빅 데이터 시각화 방법 및 시각화 프로세스

오재훈* · 김두진* · 김종달*

1. 서 론

2000년대 초에 웹의 급속한 성장으로 데이터가 폭발적으로 증가하면서 데이터 저장 장치와 CPU 기술이 생성되는 데이타를 분석하지 못하는 상황이 발생하였으며, 이를 데이터 확장성 위기(data scalability crisis)라고 부른다.

데이터 확장성 위기를 극복하기 위해 저장 장치 기술, CPU 기술, 빅데이터 처리 소프트웨어 기술이 발전되어 왔으며, 10여년간의 기술개발로 빅데이터를 처리하기 위한 플랫폼 기술들은 성숙기로 접어들고 있다.

빅데이터는 보는 관점에 따라 조금씩 다르게 정의하고 있다. 대표적인 정의를 살펴보면 다음과 같다.

- 가트너: 향상된 시사점(insight)와 더 나은 의사결정을 위해 사용되는 비용효율이 높고, 혁신적이며, 대용량, 고속 및 다양성의 특성을 가진 정보 자산
- 매킨지 : 일반적 데이터베이스 소프트웨어가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터
- IDC : 다양한 종류의 대규모 데이터로부터 낮은 비용으로 가치를 추출하고, 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처

하지만, 빅데이터의 장밋빛 전망에도 불구, 아직까지 많은 기업들이 빅데이터 분석에서 의미와 가치를 창출하는데 실패하고 있다. 지금까지는 데이터양에 압도되어 대용량의 다양한 데이터를 저장하고 관리하고 운영하기 위한 시스템 구축과 운영 이슈를 해결하는 데 급급했기 때문이다. 가트너에 따르면 빅데이터를 도입하는 기업 중 약 15% 만이 빅데이터에서 의미있는 정보를 찾아낼 것으로 예측하고 있다.

최근에는 선진국을 중심으로 빅데이터에서 의미있는 데이터를 찾아내기 위한 디스커버리 어낼리틱스(discovery analytics)에 대한 연구와 개발이 활발히 이루어지고 있다. 또한 기업에서 발생하는 데이터를 실시간으로 처리하고 실시간으로 새로운 인사이트를 찾고 실시간으로 이벤트를 처리하기 위한 실시간 분석이 빅데이터 분석의 화두가 되고 있다.

빅데이터 플랫폼을 활용하여 가치를 창출할 수 있는 새로운 통찰력을 찾기 위해서는 데이터 특성에 맞는 데이터 시각화가 핵심 기술로 떠오르고

* (주)넷스루 데이터마이닝 연구소

있다. 데이터의 패턴을 쉽게 인지할 수 있도록 숫자를 공간에 배치하여 보여주는 것을 데이터 시각화라고 한다. 정보를 시각화하면 정보 지도와 같은 풍경이 된다. 우리는 그 풍경을 눈으로 볼 수 있다. 정보에서 길을 잃고 헤메고 있다면, 정보 지도는 아주 유용해진다.

본 논문에서는 빅데이터 시각화와 관련된 문제들을 살펴보고 이를 극복하기 위한 기존 방법들을 소개한다. 또한, 새로운 방법을 제시한다.

2. 빅 데이터 분석의 특성

빅데이터의 정의는 조금씩 다르지만 빅데이터의 특성으로는 가트너가 정의한 3V를 따르고 있다.

특성	설명
Volume (양)	물리적인 크기와 개념적인 범위까지 데이타 규모가 크다.
Velocity (속도)	데이타의 생산속도가 매우 빠르다.
Variety (다양성)	정형 데이타 + 사진, 동영상, 텍스트 등의 비정형 데이터

최근에는 4번째 V 에 대한 정의들이 추가되고 있으며, 정확도(Veracity), 변동성(Variability), 가치(Value) 등이 있다. 기업이 보유한 빅데이터의 양이 가치 추출이 가능할 만한 임계치에 도달하여 가치 추출이 본격화되고 있어, 특히 가치(Value)가 주목받고 있는 상황이다.

