Graph Mining, Recommendation, and Visualization

지능정보시스템 연구실

천상진(csjin75@uos.ac.kr)

서울시립대학교
UNIVERSITY OF SEOUL

2019. 11. 25.

7. Data Mining

7.6 Graph Mining

7.7 Recommendation

7.8 Visualization

Graph Search

- 그래프 데이터베이스와 질의 그래프가 주어질 때,
  질의 그래프를 부분 그래프로 포함하는 그래프들을 데이터베이스에서 모두 찾는 문제.

- **Indexing**(인덱싱)
  - 인덱싱에 사용하는 특징: 경로, 부분 구조, 노드의 이름, **부분 그래프**

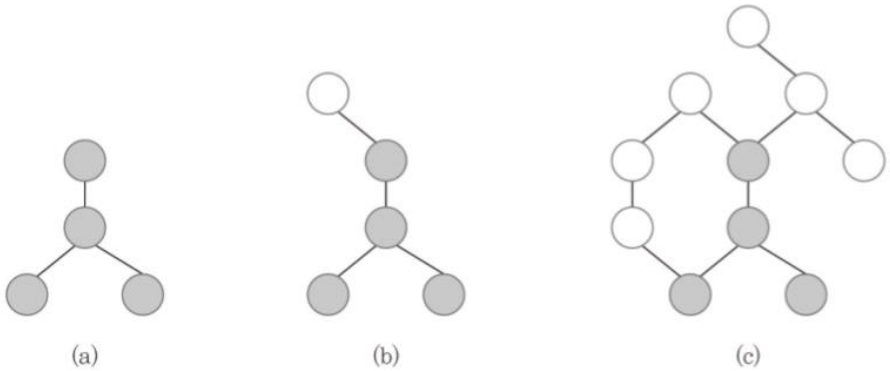


그림 7.8 부분구조
(a) 부분구조 (b) 질의 그래프 (c) 그래프 데이터

Graph Search

- **Similarity Search**(유사도 검색)
  - **Feature Vector**(특징 벡터), **Graph Edit Distance**(그래프 편집 거리)를 이용
  - 특징 벡터의 유사도가 일정값 인상인 그래프를 후보로 선택하고, 편집 거리 등으로 평가하여 최종 결정

$(f_1, f_2, \ldots, f_N)$

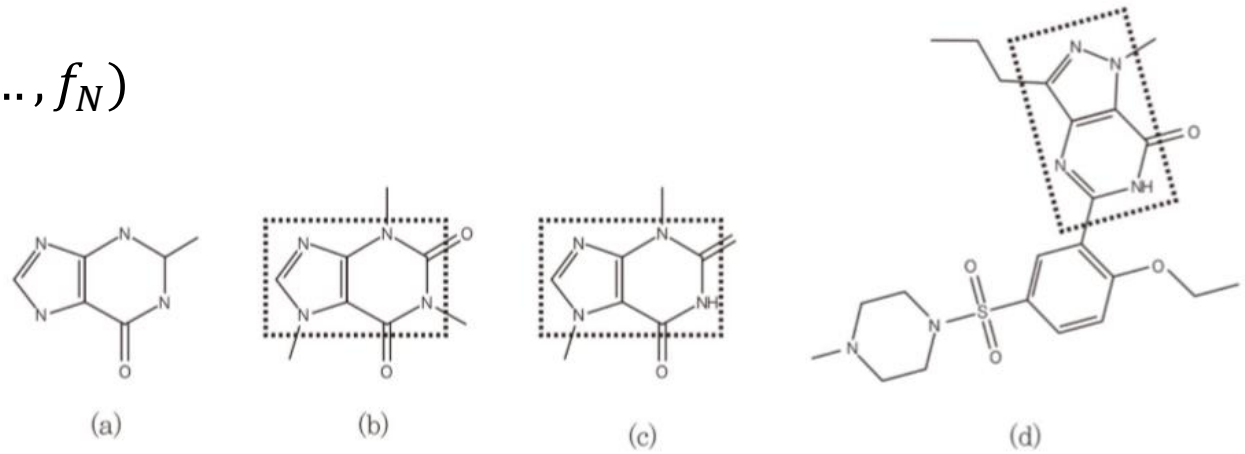
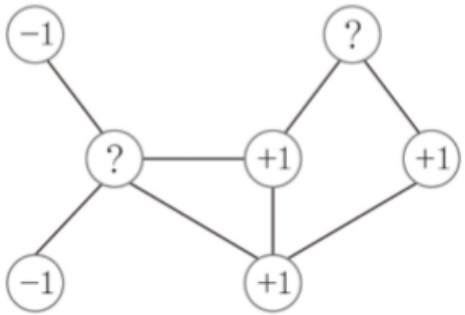


