ACADEMIC RECOMMENDER SYSTEM DESIGN

顾健喆

WHAT'S ACADEMIC RECOMMENDER SYSTEM

Similar paper to *paper*

Relevant paper to *author*

Reading suggestion to *user*

Recommendation is based on *feature* of paper.

Title, Abstract, Keyword, Reference ,User's activities…

INTRODUCTION OF RECOMMENDER SYSTEM

Two Roles:

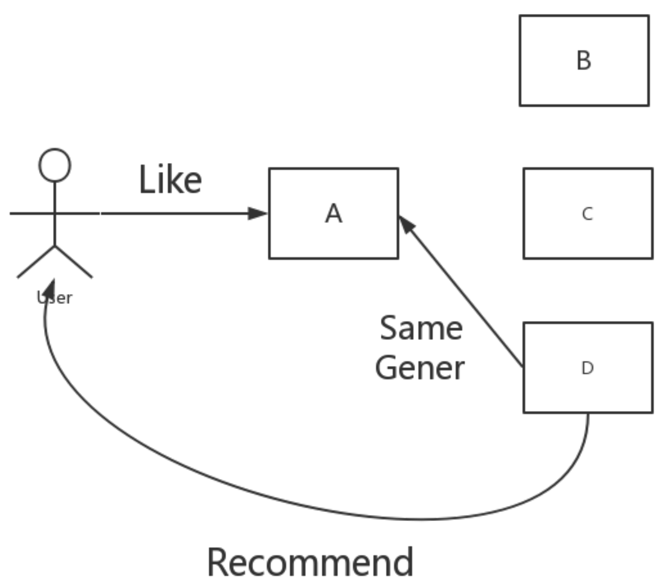
- User : Providing opinion to items
  - e.g. Rating, Thumb up, Thumbing, Star…
- Item : Providing necessary information.

Three Types:

- Content-Based Algorithm (CB)
- Collaborative Filtering Algorithm (CF)
- Hybrid Approach

CONTENT-BASED SYSTEM

Providing recommendations by comparing the representations of content contained in an item to representations of content that interests the user.



Extract item's features

COLLABORATIVE FILTERING

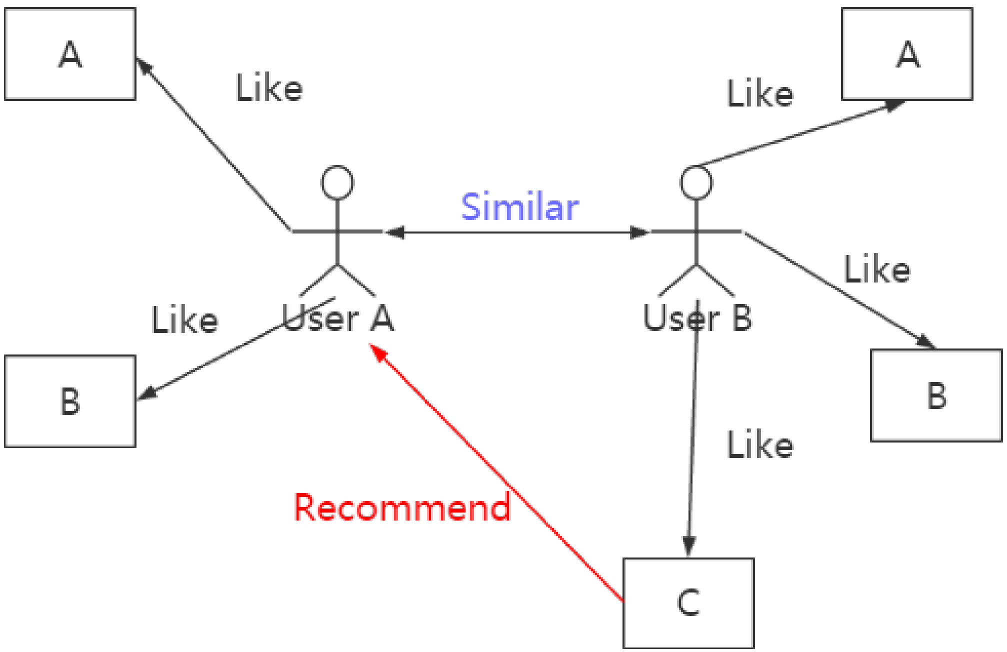
Finding a subset of users who have similar tastes and preferences to the target user and use this subset for offering recommendations.

Preferences are recorded in the *rating matrix*.

