Journal of Korean Institute of Intelligent Systems, Vol. 23, No. 5, October 2013, pp. 406-411
http://dx.doi.org/10.5391/JKIIS.2013.23.5.406

특허분석을 위한 빅 데이터학습

A Big Data Learning for Patent Analysis

전성해

Sunghae Jun

청주대학교 통계학과

Department of Statistics, Cheongju University

요 약

빅 데이터는 여러 분야에서 다양한 개념으로 사용된다. 예를 들어, 컴퓨터학과 사회학에서 빅 데이터에 대한 접근방법에 차이가 있지만, 데이터분석 관점에서는 공통적인 부분을 갖는다. 즉, 공학이든 사회과학이든 빅 데이터에 대한 분석은 반드시 필요하다. 통계학과 기계학습은 빅 데이터의 분석을 위한 대표적인 분석도구이다. 본 논문에서는 빅 데이터분석을 위한 학습도구에 대하여 알아보고 검색된 빅 데이터 원천에서부터 분석을 거쳐 최종적으로 분석결과를 사용하는 전체과정에 대하여 효율적인 빅 데이터학습 절차에 대하여 제안한다. 특히, 대표적인 빅 데이터 구조를 갖고 있는 특허문서에 대하여 빅 데이터학습을 적용하여 특허분석을 수행하고 이 결과를 기술예측에 적용하는 방법에 대하여 연구한다. 제안방법에 대한 실제적용을 위하여 전 세계 특허청으로부터 빅 데이터 관련 특허문서를 검색하여 텍스트 마이닝의 전처리와 통계학의 다중선형회귀분석을 이용한 구체적인 빅 데이터학습에 대한 사례연구를 수행하였다.

키워드 : 빅 데이터학습, 통계학, 기계학습, 텍스트 마이닝, 특허분석, 다중선형회귀분석

Abstract

Big data issue has been considered in diverse fields. Also, big data learning has been required in all areas such as engineering and social science. Statistics and machine learning algorithms are representative tools for big data learning. In this paper, we study learning tools for big data and propose an efficient methodology for big data learning via legacy data to practical application. We apply our big data learning to patent analysis, because patent is one of big data. Also, we use patent analysis result for technology forecasting. To illustrate how the proposed methodology could be applied in real domain, we will retrieve patents related to big data from patent databases in the world. Using searched patent data, we perform a case study by text mining preprocessing and multiple linear regression of statistics.

Key Words : Big Data Learning, Statistics, Machine Learning, Text Mining, Patent Analysis, Multiple Linear Regression.

1. 서론

데이터 저장장치와 클라우드 컴퓨팅 기술의 발달로 인해 대용량 데이터에 대한 간편한 접근이 가능해졌다. 이와 함께 컴퓨터 처리능력이 지속적으로 발전되어 오고 있기 때문에 대용량 데이터의 분석이 중요한 이슈가 되는 빅 데이터 시대가 되었다 [1]. 자연과학 및 공학분야 뿐만 아니라 사회과학 및 공공부분에서 빅 데이터분석에 대한 수요는 증가하고 있다. 최근 몇 년 사이 데이터 마이닝이라는 용어보다는 이제는 빅 데이터라는 개념이 더 자주 사용되고 있다. 기존의 데이터 마이닝에서 대용량 데이터 원천(legacy)은 관계형 데이터베이스로 구축되고, 이를 바탕으로 분석을 위한 데이터웨어하우스(data warehouse)을 만들어 최종적으로 분석이 이루어 졌다 [2]. 그러나 최근의 대용량 데이터의 환경은 진화되고(evolving) 데이터분석에 대한 개념도 바뀌고 있다 [3-6]. 이전에 비해 최근의 레거시 데이터는 훨씬 크고 이질적인 특성을 갖고 있다. 즉, 숫자와 문자, 그림과 그래프, 동영상 등 매우 다양한 데이터들로 이루어졌다. 데이터분석에서 고려해야 할 사항들이 더 많아졌다. 따라서 모든 레거시데이터를 관계형 데이터베이스로 구축하기에는 어려움이 있다. 사회가 빠르게 바뀌면서 데이터분석을 통한 의사결정과정의 시간도 점점 빨라지고 있다. 데이터베이스와 데이터웨어하우스를 구축할 시간도 줄여야 할 필요성이 생겼다. 그러므로 레거시데이터에서 바로 데이터분석이 이루어지고 이를 통해 즉각적인 실제적용이 가능해야 한다. 본 논문에서는 이와 같은 대용량 데이터분석방법을 빅 데이터학습(big data learning)이라 하고, 이에 대한 효율적인 방법을 연구한다. 특히, 본 연구는 제안하는 빅 데이터학습을 특허문서의 데이터분석에 적용한다. 왜냐하면 특허문서는