가치가 가장 주목을 받고 있는 이유를 4가지 관점에서 살펴보자. 첫째, 빅데이터의 활용성과와 가치창출 여부가 기업의 미래 생존을 좌우할 핵심 요소로 부상하고 있다. 비즈니스 환경에서 변화가 맹렬한 속도로 진행되고 있으며 패러다임의 변화 주기가 점점 더 짧아지고 있다. 기업이 이 변화의 흐름을 읽지 못한다면 변화를 인식하기도 전에 기회를 상실할 수 있다. 빅데이터 분석은 기업 내외부에서 생산되는 방대한 양의 데이터를 기반으로 고객과 비즈니스 환경의 변화의 흐름을 쉽고 빠르게 찾아낼 수 있는 인프라를 제공할 수 있다.

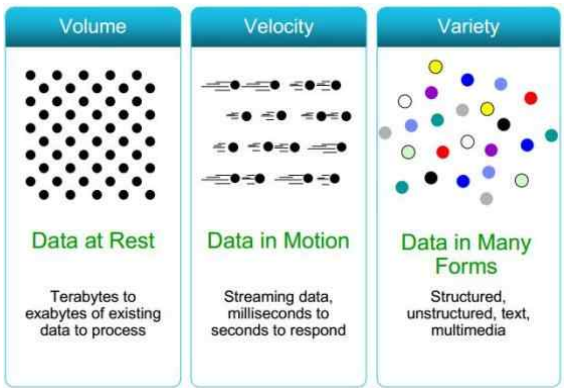

그림 1. 빅데이터의 3V 특성

둘째, 기업 의사결정의 속도와 정확성을 향상시킬 수 있다. 빅데이터 분석을 통하여 이전에는 알 수 없었던 사실들을 탐색할 수 있으며, 분석 결과를 이해함으로써 사업의 현재 상태를 좀 더 잘 이해할 수 있다.

세 번째 신속한 혁신 유발이 가능하다. 빅데이터 분석을 통해서 더 깊은 통찰력을 더 빨리 얻을 수 있으며, 새로운 기회를 창출할 수 있고, 의사결정의 정확도를 강화하고 시기 적절하게 필요한 정보를 제공할 수 있다. 이를 활용하여 기업의 혁신 속도를 가속화할 수 있다.

넷째, 고객 경험 향상이다. 고객의 행동은 지속적으로 변화하기 때문에 고객의 행동 변화를 감지하고 대처하는 것은 매우 어려운 일이다. 빅데이터 분석을 이용하면 고객의 행동변화패턴을 분석하고 예측할 수 있으며, 분석 결과를 이용하여 고

객경험을 향상시킬 수 있다.

빅데이터 시대에는 고객으로부터 발생한 빅 데이터 속에 숨어 있는 고객행동 패턴을 지속적으로 분석하여 새로운 통찰력을 찾지 못하는 기업은 점점 더 경쟁력을 상실할 수 밖에 없는 위기의 시대가 도래할 것으로 보인다.

3. 빅 데이터 분석시각화

빅데이터의 4번째 특성인 가치(Value)의 중요성에 대한 인식이 중요해지기 시작해지면서, 빅데이터 시각화가 빅데이터 분석 기술에 못지 않게 중요하게 다루어지고 시작하고 있다. 빅데이터 시각화는 빅데이터로부터 통찰력을 얻고 가치를 창출하기 위한 핵심 도구이다. 빅데이터로부터 가치를 창출하기 위해서는 데이터 특성을 고려한 다양한 종류의 시각화 도구가 제공되어야 한다.

우수한 시각화란 짧은 시간안에 작은 공간에 최소의 자료로 많은 아이디어를 주는 것이다. 좋은 시각화는 다음과 같이 사용자들을 돕는다.

- 동향이나 패턴을 즉시 식별할 수 있다.
- 새로운 아이디어를 발견할 수 있다.
- 시각화에 담긴 메시지를 명확하게 이해할 수 있다.
- 비즈니스의 질문에 즉시 답을 줄 수 있다.
- 문제영역을 효율적으로 식별할 수 있다.
- 성공으로 이끄는 의사결정을 하도록 한다.

빅데이터의 3V 특성으로 인해 빅 데이터 시각화는 일반적인 데이터 시각화와는 다른 문제를 유발시킨다. 빅 데이터를 효과적으로 시각화하기 위해서는 다음과 같은 문제를 해결해야 한다.

