

Tài liệu giảng dạy môn Khai thác dữ liệu Web

HỆ THỐNG TƯ VẤN

TS. Nguyễn Ngọc Thảo – ThS. Lê Ngọc Thành
Bộ môn Khoa học Máy tính, FIT HCMUS, VNUHCM

Thành phố Hồ Chí Minh, 02/2019

Nội dung bài giảng

- Bài toán tư vấn
- Tư vấn dựa trên nội dung
- Tư vấn dựa trên cộng tác
 - Lọc cộng tác sử dụng k-NN
 - Lọc cộng tác sử dụng luật kết hợp
 - Lọc cộng tác bằng Matrix Factorization

Hệ thống tư vấn

- Được sử dụng rộng rãi trên Web để tư vấn sản phẩm và dịch vụ đến người dùng.

Frequently Bought Together

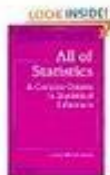


Price For All Three: \$258.02

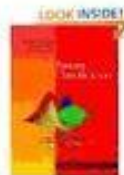
Add all three to Cart

- ☒ **This item:** The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) by Trevor Hastie
- ☒ [Pattern Recognition and Machine Learning \(Information Science and Statistics\)](#) by Christopher M. Bishop
- ☒ [Pattern Classification \(2nd Edition\)](#) by Richard O. Duda

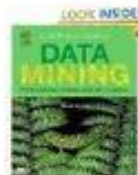
Customers Who Bought This Item Also Bought



[All of Statistics: A Concise Course in Statist...](#) by Larry Wasserman
★★★★☆ (8) \$60.00



[Pattern Classification \(2nd Edition\)](#) by Richard O. Duda
★★★★☆ (27) \$117.25



[Data Mining: Practical Machine Learning Tools an...](#) by Ian H. Witten
★★★★☆ (29) \$41.55



[Bayesian Data Analysis, Second Edition \(Texts in...](#) by Andrew Gelman
★★★★☆ (10) \$56.20



[Data Analysis Using Regression and Multilevel /...](#) by Andrew Gelman
★★★★☆ (13) \$39.59

Hệ thống tư vấn

- Hầu hết trang dịch vụ trực tuyến đều hỗ trợ hình thức này.



The New York Times



Find Movies, TV shows, Celebrities and more...

All

Movies, TV
& Showtimes

Celebs, Events
& Photos

News &
Community

Watchlist

FULL CAST AND CREW

TRIVIA

USER REVIEWS

IMDbPro

MORE

SHARE

+ Beauty and the Beast (2017)

★ 7.7
97,890

☆ Rate
This

PG | 2h 9min | Family, Fantasy, Musical | 17 March 2017 (USA)



People who liked this also liked...

[Learn more](#)



Add to Watchlist

Next »



◀ Prev 6 Next 6 ▶

Moana I (2016)

PG Animation | Adventure | Comedy

★★★★★★★ 7.7/10

In Ancient Polynesia, when a terrible curse incurred by the Demigod Maui reaches an impetuous Chieftain's daughter's island, she answers the Ocean's call to seek out the Demigod to set things right.

Directors: Ron Clements, Don Hall, ...
Stars: Auli'i Cravalho, Dwayne Johnson

Amazon.com <vfe-campaign-response@amazon.com> [Unsubscribe](#)
to me ▾

Oct 23 (11 days ago) ☆



[Your Amazon.com](#) | [Today's Deals](#) | [See All Departments](#)

Ross Beard,

Are you looking for something in our Team Sports department? If so, you might be interested in these items.

Team Sports



[Champion 6.1 oz. Cotton Jersey Shorts 6.1 oz. Cotton Jersey Champion](#)

Price: **\$5.65 - \$30.51**

[Learn more](#)

[Add to Wish List](#)

Dwight, Welcome to Your Amazon.com (If you're not Dwight K Schrute, click here.)



Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#).

Page 1 of 44



Guard Alaska™ Bear Defense Spray

★★★★☆ (8) **\$35.00**

[Fix this recommendation](#)



Pickled Beets, Sliced
by Barry Farm

★★★★☆ (1) **\$4.49**

[Fix this recommendation](#)



Battlestar Galactica - Season One

★★★★☆ (553) **\$34.99**

[Fix this recommendation](#)



Reebok 65cm Stability Ball
by Reebok

★★★★☆ (8) **\$18.78**

[Fix this recommendation](#)





MARKET SNAPSHOT

U.S.	EUROPE	ASIA
DJIA	13,342.50	+19.11 0.14%
S&P 500	1,436.88	+3.32 0.23%
NASDAQ	3,110.06	+5.53 0.18%

Canon P-215 *Scan-tini* mobile document scanner

buy now
* get a free carrying case
promotion ends 12/31/12

Bloomberg

Our Company | Professional | Anywhere

Search News, Quotes and Opinion

HOME QUICK **NEWS** OPINION MARKET DATA PERSONAL FINANCE TECH POLITICS SUSTAINABILITY TV VIDEO RADIO

U.S. Stocks Rise as Investors Weigh Stimulus Prospects

By Inyoung Hwang - Sep 12, 2012 11:14 AM ET

f t in 7 COMMENTS

+ QUEUE

Get The Market Now newsletter. Learn more >

SUBSCRIBE

HEADLINES MOST POPULAR **RECOMMENDED**

Based on your reading history you may be interested in:

Glenn Beck Returns to Television With Dish Network Agreement



The Top Ten Stocks for Wednesday, September 12

U.S. [stocks](#) rose, with benchmark indexes trading near four-year highs, as a German court cleared the way for Europe's bailout fund and investors weighed prospects for stimulus measures from the [Federal Reserve](#).

JPMorgan Chase & Co. (JPM) and Travelers Cos. rose at least 1 percent, pacing gains among financial companies. PulteGroup Inc. (PHM) advanced 8.1 percent as homebuilders rallied.

U.S. Stocks Reverse Gains as Technology Shares Retreat

Euro, Stocks Climb as German Court Approves Bailout Funds

European Stocks Rise to 14-Month High After German Ruling

Ben & Jerry's Sues Porn Seller Over Flavor-Tied Titles

U.S. Stocks Extend Gain as France Said to Press Spain

Advertisement





Recommended Pages

Get updates from your favorite businesses and brands.

+ Create Page

Search for Pages...

Recommended for You



Santa Monica Museum of Art

Museum

9,172 likes · 47 people talking about this

✓ Liked



Scott Kelby

Author

44,123 likes · 1,487 people talking about this



Brian Solis

Author

19,881 likes · 127 people talking about this



The Oatmeal

Entertainment Website

814,877 likes · 17,333 people talking about this



Slideshare

App Page

138,274 likes · 1,906 people talking about this

Home

People You May Know

12 mutual friends

Add FriendRemove

Add FriendRemove

Add FriendRemove

Add FriendRemove

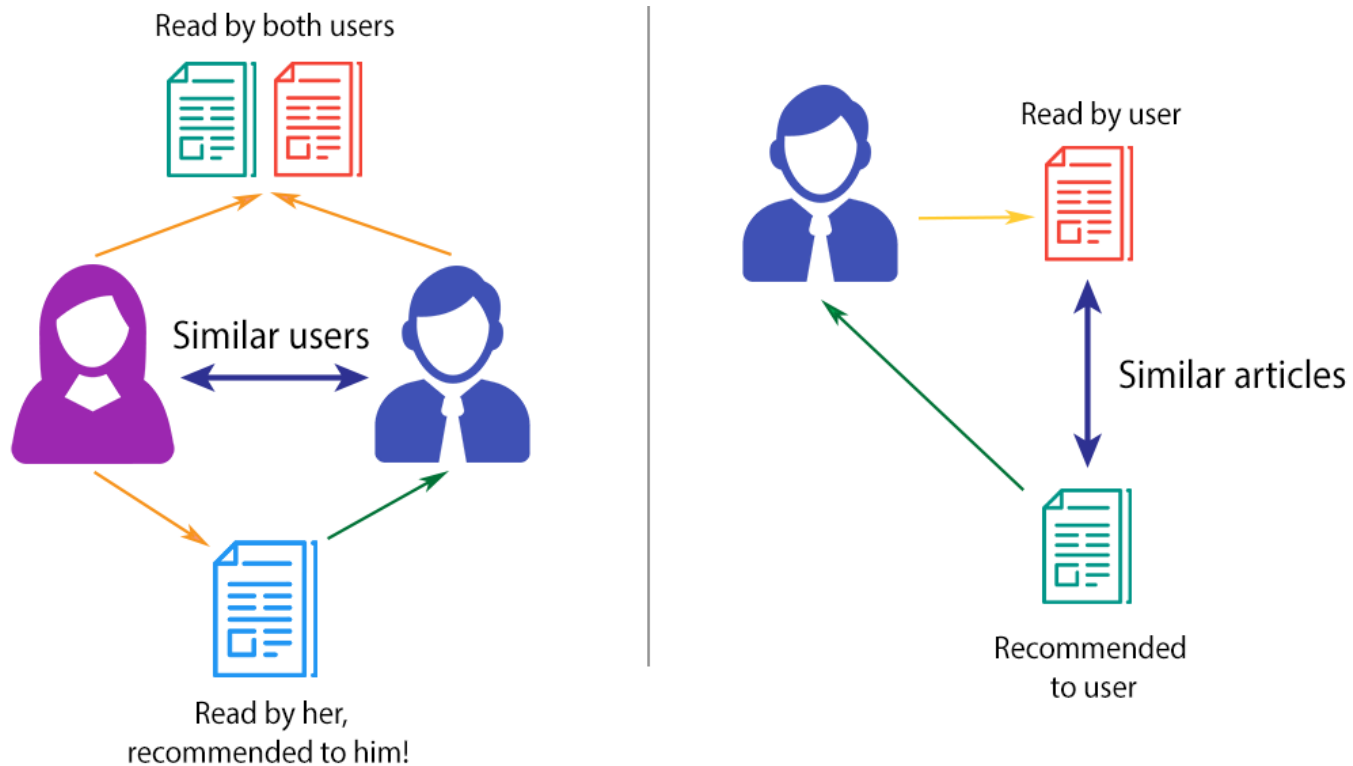
Add FriendRemove

Engaged

uzz! DE...

