

## Chapter 10. Structural Recommendations in Networks

### 10.1 Introduction

1) 다양한 웹기반 네트워크의 성장은 무수한 추천 모델을 가능케 했음.  
(웹 그 자체로 거대하고 분산된 데이터의 보존소이고, 구글과 같은 검색 엔진은 추천의 keyword 중심적 변형이라고 볼 수 있음.)

2) 검색엔진과 추천은 주로 구분됨. 검색엔진은 개인화 기능이 없기 때문임. 개인화가 되지 않는 이유는 무수한 웹 유저를 추적하는 것이 어렵기 때문임.

3) 그러나, 개인화된 추천의 개념이 떠오르기 시작했고, 개인의 취향에 맞는 웹 페이지 추천을 제공하기 시작함. 구글과 같은 검색엔진 업체는 개인화된 결과를 제공하는 능력을 갖추.

이러한 문제는 개인화된 선호를 사용한 네트워크 상의 노드를 순위 매김 하는 문제와 동일하다고 볼 수 있음.

4) 네트워크는 소셜 및 정보 네트워크를 모델링하는 범용적 툴이 되었음.  
따라서 network의 다양한 구조적 요소들을 의논하는 것이 유용할 것으로 생각됨.

각각의 구조적 추천 시스템은 각기 다른 상황에서 다른 적용법을 가질 수 있음.  
아래는 이러한 다양한 변형들의 예시임.

#### 1. 권위와 맥락에 의한 노드 추천

: 노드의 질은 그들의 유입 링크에 의해 판단되고, 노드의 개인화된 연관성은 그들의 맥락에 의해 판단됨.

-> 유입 링크가 많은 노드는 고품질의 노드임. 이런 상황은 검색 엔진의 문제와 관련됨. 기존의 검색 엔진 맥락은 다양한 고객들 사이에서 구분이 되지 않는다는 점이 문제.

-> 그러나 이를 보완하고자 맥락에 맞는 개인화 PageRank가 개발되었음.

-> PageRank와 밀접히 연관된 FolkRank와 같은 개념이 social tagging 환경에서 사용됨.

#### 2. 예시에 의한 노드 추천

: 다른 예시 노드와 비슷한 노드를 추천하고 싶을 수 있음. 이는 노드의 연속적인 분류 문제로 볼 수 있음.

-> 이 때, 개인화된 PageRank 방법론이 이러한 연속적 분류 문제의 해결책으로 사용되곤 함.

-> 따라서, 권위/맥락에 의한 추천과 예시 추천은 매우 밀접히 연관되어 있음.

-> 이러한 응용은 특정 특성의 다른 타입 노드와 유저로 구성된 정보 네트워크에서 유용할 수 있음.

#### 3. 영향과 내용에 의한 노드 추천

: 다양항 타입의 상품에 대한 지식을 전파할 수 있음. 이는 바이럴 마케팅의 상황에 대응됨. 이때 판매자는 특정 상품에 대한 그들의 의견을 전파할 것 같은 유저를 찾음.  
-> 토픽에 민감한 영향 분석에서는, 가장 특정 토픽을 전파할 것 같은 유저를 찾게 됨.

-> 영향 분석의 문제는 토픽 특정성과 바이럴 영향도에 기반해 유저를 판매자에게 추천하는 상황으로 생각할 수 있음.

#### 4. 링크 추천

: Facebook과 같은 소셜 네트워크에서는, 네트워크의 연결성을 증진시키는 것이 목적인.

-> 이러한 경우는 네트워크 상의 잠재적인 링크를 추천해주는 것과 같은 문제라고 볼 수 있음.

-> Matrix Factorization과 같은 방법이 링크 예측에 사용될수도 있음.

-> 또 연속적 분류 방법론도 사용될 수있음.

5) 이러한 구조적 추천 방법은 소셜 네트워크 도메인을 넘어섬. 이런 구조적 추천 방법론은 웹 중심 네트워크로 모델링 될 수 있는 모든 시스템의 추천에서 활용될수 있음.

가령 뉴스, 블로그 포스트, 혹은 다른 웹 기반 콘텐츠 같은 것들을 말한다.

심지어 전통적인 상품 추천도 이런 방법으로 다루어 질 수 있음. 이는 유저-상품 추천 문제가 '유저-아이템 그래프'로 모델링 될 수 있기 때문임.

## 10.2 Ranking Algorithms

1) PageRank 알고리즘은 웹 검색에서 제안됨. 검색의 질을 향상시키기 위한 알고리즘임.

