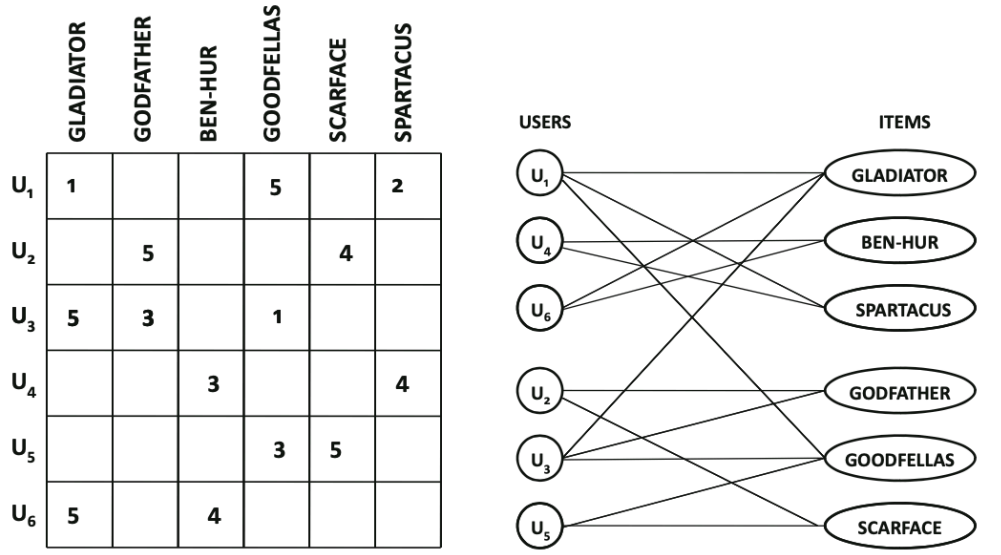
NEIGHBORHOOD-BASED COLLABORATIVE FILTERING

2.7 GRAPH MODELS FOR NEIGHBORHOOD-BASED METHODS

- Neighborhood-Based(이하 NB) 방법들의 가장 큰 문제점은 observed ratings들의 sparsity이다.
- 그래프 모델들은 structural transitivity 나 ranking techniques를 사용하여 NB methods에서의 similarity를 계산한다.
- 여러 형태의 그래프가 정의될 수 있는데 이때, random-walk나 shortest-path methods 들이 이용된다.

2.7.1 User-Item Graphs

- "neighborhoods"를 구하는 방법으로 Pearson correlation 말고, user-item graph의 sturctural measures로 구할 수 있는 방법이 있다.
- 이런 접근은 sparse ratings matrices에서 더 효과적이다. 왜냐하면 이러한 문제를 추천 process에서 edge의 구조적인 transitivity를 이용할 수 있기 때문이다.



(a) Ratings matrix (b) User-item graphs of specified ratings

Figure 2.3: A ratings matrix and corresponding user-item graph

- User-item graph은 undirected and bipartite graph이다. G = ($N_u \cup N_i$, A)
  모든 edges는 users와 items 사이에 연결된다.. users들 또는 items들 서로는 연결되지 않는다.
- Graph 기반의 방법이 가지는 주된 강점은 두 users간의 short paths가 많을 수록 두 users를 neighbors로 고려하기 위해 필요한 공통적으로 평점을 남긴 item이 많지 않아도 된다는 것이다. "indriect connectivity between nodes"
- 만약 두 users 사이에 common items가 많다면 이경우 close neighbors로 볼 수 있다.

2.7.1.1 Defining Neighborhoods with Random Walks

- random walks의 예측 빈도가 어떻게 측정될 수 있을까?
- 왜 이런 방법이 sparse matrices에서 효과적일까?
- Pearson's correlation coefficient에서 두 users가 유의미한 neighborhood로 정의되기 위해선 두 users 사이에 직접적으로 연결된 공통의 item set이 필요하다. 그런데 sparse user-itme graph에서는 이런 직접 연결된 노드들은 많지 않다.
- random-walk 방법에서는 indirect connectivity를 고려한다. 한 node에서 정해진 step동안 이어진 nodes들을 neighorhoods로 고려하는 것이다.