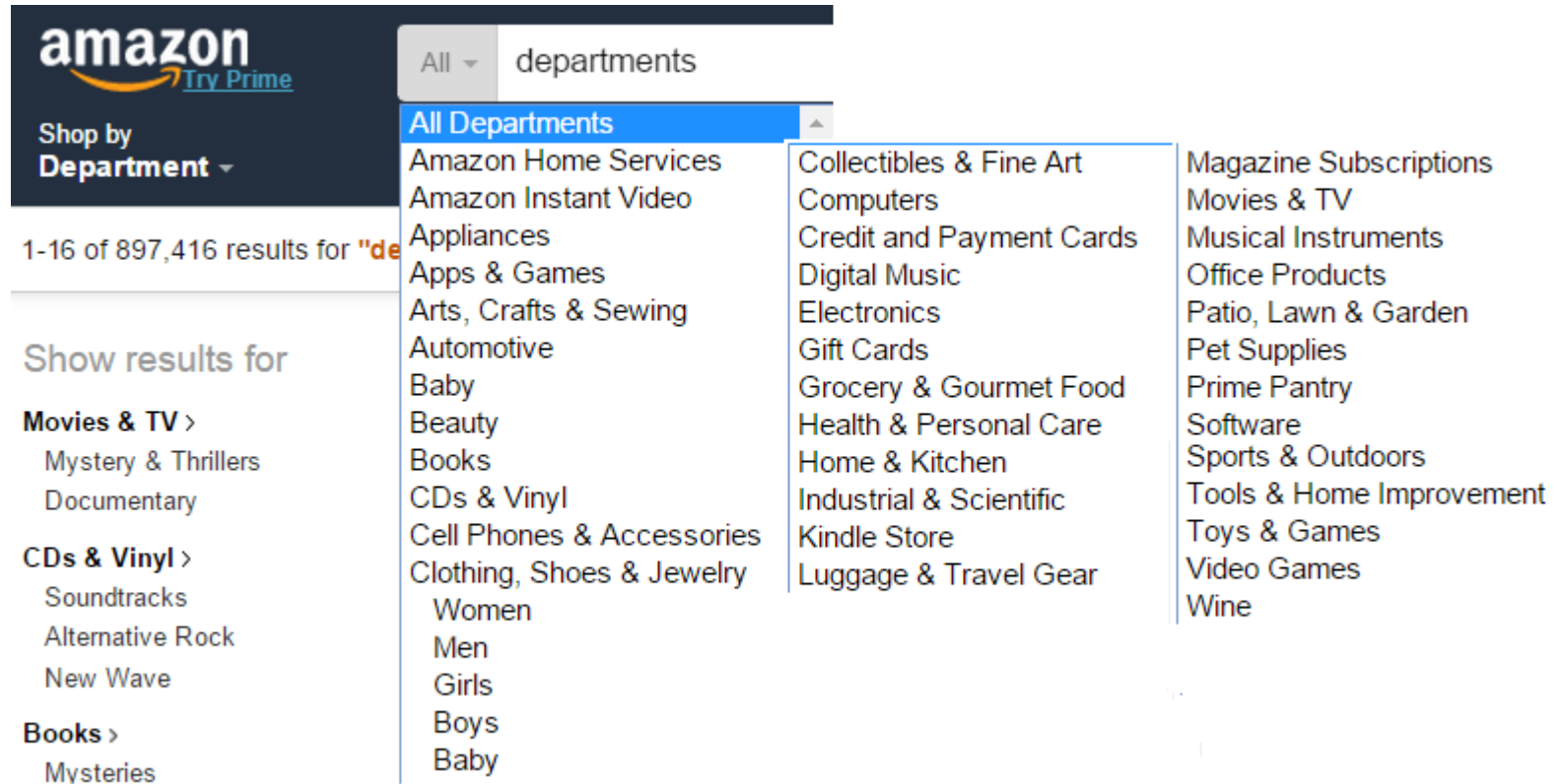
Vai trò của hệ thống tư vấn

- Hỗ trợ người dùng đối phó với lượng thông tin quá tải bằng cách đưa ra lời tư vấn cá nhân hóa.



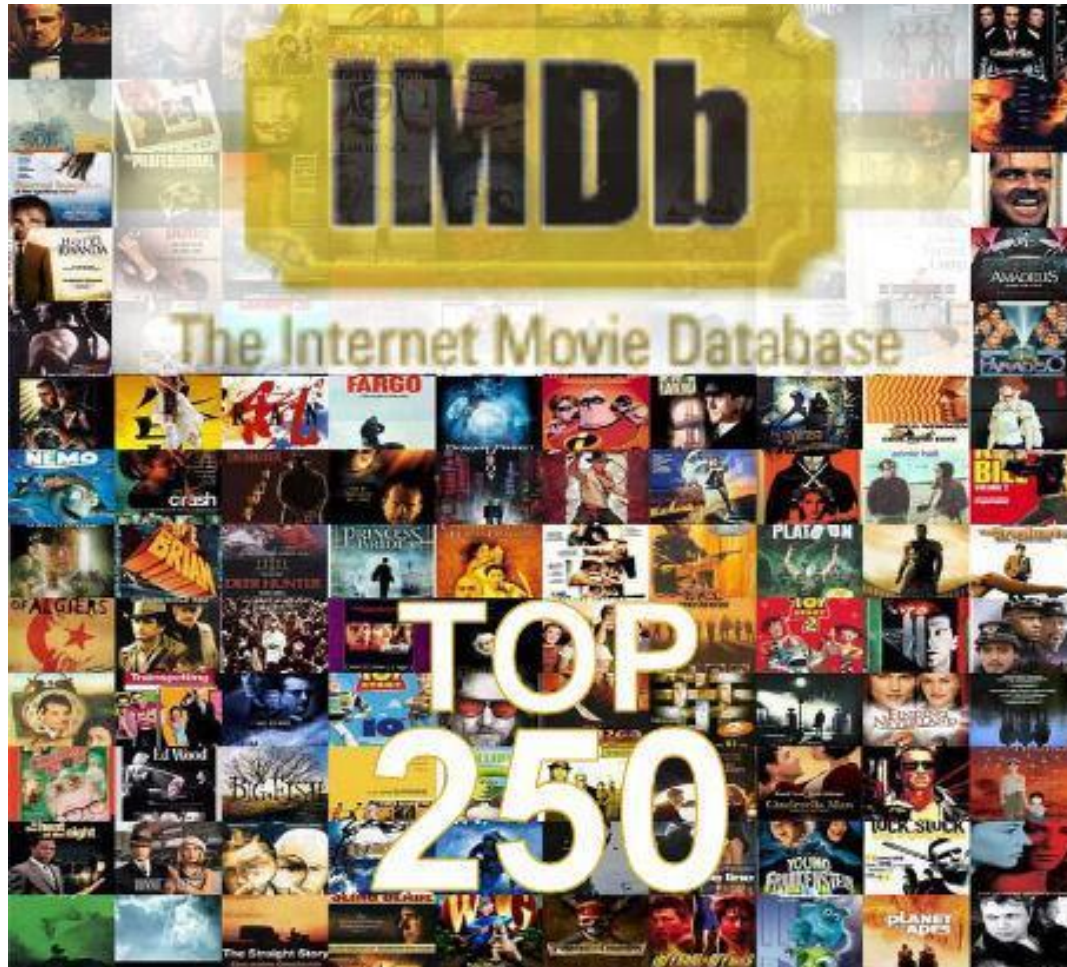
- Từ đó đem lại nhiều lợi nhuận hơn cho việc kinh doanh.

Sự đa dạng của sản phẩm



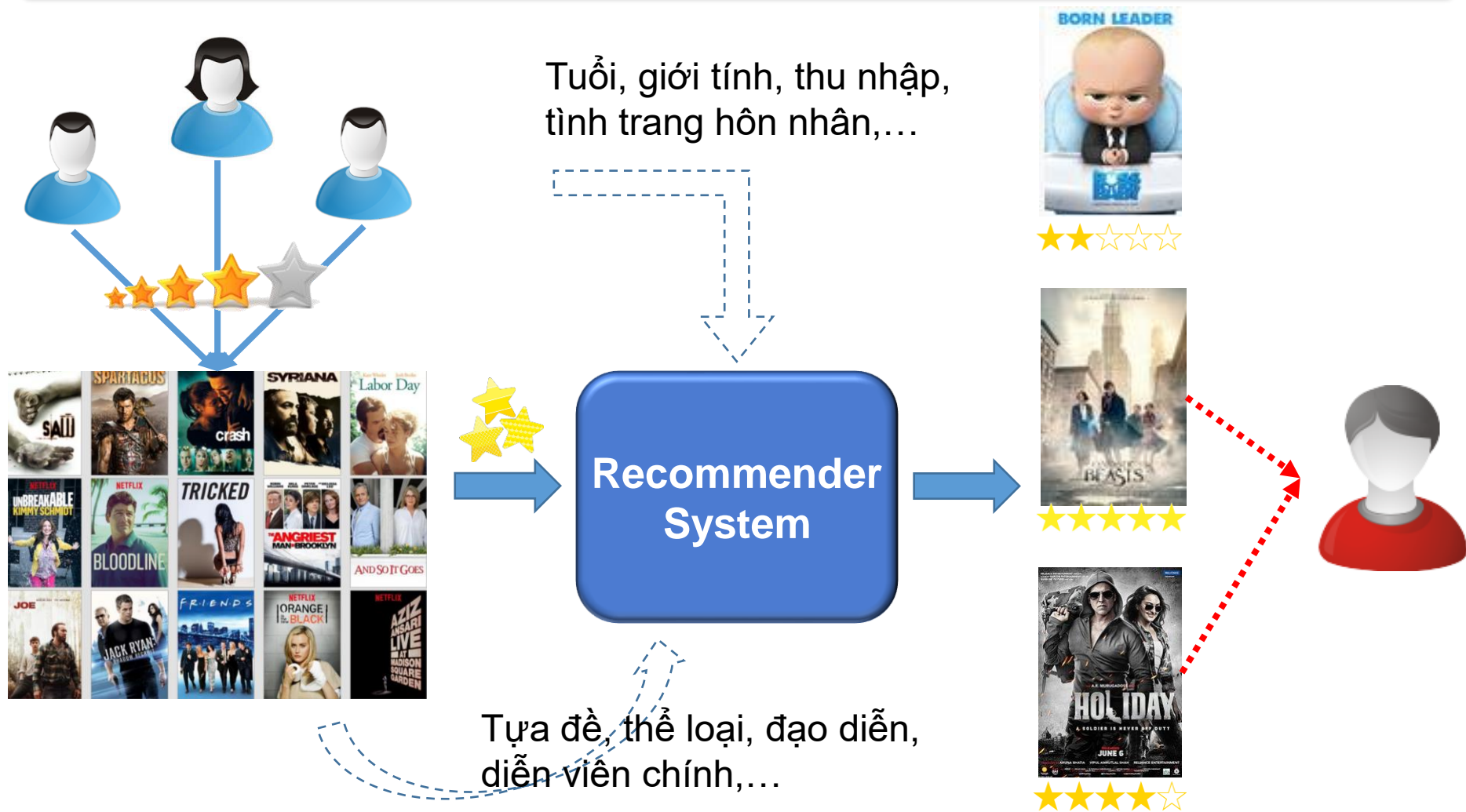
Amazon USA kinh doanh trên **480 triệu mặt hàng**, được chia thành vài chục ngành hàng khác nhau (12/2015). Trung bình mỗi ngày có khoảng 485 ngàn mặt hàng mới.

Sự đa dạng của sản phẩm



IMDB có khoảng
4.2 triệu tựa phim,
7.8 triệu nhân vật
trong cơ sở dữ liệu và
75 triệu người dùng
có đăng ký.

Hệ thống tư vấn phim



Một tập người dùng đánh giá một tập con trong cơ sở dữ liệu phim

Hệ thống dự đoán điểm cho bộ phim chưa được người dùng đánh giá để đưa ra lời tư vấn



Bài toán tư vấn

Khái niệm cơ bản

- Cho **tập người dùng** U và **tập hạng mục** S được tư vấn cho người dùng.
- Mỗi **người dùng** $u \in U$ sở hữu một **user profile**.
 - UserID, ngoài ra có thể bao gồm tuổi, giới tính, thu nhập, tình trạng hôn nhân, sở thích, nhu cầu,...
- Mỗi **hạng mục** $s \in S$ cũng được định nghĩa bằng một tập thuộc tính.
 - Ví dụ, phim được biểu diễn bằng MovieID, ngoài ra còn có tiêu đề, thể loại, đạo diễn, năm phát hành, diễn viên chính,...
- U và S thường rất lớn trong hầu hết các ứng dụng.

Nhiệm vụ của hệ thống tư vấn

- Gọi p là hàm lợi ích đo lường độ hữu dụng của hạng mục s đối với người dùng u , tức là $p: U \times S \rightarrow R$.
 - R là tập có thứ tự toàn phần (ví dụ, số nguyên không âm hay số thực trong khoảng nhất định).
- Nhiệm vụ của hệ thống tư vấn là **học hàm lợi ích p** .
 - Hàm mục tiêu để học p phụ thuộc ứng dụng, ví dụ, độ hài lòng của người dùng hay mức lãi của bên bán hàng.
- Sau đó, dùng p dự đoán giá trị lợi ích của mỗi hạng mục s ($\in S$) đối với mỗi người dùng u ($\in U$) rồi **tư vấn top- k hạng mục cho u** .
 - Ngoại trừ hạng mục đã có giá trị lợi ích đối với u từ dữ liệu đầu vào.