그림 7.9 유사 그래프 검색[출처: Han & Camber, 2011]

(a) 질의 그래프 (b) G_1 (c) G_2 (d) G_3

Graph Classification

1) 주어진 그래프에 대해서 학습 데이터를 이용하여 라벨을 모르는 노드의 라벨을 결정.



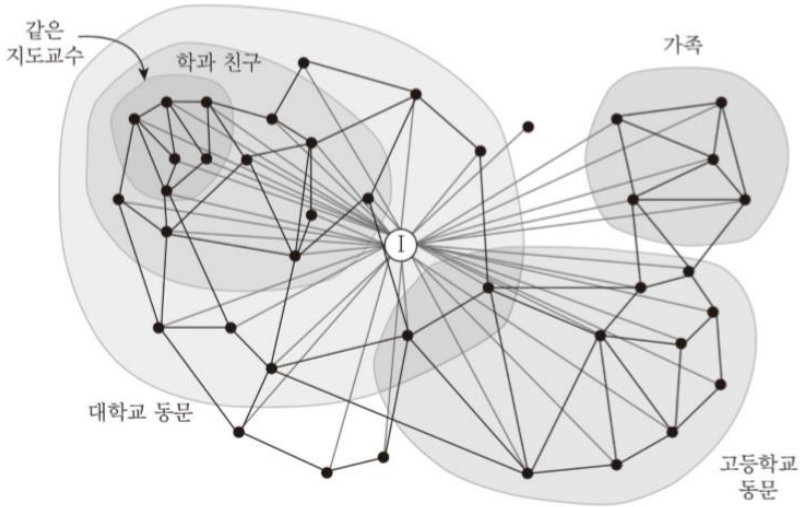
그래프 노드 라벨 결정문제

2) 그래프 데이터에 대한 부류를 결정하는 문제

- Graph convolution
  - 분류 문제의 특성에 맞는 특징을 학습을 통해 결정하여 추출하는 방법

Graph Clustering

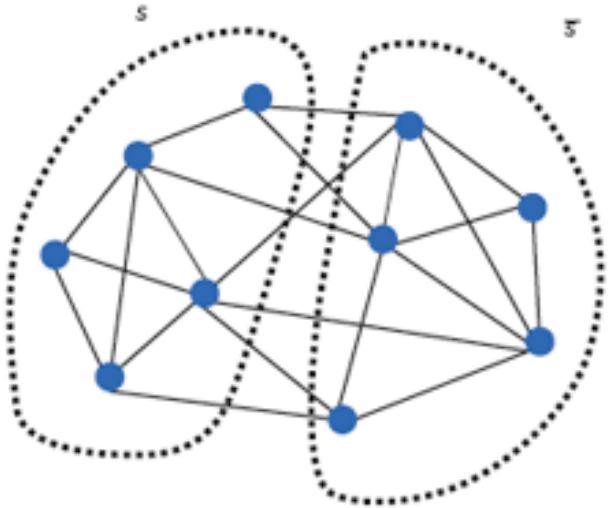
- 하나의 그래프에서 특정 성질을 만족하는 부분 그래프들을 찾거나, 많은 그래프 데이터들에서 비슷한 것들을 군집으로 묶는 것.
- **Community**(커뮤니티) : 관심사나 관계에 의해 만들어지는 집단



