Dynamic Tomorrow Ajou
Division of Information and Computer Engineering
Department of Software Convergence Technology

Big Data 분석으로 정확한 예측이 가능한가?

'과학사'와 '전자공학 운영체제' 강의 페어링

전자공학과 황용하 200921573, 정재영 교수님 지도

배경

2016 미국 대선에서 Hilary Clinton이 우세할 것이라는 여러 기관들의 예측을 뒤엎고 Donald Trump가 당선 되었다. 미국의 주류 언론사들과 여러 통계 기관들조차 Hilary Clinton의 승률을 90% 이상으로 분석했음에도 불구하고 이런 현상이 발생한 이유가 무엇일까?

Data를 통한 분석은 정치 및 사회, 경제, 문화, 과학기술과 의료 등 이미 우리 삶에서 다방면으로 활용되고 있다. 이러한 분석 결과를 이용하여 단순하게는 기상예보부터, 맞춤형 광고와 번역 서비스, 투표 결과 예측까지도 수행한다. 하지만 중요한 판단의 순간에 잘못된 분석으로 인해 틀린 결과를 얻게 되는 문제가 발생하고 있다. 본 연구에서는 이러한 문제의 원인에 대해 분석하고, 개선 방안에 대해 알아보고자 한다.

과학혁명의 구조

<패러다임의 전환>

과학의 역사란 일관적으로 연구 성과가 축적되고 진보하는 논리 축적주의의 역사가 아니다. 문제가 있을 때마다 새로운 이론이 등장해서 패러다임이 교대한다. 패러다임 이론은, 지금까지 연속적으로만 여겼던 개인과 사회의 역사를 비연속적인 다양한 문화가 병렬적으로 묶어서 이루어진 것으로 만들었다.

<과학혁명>

어느 한 과학이 체계를 잡기 전 단계에서 기득권을 잡은 과학이 정상과학이 된다. 그런데, 정상과학으로 해결할 수 없는 문제가 생기면, 그것을 극복하고자 하는 새로운 과학이 등장한다. 즉, [정상과학 - 위기 - 새로운 과학]의 구조가 성립되는 것이다. 그리고 새로운 이론도 영원히 완벽하지 않는 한 또 다른 위기에 의해 새로운 과학으로 변하게 된다.

기존 과학→패러다임 정립→정상과학→변칙 현상 발생→위기→과학혁명→경쟁적 패러다임 등장→새 정상과학

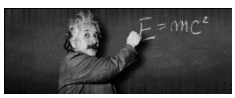
과학사

<뉴턴: 고전역학>

$$F = ma$$

물체에 작용하는 힘과 운동의 관계를 설명하는 물리학 법칙. 일상에서의 물체의 운동을 설명할 수 있는 이론이다.

<아인슈타인: 상대성이론>



뉴턴의 고전역학으로는 설명할 수 없는 광속에 얽힌 문제를 맥스웰 방정식에서 얻은 단서를 이용하여 정립한 이론. 광속에 준하는 빠른 속도에서 뉴턴 역학을 사용했을 때 발생하는 오류를 수정하였다.

상대성이론은 고전 역학의 패러다임 전환을 불러왔지만 고전 역학 자체를 부정하지 않았다.

Big data와 통계학

통계학: 데이터를 모아 분석하여 가장 올바르고 빠른 답을 제시해준다고 알려져 있는 학문

대수의 법칙: 큰 규모의 샘플 추정치가 실제 모집단 특징에 가까울 가능성이 높다. 표본의 세분화와 타게팅이 중요하다.

통계에서 사용하는 임의화의 3가지 벽(한계)

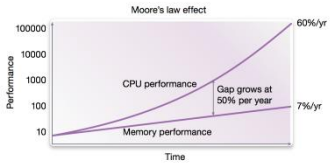
1.현실의 벽: 절대적인 표본의 수 제한

2.윤리의 벽: 도덕적인 가치의 문제

3.감정의 벽: 인간의 감성 반영의 문제

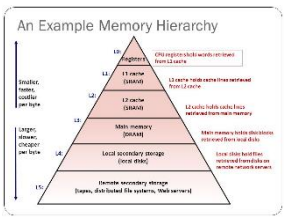
전자공학 운영체제

<Moore's Law>



반도체 직접회로의 성능이 18개월마다 2배로 증가한다는 법칙이다. 하지만 CPU와 memory의 발전속도가 달라지면서 그 성능의 gap이 커졌다.

<Memory Hierarchy>



Cpu와 memory 성능 gap을 극복하기 위해, 메모리 계층구조를 사용하였다. 빠른 memory와 느린 memory를 적절히 조합하여 performance optimization 값을 찾기 위해 노력하였다.

<Locality>

Cpu와 memory의 속도 차를 극복하기 위한 algorithm.

시간 지역성(Temporal Locality): 한 번 접근한 메모리 공간은 다시 접근할 가능성이 높다.

공간 지역성(Spatial Locality): 한 번 접근한 메모리 주변공간은 연속해서 접근할 가능성이 크다.

결론

위에서 언급한 내용들은 모든 현상을 하나의 방법으로 해결하려 하지 않았다는 점에서 유사하다.

과학사에서는 뉴턴의 고전역학으로 설명하지 못했던 광속에 얽힌 문제는 특정 조건에서 아인슈타인의 상대성 이론으로 증명해낼 수 있었으며, 이는 고전역학 자체의 부정이 아닌 하나의 패러다임의 전환이었다. 운영체제 에서는 Cpu와 memory의 발전속도 차이로 인해 성능의 문제가 생겼지만 memory hierarchy를 구성하고 locality의 개념을 도입하여 속도에 대한 문제를 해결 할 수 있었다.

Big data 분석도 마찬가지다. Data에 의한 분석은 단순히 표면에 드러난 값들에 대한 통계적 수치를 보여줄 뿐이다. 숨은 의도까지 파악하기 위해선, 단순히 데이터 분석 이외에 다른 방법을 사용해서 현상을 뒷받침하기 위한 근거로 사용해야 할 것이다.

참고문헌

토마스 쿤,『과학 혁명의 구조』, 까치글방, 2013

니시우치 히로무,『빅데이터를 지배하는 통계의 힘』, 비전코리아, 2013

Abraham Silberschartz, Peter B. Galvin, Greg Gagne,『Operating System Concepts』, WILEY, 2013