Nhiệm vụ của hệ thống tư vấn

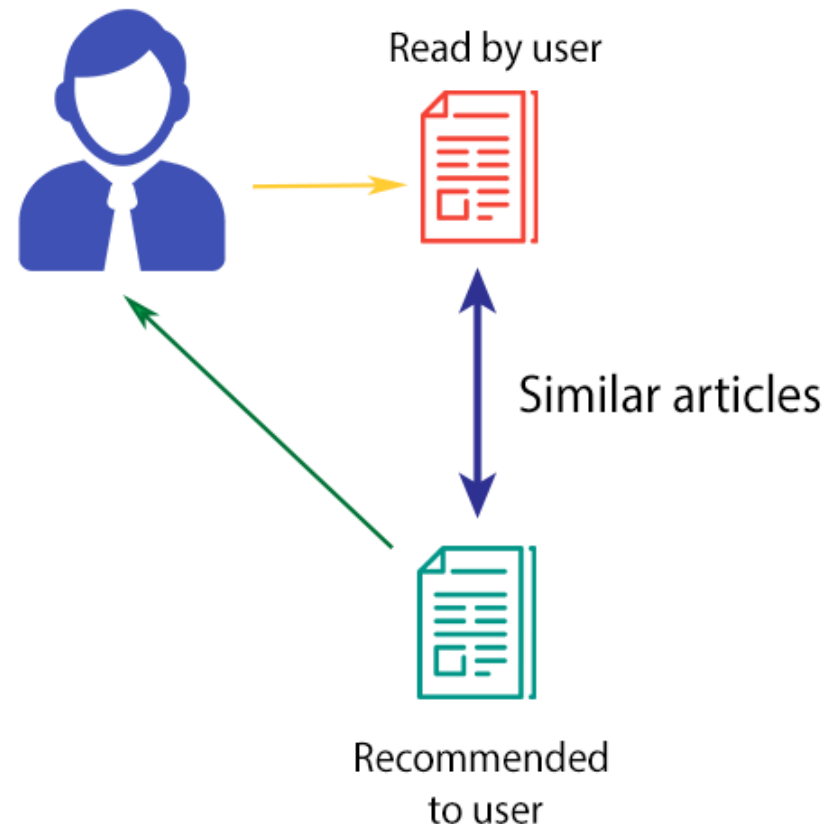
- **Dự đoán điểm đánh giá** của người dùng cho một sản phẩm mà người này chưa từng tiếp cận trước đó.
 - Ví dụ, cho điểm đánh giá về một bộ phim chưa xem.
 - Độ lợi ích của hạng mục s đối với người dùng u là điểm đánh giá của u về s .
- **Dự đoán hạng mục** mà người dùng có khả năng sử dụng.
 - Sự tương tác giữa người dùng và hạng mục thường là nhị phân (ví dụ, mua hoặc không mua) hoặc đa phân (ví dụ, điểm đánh giá 1 – 5), nhưng không quan tâm giá trị cụ thể.
 - Độ lợi ích của hạng mục s đối với người dùng u được thể hiện bằng xác suất u sẽ mua hoặc sử dụng s .

Hướng tiếp cận của bài toán

Lọc dựa trên nội dung

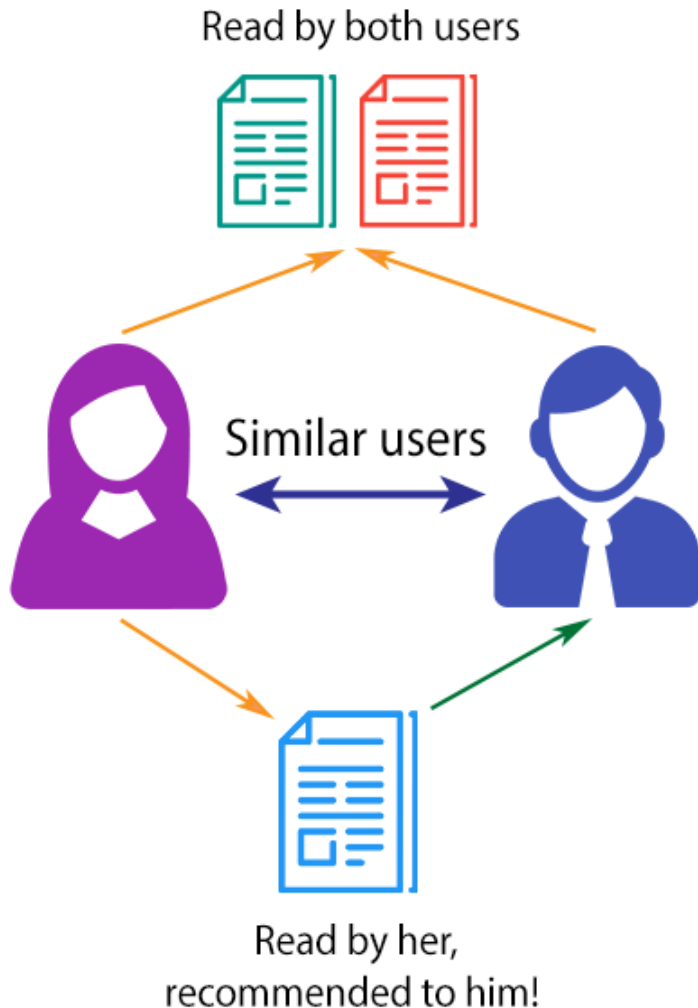
Người dùng được tư vấn các hạng mục tương tự với hạng mục mà người này quan tâm trước đây.

CONTENT-BASED FILTERING



Hướng tiếp cận của bài toán

COLLABORATIVE FILTERING

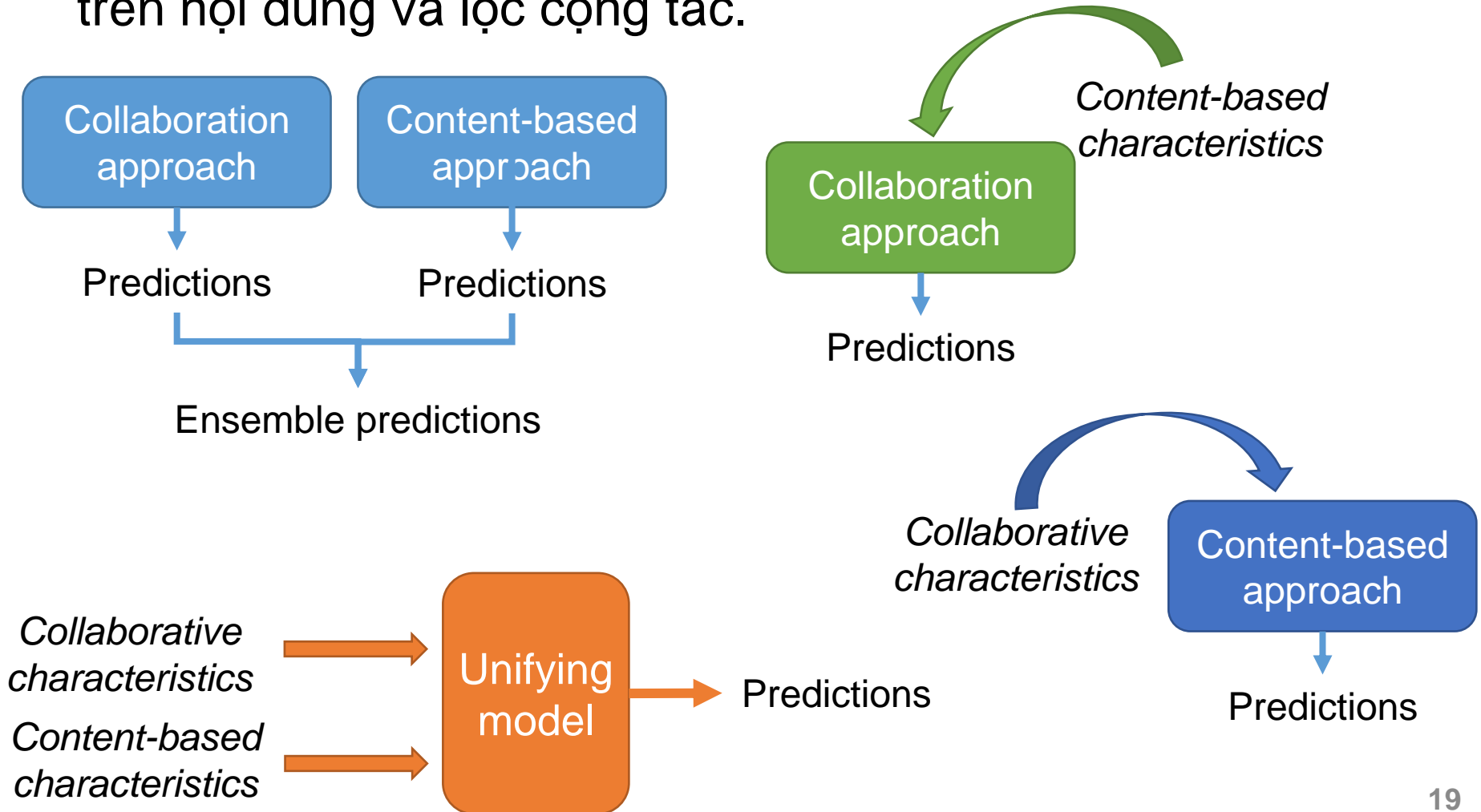


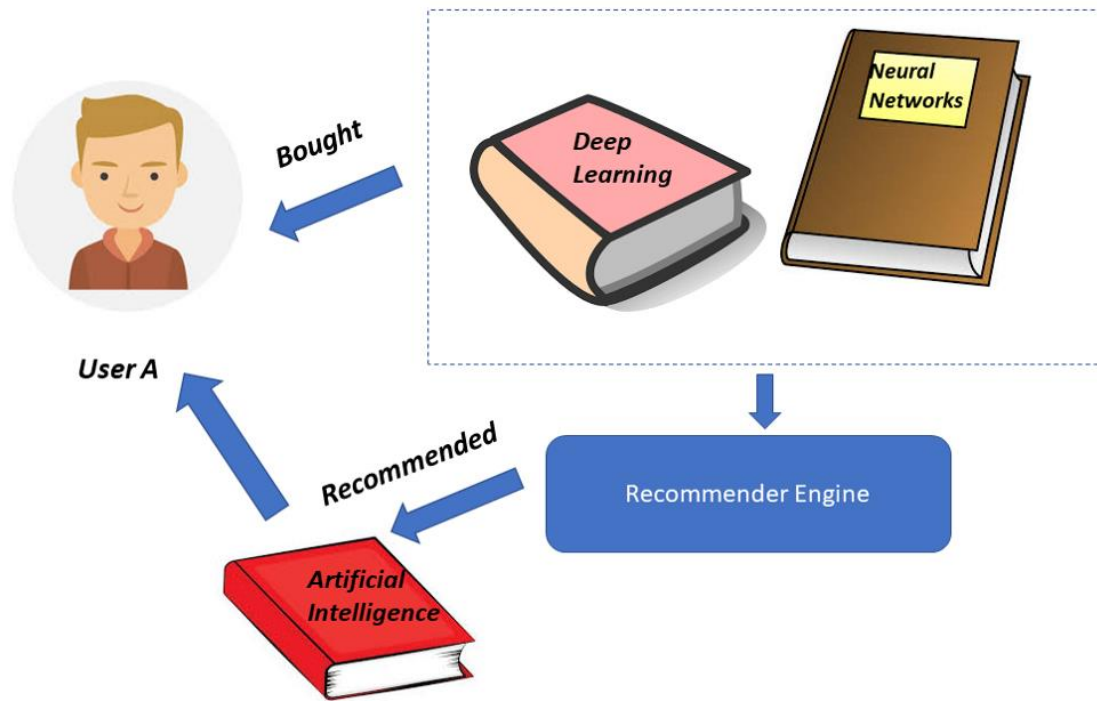
Lọc cộng tác

Người dùng được tư vấn các hạng mục mà những người dùng khác có cùng sở thích đã quan tâm trước đây.