접수일자: 2013년 8월 19일
심사(수정)일자: 2013년 9월 10일
게재확정일자 : 2013년 9월 15일

† Corresponding author

출원 및 등록날짜를 비롯하여 발명가, 특허제목, 기술요약, 인용정보, 기술상세기술, 도면, 절차도 등 숫자와 문자, 그리고 그림까지 다양한 형태의 데이터 특성을 가지고 있으며 [7] 동시에 전 세계에 출원된 특허 데이터의 크기는 매우 방대하기 때문이다. 현재 개발된 기술결과에 대하여 가장 상세하고 방대한 정보를 가지고 있는 레거시데이터는 특허이기 때문에 특허분석을 통하여 신상품개발, 기술예측(technology forecasting) 등이 가능하게 된다 [8]. 본 연구에서는 전 세계에 출원된 빅 데이터 관련 특허문서를 검색하여 빅 데이터학습을 통해 처리, 분석하고 이 분야에 대한 기술관계를 찾아내는 사례연구를 통하여 제안방법에 대한 실제적용 과정을 단계별로 나타낸다.

2. 빅 데이터

정보통신기술(information & communication technology; ICT)의 획기적인 발달과 사회의 다양성이 증가하면서 발생되는 데이터의 양은 상상을 초월하는 빅 데이터의 구조가 되었다 [1]. 빅 데이터는 데이터 자체의 크기 뿐만 아니라 데이터가 가지고 있는 복잡성에 의미를 둔다 [9]. 가트너(Gartner)는 빅 데이터를 3가지 관점에서 정의하였다 [10]. 즉, 크기(volume), 저장 및 전달 속도(velocity), 그리고 구조의 다양성과 복잡성(variety)을 가진 데이터로 빅 데이터를 나타내었다. 데이터의 크기는 데이터베이스 마케팅, 데이터 마이닝 등이 널리 알려진 1980년대부터 예상이 가능한 부분이었다. 또한 전 세계의 데이터가 서로 연결되어 누구든지 방대한 양의 데이터를 주고 받을 수 있게 되었기 때문에 데이터 전달속도의 증가도 충분히 예견되었다. 하지만 데이터의 크기와 속도의 증가에 따라 나타난 데이터의 복잡성과 다양성에 대한 문제는 상대적으로 대비가 덜 된 부분이다. 이 분야에 대하여 해결되어야 할 많은 문제점들이 존재하지만 이에 대한 연구는 상대적으로 소홀하였다. 우선 고려해야 할 문제는 이질적이고 다양한 데이터로 가득 찬 대용량 데이터로부터 어떻게 필요한 정보를 추출한 것인가 하는 것이다. 빅 데이터로부터 유용한 지식을 빠르게 추출할 수 있는 능력은 기업과 국가 모두에게 중요한 이슈가 되었다. 빅 데이터는 문자, 숫자, 그림, 음성, 동영상 등 다양한 형태의 데이터를 포함하고 있다 [11]. 이와 같은 빅 데이터의 특성 때문에 기존의 데이터 분석기법을 그대로 적용하는 데는 한계가 있다. 본 논문은 빅 데이터 환경에서 주어진 레거시데이터를 효율적으로 분석할 수 있는 빅 데이터학습 방법에 대하여 연구한다.

3. 빅 데이터학습을 이용한 특허분석

대용량 데이터베이스로부터 숨겨진 패턴(hidden pattern)을 찾아내는 데이터 마이닝은 주로 다양한 레거시근원데이터로부터 관계형 데이터베이스를 구축하고, 이를 바탕으로 데이터분석을 위한 데이터웨어하우스를 구축한 후에 본격적인 데이터분석을 진행한다 [2]. 이에 비해 빅 데이터는 데이터베이스를 구축하는 데에 시간적, 기술적으로 어려움이 존재한다. 빠르게 진화하는 데이터 환경과 의사결정 과정에서 더 빠르고 정확한 분석결과가 요구된다. 다음 그림은 데이터 마이닝과 빅 데이터에 대한 개념을 나타내고 있다.