1. 빅 데이터 분석 결과는 방대한 양의 데이터 포인트를 포함하고 있다.
2. 빅 데이터는 다양한 형식의 데이터를 포함하기 때문에 다양한 형식의 시각화 도구를 제공해야 한다.
3. 빅 데이터는 방대한 양의 데이터 포인트를 시각화해야 하기 때문에 시각화에 소요되는 시간을 최소화해야 한다.

본 절에서는 데이터 포인트를 적절하게 축소하기 위한 방법을 대해서 다룬다.

3.1 방대한 데이터 포인트

기존의 시각화 기술로는 빅 데이터 분석 결과로 발생하는 방대한 데이터 포인트를 효과적으로 표현할 수 없기 때문에 새로운 시각화 기법과 기술 개발이 필요하다.

기존의 시각화 도구들은 다음과 같은 이유로 방대한 양의 데이터 포인트를 다룰 수 없다.

- 기존 시각화에서는 데이터 포인트가 수백~수천 단위였다면, 빅데이터는 수십~수백만에 이를 수도 있다.
- 데이터 포인트가 많아질수록 시각화에 소요되는 시간이 대폭 증가하게 된다.
- 수십~수백만의 데이터를 표시해도 사용자가 이해하기 어려울 수 있다.

수백만개의 자료를 차트로 표시한다고 가정해보자. 모든 자료를 그대로 차트에 표시한다면, 차트에 모든 영역에 데이터가 표시될 가능성이 높기 때문에 시각화를 통해서 얻을 수 있는 정보가 거의 없을 것이다. 이러한 경우에 가장 많이 사용되는 방법이 데이터 비닝(data binning)과 박스 플롯(box plot)이다.

3.2 데이터 비닝(Data Binning)

데이터 비닝은 데이터 포인트가 많을 경우에 데이터 포인트를 줄임으로써 데이터의 가시성을 높이기 위한 방법이다. 데이터 구간을 여러 개의 묶음(bin)으로 나누고 하나의 묶음에 여러 데이터를 넣는다. 구간을 대표할 수 있는 값을 묶음의 대표값으로 표시한다. 동일 너비 묶음법과 동일 개수 묶음법을 많이 사용한다.

동일 너비 묶음법(Equal width binning) : 데이터를 k 개의 구간으로 나누고 싶을 때, 최대값과 최소값 사이를 동일한 간격으로 나눈다. 각 구간의 너비는 (max-min)/k 로 정한다.

동일 개수 묶음법(Equal Frequency Binning) : 각 구간에 데이터가 거의 동일하게 포함되도록 데이터를 k 개의 그룹으로 묶는다. 데이터포인트 수가 n 이라면, 각 빈에는 $\lfloor \frac{n}{k} \rfloor$ 개의 데이터를 포함하도록 한다.

다음과 같은 자료가 있다고 생각해 보자.

자료 : {1,1,4,5,6,6,7,7,8,8,9,11,12}

데이터 빈 개수를 3개로 고정하고, 동일 너비 묶음법과 동일 개수 묶음법이 어떤 차이를 보이는지 살펴보자.

먼저, 동일 너비 묶음법으로 데이터를 묶으면 결과는 다음과 같다.

빈1 : 1,1,4
빈2 : 5,6,6,7,7,8,8
빈3 : 9,11,12

이를 표로 나타내면 “표.동일 너비 묶음법“ 과 같으며, 그림으로 나타내면 ”그림.동일 너비 묶음법”과 같다.

데이터 빈	빈도
4보다 작거나 같다.	3
4보다 크고 8보다 작거나 같다.	7
8보다 크다.	3

표 4. 동일 너비 묶음법 (k=3)

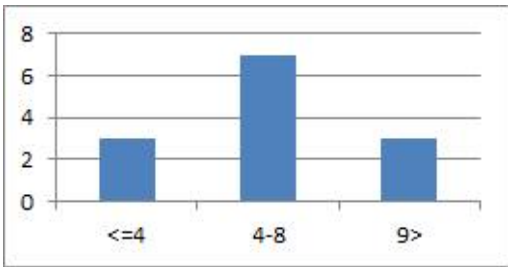

그림 2. 동일 너비 묶음법 (k=3)

동일 개수 묶음법으로 데이터를 묶으면 $\lfloor \frac{13}{3} \rfloor = 4$ 이므로, 각 빈에 4개씩의 데이터가 들어가야 한다.

