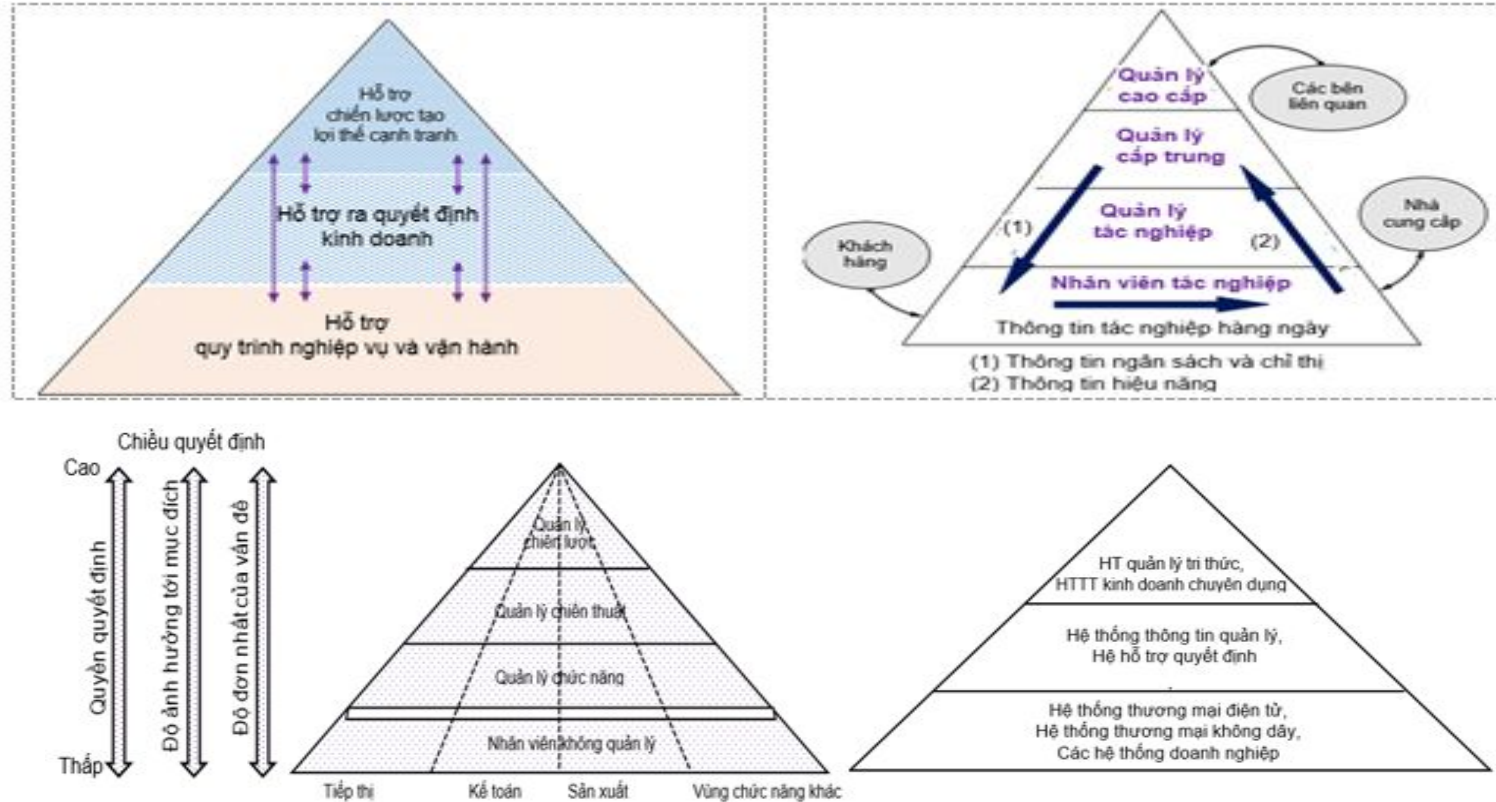


BÀI GIẢNG CSHTT

Recommender Systems

Nguyen Van Hieu
Information Technology Faculty
The University of Danang, University of Science and Technology (UD-UST)

