**연구노트**

**(수학과)이 재 봉**  2017.08.24 (목)

|  |  |  |
| --- | --- | --- |
|  | | |
| **연 구**  **주 제** | **특허분석을 통한 미래기술 예측** | |
| **관련 논문 조사** | |
| **구 분** | **연구내용** | **진행률** |
|  | 전성해 (2013). 특허분석을 위한 빅 데이터학습  빅 데이터  정보통신기술(information & communication technology;ICT)의 획기적인 발달과 사회의 다양성이 증가하면서발생되는 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 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 가능한부분이었다. 또한 전 세계의 데이터가 서로 연결되어 누구든지 방대한 양의 데이터를 주고 받을 수 있게 되었기 때문에 데이터 전달속도의 증가도 충분히 예견되었다. 하지만데이터의 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. 우선고려해야 할 문제는 이질적이고 다양한 데이터로 가득 찬대용량 데이터로부터 어떻게 필요한 정보를 추출한 것인가하는 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. 이와 같은 빅 데이터의 특성 때문에 기존의 데이터 분석기법을 그대로 적용하는 데는 한계가 있다.  빅 데이터학습을 이용한 특허분석  대용량 데이터베이스로부터 숨겨진 패턴(hidden pattern)을 찾아내는 데이터 마이닝은 주로 다양한 레거시근원데이터로부터 관계형 데이터베이스를 구축하고, 이를 바탕으로데이터분석을 위한 데이터웨어하우스를 구축한 후에 본격적인 데이터분석을 진행한다 . 이에 비해 빅 데이터는 데이터베이스를 구축하는 데에 시간적, 기술적으로 어려움이존재한다. 빠르게 진화하는 데이터 환경과 의사결정 과정에서 더 빠르고 정확한 분석결과가 요구된다. 다음 그림은 데이터 마이닝과 빅 데이터에 대한 개념을 나타내고 있다  이전에 비해 레거시데이터는 더욱 다양한 데이터 형태를 갖게 되었다. 숫자와 문자 뿐만 아니라 그림, 소리, 동영상 등이질적인(heterogeneous)이고 방대한 데이터로 진화하고있다. 따라서 이와 같은 데이터를 관계형 데이터베이스로구축하기에는 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 직접만들어야 한다. 이 과정에서 텍스트 마이닝기법 등을 이용하는 전처리작업이 필요하고 이 작업은 분석을 염두해 두고진행되어야만 빠르고 정확한 분석결과를 기대할 수 있다.이 과정은 다양한 학문분야의 전문가들에게 새로운 연구작업이 될 수 있다. 본 연구에서는 빅 데이터로서 특허문서가고려되었기 때문에 특허분석을 고려하여 다양한 전처리 과정이 이루어진다. 본 논문에서는 특허 데이터의 여러 유형중에서 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. |  |
| **차 주**  **계 획** | ■ 계획  - 관련 논문 조사 | |
| **이 슈** | ■ **해당사항 없음** | |

|  |  |
| --- | --- |
|  |  |