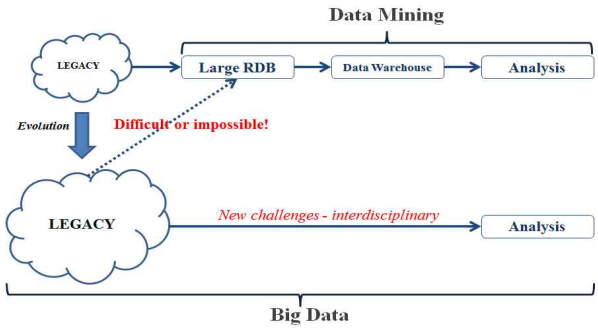

그림 1. 데이터 마이닝과 빅 데이터
Fig. 1. Data mining and big data

이전에 비해 레거시데이터는 더욱 다양한 데이터 형태를 갖게 되었다. 숫자와 문자 뿐만 아니라 그림, 소리, 동영상 등 이질적인(heterogeneous)이고 방대한 데이터로 진화하고 있다. 따라서 이와 같은 데이터를 관계형 데이터베이스로 구축하기에는 어려움이 있다. 빠르게 변화하는 데이터를 효율적으로 분석하기 위해서 별도로 데이터베이스를 구축하기에는 시간적 여유가 없다. 물론 분야에 따라 관계형 데이터베이스를 구축하여 데이터를 관리해야 하는 분야도 있지만 본 연구에서는 빅 데이터의 분석과 적용 관점에서 연구가 이루어진다. 그러므로 빅 데이터의 분석은 레거시데이터로부터 직접적인 분석 작업이 요구된다. 즉, 레거시데이터로부터 전처리(preprocessing)를 통하여 분석에 적합한 데이터를 직접 구축한다. 데이터 마이닝 과정에서는 레거시데이터로부터 관계형 데이터베이스와 데이터웨어하우스를 구축하는 작업은 분석가가 직접 하지 않고 데이터베이스 전문가에게 맡겼지만 빅 데이터 환경에서는 분석가가 레거시데이터로부터 분석을 위한 구조화된(structured) 데이터를 직접 만들어야 한다. 이 과정에서 텍스트 마이닝기법 등을 이용하는 전처리작업이 필요하고 이 작업은 분석을 염두해 두고 진행되어야만 빠르고 정확한 분석결과를 기대할 수 있다. 이 과정은 다양한 학문분야의 전문가들에게 새로운 연구작업이 될 수 있다. 본 연구에서는 빅 데이터로서 특허문서가 고려되었기 때문에 특허분석을 고려하여 다양한 전처리 과정이 이루어진다. 본 논문에서는 특허 데이터의 여러 유형 중에서 문자 데이터를 선택하여 분석한다. 특허문서를 구성하는 세부요소들 중에서 특허제목(title)과 기술요약정보(abstract)만을 선택하여 별도의 데이터 셋(data set)을 구축한다. 다음 그림은 본 연구에서 사용된 데이터 셋이다.

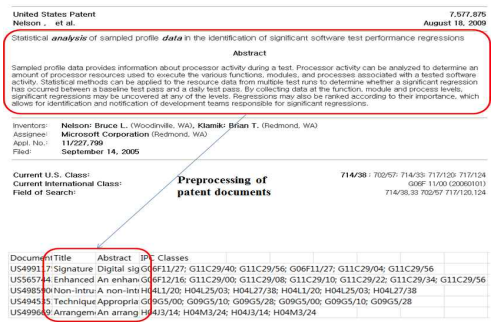

그림 2. 특허문서의 데이터 셋
Fig. 2. Data set of patent documents

개별 특허문서로부터 제목과 요약정보만을 추출하여 별도의 엑셀(Excel) 파일을 만들고, 이를 이용하여 전처리를 포함한 데이터분석을 수행한다. 이 작업은 기존의 데이터마이닝과정에서 분석을 위한 데이터베이스와 데이터웨어하우스의 구축과정과 같은 의미를 갖는다. 그러나 데이터베이스를 구축하는 과정은 데이터베이스 설계부터 시작하여 많은 시간과 비용이 필요하지만 빅 데이터 환경에서는 이 과정을 최대한 줄이고 단축해야 한다. 물론 엑셀 파일 이외에도 분석가에 따라 다양한 데이터 파일을 이용할 수 있다. 다음 그림은 본 연구에서 제안하는 빅 데이터학습을 이용한 특허분석의 과정을 나타내고 있다.

