데이터관리론 Mid Term Exam Assignment

201823869 조성우

GRAPH DATABASE

GRAPH DATABASE 기술 개요

이번 기술의 주제로 GRAPH DATABASE를 선정했습니다. 현재 데이터베이스 기술의 생태계는 SQL을 중심으로 관계형 데이터베이스가 주류를 이루고 있지만, 비정형 데이터 수집, 가공 관련 기술의 약진으로 인한 빅데이터 이슈를 충족시키기 위해 NoSQL, 즉 비관계형 데이터베이스 기술에 대한 요구와 수요가 급증하고 있는 추세입니다. 저는 이 중 강력하고 직관적인 시각화 기술과 네트워크타입의 관계를 효율적으로 다룰 수 있는 GRAPH DATABASE를 다루고자 합니다.

그래프 데이터베이스를 관통하는 단어는 시각화입니다. 그래프 데이터베이스는 이름에서부터 알 수 있듯 무엇보다도 강력한 시각화 기술을 자랑합니다. 다만 이 시각화 기술이라 함은 기존의 관계형 데이터베이스 에서처럼 테이블로 저장돼 있던 데이터를 사후적으로 가공하여 시각화 정보를 얻는 것이 아니라 수학적 그래프 이론에 토대를 두고 데이터 자체를 대상 객체를 표현하는 점(Node/Vertex)과 관계를 표현하는 선(Edge)의 형태로 시각화 된 타입으로 저장하여 객체들의 공통적 속성에 따라 형성되는 Group(Label)을 통해 관계를 따라 특정 패턴을 추적하는 것 입니다.[1]

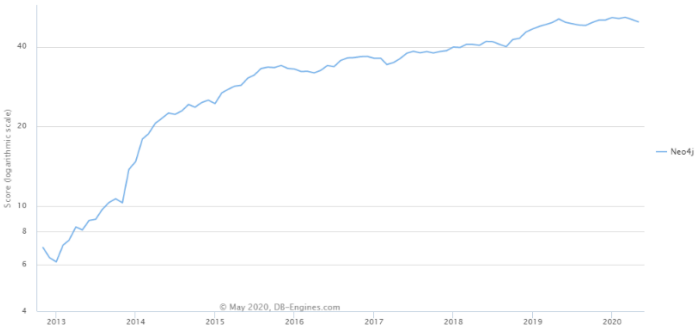
위에서 언급한 관계형 데이터베이스와 그래프데이터베이스의 구조적 차이는 데이터 저장 방법부터 분석까지 도달하는 시간과 노력, 효율성의 차이를 발생시키는데, 이는 결국 시각화 분석의 자동화에 따른 **실시간성**의 이슈로 직결됩니다. 따라서 Graph DB는 **대규모 데이터 제공 주체**의 **비정형적 데이터**를 관계적으로 분석하고 이를 통해 **집단적 패턴을 유추하고 즉각적으로 직관적인 분석결과를 반환**하여 방대하고 다양한 데이터를 신속하게 분석하기를 요구하는 빅데이터 시대에 부합하다고 할 수 있습니다.

아래의 그래프[2]는 Graph database의 대표주자 Neo4j의 시장 점유율을 시계열에 따른 추세로 보여주는 Trend Chart입니다. 스마트폰의 본격적인 보급에 따른 각종 SNS의 등장과

[1] 정보통신 기획평가원(www.iitp.kr) - ict동향정보 – 그래프 데이터베이스 기술동향 및 적용사례 (비트나인 이형주 차장 외 1인)

[2] https://db-engines.com/en/ranking (DB engine live ranking)

그로인한 비정형 데이터타입의 데이터가 효율적으로 수집되기 시작한 시기로 추정되는 2013년부터 극적으로 점유율이 증가하기 시작하였고, 2020년 5월 현재 대다수의 관계형 데이터베이스 사이에서 21위에 위치하는 기염을 토해내며 Graph DB의 실효성을 입증하고 있습니다.



GRAPH DATABASE의 대표 주자 Neo4j의 점유율 변화를 보여주는 Trend Chart

본격적인 스마트폰 보급시기와 SNS의 활성화 시기와 맞물려 2013년부터 점유율이 상승한 것으로 추정된다.

21. 21. 22. Neo4j Graph 49.76 -1.05 -1.27

[Neo4j(Graph DB) TrendChart (~2020.5)]

Graph DB for Analyzing Users' Pattern using GPS

제가 해당 Graph DB기술을 도입하고자 하는 영역은 **Marketing**입니다. 이는 특정 산업군에 국한되는 주제가 아니며 B2P를 주 영업으로 하는 기업 및 산업군이라면 모두 해당 기술을 도입하여 활용할 수 있을 것이라 기대됩니다. 또한 B2B 사업도 활용하기에 따라 해당 기술을 적용시킬 수 있을 것입니다.

Analyzing Market

시장에 물건만 내놓으면 그 상품에 대한 수요가 저절로 생기며, "공급이 수요를 창출한다*(Say's Law of Markets)*[3]는 과거의 시장분석 관점과 반대로 오히려 현재의 시장은 "수요가 공급을 만든다"는 법칙에 더욱 부합하게 변해가고 있고 이에 따라 기업들의 마케팅 전략 또한 과거 매스미디어를 통해 무작위 다수에게 광고를 노출시키던 전략과 반대로 최근엔 최대한 대상군을 세분화하는 맞춤 마케팅 전략을 시행하고있습니다. 그에 대한 예로 GOOGLE은 평소 사용자의 자사 포털사이트의 검색기록, 방문페이지의 데이터를 기반으로 특정 패턴을 분석하여 관심사를 유추하여 맞춤 광고를 노출시키는 Google Ads와 더불어 유튜브 이용자의 시청영상 패턴을 분석하고 관심사를 유추

3 Steven Horwitz, Understanding Say's Law of Markets

하여 더 적절한 영상을 제시하는 맞춤형 서비스를 제공하고 있고 또한 이와 비슷한 맥락으로 대다수의 SNS 서비스는 이미 형성된 친구목록이나 전화번호부를 통해 개인의 인간관계의 네트워크를 추론하고 맞춤형 친구추천 서비스를 제공하고 있습니다.