Two Main Approach:

- User-based
- Item-based

IDEA OF COLLABORATIVE FILTERING



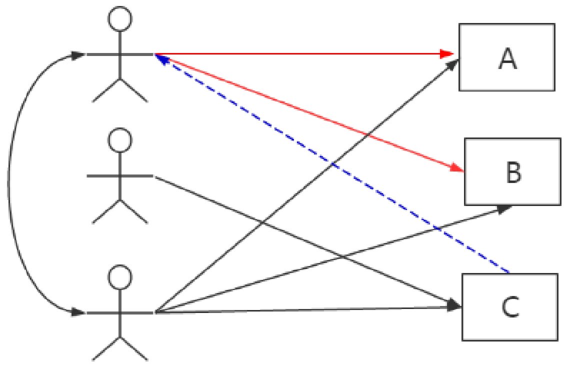
USER-BASED COLLABORATIVE FILTERING

Use user-item rating matrix

Make user-to-user correlations

Find highly correlated users

Recommend items preferred by those users



Pearson Correlation :

$$userSim(u,n) = \frac{\sum_{i \subset CR_{u,n}} (r_{ui} - \bar{r}_u)(r_{ni} - \bar{r}_n)}{\sqrt{\sum_{i \subset CR_{u,n}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \subset CR_{u,n}} (r_{ni} - \bar{r}_n)^2}}$$

Prediction Function :

$$pred(u,i) = \bar{r}_u + \frac{\sum_{n \subset neighbors(u)} userSim(u,n) \cdot (r_{ni} - \bar{r}_n)}{\sum_{n \subset neighbors(u)} userSim(u,n)}$$

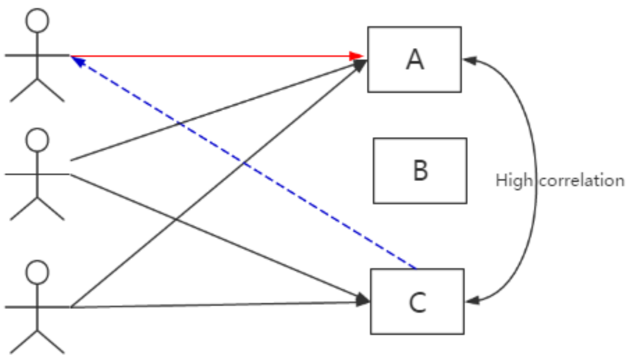
USER-BASED COLLABORATIVE FILTERING

Item User	I1	I2	I3	I4	I5
U1	5	8		7	8
U2	10		1		
U3	2	2	10	9	9
U4		2	9	9	10
U5	1	5			1
User a	2		9	10	

Recommend items preferred by highly correlated user U3
Recommend I5 to User a.

ITEM BASED COLLABORATIVE FILTERING

- Use user-item ratings matrix
- Make item-to-item correlations
- Find items that are highly correlated
- Recommend items with highest correlation



Similarity Metric :

$$itemSim(i,j) = \frac{\sum_{u \subset RB_{i,j}} (r_{ui} - \bar{r}_u)(r_{uj} - \bar{r}_u)}{\sqrt{\sum_{u \subset RB_{i,j}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{u \subset RB_{i,j}} (r_{uj} - \bar{r}_u)^2}}$$

Prediction Function :

$$pred(u,i) = \frac{\sum_{j \in ratedItems(u)} itemSim(i,j) \cdot r_{ui}}{\sum_{j \in ratedItems(u)} itemSim(i,j)}$$

ITEM BASED COLLABORATIVE FILTERING

Item User	I1	I2	I3	I4	I5
U1	5	8		7	8
U2	10		1		
U3	2		10	9	9
U4		2	9	9	10
U5	1	5			1
User a	2		9	10	