2.7.1.2 Defining Neighborhoods with Katz Measure

- 두 nodes 사이에 선호도를 고려하기 위해서 가중치가 고려된 walks를 이용할 수 있다.
- 이렇게 두 노드 사이의 weighted number of walks를 Katz measure라고 한다.

Definition 2.7. (Katz Measure)

$$Katz(i,j) = \sum_{t=1}^{\infty} \beta^t \cdot n_{ij}^{(t)}$$

n은 node i와 j 사이의 length t walks의 개수이다. (i에서 j로 한번만에 갈 수 있는 길의 수 부터 무한대 steps르으로 갈 수 있는 길의 수의 가중 합) 가중치는 steps(walks)가 길어질 수록 작아진다.

- Let K be the m × m matrix of Katz coefficients between pairs of users. If A is the symmetric adjacency matrix of an undirected network, then the pairwise Katz coefficient matrix K can be computed as follows

$$K = \sum_{i=1}^{\infty} (\beta A)^i = (I - BA)^{-1} - I$$

- The value of β should always be selected to be smaller than the inverse of the largest eigenvalue of A

  infinite summation의 수렴을 위해서 the value of β 는 항상 A의 eigenvalue 중 가장 큰 수의 역수 보다 작은 수로 정해야한다. 실제로 몇몇 collaborative recommendation methods는 diffusion kernels를 사용하기도 한다.

- A weighted version of the measure can be computed by replacing A with the weight matrix of the graph. This can be useful in cases where one wishes to weight the edges in the user-item graph with the corresponding rating.
  A를 weight matrix를 사용함으로써 weighted version이 계산될 수 있다. 이는 user-item graph에서 대응되는 rating으로 edges에 가중치를 부여할 때 이용될 수 있다.

- 다양한 기본 원칙들이 추천에 사용된다.
  - 식 2.34(Definition 2.7.1에 첫번째 식)에 maximum path length로 threshold를 걸 수 있다. 왜냐하면 path가 길어 질 수록 예측이 noisy해지기 때문이다. 그래도 discount factor β가 있기 때문에 long path를 어느정도 제한 할 수 있다.
  - 직접적으로 예측을 수행하는 한 방법으로 neighborhood methods를 사용하지 않고 user와 item 사이의 선호도를 계산할 수 있다. 이때 Katz measure로 이런 선호도를 계산할 수 있다.

2.7.2 User-User Graphs

- User-item graphs에서 user-user connectivity는 짝수 번의 hops로 정의된다.
- User-item graph를 고려하는 것 대신에, user-user graph를 2 hop 기반으로 만들 수 있다.
- user-user 그래프의 강점은 이 graph의 edge가 user-item 보다 더 informative 하다는 것이다.

- 2-hop connectivity는 edges를 형성할때 두 users 사이에 공통 아이템의 개수나 similarity를 직접적으로 반영할 수 있기 때문이다.
- **horting** and **predictability**

  `horting` to quantify the number of mutually specified ratings between two users (nodes)

  => 두 user사이에 상호 rating한 개수를 정량화

  `predictability` to quantify the level of similarity among these common ratings.

  => 공통의 ratings 사이의 similarity의 level를 정량화
- Edges are defined in this graph with the notion of horting. Horting is an asymmetric relationship between users, which is defined on the basis of their having rated similar items.