CÁC HTTT



HTTT trong các tổ chức

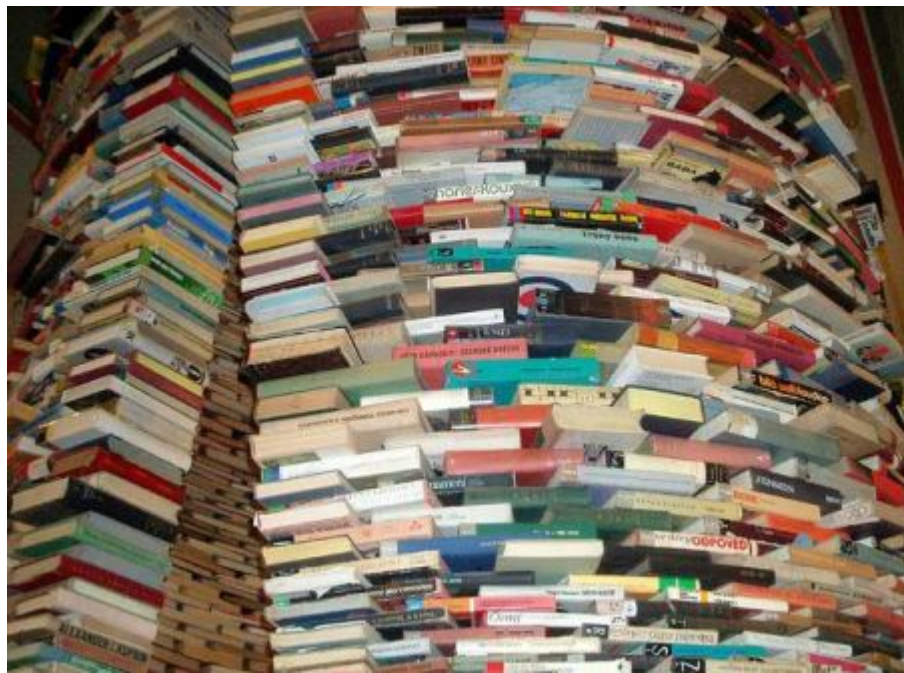
Mức trên: Hệ thống quản lý tri thức và hệ thống thông tin kinh doanh chuyên ngành . QL chiến lược

Mức giữa: HT thông tin quản lý và Hệ hỗ trợ quyết định. QL chiến thuật

Mức dưới: Thương mại điện tử, thương mại không dây (M-commerce: Mobile-commerce) và các hệ thống doanh nghiệp. QL chức năng (tác nghiệp)

Giới thiệu về RS

- Sự quá tải thông tin (Information overload)



Giới thiệu về RS

- Sự quá tải thông tin (Information overload)



- Vấn đề: Cần hệ thống hỗ trợ ra quyết định(DSS)
Cần hệ thống gợi ý (RS)

Giới thiệu về RS

- Sự quá tải thông tin(**Information overload**)
- Phần thưởng của Netflix là 1 triệu USD, “BellKor’s Pragmatic Chaos” đã giành chiến thắng hồi năm 2009.
- Thuật toán của nhóm này hiệu quả hơn 10% so với dịch vụ “khuyến dùng” của Netflix



The screenshot shows the Netflix Prize Leaderboard interface. At the top, there is a yellow banner with the text "Netflix Prize". Below the banner is a navigation bar with links: Home, Rules, Leaderboard, Register, Update, Submit, and Download. The "Leaderboard" link is highlighted. Below the navigation bar, the word "Leaderboard" is displayed in large blue text, followed by "10.05%" in bold black text. To the right of "10.05%" is a small input field with the number "20" and the text "Display top 20 leaders.". Below this, a table lists the top teams. The first team is "BellKor's Pragmatic Chaos" with a rank of 1, a best score of 0.8558, a 10.05% improvement, and a last submit time of 2009-06-26 18:42:37. A yellow arrow points from the "10.05%" text to the "% Improvement" column header. At the bottom of the table, a red banner displays the text "Grand Prize - RMSE <= 0.8563".