2) 초기 검색엔진은 스팸, 잘못된 정보, 페이지 내의 부정확한 내용물, 변동이 심한 문서의 질 등의 문제에 심각하게 노출됨.

따라서 웹의 인용 구조에 따라, 웹 페이지의 명성/질을 결정하는 방법론이 사용됨. 그러나 링크의 수 자체로는 페이지의 질을 정확히 판단할 수 없었음.

이를 개선하기 위해 PageRank 알고리즘이 재귀적 방식의 인용 기반 랭킹의 개념을 일반화시킴.

3) PageRank 알고리즘은 직접적인 추천 방법론은 아니지만, 추천 분석의 주제에 밀접하게 연관되어 있음.

4) PageRank의 다수의 변형 알고리즘들이 개인화된 추천 메커니즘에 사용됨. 이는 많은 경우에 추천은 연결 네트워크의 형태로 표현될 수 있기 때문임.

(유저-아이템 시나리오 포함)

5) 이 장에서는 검색과 추천의 밀접히 연관된 관계를 탐구하고, 다양한 추천 시나리오 내에서의 PageRank 알고리즘의 응용을 살펴봄.

### 10.3 Recommendations by Collective Classification

#### 1) 연속적 분류 방법 (collective classification)

: 추천 프로세스에 콘텐츠를 결합시키는 데 특히 효과적인 방법임.

-> 가령 골프 용품 제작자가 골프에 관심있는 개인을 알고 싶다고 할 때, 골프에 관심있는 개인들의 몇 예시를 알고 있다고 하자.

이 경우 소셜 네트워크의 유저 프로필, 골프 관련 게시물 좋아요 기록 등을 사용해 그러한 개인을 찾을 수 있을 것.

-> 이처럼 소비자의 피드백을 사용할 수 있는 경우, 네트워크 내의 다양한 노드들(고객)의 선호/불호 정보를 얻을 수 있음.

2) 노드의 일부는 라벨과 연관됨. 즉 일부 고객이 학습데이터로써 사용되고, 이를 활용해 새로운 노드를 예측하는 데 사용됨.

3) 협업 필터링 문제처럼, 불완전한 데이터 추정 문제임. 차이점은, 네트워크 구조의 맥락에서 이루어진다는 점.

4) 이러한 문제의 해결책은 동종선호의 개념에 의존함 (homophily, 자신과 비슷한 사람과 어울리려는 성향)

이 개념은 이웃을 이용한 소셜 네트워크 추론의 관점에서 이해될 수 있음.

비슷한 특성을 갖는 노드들은 보통 연관되어 있고, 따라서 노드 라벨 역시 연관되어 있으리라 예상하는 것이 합리적임.

5) 이 문제에 대한 간단한 해결책은, 근접한 k개의 라벨된 노드를 조사하여 다수결 원칙에 따라 라벨을 매기는 것임.

이 방법은 실제로 최근접 이웃 분류기의 네트워크 버전이라고 할 수 있음.

그러나 이 방식은 노드 라벨의 희소성 sparsity 때문에 연속적 분류 문제에서는 일반적으로 사용할 수 없음.

6) 따라서 희소성의 문제는 평점 기반 데이터에서와 마찬가지로 네트워크 기반 예측의 맥락에서도 문제가 됨.

7) 이를 해결하기 위해서는, 라벨된 노드와의 직접적인 연결 뿐만 아니라 라벨이 없는 노드에 걸친 간접적인 연결 역시 활용해야함.

8) 이 장에서는, 반복적 분류 알고리즘과 랜덤 워크 기반 방법들 간단히 살펴볼 것.

### 10.4 Recommending Friends: Link Prediction

1) 페이스북과 같은 SNS에서, 친구를 추천해줌으로써 네트워크를 확장시킬 수 있게끔 도움.

2) 노드간의 공통적인 이웃의 개수를 측정함. 따라서 공통 이웃이 많은 경우 효과적인 방식.

3) Common Neighbor

장단점

-> 장점: 공통 이웃 개수를 반영하여, 가능한 이웃 추천 가능

-> 단점: 절대적인 개수를 이용하므로 네트워크 내에서의 비중을 반영하지

못함.

#### 4) Jaccard Measure

-> 장점: node degree에 따라 상대적인 크기가 조절됨

-> 단점: intermediate neighbor(많은 사람과 이웃인 노드, 연예인 같은 경우)를 반영하지 못함.

(A와 B의 이웃은 node degree가 아주 큰 인기가 많은 연예인일 수 있음에도 추천이 됨.)