커뮤니티의 예시

Graph Clustering

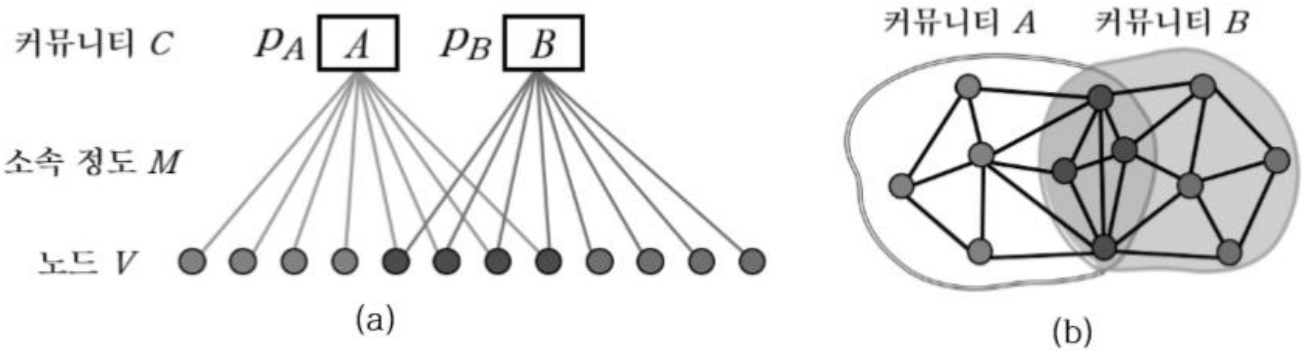
- 그래프 분할
  - 분할될 때 제거되는 간선의 개수나 간선의 가중치의 합이 최소가 되도록 등



그래프 분할의 예시

Graph Clustering

- **Community Affiliation Graph (커뮤니티 소속 그래프)**
  - 노드와 커뮤니티의 소속 관계를 확률적으로 표현하는 그래프
  - 커뮤니티 소속 그래프 모델이 주어진 그래프를 생성할 확률이 최대가 되도록 커뮤니티 소속 그래프 모델의 파라미터를 결정



커뮤니티 소속 그래프(a)와 생성된 그래프(b)

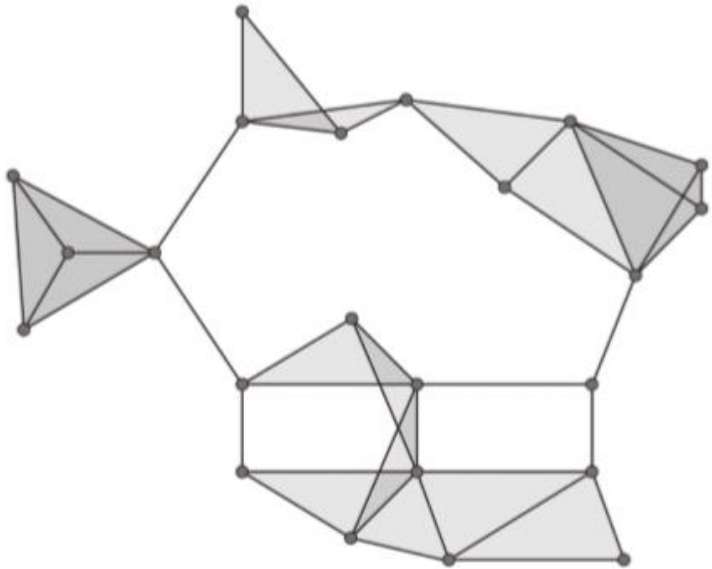
Keyword Search on Graph

- 데이터가 그래프로 표현되어 있을 때, 키워드를 사용하여 필요한 정보를 검색

"Rusell Crowe가 출연한 영화의 감독이 연출한 영화 중에 Helena Carter가 주연한 것"