I5 is highly correlated to preferred items I4

HYBRID RECOMMEND APPROACH

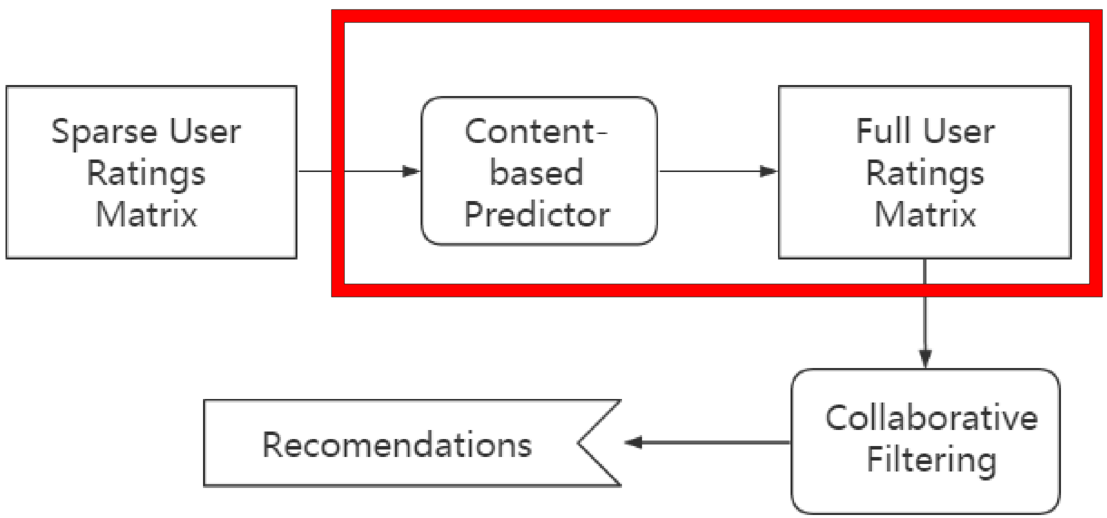
The problem of the Collaborative Filtering:

- Sparsity: Most users do not rate most items and hence the user-item rating matrix is typically very sparse.
- Cold Start: An item cannot be recommended unless a user has rated it before.

Hybrid Recommend Approach can overcome these shortages.

CONTENT-BOOSTED COLLABORATIVE FILTERING

Adding Content-based Predictor before Collaborative Filtering



pseudo user-ratings vector:

$$v_{u,i} = \begin{cases} r_{u,i} : & \text{if user } u \text{ rated item } i \\ c_{u,i} : & \text{otherwise} \end{cases}$$