Hướng tiếp cận của bài toán

- **Phương pháp lai** kết hợp cả hai hướng tiếp cận, lọc dựa trên nội dung và lọc cộng tác.

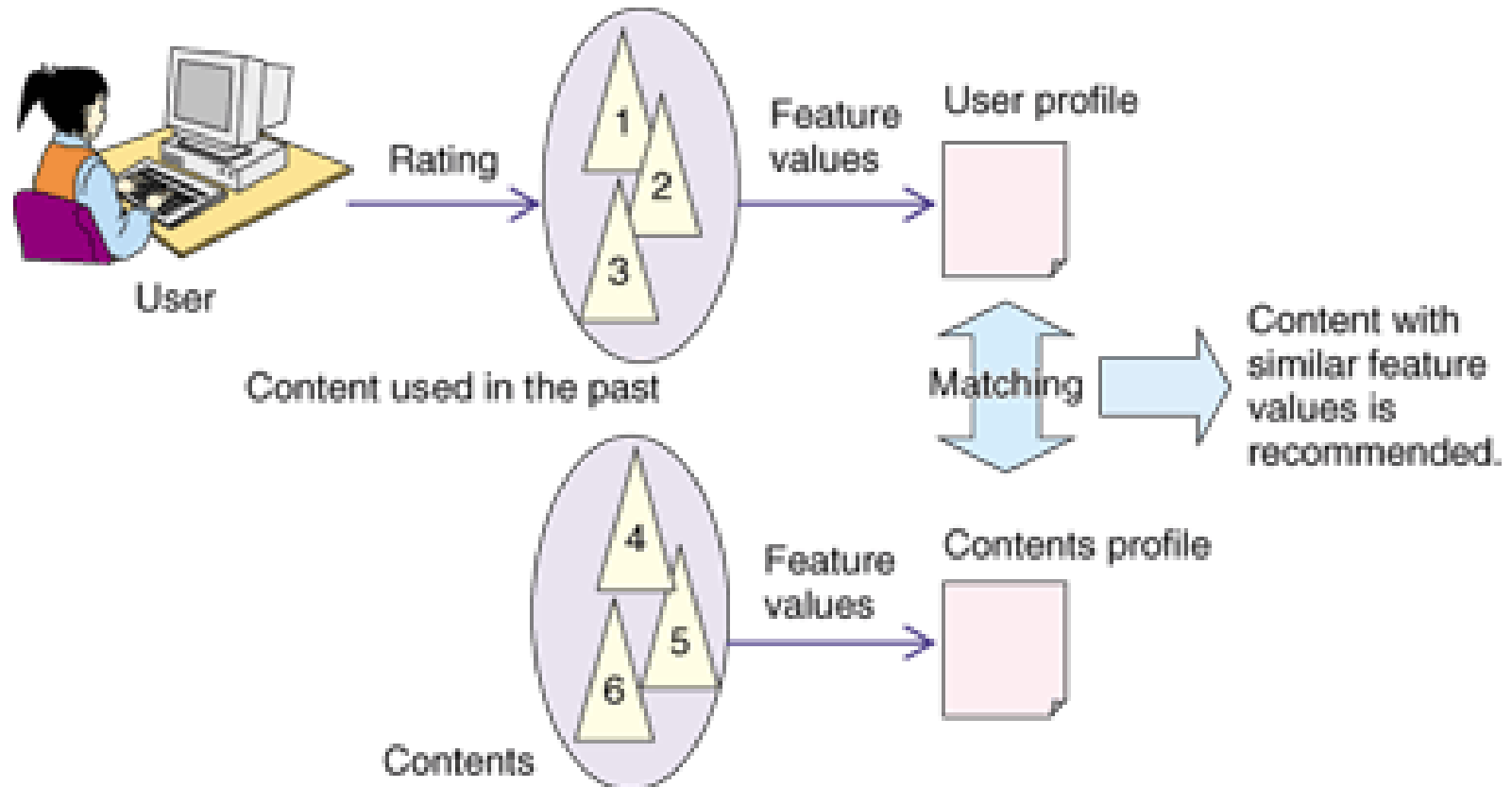




Tư vấn dựa trên nội dung

Tư vấn dựa trên nội dung

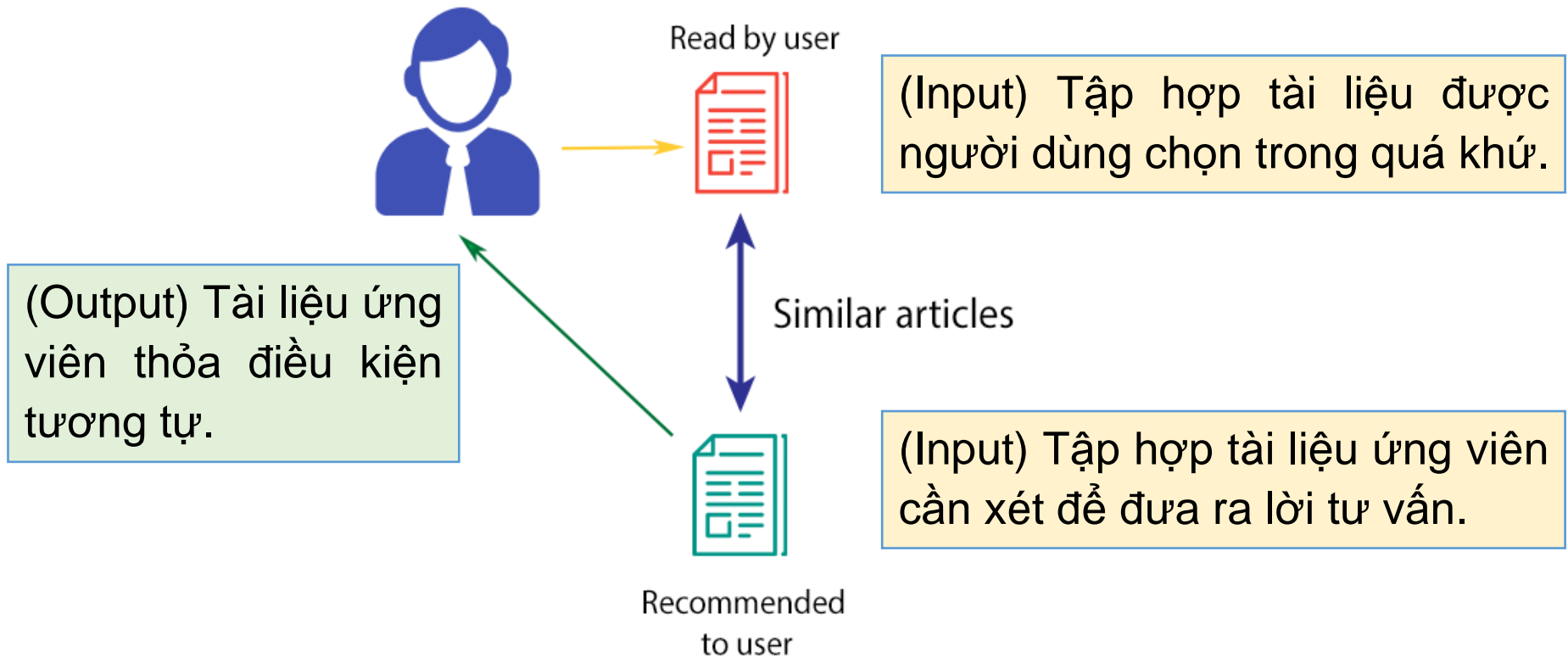
- Dự đoán độ lợi ích của hạng mục đối với một người dùng cụ thể dựa trên sự tương tự của hạng mục với những hạng mục mà người dùng đã chọn trong quá khứ.



Tư vấn dựa trên nội dung

- Mỗi hạng mục được biểu diễn bằng một **tập đặc trưng**.
 - Ví dụ, biểu diễn bộ phim bằng diễn viên, đạo diễn, thể loại, tình tiết chính,...
- Sở thích của người dùng cũng được biểu diễn bằng cùng tập đặc trưng, gọi là **user profile**.
 - Profile được hình thành một cách tường minh từ bản thăm dò ý kiến hoặc một cách hàm ý thông qua hành vi giao dịch theo thời gian.
- Hệ thống so sánh user profile với hạng mục ứng viên trên cùng tập đặc trưng để đưa ra tư vấn gồm **top-k** hạng mục trùng khớp nhất.

Tư vấn trong lĩnh vực văn bản



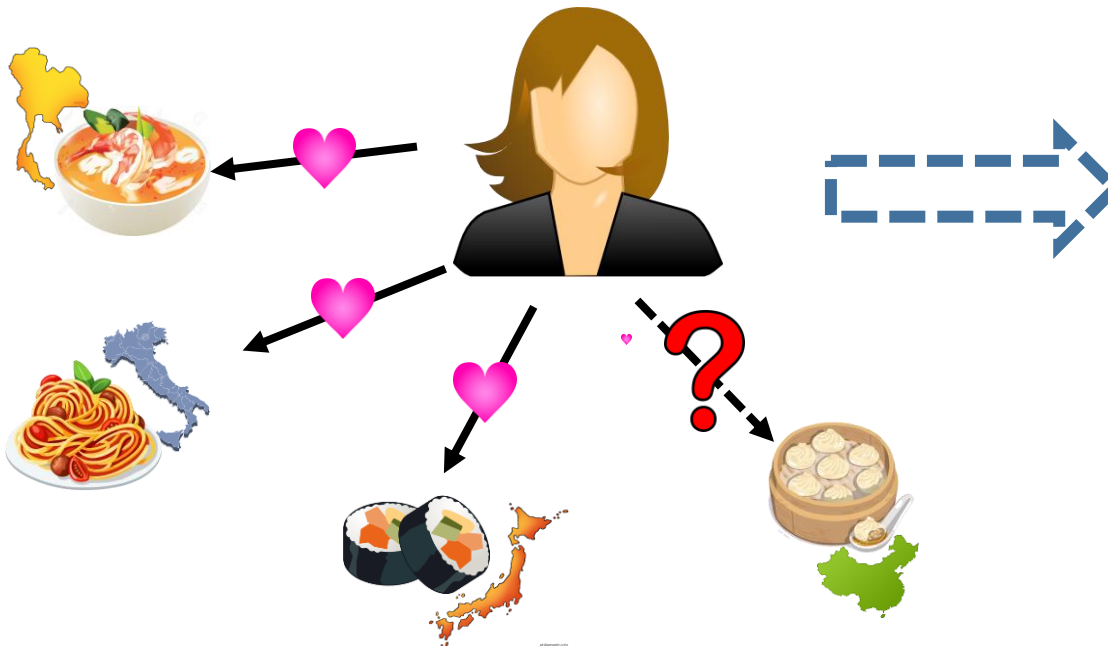
- Biểu diễn văn bản bằng các kỹ thuật truy vấn thông tin.
 - Ví dụ, mô hình không gian vector với lược đồ TF-IDF và độ đo cosine.
- User profile có thể được biểu diễn bằng vector trung bình (hoặc prototype) của các tài liệu liên quan, ví dụ, phương pháp Rocchio.