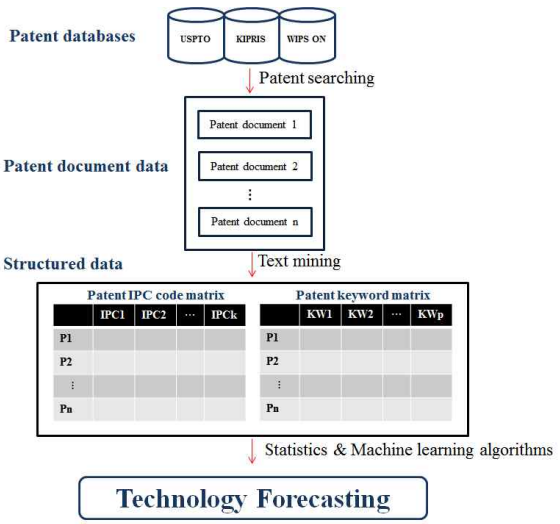

그림 3. 빅 데이터학습을 이용한 특허분석
Fig. 3. Patent analysis using big data learning

특허문서 데이터는 전 세계에 존재하는 각국의 특허데이터베이스가 레거시데이터가 되며 이곳으로부터 특정기술에 대한 특허문서를 검색하고 엑셀파일과 같은 특허데이터 셋을 구축한다. 이와 같이 구축된 n개의 특허문서들은 텍스트마이닝의 전처리 과정을 통하여[12-14] 분석모형에 적합한 데이터 구조를 같는 특허-IPC 코드 행렬, 또는 특허-단어 행렬로 변환된다.

구축된 데이터행렬을 위한 빅 데이터학습 도구는 크게 3가지 이루어진다. 첫 번째 분석도구는 통계학이다. 로스(S. M. Ross)는 통계학을 다음과 같이 정의하였다. “Statistics is the art of learning from data” [15]. 즉, 통계학은 주어진 데이터로부터 다양한 학습기법을 이용하여 의사결정을 필요한 지식을 추출하는 기술이다. 그러므로 통계학은 빅 데이터 분석을 위한 좋은 도구가 된다. 두 번째 분석도구는 기계학습(machine learning) 알고리즘이다 [16]. 인공신경망(artificial neural networks) 모형을 비롯한 대부분의 기계학습 알고리즘은 빅 데이터학습을 위한 분석도구가 된다. 마지막으로 자료구조(data structure)와 컴퓨터 알고리즘(computer algorithm)도 빅 데이터학습을 위한 분석도구를 만드는 이론적 근거가 된다. 예를 들어 트리이론(tree theory)을 통하여 의사결정나무모형(decision tree model)을 구축할 수 있고, 그래프이론(graph theory)을 이용하여 사회네트워크분석(social network analysis) 모형을 만들 수 있다. 모두 빅 데이터의 효과적인 분석도구가 된다. 제안방법은 빅 데이터학습을 위한 3가지 분석도구를 이용하여 구조화된 특허데이터를 분석하고 이 결과를 이용하여 기술예측을 위한 유용한 패턴을 찾는 것이다. 다음 그림은 레거시데이터로부터 빅 데이터학습과정을 일반화하여 보여준다.

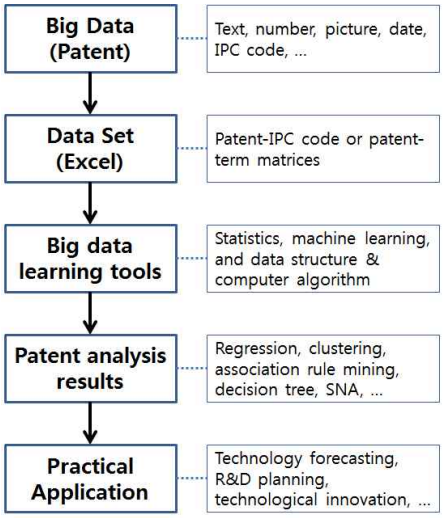

그림 4. 빅 데이터 학습과정
Fig. 4. Learning processing of big data

그림4는 빅 데이터의 레거시데이터로부터 최종적인 실제 적용까지 전 과정에 대하여 세부적으로 보여준다. 본 연구에서는 통계학, 기계학습, 자료구조 및 컴퓨터알고리즘의 3가지 학습도구 중에서 통계학에서 제공하는 회귀분석모형(regression model)을 사용한다 [17]. 회귀분석을 통하여 검색된 특허데이터에 포함된 단어들 간의 인과관계를 모형화하여 기술들 간의 연관관계를 찾아낸다. 다음 식은 본 연구에서 사용되는 다중선형회귀모형이다.