빈1 : 1,1,4,5
빈2 : 6,6,7,7
빈3 : 8,8,9,11,12

데이터 빈	빈도
5보다 작거나 같다.	4
5보다 크고 8보다 작다	4
8보다 크거나 같다.	5

표 5. 동일 개수 묶음법 (k=3)

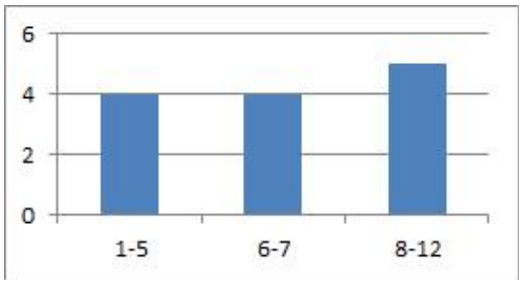

그림 3. 동일 개수 묶음법 (k=3)

데이터 빈의 숫자를 결정하는 것도 해결해야 할 문제다. 해상도와 오류 사이의 트레이드 오프 관계가 발생하기 때문이다. 빈 사이즈가 작아지면 해상도가 높아지고, 각각의 빈에 포함된 데이터 수가 줄어들면서 통계적인 에러가 증가한다.

3.3 박스 플롯(Box Plot)

데이터 비닝과 같은 방법으로 데이터를 묶게 되면, 빈 안에 여러 개의 값들이 들어 있기 때문에 빈 안의 데이터 분포를 파악할 필요가 있다. 박스 플롯(Box Plot)은 같은 빈(bin)에 있는 값의 분포를 5개의 값으로 표시하여 빈에 속한 데이터를 시각화하는 방법이다.

- 최대 : 해당 구간에서 발생했던 최대값을 표시한다.
- 4/3 지점 : 최소값에서 3/4 위치에 있는 값을 표시한다.
- 중간값: 최소값에서 1/2 위치에 있는 값을 표시한다.
- 1/4 지점 : 최소값에서 1/4 위치에 있는 값을 표시한다.
- 최소 : 해당 구간에서 발생했던 최소값을 표시한다.

1/4지점에서 3/4 지점 사이를 IQR(Inter Quartile Range)라고 한다. 어떤 데이터를 분석한 결과 분포가 “표. 구간별 데이터 분포” 와 같다고 하자. 이를 박스 플롯 형식으로 그래프로 표시하면 “그림. 박스 플롯 예제”과 같다.

	구간1	구간2
최대	1000	951
4/3 지점	950	850
중간값	940	849
1/4 지점	850	811
최소	600	760

표 6. 구간별 데이터 분포

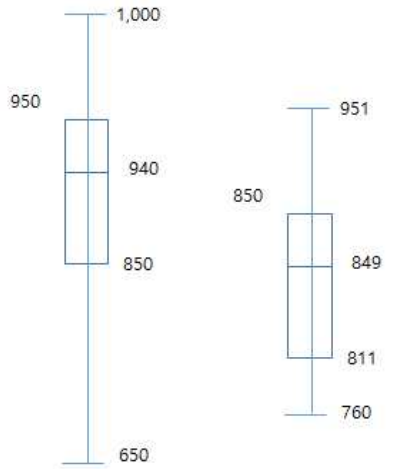

그림 4. 박스 플롯 예제

3.4 시각화 연산

빅 데이터 시각화에서는 수백만개 이상의 데이터포인트를 시각화하기 위해서 근처에 위치한 데이터를 묶는 그루핑 방법이 가장 많이 사용된다. 데이터 비닝과 같은 그루핑 방법을 사용하여 데이터를 그루핑한 경우에 전체적인 데이터를 파악하고 난 후에 세부 데이터를 조회하기 위한 방법들이 제공되어야 한다.

세부 데이터를 조회하기 위해 많이 사용되는 방법에는 확대(zoom in)/축소(zoom out) 방법과 차원 축소(drill-down)/차원 확대(roll-up) 방법이 있다.

확대/축소 방법은 데이터의 차원을 그대로 유지하면서 특정 영역을 자세히 보기 위해서 방법이다. “그림. 확대/축소 연산” 은 문서에 나타나는 단어의 빈도를 점으로 표시한 것이다. 같이 나타나는 단어들끼리 뭉쳐있도록 시각화하였다. 위쪽의 둥근 원안에 있는 특정 점을 클릭하면, 해당 영역에 들어 있는 단어들이 확대되면서 아래쪽 그림과 같이 확대된다. 확대/축소 방법은 덩어리로 요약된 데이터를 상세화하는 방법으로, 사용자가 선택한 데이터에만 집중할 수 있도록 도와준다.