위의 서비스들은 유의미한 성과를 거두고 있으며 그 자체로 충분히 훌륭한 전략입니다.

다만 위에 나열한 기술들의 경우 통칭하여 'Online behavioral advertising'으로 불리우는데 이름에서 알 수 있다시피 쿠키(Cookies)와 같은 정보수집 장치를 사용하여 이용자의 온라인 검색기록이나 브라우징 정보 등 [4]대개는 **온라인상의 활동만을 데이터로 수집**할 수 있기 때문에, 온라인 기반의 활동(online behavior)이 마케팅 대상의 **실제 라이프스타일, 실제 생활환경, 실제 인간관계와 부합한지에 대한 의문점**이 남아있고 **마케팅 대상의 특정 주제에 관한 관심사의 정도를 어떻게 정확히 측정할 것인**가에 관한 의문점도 존재합니다. 이러한 의문점들에 대한 아래의 답이 아직 규명되지 않은 해당 마케팅전략의 비효율을 설명할 수 있을 것이라 기대합니다.

현재의 약점을 보완하기 위해 마케팅 대상의 Mobile 기기에 탑재된 GPS를 통해 데이터를 수집하고 이 데이터를 Graph DB로 분석하여 기존에 수집,분석 했던 Online behavior에 추가적으로 Offline behavior정보를 더해 더욱 정확하고 실질적인 Personal Marketing 전략을 도모 하고자 합니다. 이는 독립적인 기술로써 존재하는 것 보다 기존에 온라인에서 수집, 분석되던 Online Behavior와 보완적으로 사용할 때 더 큰 시너지 효과를 발휘할 것으로 기대됩니다.

Why GPS?

GPS는 이용자가 실제 생활에서 가지는 위치정보에 기반하여 활동지역, 관심사, 인간관계 등 오프라인 활동에 관한 데이터를 제공합니다. 기업은 이를 기반으로 실제 생활에 반영되고 있는 더욱 실용적이고 정확한 라이프스타일에 관한 정보를 분석할 수 있을 것이고 또한 실생활에 존재하는 셀 수 없이 많은 객체와의 상호작용(특히 온라인상으로 수집할 수 없는)을 파악하고 이를 분석하여 실생활에 존재하는 매우 복잡한 네트워크에서 특정한 패턴을 찾아 낼 수 있을 것입니다.

바로 여기에서 GPS에서 위치로 표현되는 수많은 객체-객체 간, 이용자-객체 간, 이용자-이용자

---

4 온라인 맞춤형 광고(Online Behavioral Advertising)와 개인정보보호 Online Behavioral Advertising and Privacy 안정민(한림대학교) 2013.12

https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE02330914

간, 관계-관계 간의 상호작용을 통해 형성된 매우 복잡한 네트워크 타입으로 연결된 데이터를 다루기 위해 GRAPH DB의 사용이 요구됩니다.

▌Why GRAPH DB for Analyzing GPS Data?

1) 기본적으로 GPS로 수집되는 데이터는 Map을 Background로 가지고 이용자의 활동이 좌표로써 그 위에 출력되며 구성됩니다. Map을 기반으로 하는 데이터를 만약 Relational DB를 통해 관리한다면 기록하고자 하는 Event가 발생할 때 좌표로 저장하고 그것을 필요에 따라 시각화 시키는 불필요한 작업을 반복해야 하며 이는 매우 방대한 데이터를 다뤄야 하는 해당 프로젝트에 방대한 구조적 비효율을 누적시킬 것입니다.

2) 이 프로젝트에서 다룰 데이터는 관계형 데이터베이스처럼 모델링할 시 사전에 정의된 객체나 관계를 다루기 보다 객체간 상호작용으로 누적되는 관계간, 집단간의 데이터를 패턴으로 유추해내는 작업이 될 것이기 때문에 기술개요에서 이미 언급했던 것처럼 Graph DB가 적합합니다.

Technical Requirement

◆ User와 유의미한 위치정보(ex.특정 가게의 좌표)는 한 점(Vertex)으로 저장되며 User-User간 또는 User-특정위치 간 유의미한 정도의 시간을 가질 때 두 Vertex간 관계(Edge)가 기록됩니다(빈도/보낸 시간의 구간에 따라 숫자로 관계의 정도가 누적될 것입니다).

관계의 정도를 측정하려면 특정 좌표내의 체류시간을 알아야 하는데 이를 위해서 DB는 데이터를 시계열에 기반하여 기록, 분석해야 할 것입니다.

◆ 지도상 존재하는 위치정보(ex.가게)등을 가장 기본단위의 객체로 사용할 것입니다. 이를 위해 위치정보를 가장 많이 가지고 Update가 빠른 Map데이터를 사용해야 할 것입니다. 또한 활동반경 및 거주지의 고도에 따른 생활양식의 패턴도 분석하기 위해 z좌표까지 포함한의 3차원 Map을 사용하길 권장합니다.

◆ 앞서 말한 대로 Online Behavior과 비교한 실제 생활의 관계를 알기 위해 기존의 Online Behavior을 관리하는 DB의 Data와 연동시켜야 합니다.

◆ Privacy이슈와 관련하여 User에게 GPS정보 수집동의를 받을 때 민감지역을 등록하도록 하고, 민감지역 주변부터의 Data는 수집되지 않도록 배제시키는 방법이 요구됩니다.