빅데이터의 이해 HW7

영어영문학과(휴먼ICT) 박도준

1. 빅데이터란 무엇인가?

빅데이터(big data)란 그 단어가 말해주듯 기존의 분석 역량을 넘어서는 크기의 대용량 데이터를 말한다. 빅데이터라는 용어는 2001년 시장조사업체 더그 레이니가 발표한 연구보고서의 의해 처음 정의되었다. 이 보고서에 따르면, 빅데이터는 3V - Volume, Velocity, Variety의 특징을 가진다. 첫 번째 특징인 Volume은 데이터의 크기를 의미한다. 기존의 데이터는 CD나 USB에 담을 수 있는 MB, GB 단위에 머물렀지만, 빅데이터는 그 규모를 넘어서서 TB에서 PB까지 이르는 크기를 가지기도 한다. 두 번째 특징은 Velocity로 빠르게 생성되고, 실시간으로 분석 및 저장할 수 있는 속도를 의미한다. 마지막 세 번째 특징은 Variety로 다양성을 의미한다. 데이터는 정형화의 정도에 따라 정형, 반정형, 비정형 데이터로 구분이 가능한데, row와 column으로 구성되어 테이블 형태를 갖는 관계형 데이터베이스, 스프레드시트, csv가 정형 데이터이며, 고정된 필드로 저장되지는 않지만, html, xml, json과 같은 메타데이터를 포함하는 데이터를 반정형 데이터, 그리고 형식이 정해져 있지 않은 사진, 영상, 글, 음성과 같은 데이터를 비정형 데이터라고 부른다. 정형 데이터만을 다루는 기존의 통계학과는 다르게, 빅데이터는 정형데이터 뿐만 아니라, 반정형, 비정형 데이터를 모두 포함한다.

2. 빅데이터가 어디에 활용되고 있는가?

빅데이터는 대량의 고객 데이터를 보유하고 있는 글로벌 기업에 의해 활용되고 있다. 아마존은 고객들의 구매 도서 구매 내역을 분석하여 고객의 취향과 근접한 신간 도서를 추천하는 차별화된 추천 시스템을 적용하고 있고, 볼보와 도요타는 차량에 부착된 정보수집장치(RFID)의 센서 정보를 기반으로 운전자의 위급 상황을 판단하여 긴급 서비스를 제공하고 있으며, 차량 결함 문제를 미리 예측해 서비스의 질을 높이고 있다. 세계 최대 인터넷 검색 엔진 서비스 업체인 구글은 감기 증상과 관련된 키워드가 많이 검색된 지역을 독감 유행 가능 지역으로 예측하는 감기지도(Flu-map) 서비스를 시행하였다. 그러나 2012년 미국에서 실제 독감 발생율보다 2배 높게 예측하면서 해당 서비스를 중단한 바 있다.

기업 뿐만 아니라 정부 및 의료기관에서도 빅데이터를 적극 활용하는 추세이다. 삼성의료원은 환자들의 정보를 통합 분석하여 자살 가능성이 높은 환자를 미리 예측하는 자살예보서비스를 선보였고, 질병관리본부는 국립중앙인체자원은행과 전국 17개 병원으로 구성된 한국인체자원은행네트워크(Korea Biobank Network, KBN)을 구성하여 36만명의 인체자원정보로 질병 지표를 만들었고,

이를 질병을 조기 진단하는 데 활용하고 있다. 또한 미국의 샌프란시스코 경찰청은 과거 8년 간의 범죄 발생한 지역 및 유형 데이터를 분석하여 후속 범죄가 발생한 가능성이 높은 지역과 시각을 예측하는 예보시스템을 만들어 안전한 지역 사회를 조성하는 데 빅데이터를 활용하고 있다.

3. 자료를 모아 보고서를 작성하려고 한다. 진행 절차를 설명하시오.

가장 먼저 선행되어야 할 것은 연구 문제의 설정이다. 연구 문제는 기존에 없던 새로운 문제여야 하며, 경험적 검증이 가능해야 하고, 해결 가능한 문제여야 한다. 연구 문제를 설정하기 위해서는 주제와 관련된 책, 기사, 논문 등을 종합적으로 살펴보는 심도 있는 문헌 연구가 필요하다. 문헌 연구를 통해 기존에 발견된 사실, 연구 설계 방법, 측정 방법, 통계 분석 기법 등을 참고할 수 있다.

다음 단계는 조사 설계이다. 조사 설계 단계에서는 우선 연구 문제를 실증적으로 검증할 수 있는 연구 가설을 세우고, 이를 통계적 검증을 위한 변수로 구분 짓는다. 변수의 종류로는 연구자의 주된 관심이 되는 종속변수, 종속변수에 영향을 미치는 독립변수, 그리고 종속변수에 영향을 줄 수 있기에 실험 중 통제되어야 하는 외생 변수가 있다. 연구 가설 및 변수 설정이 완료되었다면 구체적인 조사 방법, 자료 수집 방법, 그리고 자료 분석 기법 등을 결정하고, 결정된 사항에 따라 예산 및 조사 일정을 계획한다. 마지막으로 설계된 조사 방법을 신뢰성, 타당성, 일반화 가능성 등을 기준으로 평가함으로 조사 설계의 단계를 마무리 짓는다.

다음은 자료 수집 단계이다. 자료 수집은 연구에 직간접적으로 필요한 정보를 수집하는 단계로 좋은 연구 결과를 얻기 위한 중요한 부분이다. 자료는 수집 방법에 따라 조사자가 직접 수집하는 1차자료와 기관에서 제공하는 2차자료로 구분된다. 공공데이터포털, 통계청, 한국은행 사이트 등을 이용하면 정부가 보유한 각종 데이터를 손쉽게 접근하여 이용이 가능하다.

자료 수집을 마쳤다면, 다음으로는 수집된 자료를 분석하고 해석해야 한다. 이 단계에서는 구체적인 통계 분석이 시행되는데, 표본 자료를 이용하여 모집단의 모수를 예측하는 추정(estimation), 그리고 기존의 주장이나 상식을 귀무가설로 설정하고, 입증하고자 하는 연구가설을 대립가설로 설정하여 귀무가설의 기각 여부를 결정하는 방법인 가설검정(test of hypothesis) 등의 통계적 추론 방법이 존재한다.

마지막 단계는 보고서의 작성이다. 이 단계에서는 연구 주제의 설정부터 연구 설계 및 연구 결과에 이르기까지 전체적인 내용을 요약 및 정리한다. 일반적으로 서론, 본론, 결론의 구성을 따르며, 보고서의 목적에 따라 세부적인 항목은 달라질 수 있다. 연구 성과를 대변하는 최종 결과물이기에 문체의 일관성, 사실 관계, 맞춤법 등의 세밀한 검토가 필요하다.