 있다.

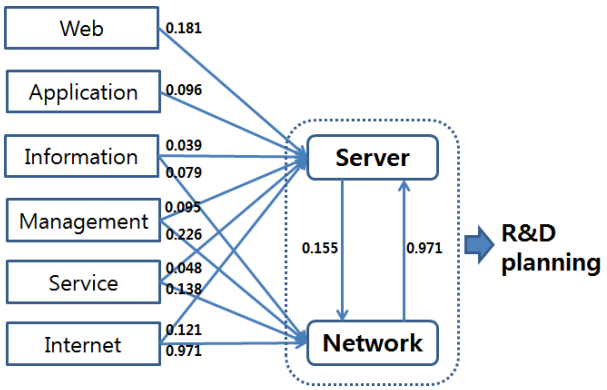

그림 12. 특허 빅데이터분석에 의한 R&D 기획
Fig. 12. R&D planning using patent big data analysis

HP는 Server와 Network 기술을 기반으로 연구개발을 진행하고 이와 같은 연구개발이 진행되기 위해서는 Web, Application, Information, Management, Service, 그리고 Internet 기반의 세부 기술이 필요하게 됨을 알 수 있다. 각 기술들 간의 연관정도는 연결선의 가중치를 통하여 확인할 수 있다.

5. 결론 및 향후 연구과제

본 논문은 빅 데이터의 효과적인 분석과 분석 결과의 활용을 위한 통계적 텍스트 마이닝 방법론에 대하여 연구하였다. 제안된 통계적 텍스트 마이닝 절차는 통계분석, 자연어처리, 텍스트 마이닝의 기법들과 R 데이터 언어를 이용하여 수행되었다. 새로운 분석 알고리즘보다는 방법론에 대한 소개이기 때문에 기존의 분석 기법들과의 비교보다는 실제 적용이 가능한 사례연구를 수행하였다. HP가 지금까지 출원, 등록한 전체 특허문서 데이터를 이용한 사례분석을 수행하였다. 초기에 특허 빅 데이터의 수집에서부터 정형화된 데이터 구축, 시각화, 통계분석을 통하여 최종적으로 HP의 R&D 전략수립을 위한 결과에 이르기까지 전 과정을 진행하였다. 최종 결과에 대한 실제 적용은 R&D 현업 실무자의 몫이 될 것이다.

제안된 HP의 특허데이터에 대한 STM 결과가 HP의 모든 R&D 전략을 결정하는 것은 아니지만 본 연구의 결과가 기업의 R&D 기획 과정에 사용되어 더 좋은 전략수립에 기여하게 될 것이다. 본 논문에서는 STM 절차를 기업의 R&D 전략수립을 위한 과정에 사용되었지만 기업의 마케팅, 바이오 신약 개발, 사회연결망 분석 등 다양한 분야에서 여러 가지 목적을 위하여 응용될 수 있으리라 기대된다. 본 연구에서는 선형회귀모형과 SNA 그래프 등 방대한 통계분석의 일부 기법만을 사용하였지만 앞으로 더 많은 분석 기법을 사용하여 더 정교하고 다양한 STM 방법론에 대한 연구가 기대된다.

References

[1] IBM, "*What is big data?*" www-01.ibm.com/software/data/bigdata, 2015.

[2] Gartner, "*Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data*," www.gartner.com/newsroom/id/1731916, 2015.

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, 2011.

[4] B. Choi, J. Kong, and M. Han, "The Model of Network Packet Analysis based on Big Data", *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 392-399, 2013.

[5] K. Kim, J. Jeong, and G. Park, "Assessment of External Force Acting on Ship Using Big Data in Maritime Traffic", *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 379-384, 2013.

[6] S. Hong, and M. Han, "The Efficient Method of Parallel Genetic Algorithm using MapReduce of Big Data", *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 385-391, 2013.

[7] H. Yoon, S. Park, "Pattern and Instance Generation for Self-knowledge Learning in Korean", *Journal of Korean Institute of Intelligent Systems*, Vol. 25, No.