#### 5) Adamic-Adar Measure

이웃의 weight를 계산해서, 유명인의 경우 페널티를 가함으로써 진짜 친구에 중요도를 부여함.

#### 6) Katz measures (Likelihood measure 중 하나)

: 이웃 기반 방법은 robust하지만, 노드간의 공통된 이웃이 적을 경우 효과적이지 않음.

-> 따라서 관계의 길이를 사용, 멀 수록 페널티를 주되 먼 관계까지 반영할 수 있도록 구성

-> 각 이웃들의  $n \times n$  행렬로 확장할 수 있음. adjacency matrix 이용 (관계가 있으면 1, 없으면 0)

#### 7) Link prediction과 Collaborative Filtering

-> unary rating/implicit feedback 문제와 유사함.

-> 링크의 특성이 없는 **unsigned link prediction**

-> 노드간의 likelihood 예측값을 유저가 item을 좋아할 선호에 대한 예측값으로 해석할 수 있음.

-> **signed link prediction**:

유저가 있을 때, 아이템을 좋아할지/싫어할지를 유저와 아이템의 관계로 도식화한 link 형태로 prediction하는 것

Link의 특성이 있는 경우(binary rating)이기 때문에, 앞서 설명한 Link prediction과는 조금 다름. 관련 문서를 참고하라고 되어 있음

### 10.5 Social Influence Analysis and Viral Marketing

#### 1) 모든 사회적 상호작용은 개인간의 다양한 수준의 영향을 미침.

전통적인 사회적 상호작용에서, 이는 "word-of-mouth" 영향이라고 불리었음.

이는 온라인 소셜 네트워크 상에서도 일반적인 원리임.

가령 인플루언서가 메시지를 트윗을 하면 팔로워가 모두 이 메시지에 노출되고, 이를 재전파할 수 있음.

결과적으로 소셜 네트워크 내 정보, 아이디어, 의견의 전파가 이루어짐.

2) 이러한 정보의 확산을 중요한 광고 채널로 사용함.

3) 시장 참가자들에게 영향을 미치는 방법론을 바이럴 마케팅이라고 함.

다른 행위자는 그들의 동료들에게 영향을 끼치는 각자 다른 능력을 가지고 있음.

4) 한 행위자의 영향을 조절하는 가장 일반적인 두가지 요소는 다음과 같음.

1. 소셜 네트워크 구조에서 행위자의 중심성은 영향력 정도의 결정적인

요소임.

-> 높은 중심성을 가질수록 더 강력한 영향력을 행사할 가능성이 높음.

-> 단방향 네트워크에서, 높은 위신을 갖는 행위자는 영향력이 높을

것임.

-> PageRank는 중심성과 위신의 척도로 사용될 수 있음.

2. 네트워크의 엣지(간선)는 해당 쌍이 서로 영향을 주고받는 정도에 상응하는 가중치를 의미하곤 함.

-> 확산(diffusion) 모델에 따라, 이러한 가중치는 영향 전파 확률로 직접적으로 해석될 수 있음.

-> 오랜 기간동안 친구였던 사람끼리는 높은 영향을 주고 받을 것으로 예상할 수 있음.

5) 영향 전파 모델은 앞서 언급한 요소들의 정확한 효과를 정량화하기 위해 사용됨.

이러한 모델들은 전파 모델로도 알려져 있음.

-> 이 모델의 주된 목표는 정보 전파의 영향력을 극대화하는 씨앗 노드 집합을 찾는 것임.

-> 이러한 맥락에서, 영향 극대화 모델은 판매자에게 중요한 사회적 행위자를 추천하는 모델로 볼 수 있음.

6) 영향력 함수  $f(S)$ 의 가장 흔한 두가지 접근은 선형 역치 모델(Linear Threshold Model)과 독립 폭포수 모델(Independent Cascade Model)이 있음.

-> 이 모델들은 사회적 영향력 분석의 초기 연구에서 고안된 방법론임.

-> 이 확산 모델들의 일반적인 작위적 가정은 노드들이 활동 혹은 비활동 상황이라는 것임.

-> 직관적으로, 활동 노드는 바람직한 행동 집합에 의해 영향을 이미 받은 노드들이고, 한번 active되면 다시 비활성화 되는 일이 없음.

-> 모델에 따라서 활동 노드는 이웃 노드들의 활성화를 단기간 혹은 장기간 동안 활성화 시킴.

-> 노드는 순차적으로 활성화되고 주어진 반복 횟수에서 더이상 활성화되는 노드가 없어질 때까지 계속됨.

->  $f(S)$ 의 값은 종료 시점의 활성화된 노드의 수에 의해 평가됨.