Graph Data

- 대규모 그래프에는 기존의 그래프 알고리즘을 그대로 적용하기 힘듦.
- **Clique**(클릭) : 모든 노드 쌍 사이에 간선(edge)가 존재하는 부분 그래프

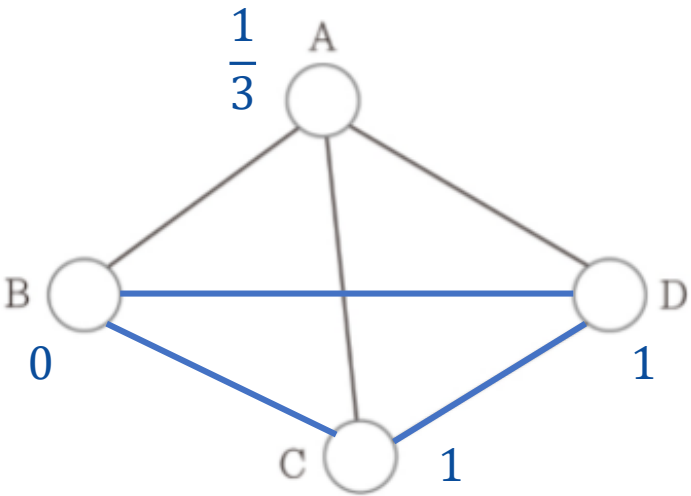


클릭을 음영으로 나타낸 예시

Graph Data

- **Clustering Coefficient (군집화 상수)**
  - 그래프에서 노드들이 뭉쳐 있는 정도가 얼마나 강한지 측정

$$c(v) = \frac{\text{노드 } v\text{의 인접 노드 간의 간선의 개수}}{\text{노드 } v\text{의 인접 노드 간에 만들 수 있는 간선의 개수}}$$



Graph Data

- **Structural Similarity**(구조 유사도)
  - 노드 쌍에 대해서 측정하며,
    값이 클수록 해당 노드들이 동일한 클릭이나 커뮤니티에 속할 가능성이 높음

$$\sigma(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{\sqrt{|\Gamma(u)||\Gamma(v)|}}$$

$\Gamma(u)$: 노드 u에 이웃한 노드들의 집합

Recommendation

- 사용자에게 맞춤형 정보를 제공하여 정보 검색의 부하를 줄임

Recommendation

서울시립대학교
UNIVERSITY OF SEOUL

Page Rank Algorithm

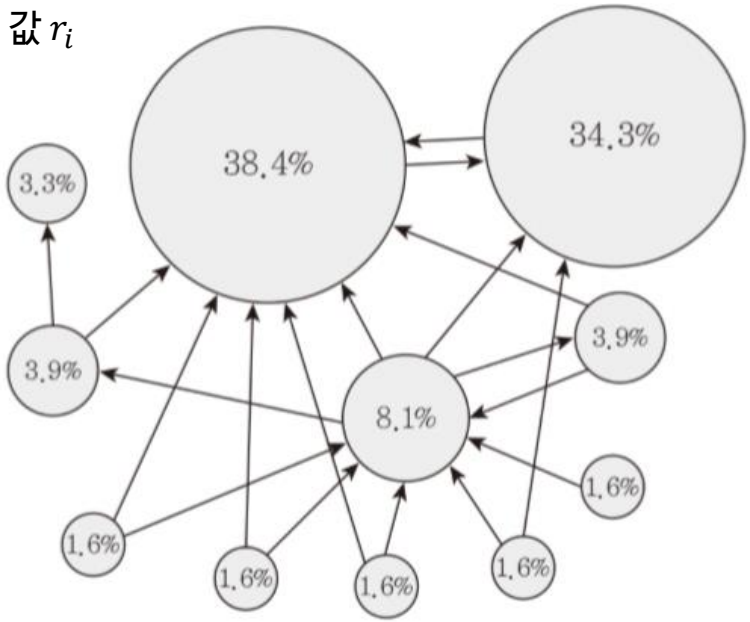
➢ 다른 페이지로부터 들어오는 링크가 많은 페이지일수록 중요도가 높다.

j의 중요도 값

$$r_j = \sum_{i \to j} \frac{r_i}{n_i}$$

r_i → j로 오는 링크를 가지고 있는 페이지 i의 중요도 값 r_i

n_i → i의 링크 수



➢ Random Surfer

- 무작위로 웹 페이지를 돌아다니는 서퍼(surfer)

➢ Teleport

- 일정 확률로 무작위로 임의의 페이지로 갈 수 있게 함.