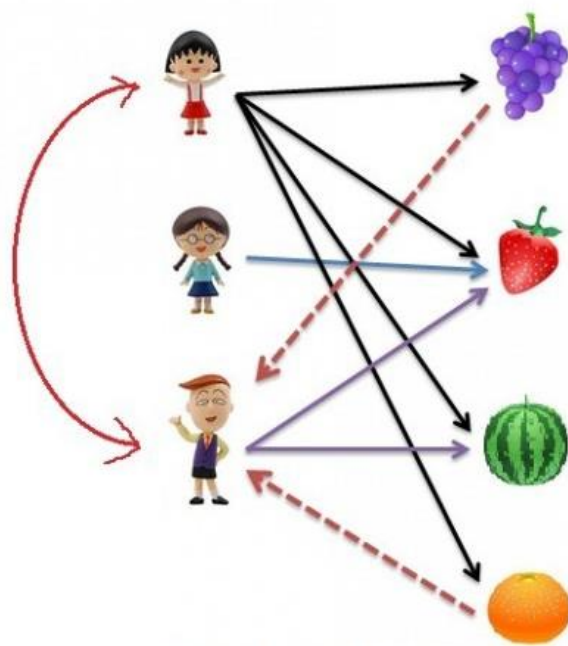
Tư vấn trong lĩnh vực văn bản

- Đôi lúc không thể tính user profile thành vector trung bình mà cần “học máy” trên các tài liệu liên quan mà người dùng cung cấp.
- **PU learning**: xem các tài liệu người dùng cung cấp thuộc lớp dương và tập tài liệu ứng viên là mẫu chưa gán nhãn.
 - EM+Naïve Bayesian, Co-training, Self-training,...
- **Supervised learning**: có thể áp dụng nếu người dùng cung cấp cả tài liệu liên quan và không liên quan.
- **LU learning**: nếu tập hợp liên quan và không liên quan nhỏ, áp dụng học bán giám sát.

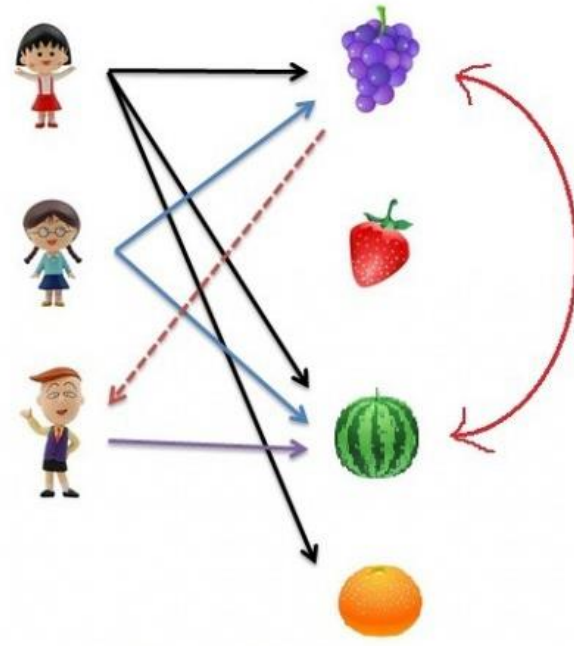
Tư vấn dựa trên nội dung: Khuyết điểm

- Không thể giới thiệu các hạng mục không tương tự với những hạng mục người dùng đã chọn trong quá khứ.
- Người dùng sẽ không bao giờ thấy được cái gì hoàn toàn mới lạ mà họ có thể quan tâm → đem lại ít lợi nhuận cho việc kinh doanh.





User-based filtering



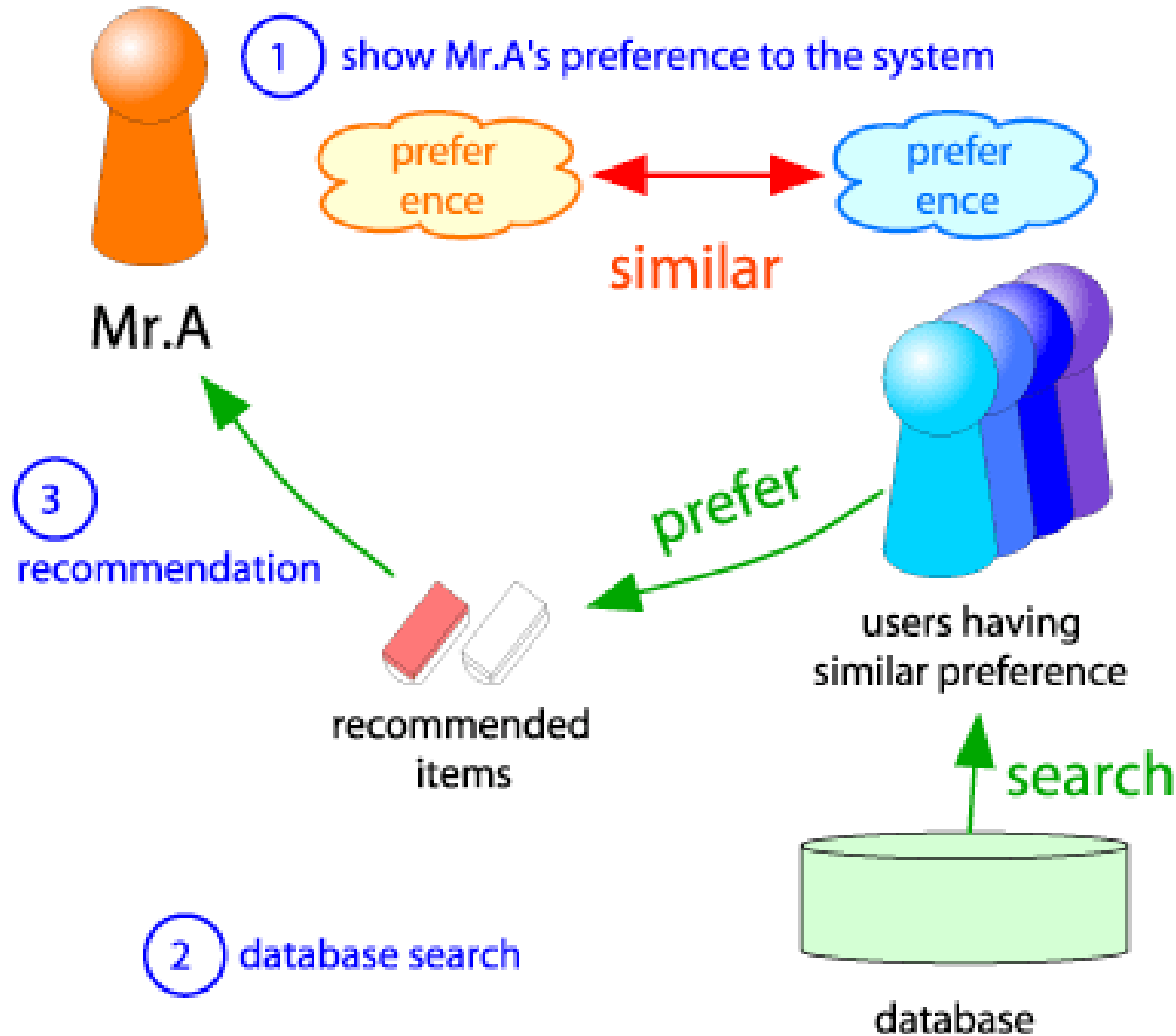
Item-based filtering

Tư vấn dựa trên cộng tác

Tư vấn dựa trên cộng tác

- Dự đoán độ lợi ích của hạng mục đối với một người dùng cụ thể dựa trên các hạng mục đã được đánh giá bởi những người dùng khác có sở thích tương tự.
- Sử dụng dữ liệu tương tác khách hàng – sản phẩm, bỏ qua thuộc tính của khách hàng và sản phẩm.
- Hướng tiếp cận tư vấn được nghiên cứu nhiều nhất và được áp dụng rộng rãi nhất.
- Từ khóa: **collaboration filtering**.

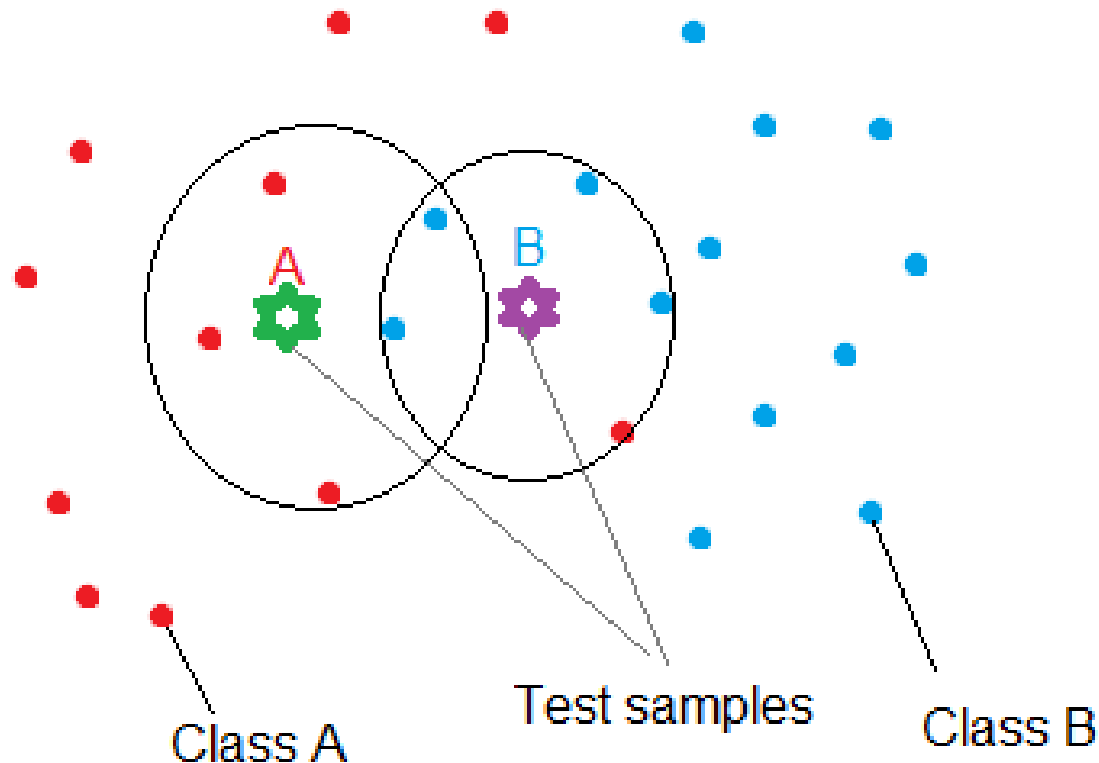
Tư vấn dựa trên cộng tác



Lọc cộng tác sử dụng k-NN

Phương pháp k -nearest neighbors

- k -nearest neighbors (k -NN) phát sinh dự đoán trực tiếp từ toàn bộ cơ sở dữ liệu người dùng – hạng mục mà **không** xây dựng mô hình.



Áp dụng k -NN cho lọc cộng tác

- Sử dụng bộ phân lớp k -NN để dự đoán điểm đánh giá hoặc tiềm năng mua sản phẩm của người dùng.
- Xác định độ tương quan giữa profile của người dùng mục tiêu và profile của những người dùng khác trong cơ sở dữ liệu để tìm người dùng có chung đặc điểm hoặc sở thích.
- Đưa ra dự đoán dựa trên việc kết hợp giá trị từ k người dùng gần nhất.
- Một quy trình tiêu biểu bao gồm hai pha chính: **hình thành vùng láng giềng** và **tư vấn**.

Pha 1: Hình thành vùng láng giềng

- So sánh profile của người dùng mục tiêu với tập profile theo lịch sử T của các người dùng khác để tìm ra top- k người dùng có chung sở thích.
- Dựa trên độ tương tự về điểm đánh giá sản phẩm, nội dung trang truy cập, hoặc sản phẩm được mua.



- Trong đa số ứng dụng lọc cộng tác, user profile thường là một tập điểm đánh giá cho một tập con sản phẩm.