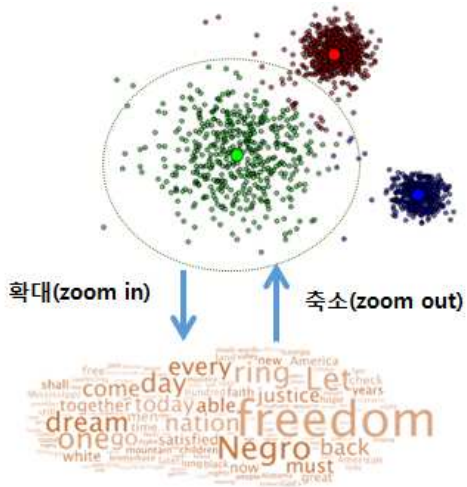

그림 5. 확대/축소 연산

차원 축소(drill-down)/차원 확대(roll-up) 방법은 다차원 데이터를 조회하는 경우에 특정 차원만으로 데이터를 조회하기 위한 방법이다. 차원을 축소하는 것을 드릴다운(drill-down)이라 하고, 원래 차원으로 돌아가는 방법을 롤업(roll-up)이라고 한다.

4. 데이터 시각화 프로세스

어떤 시각화 방법을 택하느냐에 따라 시각화의 효율성과 효과성이 완전히 달라지기 때문에 시각화 방법을 선택하는 것은 빅 데이터 분석에 있어서 아주 중요한 고려 요소이다. 하지만, 데이터의 특성에 맞는 데이터 시각화 방법을 선택하는 것은 쉽지 않다.

본 절에서는 필드 특성에 따라 분석 방법과 시각화 방법을 가이드하는 시각화 처리 절차를 제시한다. 시각화 처리 단계는 데이터 탐색, 분석도구 제시, 샘플 데이터 분석, 데이터 분석, 시각화 단계를 거친다.

데이터 탐색단계: 이 단계에서 사용자는 분석할 데이터를 탐색한다. 사용자의 탐색 단계를 지원하기 위해서 사용자가 선택한 데이터의 특성을 요약하여 제시한다. 데이터 요약기가 사용자가 선택한 데이터에서 데이터 요약정보를 추출한다. 이때, 문제가 되는 부분은 데이터를 구성하는 컬럼의 타입과 도메인을 자동으로 인식하는 것이다.

특성	설명
레코드 수	선택한 데이터의 레코드 수
컬럼 수	선택한 데이터의 컬럼수
크기	선택한 데이터의 크기(byte 단위)
컬럼 목록	컬럼의 타입과 도메인 컬럼의 카디날러티(Cardinality)

표 7. 데이터 요약 정보

컬럼 타입은 컬럼 데이터의 형식으로 문자열, 정수, 실수 등과 같은 것이다. 컬럼 도메인은 나이, 키, 몸무게와 같은 의미를 가진다. 컬럼 데이터 형식이 숫자라 하더라도 도메인은 달라질 수 있다. 예를 들어, 나이와 키 필드는 모두 정수형이지만 나이가 가질 수 있는 값과 키가 가질 수 있는 값의 범위가 다르므로 둘은 도메인이 다르다.

컬럼의 데이터 형식을 식별하는 것은 어렵지 않지만, 컬럼의 도메인을 결정하는 것은 아주 어려운 일이다. 도메인 식별기는 데이터 필드의 패턴을 해석하여 적합한 필드의 도메인을 자동으로 인식하는 기능을 담당한다. 필드 도메인을 자동으로 인식하기 위해서는 필드 도메인 데이터를 구축하여야 한다.

컬럼	설명
도메인 이름	도메인의 이름 (예, 나이, 몸무게)
도메인 형식	문자열, 정수, 실수
최대값	해당 도메인이 가질 수 있는 최대값
최소값	해당 도메인이 가질 수 있는 최소값
데이터 포맷	정규표현식을 이용하여 정의한 데이터 형식

표 8. 도메인 자동 인식을 위한 테이블

분석도구 제시 단계 : 데이터 요약이 완료되면, 요약된 데이터 정보를 이용하여 분석도구를 추천한다. 분석도구 추천시에는 도메인 목록을 이용하여 적합한 분석도구를 선택하고 사용자에게 제시한다.