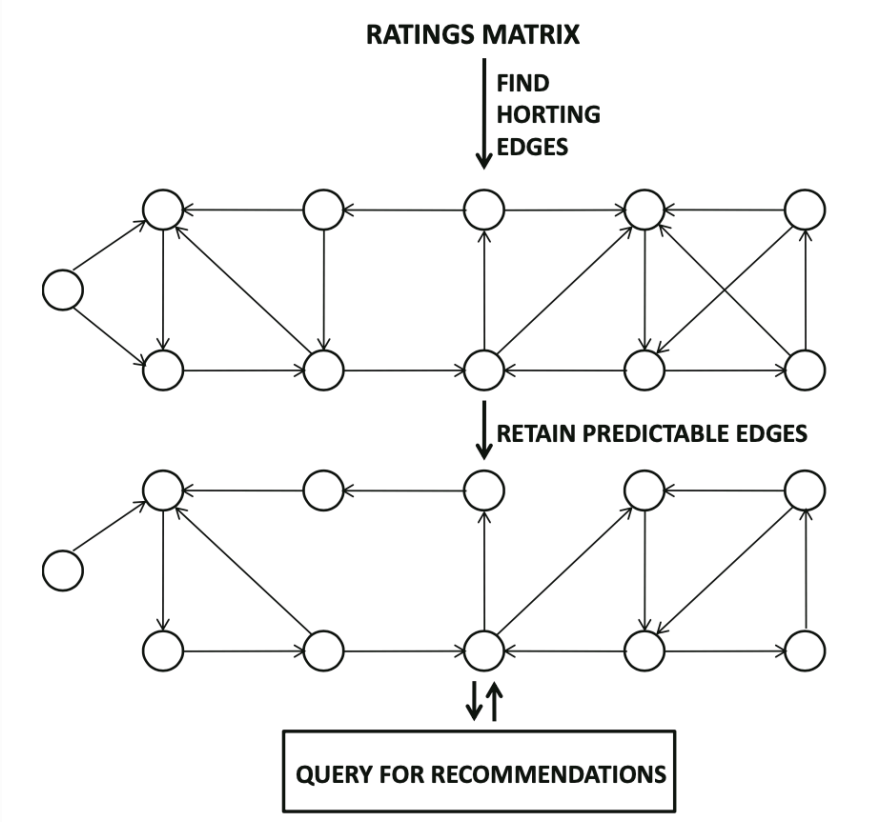


Figure 2.4: The user-user predictability approach

Definition 2.7.2 (Horting)

- Edge 연결 조건 : A user u is said to hort user v at level (F,G) => (G : u가 v를 follow 하는 정도?)

$$|I_u \cap I_v| \geq F$$
$$|I_u \cap I_v|/|I_u| \geq G$$

Definition 2.7.3 (Predictability)

- Edge 연결 조건 : The user v predicts user u, if u horts v and there exists a linear transformation function f(·) such that the following is true:

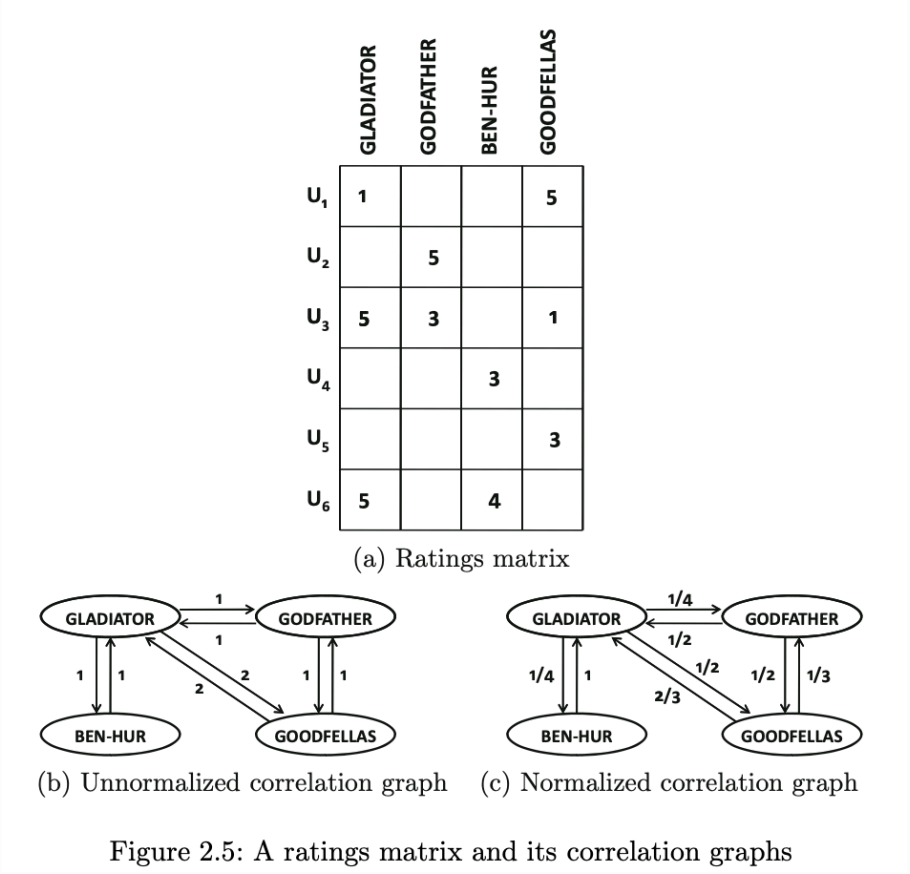
$$\frac{\sum_{k \in I_u \cap I_v} |r_{uk} - f(r_{vk})|}{|I_u \cap I_v|} \leq U$$

