데이터 과학 입문 10장

-소셜네트워크와 데이터 저널리즘-

2015.07.11

10장 소셜네트워크와 데이터 저널리즘

1. 소셜 네트워크의 기본 이론

2. 모닝사이드 애널리틱스의 네트워크 데이터 시각화 예제

3데이터 저널리즘의 한 형태에 대해 알아보기

사례-속성 데이터 대 소셜네트워크 데이터

사례-속성 데이터란?

- 모형에 다양한 사례를 적용시켜 보는 것.
- 여기서 사례는 사람 혹은 사건이며, 속성은 나이. 운영시스템 검색 기록 등
- 쉬운 방법이나 문제의 핵심 포인트를 놓칠 수 있음

-> 소셜네트워크 분석을 통해 개인만이 아니라 사람들 간의 관계를 기반으로 하는 접근법 개발

네트워크 분석이 사례-속성 분석보다 우수한가에 대한 예

미국이 **아프가니스탄의 미래가 어떻게 될지 예측하기 위해** 시민들을 대상으로 여론 조사를 함.

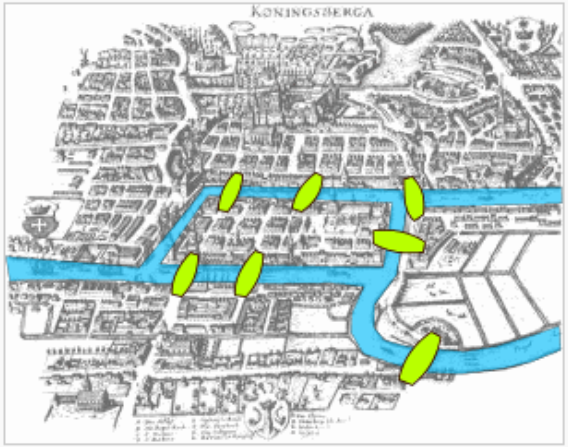
-> 어떤 일이 벌어질 것인가는 개개인이 생각하는 것의 단순한 함수가 아님

-> 누가 권력을 쥐고 있고, 그들이 어떤 생각을 가지고 있는가의 문제

소셜네트워크 분석

소셜 네트워크 분석은 두 분야에서 출발

- 오일러가 7개의 쾨니히스베르그 다리 문제를 풀었던 것에서 출발한 그래프 이론
- 사회관계나 인간관계를 측정하는 이론 및 기술을 의미하는 제이콥 모레노의 계량 사회학



오일러 시절 쾨니히스베르크의 지도. 프레겔 강과 일곱 다리는 색으로 구분하였음.

소셜네트워크 초기 시작 아이디어

- 사람들의 행동은 그들의 속성과 연관되어 있지만, 그 사람들을 정말로 이해하기 위해서는 사람들로 하여금 무언가를 할 수 있도록 만드는 네트워크를 검토할 필요가 있다는 것

소셜 네트워크의 용어들

용어	설명
행위자, 노드	네트워크의 기본 단위 (사람, 웹사이트, 생각하는 어떤 것)
관계적 유대 또는 연결선	행위자들 사이의 관계 (어떤 사람을 좋아하거나 친구가 된 경우에는 연결선으로 표현)
관계	행위자들 사이에 어떤 관계적 유대를 갖는 방식
소셜네트워크	어떤 행위자들과 관계들의 집합체
다이야드(dyad)	연결된 한 쌍의 행위자
트라이애드(triad)	연결된 세명의 행위자는 A-B 간의 유대 관계, B-C간의 유대 관계 → A-C사이에도 연결선 존재
하위 네트워크 또는 하위그룹	관계적 유대를 갖는 전체 행위자 집합의 부분집합

소셜 네트워크의 용어들

이분 그래프의 연결

- 서로 분리된 두 개의 객체집단 사이에서만 존재
- 사람들과 회사, 사람들과 그들이 관심가질 만한 대상들의 집단 사이의 연결

자아 네트워크

- 어떤 한 사람을 둘러싸고 있는 네트워크의 일부분으로 구성
- 페이스북에서 어떤 상황에서 서로 알 수도 있는 내 친구의 하위 네트워크
- 높은 사회경제적 지위를 가진 사람일수록 더 복잡한 자아 네트워크를 가지고 있다는 연구 결과가 있음

이분 그래프: 그래프의 정점의 집합을 둘로 분할하여, 각 집합에 속한 정점끼리는 서로 인접하지 않도록 분할할 수 있을 때, 그러한 그래프를 특별히 이분 그래프 (BIPARTITE GRAPH) 라 부른다.

중심성 척도

소셜 네트워크에 대해 사람들이 하는 첫 질문

여기서 누가 중요한가?

중심성 척도

연결정도: 얼마나 많은 사람이 연결되어 있는지 세는 것

근접성 : 얼마나 많은 사람과 가까운가를 계산.

연결 그래프에서 노드 사이의 거리 개념을 이용.

$$C(x) = \sum 2^{-d(x,y)}$$ d(x,y) = 노드 x와 y의 거리

매개성 : 네트워크 내 사람들이 어느 특정인을 통해 서로를 아는 정도를 측정하는 것.