$$W_T = b_0 + b_1 W_1 + \cdots + b_r W_r \quad (1)$$

이 식에서 W_T는 목표변수(target variable)를 나타내는 단어이고, W_1, ..., W_r은 목표변수에 영향을 주는 설명변수(exploratory variable)를 나타내는 단어들이다. b_0는 편이(bias)를 나타내는 절편(intercept)이고, $b_1, b_2, \ldots, b_r$은 해당되는 각 설명변수(단어)의 목표변수(단어)에 대한 가중치인 모수(parameter)이다. 각 모수에 대한 유의확률(probability value, p-value)값이 0.1보다 작으면 90% 신뢰수준에서 해당되는 설명변수가 목표변수에 유의한 영향을 미친다고 통계적으로 판단한다 [17]. 이 결과를 이용하여 목표기술에 영향을 미치는 세부기술을 찾을 수 있다. 즉 특허분석을 통하여 기술예측을 위한 유용한 패턴을 찾는다 [18-19].

4. 실험 및 결과

본 논문에서는 제안방법의 성능평가를 위하여 특허데이터를 이용한 사례분석을 수행하였다. 특허의 제목에 “빅 데이터”를 포함하는 전 세계 모든 특허를 검색하였다 [20]. 검

색이 가능한 12개 국가들 중에서 일본, 영국, 독일, 프랑스, 호주, 캐나다, 러시아, 대만은 출원, 등록된 특허가 없었고, 미국이 17건, 유럽이 1건, 그리고 중국이 15건이었다. PCT(Patent Cooperation Treaty) 국제특허는 3건이었다. 미국과 중국을 제외한 대부분의 나라에서는 아직 빅 데이터에 관한 기술특허는 제대로 이루어지지 못하고 있음을 알 수 있다.빅 데이터에 관한 연구논문이 활발히 발표되고 있기 때문에 앞으로 전 세계적으로 빅 데이터관련 기술특허에 대한 출원, 등록의 빠른 증가가 기대된다. 다음 그림은 전체 특허에 대한 연도별 특허건수를 나타낸다.

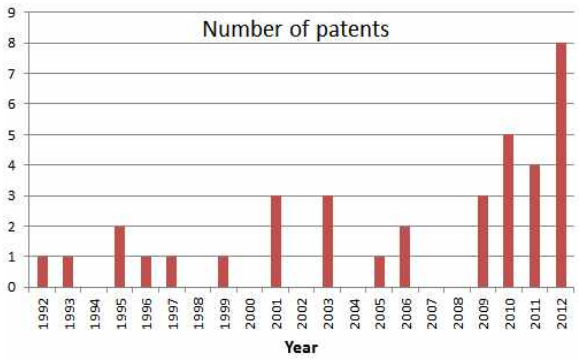

그림 5. 연도별 빅 데이터 특허건수
Fig. 5. Number of patents related to big data by year

2010년 전후로 빅 데이터기술관련 특허의 출원 건수가 증가하고 있음을 알 수 있다. 즉, 빅 데이터가 적극적으로 알려진 시점에 이와 관련된 기술특허의 출원도 함께 증가하고 있음을 알 수 있다. 기울기가 매우 큼에 따라 앞으로 증가의 정도가 매우 클 것으로 판단된다. 다음은 본 논문에서 사용한 빅 데이터관련 특허데이터에 대한 엑셀파일의 일부이다.

국가	출원번호	출원일자	IPC분류	출원인	발명자/고	발명의명칭	초록	청구항
US	11492151	2006.07.25	G11C 7/0(	Samsung	Lee Jin-Yu	Flash memor	There is p	1. A flash
US	13288950	2011.11.03	G06F 17/3	Roger Bai	Roger Bai	SYSTEMS AN	Data mar	1. A data
US	13415622	2012.03.08	G06F 17/0	Wenwey	Wenwey	Apparatus a	An appar	1. An ap
WO	US201206	2012.11.01	G06F 17/0	MICROSO	BARGA, R	SYSTEMS AN	Data mar	
CN	2.01E+11	2012.08.27	G06F 17/3	Nanjing Ir	Zhang Zh	Method achi	The invention disclo	
CN	2.01E+11	2010.11.18	G06Q 40/(	Zheng Shi	Zheng Shi	Data process	The invention relates	
CN	2.01E+11	2010.12.30	G06F 17/3	Chanjet Si	Huang Ha	Method and	The invention provid	
US	8497492	1995.06.30	G06F 13/0	Thomson	Willard; Pi	Apparatus fc	A distribu	What is cl