Data on Recommendation

- **Sparse Matrix (희소 행렬)**
  - 극히 일부의 행렬 원소만 값을 갖는 행렬

	1	2	3	4	5	6	7	8	9	10	11	12
1	1		3			5			5		4	
2			5	4			4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2					2	5
6	1		3		3			2			4	

Table 2: Description of the datasets after pre-processing.

	# users	# items	nnz	side information
ML-1M	6038	3522	2.7%	genres, cast,
ML-10M	69797	10258	0.7%	directors, writers
BX	7160	16273	0.18%	publishers, authors
AMZe	124895	44483	0.02%	categories,
AMZvg	14251	6858	0.13%	brands
YaMus	183003	134059	0.1%	artists, genres, albums

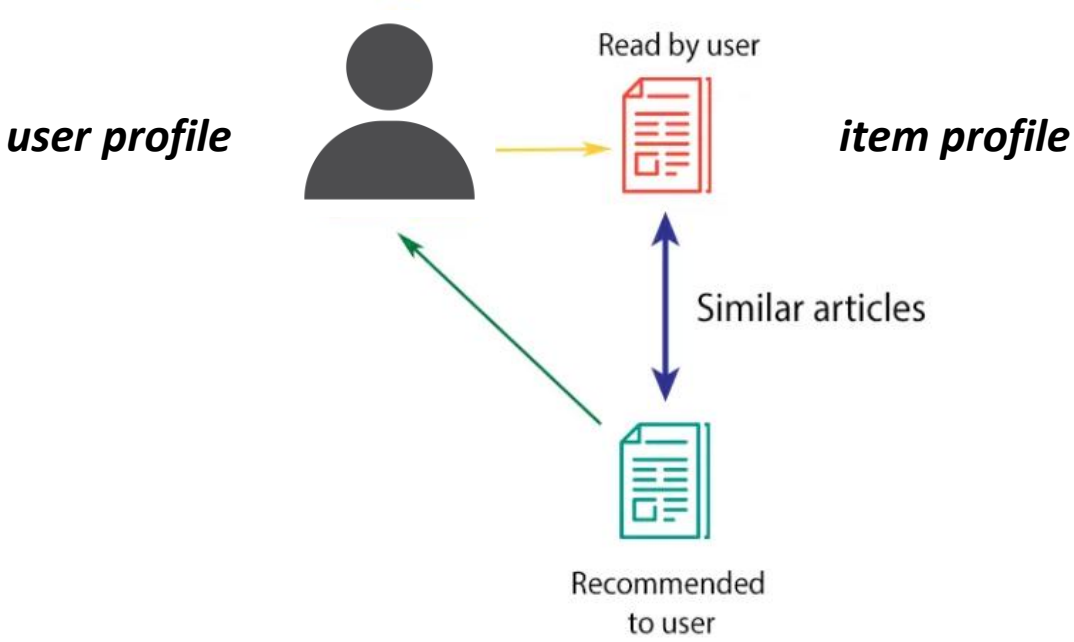
Table 1: Statistics of the evaluation datasets.