1, pp. 63-69, 2015.
[8] S. Jun, "A Big Data Learning for Patent Analysis", *Journal of Korean Institute of Intelligent Systems*, Vol. 23, No. 5, pp. 406-411, 2013.
[9] S. Choi, and S. Jun, "Vacant technology forecasting using new Bayesian patent clustering," *Technology Analysis & Strategic Management*, Vol. 26, Iss. 3, pp. 241-251, 2014.
[10] S. Park, and S. Jun, "A Technology Forecasting Model Using Support Vector Clustering and Voting Approach," *Information – An International Interdisciplinary Journal*, Vol. 16, No. 2(B), pp. 1523-1528, 2013.
[11] H. Kim, J. Kim, J. Lee, S. Park, D. Jang, "A Novel Methodology for Extracting Core Technology and Patents by IP Mining", *Journal of Korean Institute of Intelligent Systems*, Vol. 25, No. 4, pp. 392-397, 2015.
[12] S. Jun, "Technology Forecasting of Intelligent Systems using Patent Analysis", *Journal of Korean Institute of Intelligent Systems*, Vol. 21, No. 1, pp. 100-105, 2011.
[13] D. Hunt, L. D. Nguyen, and M. Rodgers, *Patent Searching Tools & Techniques*, Wiley, 2007.
[14] A. T. Roper, S. W. Cunningham, A. L. Porter, T. W. Mason, F. A. Rossini, and J. Banks, *Forecasting and Management of Technology*, Wiley, 2011.
[15] S. Jun, and J. Choi, "Patent and Big Data, What's the Connection?", *Proceedings of KIIS Autumn Conference 2014* Vol. 24, No. 2, pp 183-184, 2014.
[16] J. Han, and M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001.
[17] R Development Core Team, *R: A language and environment for statistical computing, R Foundation for Statistical Computing,* Vienna, Austria, 2015.
[18] K. Hornik, *Package 'NLP' - Natural Language Processing Infrastructure*, CRAN R Project, 2015.
[19] I. Feinerer, K. Hornik, and D. Meyer, "Text mining infrastructure in R", *Journal of Statistical Software,* Vol. 25, No. 5, pp. 1-54, 2008.
[20] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C. C. Chang, and C. C. Lin, *Package 'e1071' - Misc Functions of the Department of Statistics, Probability Theory Group*, CRAN R Project, 2015.
[21] C. T. Butts, "Social Network Analysis with sna", *Journal of Statistical Software*, Vol. 24, Iss. 6, pp. 1-51, 2008.
[22] USPTO, The United States Patent and Trademark Office, http://www.uspto.gov, 2015.
[23] WIPSON, 'WIPS Corporation'. http://www.wipson.com, 2015.
[24] V. Nagali, J. Hwang, D. Sanghera, M. Gaskins, M. Pridgen, T. Thurston, P. Mackenroth, D. Branvold, P. Scholler, and G. Shoemaker, "Procurement Risk Management (PRM) at Hewlett-Packard Company", *Interfaces*, Vol. 38, Iss. 1, pp. 51-60, 2008.
[25] HP Office Site, http://www.hp.com, 2015.
[26] Hewlett-Packard from Wikipedia, https://en.wikipedia.org/wiki/Hewlett-Packard, 2015.
[27] Hewlett-Packard on Forbes Lists, http://www.forbes.com/companies/hewlett-packard, 2015.
[28] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*, Elsevier, 2012.

저 자 소 개

전성해(Sunghae Jun)
1993년 : 인하대 통계학과 (학사)
1996년 : 인하대 통계학과 (이학석사)
2001년 : 인하대 통계학과 (이학박사)
2007년 : 서강대학교 컴퓨터공학과 (공학박사)
2013년 : 고려대학교 정보경영공학과 (공학박사)
2003년~현재 : 청주대학교 통계학과 교수

관심분야 : 데이터과학, 인공지능, 기술경영
Phone : +82-43-229-8205
Fax : +82-43-229-8432
E-mail : shjun@cju.ac.kr