그림 6. 데이터 셋: 엑셀파일
Fig. 6. Data set: Excel file

즉, 검색된 특허문서들로부터 빅 데이터학습에 필요한 항목들만 별도로 추출하여 엑셀파일 형식의 데이터 셋을 만들었다. 미국특허청(USPTO)에 비해 한국특허청(KIPRIS)의 검색사이트에서는 검색된 특허문서에 대한 엑셀파일을 자동으로 제공해 준다. 만약 이와 같은 기능이 없을 경우에는 검색된 특허문서로부터 엑셀파일을 만드는 별도의 작업이 필요하게 된다. 다음 그림은 빅 데이터학습 도구를 사용하기 위하여 엑셀파일로부터 전처리 과정을 거친 구조화된 데이터인 특허-단어행렬을 나타낸다.

	able	access	⋯	work	writing
$patent_1$	0	1	⋯	0	0
$patent_2$	4	0	⋯	0	3
⋮	⋮	⋮	⋱	⋮	⋮
$patent_{36}$	1	0	⋯	1	0

그림 7. 구조화된 데이터: 특허-단어 행렬
Fig. 7. Structured data: patent-word matrix

위 행렬의 차원은 (36*1009)이다. 즉 36개의 특허문서와 전체특허문서에 적어도 1번 이상 나타난 1009개의 단어로 이루어졌다. 각 셀은 각 특허에 나타난 해당 단어의 빈도값을 나타낸다. 특허데이터셋의 전처리를 위하여 본 논문에서는 R Project와 'tm' 패키지를 사용하였다 [13, 21]. 이 패키지는 공개 소프트웨어로서 텍스트 마이닝 기법을 이용한 다양한 전처리 기능을 제공한다 [14]. 상위 출현빈도를 갖는 단어를 찾기 위하여 tm 패키지는 findFreqTerms 함수를 제공한다. 이 함수를 이용하여 빈도별 순위를 갖는 의미 있는 단어의 목록을 다음 표와 같이 구하였다.

표 1. 빅 데이터특허의 상위 출현단어
Table 1. High ranked words of big data patents

Rank	Extracted words
1	endian
2	memory, method, system
3	bit, block, query
4	control, processing, device, storage
5	interface, network, window, circuit, semiconductor, initiator, mobile, sum, switching, width, address, client, sliding, writing, attribute, communication, length, parallel, compression, line, computer, flash, unaligned, location, information

모든 문서에 나타나는 "big"과 "data", 그리고 "and", "the", 등 공통단어(common word)는 제외했다. 이 표는 전체 특허문서 데이터에서 50번 이상 발생한 단어를 1위로 하고, 마지막으로 10번 이상을 5위로 한 결과이다. 메모리 속에서 데이터의 순서를 정하는 방법을 나타내는 단어인 "endian"이 가장 많은 빈도를 나타내고 있다. 전체적으로 현재까지 빅 데이터 관련 특허에서는 소프트웨어보다는 하드웨어 관련 기술이 주로 개발되고 있음을 알 수 있다. 1992년에 빅 데이터 관련 기술특허가 처음 출원, 등록되었고 비교적 최근에 증가하는 경향을 나타내기 때문에 향 후 더 많은 관련기술의 개발이 기대된다. 다음으로 빅 데이터에 구체적으로 유의한 영향을 미치는 기술단어가 무엇인지 찾기 위하여 다중선형회귀분석을 수행한다. 빅 데이터 관련 특허문서에 대한 분석이기 때문에 "big" 과 "data"를 목표변수로 하고 20번 이상 나타나 단어들을 설명변수로 하였다. 다음 표는 회귀분석모형에 대한 결과를 나타낸다.