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37

Grand Prize - RMSE <= 0.8563

Giới thiệu về RS



Nguồn: Lester Mackey, 2009

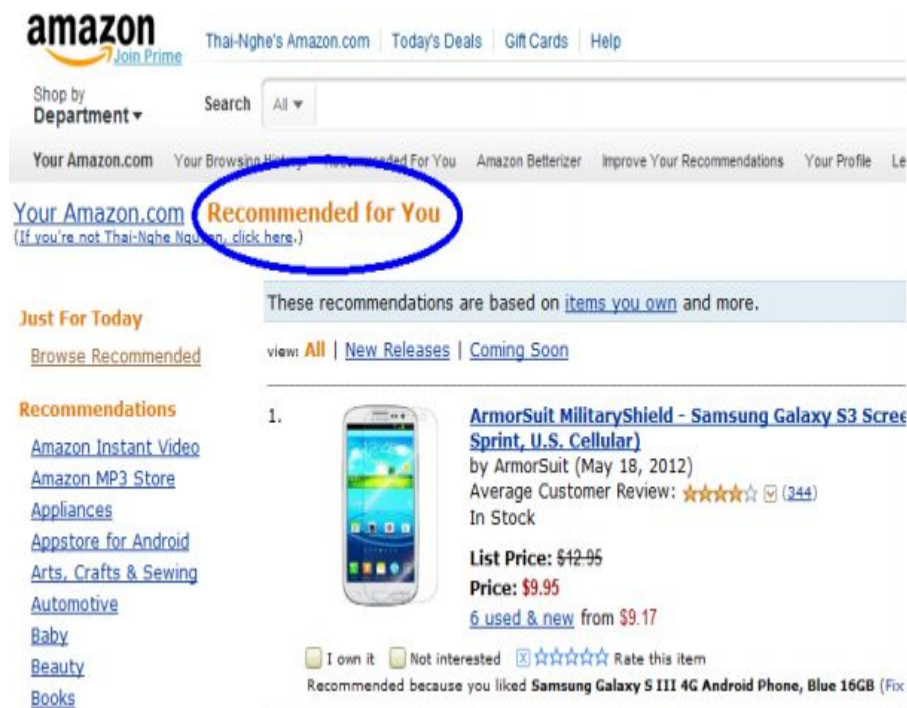
- Tìm hiểu sở thích trong quá khứ của người dùng
- Dự đoán sở thích mới: Bob có thích dâu tây không?

Giới thiệu về RS

- Giả định : người dùng có mối quan hệ "liên quan"
- Ví dụ:
 - Nếu Jack thích A, B, C
 - Nếu John thích A, B
 - Thì khả năng John thích C là rất cao
- Dự đoán sở thích dựa vào
 - Thông tin người dùng
 - Thông tin sản phẩm
 - Thông tin quá khứ,
 - Xếp hạng, số lần kích chuột

Giới thiệu về RS

- Gợi ý bán hàng của Amazon



The screenshot shows the Amazon.com homepage for a user named 'Thai-Nghe'. The 'Recommended for You' section is circled in blue. It features a product recommendation for the 'ArmorSuit MilitaryShield - Samsung Galaxy S3 Screen Protector' by ArmorSuit, which is priced at \$9.95. The recommendation is based on the user's purchase history, specifically mentioning the 'Samsung Galaxy S III 4G Android Phone, Blue 16GB'.

amazon.com Thai-Nghe's Amazon.com Today's Deals Gift Cards Help

Shop by Department Search All

Your Amazon.com Your Browsing History Items Recommended For You Amazon Betterizer Improve Your Recommendations Your Profile

Recommended for You
(If you're not Thai-Nghe Nguyen, click here.)