Dataset	Interaction#	Item#	User#	Sparsity
MovieLens	1,000,209	3,706	6,040	95.53%
Pinterest	1,500,809	9,916	55,187	99.73%

[1] Evgeny Frolov and Ivan Oseledets, "HybridSVD-When Collaborative Information is Not Enough," RecSys'19
[2] X. He, L. Liao, H Zhang, L. Nie, X. Hu and T.S. Chua, "Neural Collaborative Filtering," WWW'17

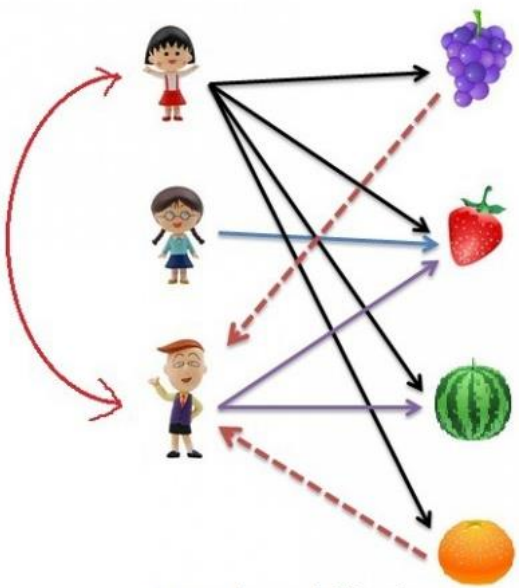
Content-based Recommendation(CB)

- 사용자가 이전에 높게 평가했던 것과 유사한 내용을 갖는 대상을 추천

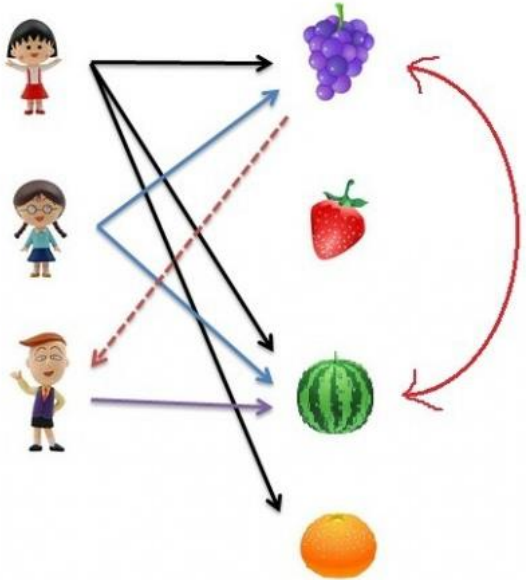


Collaborative Filtering(CF)

- User-based CF
  - similarity between user vectors
- Item-based CF
  - similarity between item vectors