ACADEMIC RECSYS DATA

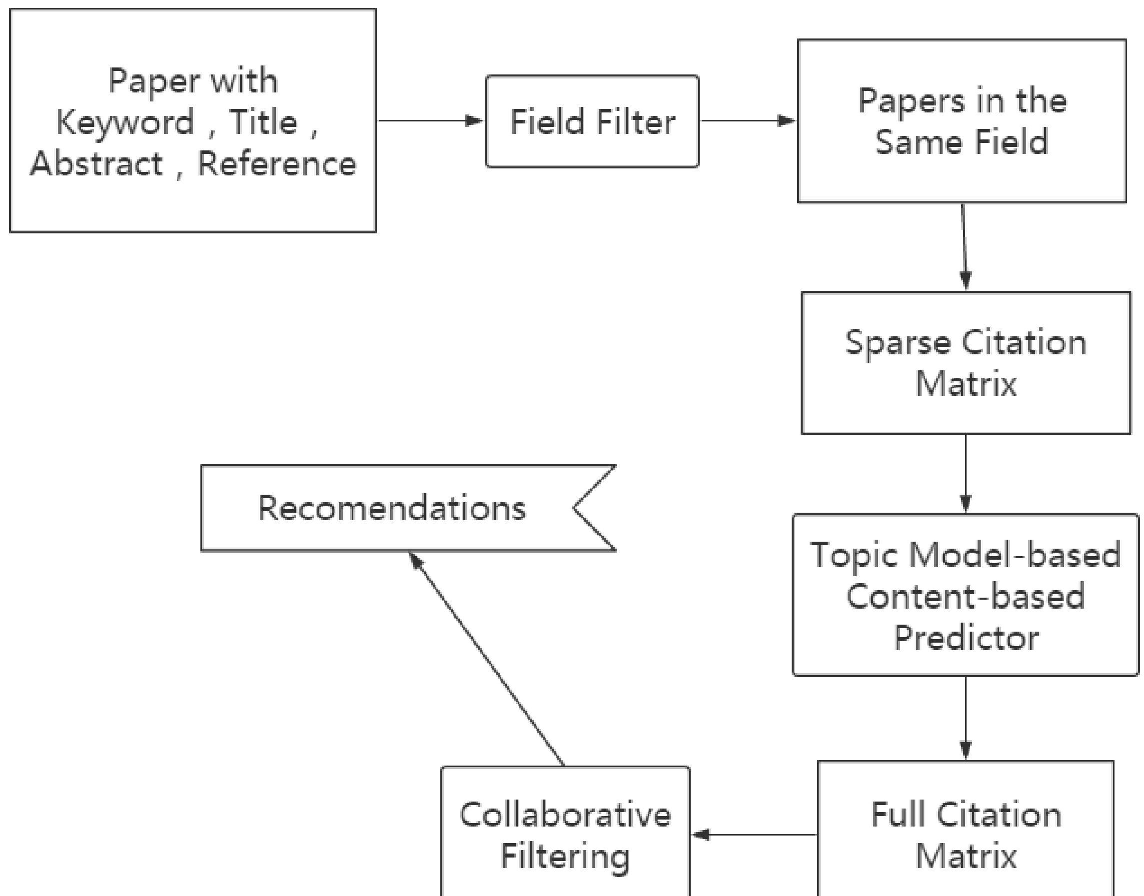
Content-based Recommender system

- Title
- Abstract
- Keyword

Collaborative Filtering Recommender System

- Reference

HYBRID ACADEMIC RECSYS DESIGN



ACADEMIC COLLABORATIVE FILTERING RECSYS

Integrating CF into the domain of research papers

- CF works with *ratings matrix*
- *Columns represent 'users'*.
- *Rows represent 'item'*
- Maping citation web onto ratings matrix.

	Item 1	Item 2
User 1	R1,1	R1,2
User 2	R2,1	R2,2

MAPPING CITATION WEB ONTO CF RATINGS MATRIX(1)

'Item': Citations

'User': Real Users

'Rating': Users' activities: Thumb Up, Thumb down, Rating etc.

Problem:

- *Startup problem*
  - *Not enough users and users activities in the dataset*

MAPPING CITATION WEB ONTO CF RATINGS MATRIX(2)

‘*Item*’*:* Citations

‘*User*’*:* Paper authors

‘*Rating*’:”Vote” for the papers if he has cited

Advantage: No startup problems

Disadvantage:

- Many authors have written papers in several different fields over their careers.
  - Serendipity is not useful in academic recsys.

MAPPING CITATION WEB ONTO CF RATINGS MATRIX(3)

'Item': Citations

'User': Paper

'Rating': Each paper would then vote for the citations found in its references list.

	Ciation1	Citation2	Citation3	Citation4	Citation5
Paper1	♥		♥	♥	
Paper2		♥			♥
Paper3	♥		♥	♥	♥

COLLABORATIVE FILTERING ALGORITHMS

Co-Citation Matching

- Co-citation Matching works by counting co-citations

User-Item CF

- User-Item algorithm compares papers (rows) in the matrix to create a neighborhood of the most similar papers to the target paper.

Item-Item CF

- The Item-Item algorithm compares citations (columns) in the ratings matrix to create a neighborhood

ACADEMIC CONTENT-BOOSTED RECSYS

Data Sparsity

	Ciation1	Citation2	Citation3	…………….	Citation n	Citation n+1
Paper1	1	Empty	1	Empty	1	1
Paper2	Empty	1	Empty	Empty	Empty	Empty
Paper3	1	Empty	1	Empty	1	Empty

Serendipity is not useful

The Long Tail

FIELD FILTER

Serendipity is not useful

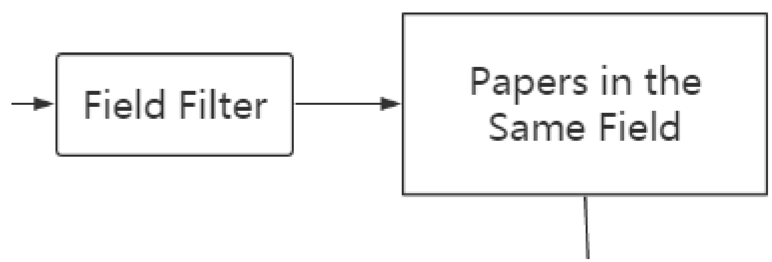
- Recommending paper in its filed.

Using keyword and keyword hierarchy to extract paper’s field.

Using PaperRank to find

the important paper in

fields.



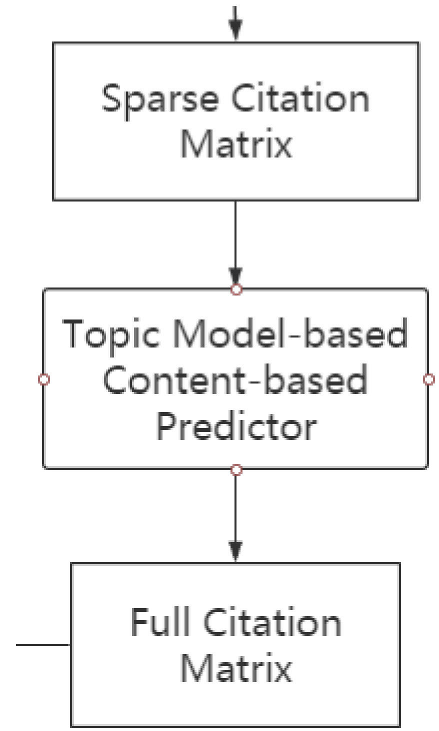
TOPIC MODEL-BASED CONTENT-BASED PREDICTOR

Using Topic Model to analyze the similarity of papers.