사람들 사이의 최단경로가 얼마나 이 특정인을 통해 가는지 측정하는 것

$$B(v) = \sum \frac{\sigma_{x,y}(v)}{\sigma_{x,y}}$$

$\sigma_{x,y}(v)$ 노드 v를 통과하는 노드 x와 y의 최단거리 개수

$\sigma_{x,y}$ 노드 x, y 사이의 최단거리의 개수

(합은 v가 아닌 모든 두 노드 x와 y 쌍들에 걸친 합)

고유벡터 중심성: 어떤 사람이 인기 있는 아이들에게 인기 있다면 고유벡터 중심성이 높음

(구글의 페이지랭크)

중심성 척도 사용 주의점

올바른 네트워크 혹은 올바른 하위네트워크를 볼 줄 알아야 한다는 것

문) 무슬림 형제단에서 대단히 영향력 있는 블로거가 누구인가?

- 큰 블로거 그래프에서 상위 100명의 블로거를 뽑고 그 목록에서 아래로 훑어가면서 무슬림 형제단 블로거가 있는지 찾는 방법은 성공하지 못함
- 무슬림 형제단보다는 거대 네트워크 내에서 초국가적 엘리트로서의 영향을 발휘하는 사람일 것

→ 그래프의 국지적 근접성을 항상 염두에 두고 있어야 함

서로 다른 상황에서는 각기 다른 도구가 필요함

- 중심성 척도를 가지고 노는 방법이 다름

다양한 중심성 지표들을 계산해 주는 패키지

Some network packages exist already and can compute the various centrality measures mentioned previously. For example, see NetworkX or igraph if you use Python, or statnet for R, or NodeXL, if you prefer Excel, and finally keep an eye out for a forthcoming C package from Jure Leskovec at Stanford.

사고 실험

사고의 전제

- 워싱턴 DC의 엘리트고, 자금이 풍부한 씽크탱크의 일원
- 해야 할 일은 경험을 토대로 이집트의 매리 정치 상황을 예측하는 것
  어떤 정당이 나타날까?, 5년, 10년, 20년 후의 이집트는 어떤 모습일까?
- 모든 이집트 국민에 대해 다음 데이터들 중 오직 두 개만 접근이 가능
  페이스북 또는 트위터 네트워크, 누가 누구와 함께 학교에 가는가에 대한 완전한 기록,
  모든 사람의 문자와 통화 기록, 모든 사람의 주소,
  또는 모든 공공의 정치 조직과 사기업의 구성원들에 관한 네트워크 데이터

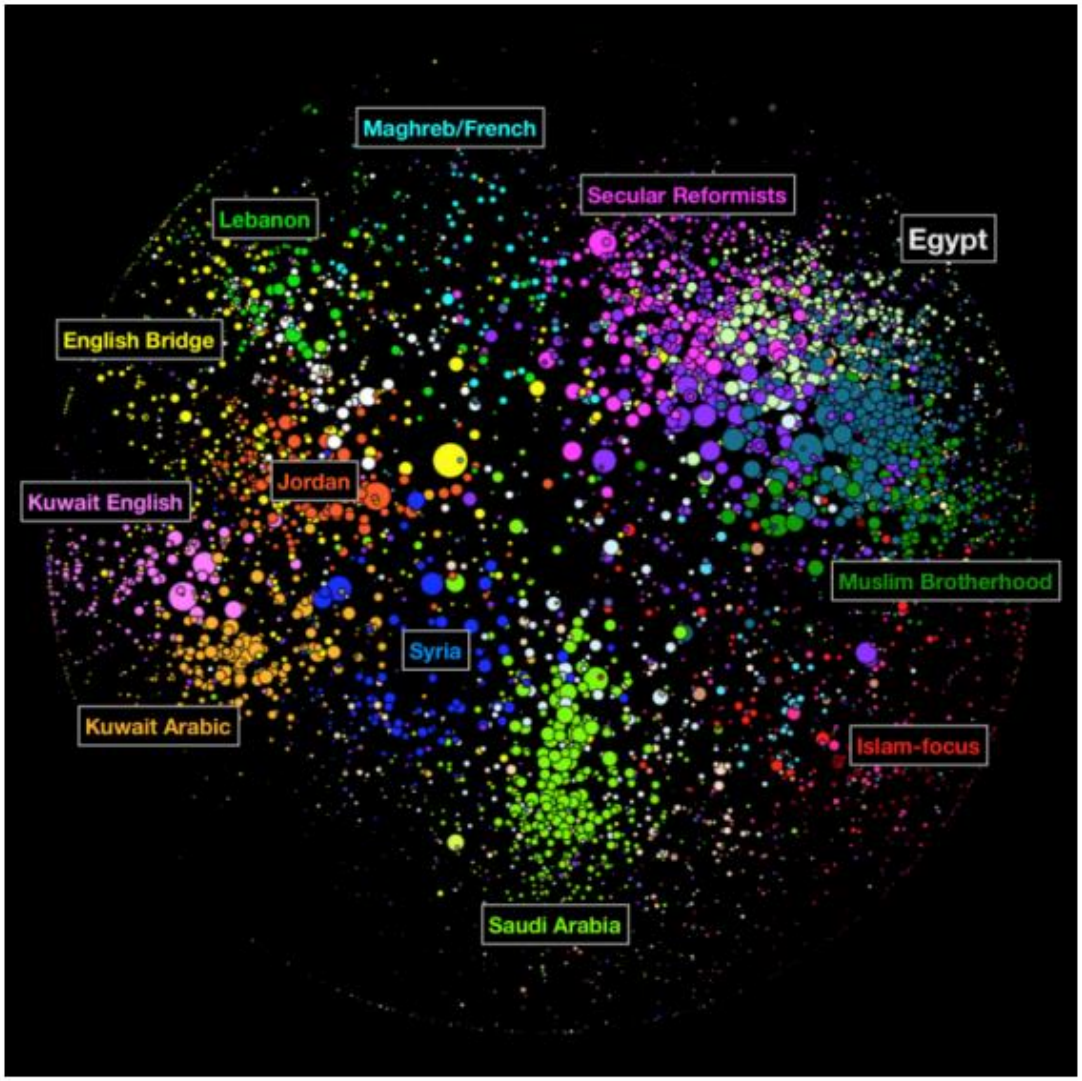
사고의 표준 방향을 바꿀 것을 고려해야 함.