Pha 1: Hình thành vùng láng giềng

- Gọi \mathbf{u} và \mathbf{v} ($\mathbf{v} \in T$) là vector biểu diễn profile của người dùng mục tiêu và một người dùng khác trong cơ sở dữ liệu.
- Độ tương tự giữa \mathbf{u} và \mathbf{v} được xác định bằng công thức **Pearson's correlation coefficient**

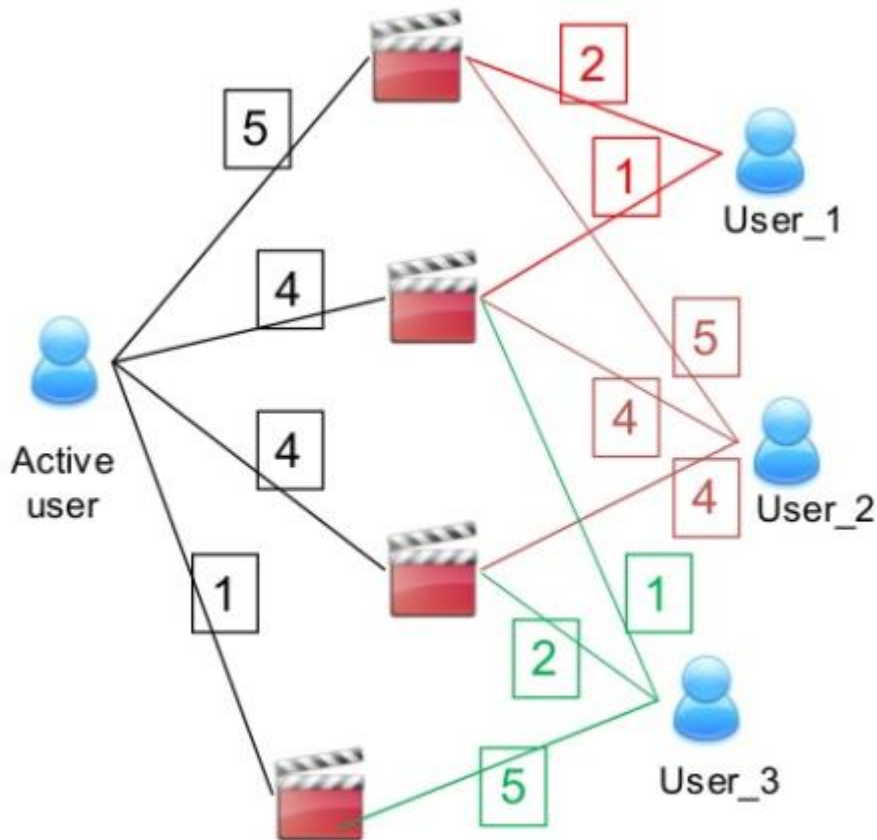
$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\sum_{i \in C} (r_{\mathbf{u},i} - \bar{r}_{\mathbf{u}})(r_{\mathbf{v},i} - \bar{r}_{\mathbf{v}})}{\sqrt{\sum_{i \in C} (r_{\mathbf{u},i} - \bar{r}_{\mathbf{u}})^2} \sqrt{\sum_{i \in C} (r_{\mathbf{v},i} - \bar{r}_{\mathbf{v}})^2}}$$

trong đó

- C là tập hạng mục được đánh giá bởi cả \mathbf{u} và \mathbf{v} .
- $r_{\mathbf{u},i}$ và $\bar{r}_{\mathbf{u}}$ là điểm đánh giá cho hạng mục i và điểm đánh giá trung bình của \mathbf{u} , tương tự cho $r_{\mathbf{v},i}$ và $\bar{r}_{\mathbf{v}}$ của \mathbf{v} .

Pha 1: Hình thành vùng láng giềng

$$sim(\mathbf{u}, \mathbf{v}) = \frac{\sum_{i \in C} (r_{\mathbf{u},i} - \bar{r}_{\mathbf{u}})(r_{\mathbf{v},i} - \bar{r}_{\mathbf{v}})}{\sqrt{\sum_{i \in C} (r_{\mathbf{u},i} - \bar{r}_{\mathbf{u}})^2} \sqrt{\sum_{i \in C} (r_{\mathbf{v},i} - \bar{r}_{\mathbf{v}})^2}}$$



	active user
user_1	0.4472136
user_2	0.49236596
user_3	-0.91520863

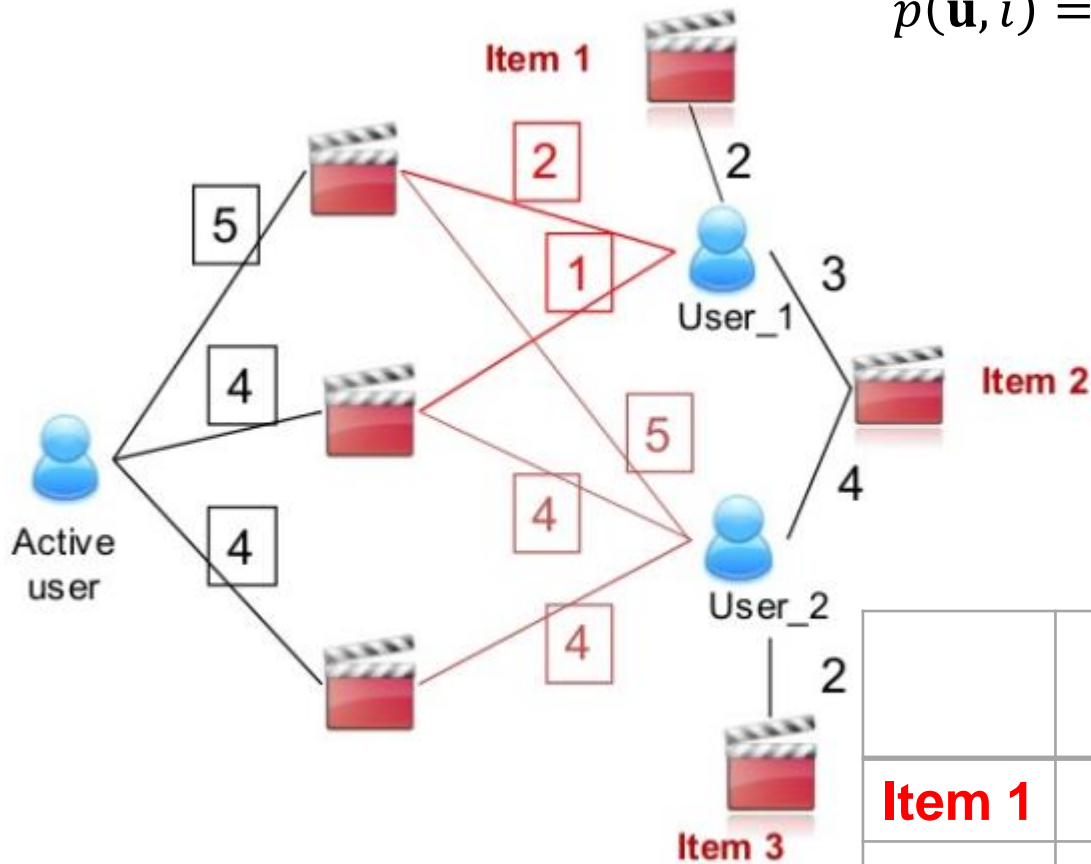
Pha 2: Tư vấn

- Dự đoán điểm đánh giá cho hạng mục i của người dùng \mathbf{u}

$$p(\mathbf{u}, i) = \bar{r}_{\mathbf{u}} + \frac{\sum_{\mathbf{v} \in V} sim(\mathbf{u}, \mathbf{v}) \times (r_{\mathbf{v}, i} - \bar{r}_{\mathbf{v}})}{\sum_{\mathbf{v} \in V} |sim(\mathbf{u}, \mathbf{v})|}$$

- Trong đó V là tập hợp k người dùng tương tự
- Chọn tư vấn cho \mathbf{u} những hạng mục có điểm đánh giá cao.

Pha 2: Tư vấn

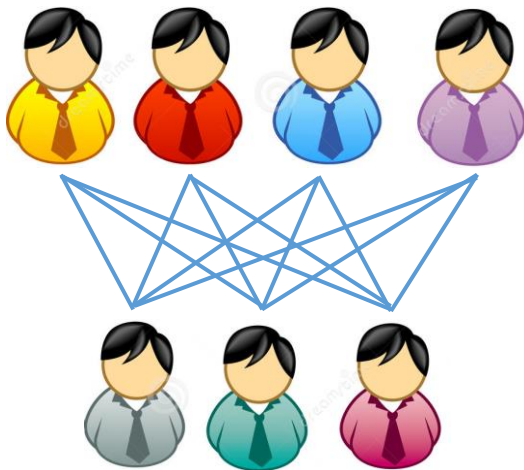


$$p(\mathbf{u}, i) = \bar{r}_{\mathbf{u}} + \frac{\sum_{\mathbf{v} \in V} \text{sim}(\mathbf{u}, \mathbf{v}) \times (r_{\mathbf{v}, i} - \bar{r}_{\mathbf{v}})}{\sum_{\mathbf{v} \in V} |\text{sim}(\mathbf{u}, \mathbf{v})|}$$

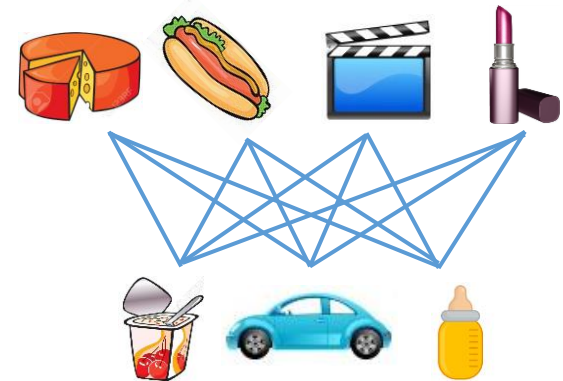
	User 1	User 2	predicted interest
Item 1	2	—	3.737986018
Item 2	3	4	4.039282065
Item 3	—	2	2.277268083

Lọc cộng tác dựa trên hạng mục

- Lọc cộng tác dựa trên người dùng **thiếu tính mở rộng**: làm thế nào để so sánh người dùng mục tiêu với những người dùng khác trong **thời gian thực** để phát sinh lời tư vấn.



VS.



- Trong khi đó, lọc cộng tác dựa trên hạng mục có thể **tính trước độ tương tự** của mọi cặp hạng mục.

Lọc cộng tác dựa trên hạng mục

- So sánh hạng mục theo mẫu đánh giá của mọi người dùng.
- Xác định độ tương tự của các cặp hạng mục được cùng đánh giá bởi những người dùng khác nhau.

$$\text{sim}(i, j) = \frac{\sum_{\mathbf{u} \in U} (r_{\mathbf{u},i} - \bar{r}_{\mathbf{u}})(r_{\mathbf{u},j} - \bar{r}_{\mathbf{u}})}{\sqrt{\sum_{\mathbf{u} \in U} (r_{\mathbf{u},i} - \bar{r}_{\mathbf{u}})^2} \sqrt{\sum_{\mathbf{u} \in U} (r_{\mathbf{u},j} - \bar{r}_{\mathbf{u}})^2}}$$

trong đó

- U là tập hợp gồm mọi người dùng, i và j là các hạng mục.
- $r_{\mathbf{u},i}$ và $\bar{r}_{\mathbf{u}}$ là điểm đánh giá cho hạng mục i và điểm đánh giá trung bình của người dùng $\mathbf{u} \in U$.
- Chọn k hạng mục tương tự nhất với hạng mục mục tiêu i .

Lọc cộng tác dựa trên hạng mục

- Dự đoán điểm đánh giá của người dùng u đối với hạng mục mục tiêu i như sau

$$p(\mathbf{u}, i) = \frac{\sum_{j \in J} r_{\mathbf{u}, j} \times \text{sim}(i, j)}{\sum_{j \in J} \text{sim}(i, j)}$$

trong đó

- J là tập gồm k hạng mục tương tự với hạng mục i .
- Những hạng mục có độ tương tự âm đối với i thường được loại bỏ.
- Sử dụng chính điểm đánh giá của người dùng cho các hạng mục tương tự để ngoại suy dự đoán cho hạng mục mục tiêu.