Content: Title and Abstract

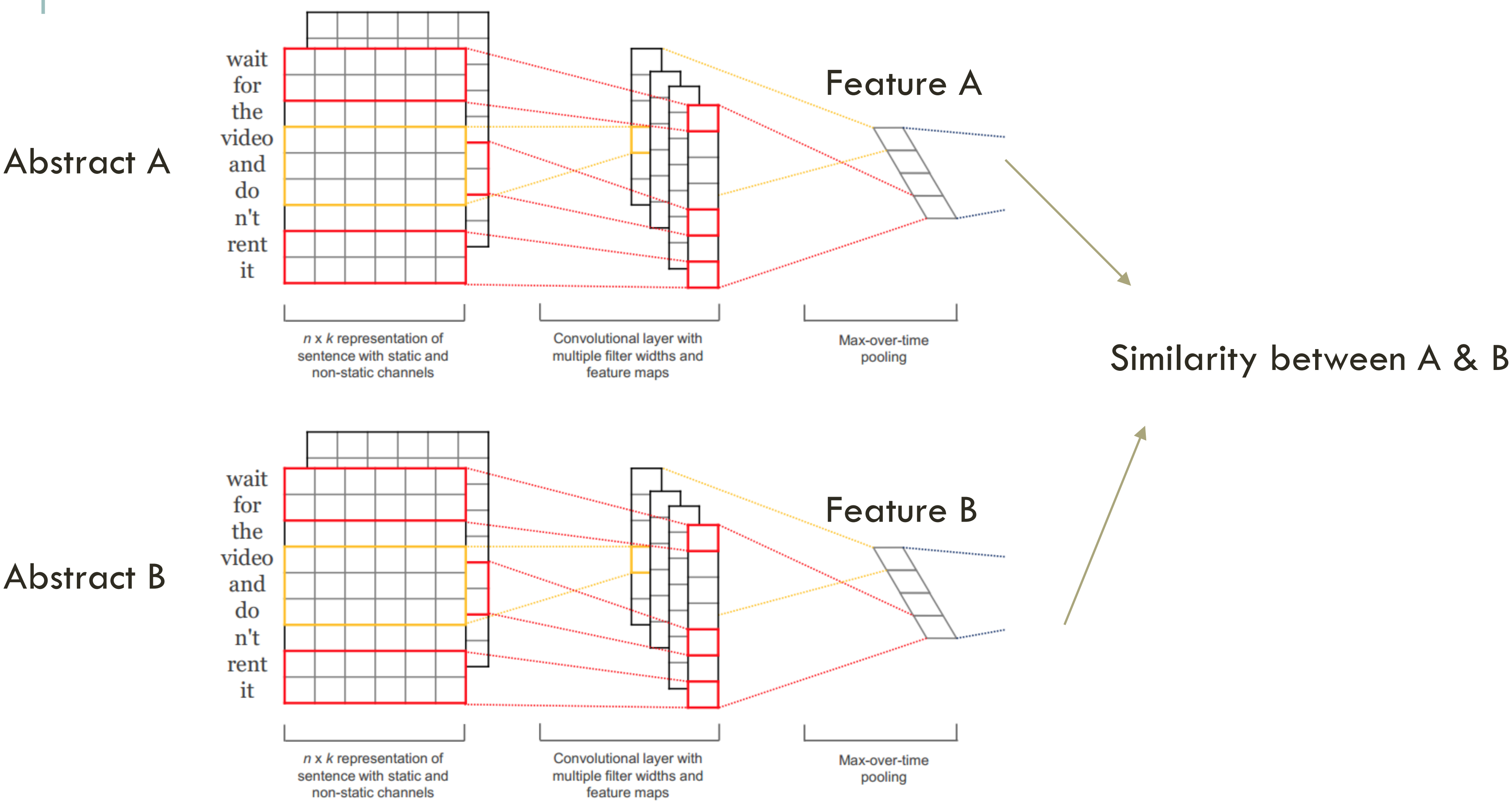
- ‘Title’ has more weight than ‘abstract’

Giving the top similar paper rating in the “Citation Matrix”



	Ciation1	Citation2	Citation3	Citation4	Citation5
Paper1	5	3	5	5	
Paper2		5		3	5
Paper3	5		5	5	5

TEXT-CNN-BASED CONTENT-BASED PREDICTOR



Using TextCNN to analyze the similarity of papers.

End.