샘플 데이터 분석 단계 : 사용자는 제시된 분석도구 중 하나를 고르거나 자신이 원하는 분석 도구를 선택한다. 사용자가 분석 도구를 선택하면 선택한 데이터에서 샘플 데이터를 분석하여 시각화 샘플을 제공한다. 사용자는 데이터분석 방법을 바꿔가면서 위해 샘플 데이터 분석 단계를 반복하여 원하는 데이터 분석 방법을 찾는다.

데이터 분석 및 시각화 단계 : 사용자가 원하는 데이터 분석 방법을 확정하여 데이터 분석을 실행시킨다. 분석 결과는 분석 방법과 연결된 시각화 방법을 이용하여 제시된다. 분석이 종료된 후에라도 분석 결과를 다양한 형태의 시각화 방법을 변형하면서 분석 결과를 조회할 수 있다.

5. 결론

본 논문에서는 빅데이터 분석 시각화시에 발생할 수 있는 문제들을 살펴보고 각 문제들을 해결하기 위한 방법을 조사 정리하였다. 빅 데이터 분석 결과 데이터의 데이터 포인트가 수백만에 이를 때, 데이터 포인트를 그루핑하기 위한 여러 가지 데이터 비닝 방법을 살펴보았고, 동일한 빈에 포함된 데이터의 분포 정보를 제시할 수 있는 박스플롯 방법도 정리해 보았다.

또한 빅 데이터 시각화를 위해 데이터의 특성 파악에서부터 최종 시각화에 이르는 빅 데이터 분석 프로세스를 살펴보았고, 각각의 단계에서 사용자가 어려움을 겪을 수 있는 어려움을 가이드할 수 있는 방법을 제시하였다.

참 고 문 헌

[1] William Cleveland, Robert McGill, Graphical Perception: Theory Experimentation, and Application to the Development of Graphical Methods, Journal of the American Association, Vol. 79, No. 387, 1984

[1] Robert Amar, James Eagan, and John Stasko, Low Level Components of Analytic Activity in Information Visualization, Information Interfaces Reasearch Group

[2] Zhao Kaidi, Data Visualization, Ph.D. Thesis, School of Computing

[3] Justin Talbot, John Gerth, and Pat Hanrahan, An Empirical Model of Slope Ratio Comparisons

[4] Nicholas Kong and Maneesh Agrawala, Graphical Overlays : Using Layered Elements to Aid Chart Reading

[5] Data Visualization: Making Big Data Approachable and Valuable, White Paper

A Survey on Information Visualization: Recent Advances and Challenges

[6] SAS Data Visualization Techniques From Basics to Big Data with SAS Analytics

[7] Ben Schneiderman, The Eyes Have It: A Task by Data Type Taxonomy for Information Visualization basing on integral imaging," Opt. Commun. 277, 40-49 (2007)

[20] Y. Piao, D.-H. Shin and E.-S. Kim "Robust image encryption by combined use of integral imaging and pixel scrambling techniques," Optics and Lasers in Engineering 47(11), pp. 1273-1281, 2009

오 재 훈

• 1995년 포항공과대학교 컴퓨터공학과 (공학석사)
• 1995년-2000년 포항공과 자동화센터 전임연구원
• 2000년-2008년 ㈜넷스루 개발팀장
• 2008년-현재 ㈜넷스루 데이터마이닝 연구소장
• 관심분야: 클릭스트림 분석, 빅 데이터 분석, 데이터 마이닝, 개인화, 추천

김 종 달

• 2002년 포항공과대학교 컴퓨터공학과(공학석사)
• 2000년-현재 ㈜넷스루 데이터마이닝 연구소
• 관심분야: 빅데이터 분석, 데이터 마이닝, 클릭 스트림 데이터 분석

김 종 달

• 1995년 포항공과대학교 컴퓨터공학과(공학석사)
• 1994년-2000년 포스데이타
• 2000년-현재 ㈜넷스루 데이터마이닝 연구소
• 관심분야: 빅데이터 분석, 데이터 마이닝, 클릭 스트림 데이터 분석, 개인화 추천