User-based filtering

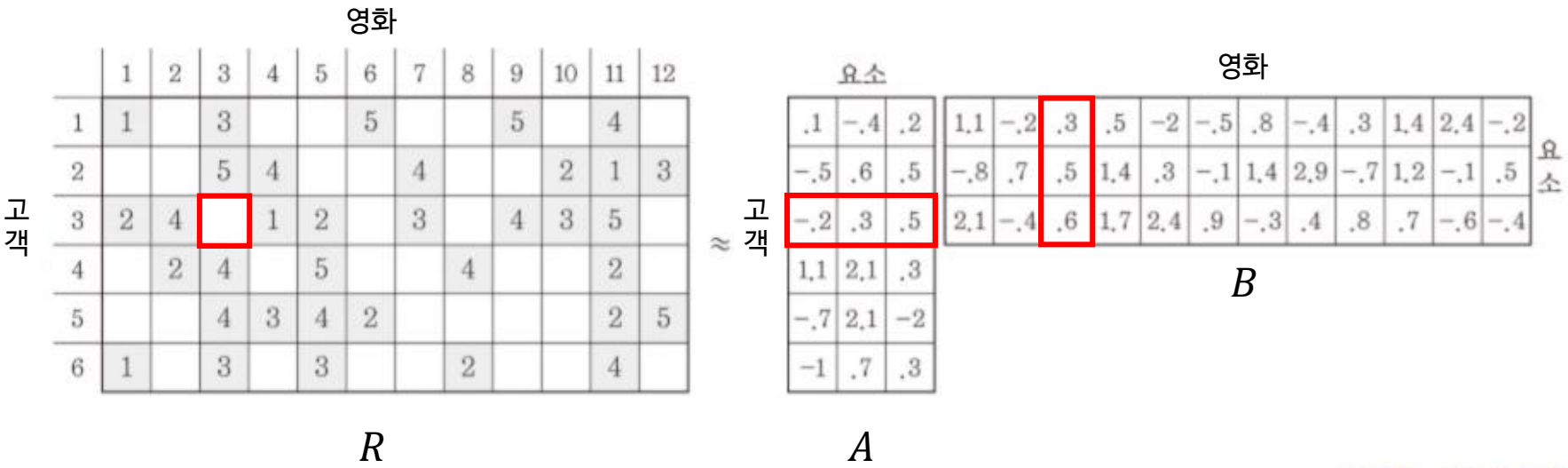


Item-based filtering

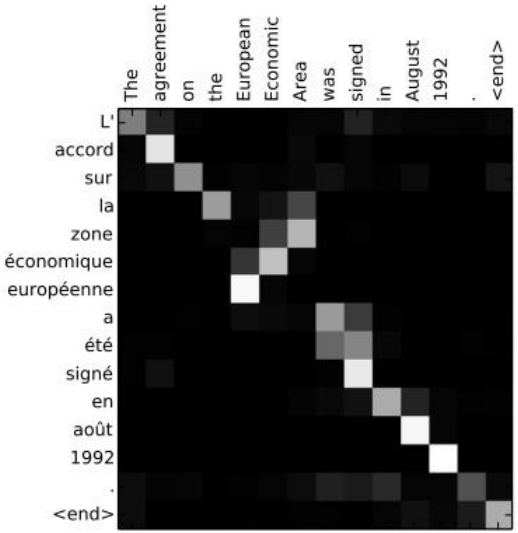
Latent Factor Model

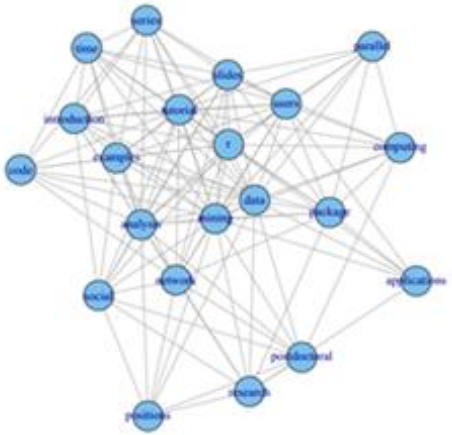
- Matrix Factorization (MF)
  - 주어진 행렬을 2개의 행렬로 나누어 곱으로 나타냄.
  - 분해된 행렬의 곱으로 기존 행렬의 빈 원소를 결정.

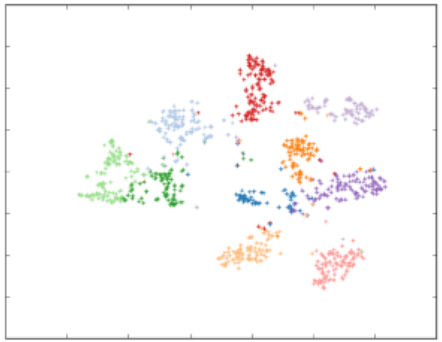
$$r_{ij} = \sum_{k=1}^{K} a_{ik} b_{kj}$$

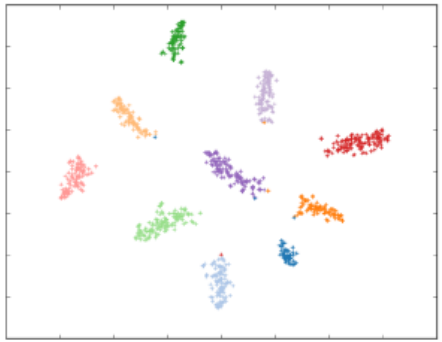


Visualization









(a) Epoch 0

(e) Epoch 12

[1] J. Xie, R. Girshick and A. Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," ICML'16