- user u의 ratings와 user v의 transformed ratings의 Manhattan distance를 구하게 된다. 여기서 한 가지 특징은 공통적으로 ratings한 item의 개수로 normalized 한다는 것이다. 이를 **Manhattan segmental distance** 라고 한다.
- The directions of horting and predictability are opposite one another. In other words, for user v to predict user u, u must hort v. (내가 이영자가 추천한 맛집을 신뢰한다면, 이영자가 돌아본 맛집들로 내가 가보지 않은 식당에 대한 평점을 예측할 수 있다.)
- edge에서 head의 rating은 tail의 rating을 예측하는데 사용 가능하다.
- 게다가, 한가지 가정이 존재하는데, path의 종착지점에서의 rating에서 source의 rating를 예측 하기위해서 이런 방향성이 있는 path 전반에 걸친 transitive 방식에서 linear transformations를 적용할 수 있다.
- target user u의 item k에 대한 rating은 u에서 item k를 평가한 모든 다른 users 까지 모든 directed shortest paths로 계산되어 진다.
- 예를 들어, u에서 item k를 평가한 v까지 directed path 가 r인 경우를 생각해보자. u의 k에 대한 예측 평점 $\hat{r}_{uk}^{(v)}$ 은 아래와 같이 linear mappings의 합성으로 구해진다.

  $\hat{r}_{uk}^{(v)} = (f_1 \circ f_2 \circ \cdots \circ f_r)(r_{vk})$
- k를 평가한 유저 v는 다수일 수 있으니 최종 예측 $\hat{r}_{uk}$ 는 위 예측 평점의 평균으로 계산한다. 이때 r을 최대 길이 D의 thershold distance를 두고 계산한다. (r이 1~D ?)

- 그렇다면 어떻게 r 만큼 나갈 것인가를 정해줘야한다. 즉 path를 나아가는 방법 또는 hop을 넘어가는 방법을 정해야한다.
  shortest path를 구하는 방법으로, breadth-first algorithm이 꽤 효과적이다.

2.7.3 Item-Item Graphs



(a) Ratings matrix

(b) Unnormalized correlation graph (c) Normalized correlation graph

Figure 2.5: A ratings matrix and its correlation graphs

- item-item graph는 correlation graph라고도 한다.
- 위 그림을 보면 어떻게 item-item 그래프가 만들어지는지 알 수 있다.

  node는 item이 된다. edges는 user가 된다. i를 소비한 사람이 j를 소비 했다면 i->j, j->i 각 방향의 edge가 만들어질 것이다.
- Normalized correlation graph는 node의 outgoing edge 개수로 outgoing edge를 나누어 준다. 즉 임의의 노드의 outgoing edges의 합이 1이 된다.
  확률로 생각할 수 있다.
- 이렇게 계산된 edges의 weights는 random-walk의 probabilities가 된다.
- correlation graph의 한가지 특징은 rating value가 사용되지 않고 오로지 연결성만 고려된다는 것이다.

- 물론 두 item간의 rating vectors를 cosine function을 이용하여 계산한 correlation graph를 정의할 수도 있다.