표 2. 회귀분석결과: 유의확률
Table 2. Regression result: p-value

Exploratory word	Target word	
	big	data
bit	0.6415	0.6723
block	0.0835	0.0022
control	0.0713	0.1461
device	0.4931	0.4314
endian	0.0205	0.8302
memory	0.2433	0.0242
method	0.0503	0.6722
processing	0.0420	0.8589
query	0.3899	0.1819
storage	0.3727	0.7825
system	0.1736	0.9470

목표단어가 “big”인 경우 p-value가 0.1보다 작은 유의한 설명단어는 “block”, “control”, “endian”, “method”, 그리고, “processing”이었다. 즉, 이들 설명단어들로 이루어진 기술이 빅 데이터의 최종기술개발에 영향을 주게 됨을 알 수 있다. 또한 “data”에 영향을 미치는 단어는 “block”과 “memory”임을 알 수 있다. 따라서 “block”과 “memory”에 기반한 기술도 역시 빅 데이터의 최종기술개발에 영향을 끼침을 알 수 있다. 다음 그림은 제안 방법에 의한 빅 데이터 관련 특허분석의 전체과정을 나타내고 있으며, 마지막으로 기술예측을 위한 결론 도출까지를 포함하고 있다.

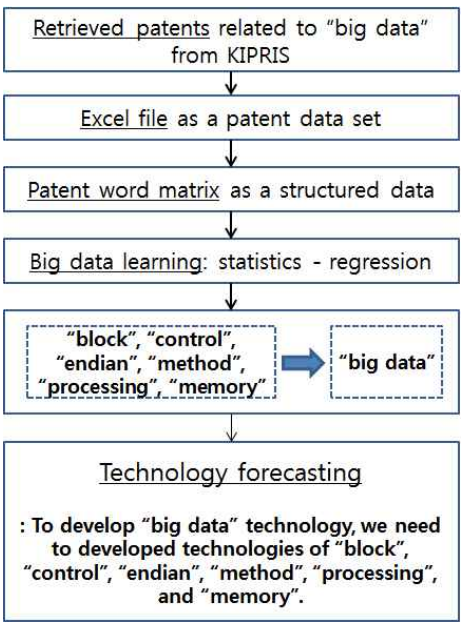

그림 8. 빅 데이터 관련 특허분석
Fig. 8. Patent analysis related to big data

즉, 빅 데이터관련 특허문서를 검색하고 기술예측이라는 주제에 필요한 항목만 별도로 엑셀파일에 저장하였다. 텍스트 마이닝을 이용한 전처리과정을 통하여 특허-단어 행렬을 구축하고 이를 이용하여 특허분석을 하였다. 본 연구에서는 빅 데이터학습 도구로 통계학의 회귀분석모형을 이용하였다. 물론 통계학과 기계학습에서 제공하는 대부분의 분석기법들은 빅 데이터학습 도구로 사용될 수 있다. 다중선형회귀모형에서 "big data"가 목표변수가 되고, 나머지 상위출현단어들을 설명변수로 하였다. 회귀분석결과 “big data"에 통계적으로 유의한 영향을 미치는 단어들로 “block”, “control”, “endian”, “method”, “processing”, 그리고 “memory”을 찾았다. 이를 통해 기술예측의 실제 적용에 대한 지식으로 빅 데이터 관련 기술의 개발을 위해서 통계적으로 유의한 판정을 받은 설명변수에 해당되는 단어들로 이루어진 기술들에 대한 선행개발이 필요하다는 결론을 얻게 되었다.

5. 결론 및 향후 연구과제

본 논문은 빅 데이터에 대한 효율적인 분석을 위하여 빅 데이터학습에 대한 방법을 제안하였다. 관계형 데이터베이스를 구축하고 분석을 위한 데이터웨어하우스를 추가적으로 만들어야 하는 기존의 데이터 마이닝 과정에 비해 제안하는 빅 데이터학습 방법은 레거시데이터로부터 직접 필요한 항목만을 선별적으로 찾아내어 데이터 셋을 구축하고 텍스트 마이닝의 전처리과정을 통하여 분석에 필요한 구조화된 데이터를 최종적으로 구축하였다. 기존의 방법에 비해 시간과 비용의 측면에서 효율적 빅 데이터분석이 가능하게 되었다. 빅 데이터분석을 위한 학습도구로 본 연구에서는 통계학, 기계학습, 그리고 자료구조 및 컴퓨터 알고리즘으로부터의 개념 및 분석기법을 사용하였다. 제안방법의 실제적용을 보이기 위하여 빅 데이터관련 특허문서를 수집하여 제안한 빅 데이터학습 절차에 따라 분석하였다. 최종적으로 특허분석을 이용한 기술예측의 결론을 도출하였다. 제안방법은 다른 기술의 모든 특허분석에 적용될 수 있을 뿐만 아니라 특허를 포함한 다른 빅 데이터의 분석에도 사용될 수 있다.