Áp dụng k -NN cho lọc cộng tác

- k -NN dựa trên người dùng hay dựa trên sản phẩm đều gặp vấn đề số chiều lớn \rightarrow trong thực tế không thể tính độ tương tự giữa các cặp người dùng hay hạng mục.
- Áp dụng các kỹ thuật giảm chiều để thu gọn kích thước user profile và item profile.
 - Chiếu ma trận người dùng - hạng mục vào không gian nhỏ hơn, ví dụ, Principal Component Analysis (PCA).
 - Factorize ma trận người dùng – hạng mục thành các biểu diễn thứ hạng thấp hơn về người dùng (hoặc hạng mục), ví dụ Singular Value Decomposition (SVD).
- Nhận diện đối tượng tương tự trong không gian con.

Lọc cộng tác sử dụng LKH

Áp dụng LKH cho lọc cộng tác

- Các hạng mục được mỗi người dùng thanh toán có thể được xem như một giao dịch.
- Trong lọc cộng tác, luật $X \rightarrow Y$ có **tiền đề chứa một hay nhiều hạng mục** trong khi **hệ quả chỉ có một hạng mục**.

Users
who
watched

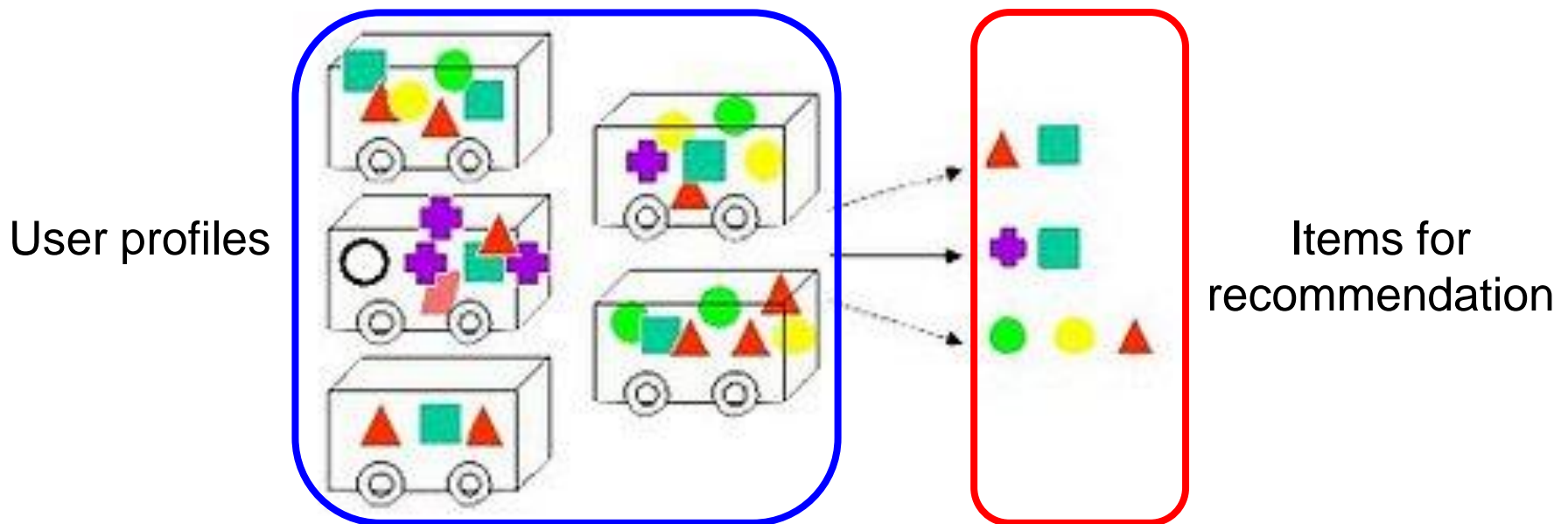


may
also
watch



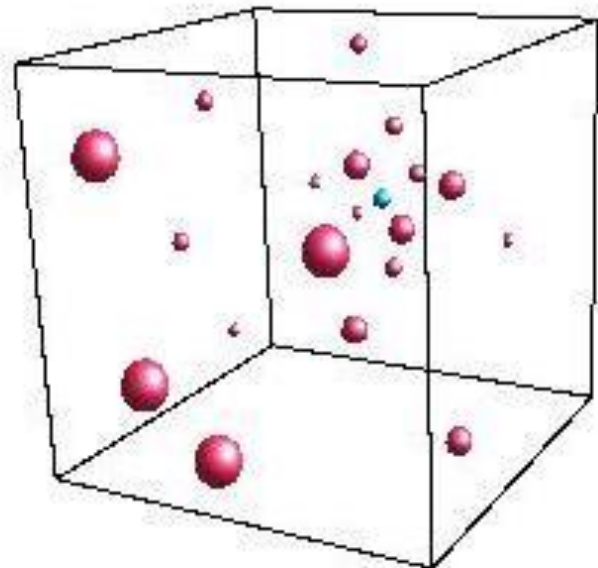
Áp dụng LKH cho lọc cộng tác

- Sở thích của người dùng mục tiêu được so khớp với các hạng mục trên vế trái X .
- Hạng mục trên vế phải của các luật thỏa được sắp theo độ tin cậy và chọn **top N hạng mục** để tư vấn cho người dùng.



Vấn đề gặp phải với LKH

- Các hệ thống tư vấn sử dụng luật kết hợp gặp khó khăn khi cơ sở dữ liệu thưa.
- Hầu hết người dùng chỉ xem (hoặc đánh giá) một tỉ lệ rất nhỏ trong số các hạng mục sẵn có, do đó khó tìm đủ số hạng mục chung giữa các user profile.



Giải pháp của Sarwar et al. 2000.

- Làm nhẹ vấn đề bằng một số kỹ thuật giảm chiều cơ bản.
- Khuyết điểm: một vài hạng mục hữu dụng hoặc thú vị có thể bị loại bỏ và do đó không thể xuất hiện trong mẫu kết quả.
- Ref: Sarwar, B., G. Karypis, J. Konstan, and J. Riedl. Application of Dimensionality Reduction in Recommender Systems: a case study. In Proceedings of WebKDD Workshop at the ACM SIGKDD, 2000.

Giải pháp của Fu et al. 2000.

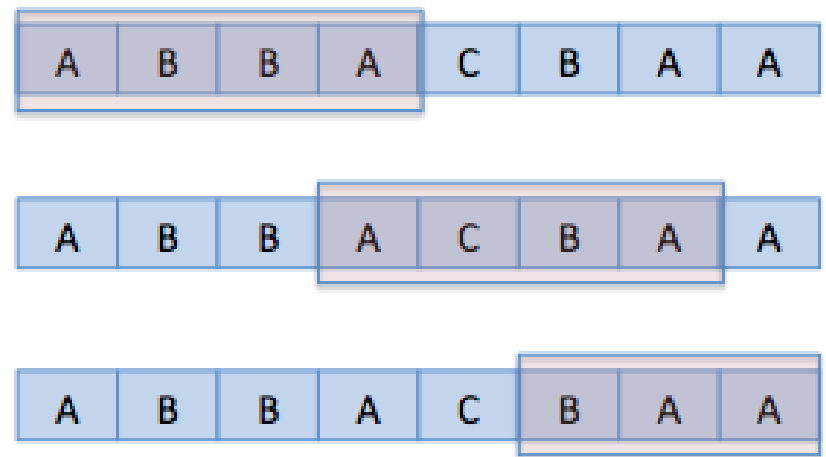
- Giải pháp 1: xếp hạng mọi luật tìm được theo độ giao giữa tiền đề của luật và profile của người dùng mục tiêu rồi phát sinh top- k tư vấn.
- Giải pháp 2: tìm các người dùng “láng giềng gần” có sở thích tương tự với người dùng mục tiêu và đưa ra tư vấn dựa trên lịch sử của những người này.
- Ref: Fu, X., J. Budzik, and K. Hammond. Mining navigation history for recommendation. In Proceedings of Intl. Conf. on Intelligent User Interfaces, 2000.

Giải pháp của Lin et al.

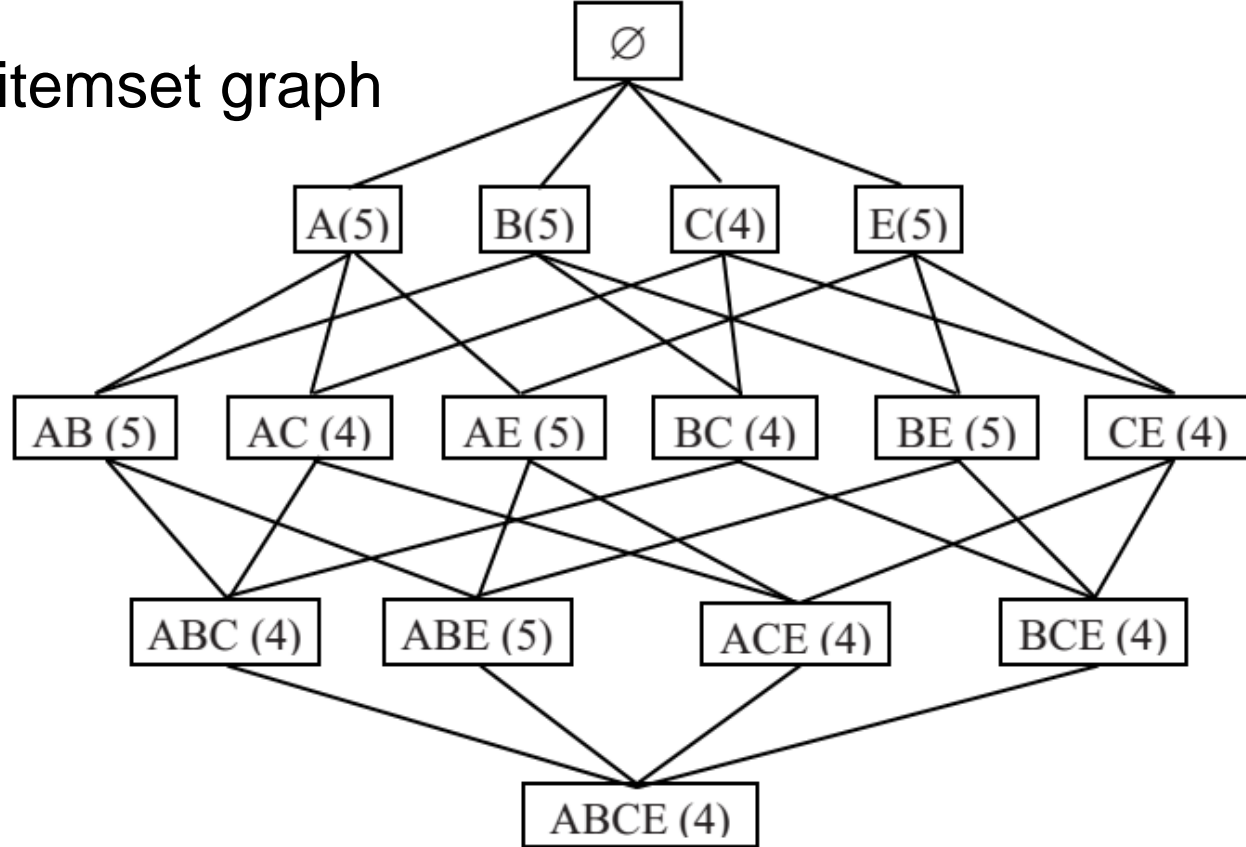
- Phát sinh luật kết hợp giữa các người dùng (user associations) và giữa các hạng mục (item associations).
- Tự động chọn minimum support để xác định đủ lượng luật cho mỗi người dùng mục tiêu.
- Nếu user minimum support lớn hơn ngưỡng, hệ thống sẽ tư vấn dựa trên user associations, ngược lại, sử dụng item associations.
- Ref: Lin, W., S. Alvarez, and C. Ruiz. Efficient adaptive support association rule mining for recommender systems. Data Mining and Knowledge Discovery, 2002, 6(1): p. 83-105.