- 사회에서 정치를 예측하는 것은 무엇을 의미 하는가?
- 그것을 알려면 어떤 종류의 데이터가 필요한가?

즉, 먼저 질문을 파악하고 나서 대답을 하기 위한 데이터를 찾으라는 것!

모닝사이드 애널리틱스

아라비아 블로그 방문자들의 예



각가의 색: 블로그의 국가와 군집
점의 크기: 네트워크 내에 다른 블로그와의 연결 정도를 통한 중심성

12개의 블로그 스피어는 서로 다르게 보임

-> 각 사회들이 서로 다른 이해관계를 가지고 있으면서 제각각의 패턴을 형성함

시각화가 어떻게 물고기 떼를 발견하는데 도움을 주는가?

시각화가 필요한 이유

본 것을 믿으려면 게임에 앞서가야 하는데 이를 위해서는 게임이 어떻게 작동 하는지 이해해야 하고 또 게임을 해독해야 한다.

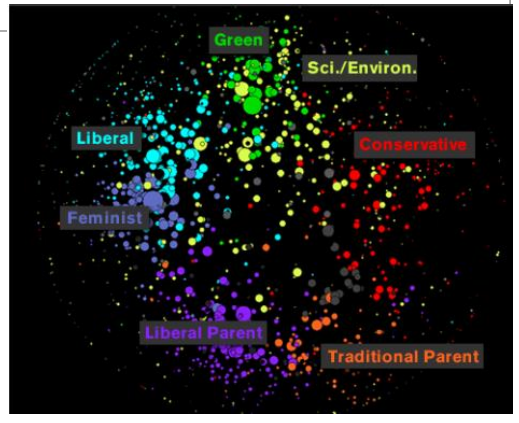
예) 마틴 루터 킹 목사의 "나는 꿈이 있습니다." 연설 동영상과 롬니(Romney) 선거 운동 동영상에 대해 다양한 형태의 링크분석

① 마틴 루터 킹 목사 사례의 경우, 모든 플로고스피어 곳곳에서 선거 기간 내내 많은 포스팅 발생

② 롬니 선거 동영상의 경우, 보수적인 블로거들이 단결하여 일제히 이 동영상을 게시

③ 링크들의 횟수만 기록한 히스토그램만을 본다면 롬니의 비디오는 마치 전염병처럼 확산되는 것처럼 보임

> 확상성 척도를 가지고 장난을 친 치밀하게 계획된 작전



데이터 저널리즘 역사

- 데이터 저널 리즘의 보조 리포팅 수단이 최근 까지도 엑셀 고급 사용자의 영역이었음
- 현재는 API 형태로 더 많은 데이터를 활용할 수 있음
- 누구나 랩탑등을 이용해 큰 용량의 데이터를 분석할 수 있음

조직의 크기에 따라 데이터 저널리즘의 역할 세분화

- 뉴욕타임즈는 그래픽 대 상호작용적 특징, 연구, 데이터베이스 공학자 등의 세분화된 영역으로 나눔
- 작은 조직 에서는 모든 것을 스스로 해야 할 수도 있음

집필 기법 데이터 저널리즘:전문가의 조언

데이터 저널리즘의 필요 요소

- 데이터세트의 핵심을 묘사하는 데이터 시각화
- 빠듯한 마감시간과 지저분한 데이터를 다루기 위해 컴퓨터과학 스킬이 중요

  (예: 파이썬, SQL, 몽고DB 등)
- 세상에 대해 생각하는 법을 알려주는 통계학
- 커뮤니케이션과 프레젠테이션
- 복잡한 이야기를 독자들이 이해할 수 있는 의미로 바꾸는 번역 능력