앞으로도 데이터의 크기는 빠른 속도로 증가할 것이고 빅 데이터에 포함된 데이터의 형태는 더욱 이질적인 특징을 갖게 될 것으로 기대된다. 이에 따라 더욱 빠르고 정확한 빅 데이터학습 방법이 연구되어야 할 것이다. 이를 위하여 통계학과 기계학습, 그리고 자료구조 및 컴퓨터 알고리즘에 대한 다양한 분석기법들이 빅 데이터학습에 적용될 때 발생될 수 있는 문제점과 이에 대한 해결방안에 대한 연구가 진행되어야 할 것이다.

References

[1] H. Yang, *Technology Planning Methodology Using Big Data*, Issue paper 2012-14, Korea Institue of Science & Technology Evaluation and Planning, 2012.

[2] J. Han, M. Kamber, *Data Mining Concepts and Techniques*, Morgan Kaufmann, 2001.

[3] H. Shin, H. Jung, K. Cho, J. Lee, “A Prediction Method of Learning Outcomes based on Regression Model for Effective Peer Review Learning,” *Journal of Korean Institute od Intelligent Systems*, vol. 22, no. 5, pp. 624-630, 2012.

[4] Y. Park, K. Park, “Estimation of Project Performance Using Fuzzy Linear Regression,” *Journal of Korean Institute od Intelligent Systems*, vol. 18, no. 6, pp. 832-836, 2008.
[5] S. Kang, J. Kim, “Intelligent Spam-mail Filtering Based on Textual Information and Hyperlinks,” *Journal of Korean Institute od Intelligent Systems*, vol. 14, no. 7, pp. 895-901, 2004.
[6] K. Kim, S. Lim, “Building Domain Ontology Based on Linguistic Patterns,” *Journal of Korean Institute od Intelligent Systems*, vol. 16, no. 6, pp. 766-771, 2006.
[7] D. Hunt, L. D. Nguyen, M. Rodgers, *Patent Searching Tools & Techniques*, Wiley, 2007.
[8] A. T. Roper, S. W. Cunningham, A. L. Porter, T. W. Mason, F. A. Rossini, J. Banks, *Forecasting and Management of Technology*, Wiley, 2011.
[9] IBM, *"What is big data?"* www-01.ibm.com/software/data/bigdata, 2013, [Accessed: July 11, 2013]
[10] Gartner, *"Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data,"* www.gartner.com/newsroom/id/1731916, 2013, [Accessed: July 22, 2013]
[11] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers, *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, 2011.
[12] I. Feinerer, *A Text Mining Framework in R and Its Applications*, PhD Dissertation, Department of Statistics and Mathematics Vienna University of Economics and Business Administration, 2008.
[13] I. Feinerer, K. Hornik, *Package 'tm', Text Mining Package*, R Project CRAN, 2013.
[14] I. Feinerer, K. Hornik, D. Meyer, “Text mining infrastructure in R,” *Journal of Statistical Software*, vol. 25, no. 5, pp. 1-54, 2008.
[15] S. M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists*, Elsevier, 2009.
[16] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, Springer, 2001.
[17] B. L. Bowerman, R. T. O'Connell, A. B. Koehler, *Forecasting, Time Series, and Regression, An Applied Approach*, Brooks/Cole, 2005.
[18] S. Jun, “Technology Forecasting of Intelligent Systems Using Patent Analysis,” *Journal of Korean Institute od Intelligent Systems*, vol. 21, no. 1, pp. 1-6, 2011.
[19] S. Jun, "Vacant Technology Forecasting Using Ensemble Model," *Journal of Korean Institute od Intelligent Systems*, vol. 21, no. 3, pp. 341-346, 2011.
[20] KIPRIS, "Korea Intellectual Property Rights Information Service," www.kipris.or.kr, 2013, [Accessed: July 5, 2013]
[21] R Development Core Team, *R: A language and environment for statistical computing, R Foundation for Statistical Computing*, Vienna, Austria, 2013.

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 부교수

관심분야 : 기술예측, 인공지능, 데이터마이닝, 빅 데이터
Phone : +82-43-229-8205
Fax : +82-43-229-8432
E-mail : shjun@cju.ac.kr