So khớp user profile và tiền đề luật

- Định nghĩa một sliding window có kích thước w di chuyển trên profile của người dùng mục tiêu.
 - w giảm dần đến khi so khớp chính xác với tiền đề của một luật.
- Cửa sổ biểu diễn một phần hành vi lịch sử của người dùng mục tiêu để dự đoán hành vi của người này trong tương lai.



Frequent itemset graph

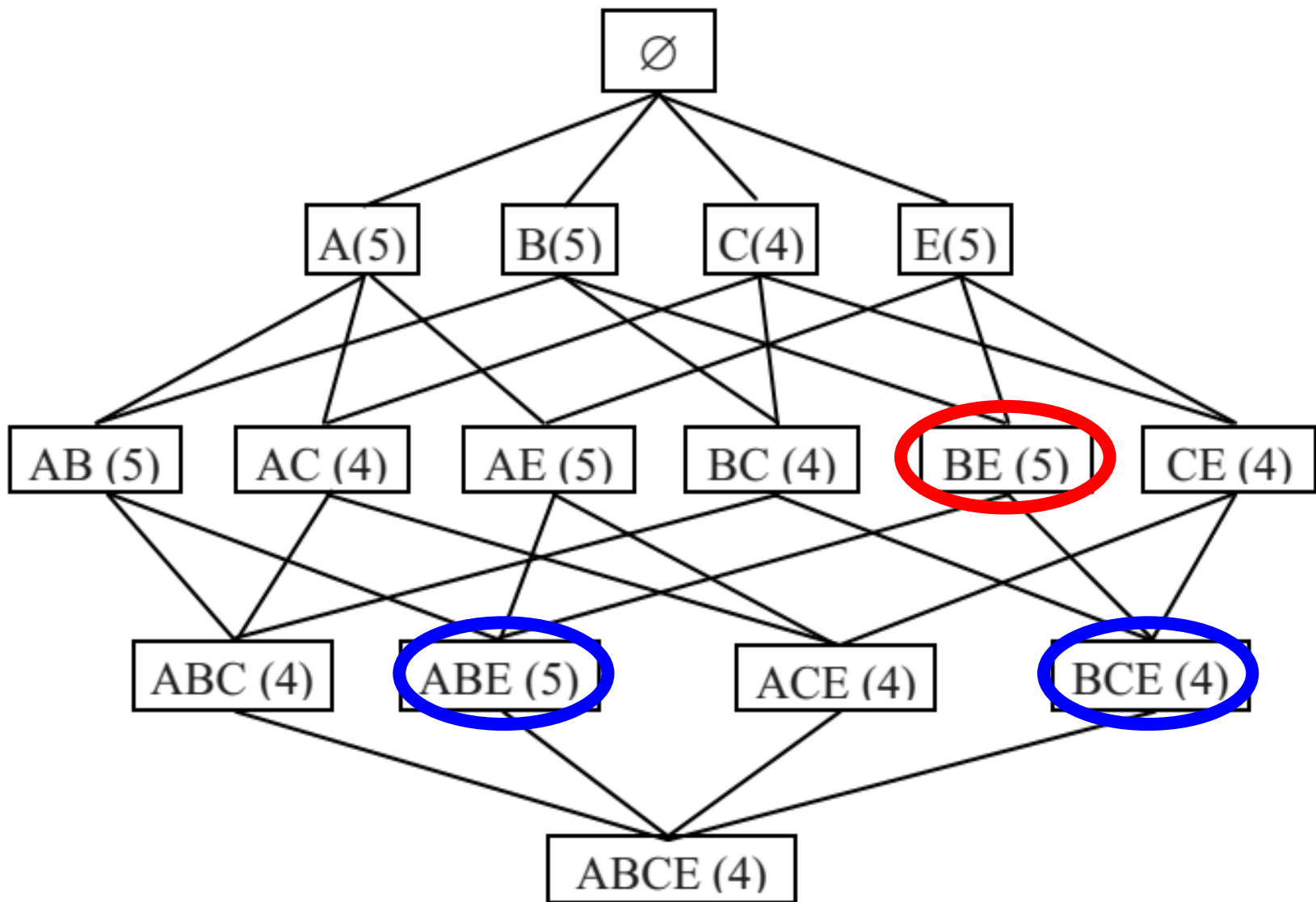


Size 1		Size 2		Size 3		Size 4	
Itemset	Supp.	Itemset	Supp.	Itemset	Supp.	Itemset	Supp.
A	5	A,B	5	A,B,C	4	A,B,C,E	4
B	5	A,C	4	A,B,E	5		
C	4	A,E	5	A,C,E	4		
E	5	B,C	4	B,C,E	4		
		B,E	5				
		C,E	4				

Web transactions and resulting frequent itemsets (minsup = 4)

So khớp user profile và tiền đề luật

- Profile của người dùng mục tiêu được so khớp với các mẫu phổ biến đã tìm thấy để xác định hạng mục ứng viên tư vấn.
- Cho trước cửa sổ profile kích thước w và một nhóm mẫu phổ biến.
- Thuật toán duyệt theo chiều sâu trên Frequent Itemset Graph đến mức $|w|$ và phát sinh ứng viên từ các nút con của nút n thỏa việc so khớp.
 - Giá trị tư vấn của ứng viên dựa trên độ tin cậy của luật tương ứng chứa duy nhất ứng viên ở phần hệ quả.



- User profile window $\langle B, E \rangle$
- Điểm tư vấn của ứng viên A là $1/5$ ($B, E \rightarrow A$) và của ứng viên B là $4/5$ ($B, E \rightarrow C$)

Áp dụng LKH cho lọc cộng tác

- Mẫu phổ biến không thể chứa các hạng mục “hiếm nhưng quan trọng”.
 - Những hạng mục này xuất hiện không thường xuyên nên bị loại bỏ bởi ngưỡng minimum support toàn cục.
- Đối với dữ liệu hành vi sử dụng Web, tham chiếu đến các trang có nội dung chuyên sâu xuất hiện rất ít so với các trang dẫn đường.
- Giải pháp **multiple minimum supports** cho phép người dùng xác định riêng ngưỡng support cho từng hạng mục.

Lọc cộng tác sử dụng Matrix Factorization

Matrix factorization (MF)

- Phân rã ma trận M thành tích của một vài ma trận hệ số

$$\mathbf{M} = \mathbf{F}_1 \mathbf{F}_2 \dots \mathbf{F}_n$$

- n có thể mang giá trị bất kỳ nhưng thường là 2 hoặc 3.
- Phổ biến trong lọc cộng tác do tính ưu việt về chất lượng tư vấn và khả năng mở rộng.
- timeSVD++ (Netflix Prize contest 2006), Nonnegative Matrix Factorization, MaximumMargin Matrix Factorization, và Probabilistic Matrix Factorization

Áp dụng MF cho lọc cộng tác

- Matrix factorization thuộc về nhóm **mô hình hệ số tiềm ẩn** (latent factor models).
- **Biến tiềm ẩn** (latent variables) biểu diễn lý do nền tảng của việc mua/sử dụng sản phẩm của một người dùng.
 - Còn gọi là đặc trưng (feature), khía cạnh (aspect), hay hệ số (factor).
- Pha huấn luyện thiết lập liên kết giữa biến tiềm ẩn và biến quan sát được (người dùng, sản phẩm, điểm đánh giá,...).
- Tương tác có thể xảy ra giữa người dùng với sản phẩm thông qua biến tiềm ẩn được tính toán để đưa ra tư vấn.

Phương pháp SVD trong lọc cộng tác

- Đọc mục 12.4.5, trang 586, tài liệu tham khảo Web Data Mining, 2nd edition, Bing Liu.

Bài tập 1: Chiến lược lọc cộng tác

- John cần thiết kế hệ thống tư vấn cho một cửa hàng sách trực tuyến mới khai trương gần đây. Cửa hàng có hơn 1 triệu tựa sách nhưng cơ sở dữ liệu đánh giá mới chỉ có 10,000 đánh giá.
- Chiến lược nào sẽ giúp John có được hệ thống tư vấn tốt, *content-based recommendation*, *user-based collaborative filtering*, hay *item-based collaboration filtering*? Giải thích.
- Một khách hàng đã đánh giá 5/5 sao cho cả hai cuốn sách “Linear Algebra” and “Differential Equations”. Quyển sách nào sau đây ít có khả năng được giới thiệu nhất theo hệ thống tư vấn trên? Giải thích.
 - a) “Operating Systems”
 - b) “Convex Optimization”
 - c) “Harry Potter: The Goblet of Fire”
 - d) Không thể xác định được vì còn tùy thuộc vào đánh giá của những người dùng khác.

Bài tập 2: k-NN user-based CF

- Bảng bên thể hiện độ yêu thích (1 – 5) của ba nhân vật, Mark Zuckerberg, Bill Gates, và Guido van Rossum, đối với bốn công nghệ, PHP, Spark, Microsoft .NET và Python.

				
	4.5	4.0	1.5	4.5
	3.0	1.0	4.0	2.0
	4.5		2.0	5.0

- Áp dụng lọc cộng tác theo người dùng với k-NN ($k = 1$)
- Xác định độ tương tự về sở thích giữa Guido van Rossum với các nhân vật còn lại.
- Dự đoán điểm yêu thích của Guido van Rossum đối với Spark.

Bài tập 3: k-NN item-based CF

- Bảng bên thể hiện độ yêu thích (1 – 5) của ba nhân vật, Mark Zuckerberg, Bill Gates, và Guido van Rossum, đối với bốn công nghệ, PHP, Spark, Microsoft .NET và Python.

				
	4.5	4.0	1.5	4.5
	3.0	1.0	4.0	2.0
	4.5		2.0	5.0

- Áp dụng lọc cộng tác theo hạng mục với k-NN ($k = 1$)
- Xác định độ tương tự điểm đánh giá giữa Spark và các sản phẩm khác
- Dự đoán điểm yêu thích của Guido van Rossum đối với Spark.

Bài tập 4: Tư vấn theo lọc cộng tác

- Bảng bên cạnh thể hiện tập dữ liệu đánh giá của 4 người dùng đối với 5 sản phẩm. Thang điểm đánh giá từ 1 (nhỏ nhất) đến 5 (lớn nhất). Dấu ? nghĩa là người dùng chưa xem hoặc chưa đánh giá sản phẩm này

	Book1	Book2	Book3	Book4	Book5
Alice	1	2	5	?	1
George	5	?	1	?	5
Mary	?	?	4	3	4
Tom	1	1	5	4	?

- Dự đoán điểm đánh giá của Tom đối với Book 5 bằng lọc cộng tác theo người dùng với k-NN ($k = 1$).
- Tương tự, dự đoán bằng lọc cộng tác theo hạng mục với k=NN ($k = 1$)

Tài liệu tham khảo



- Bing Liu. 2007. *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data*. Springer Series on Data-Centric Systems and Applications. **Chapter 12.4**.

Bài tập 1: Chiến lược lọc cộng tác

- Content-based recommendation vì tỉ lệ đánh giá quá ít so với số lượng tựa sách.
- Người dùng đã chọn hai tựa sách về chủ đề toán học.
- b) có cùng chủ đề nên khả năng được giới thiệu cao.
- a) có chủ đề máy tính và c) có chủ đề văn học. Nên giữa c) và a) thì c) sẽ ít có khả năng giới thiệu hơn.

Bài tập 2,3: Hệ thống tư vấn

- Điểm của Guido đối với Spark theo user-based = $3.833 + 0.984 \cdot (4 - 3.625) / 0.984 = 4.208$

	\bar{r}	$\sqrt{\sum_{i \in C} (r_i - \bar{r})^2}$	$\sum_{i \in C} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)$	sim
Mark	3.625	2.459	5.5	0.984
Bill	2.5	1.658	-3	-0.796
Guido	3.833	2.273		

- Điểm của Guido đối với Spark theo item-based = điểm đối với Python

sim	php	.NET	Python
Spark	-0.271	-0.758	0.692