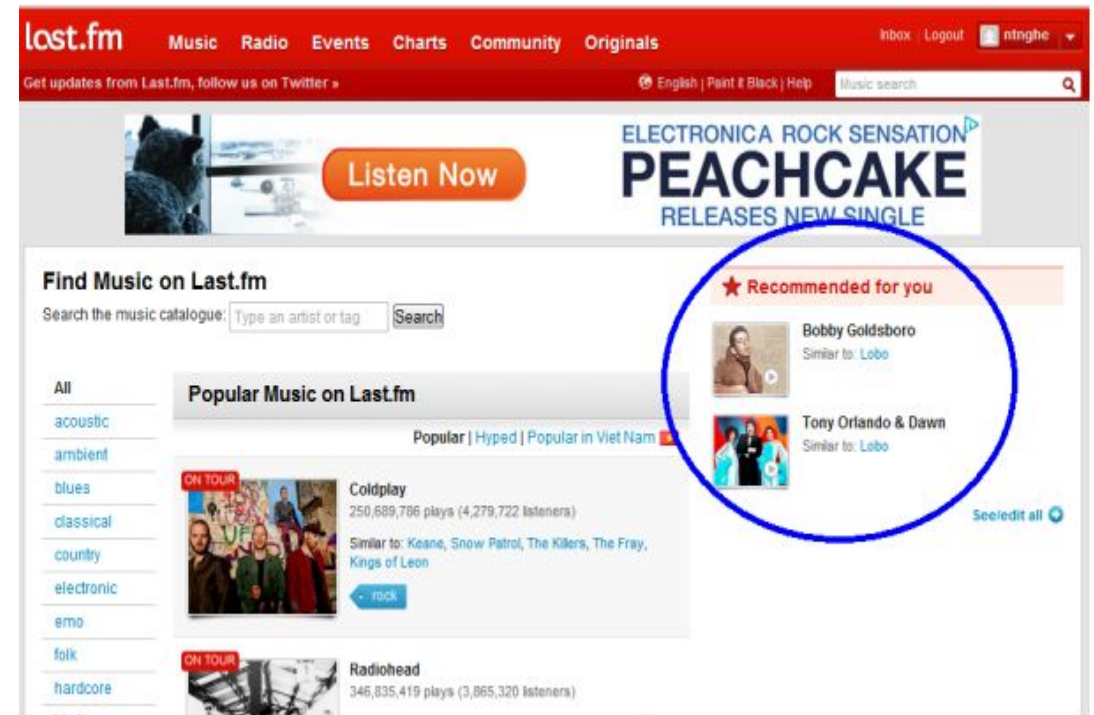
These recommendations are based on [items you own](#) and more.

view: All | New Releases | Coming Soon

1. **ArmorSuit MilitaryShield - Samsung Galaxy S3 Screen Protector**
by ArmorSuit (May 18, 2012)
Average Customer Review: ★★★★★ (344)
In Stock
List Price: \$42.95
Price: \$9.95
6 used & new from \$9.17

I own it Not interested ☆☆☆☆ Rate this item
Recommended because you liked **Samsung Galaxy S III 4G Android Phone, Blue 16GB** (Fix)

- Gợi ý giải trí



The screenshot shows the Last.fm homepage. The 'Recommended for you' section is circled in blue. It features two music recommendations: 'Bobby Goldsboro' and 'Tony Orlando & Dawn', both similar to the user's listening history. The page also displays popular music on Last.fm, including Coldplay and Radiohead.

last.fm Music Radio Events Charts Community Originals Inbox Logout ntinghe

Get updates from Last.fm, follow us on Twitter » English | Paint & Black | Help Music search

Listen Now

PEACHCAKE
RELEASES NEW SINGLE

Find Music on Last.fm
Search the music catalogue: Type an artist or tag Search

Recommended for you

Bobby Goldsboro
Similar to: Lobo

Tony Orlando & Dawn
Similar to: Lobo

See/edit all

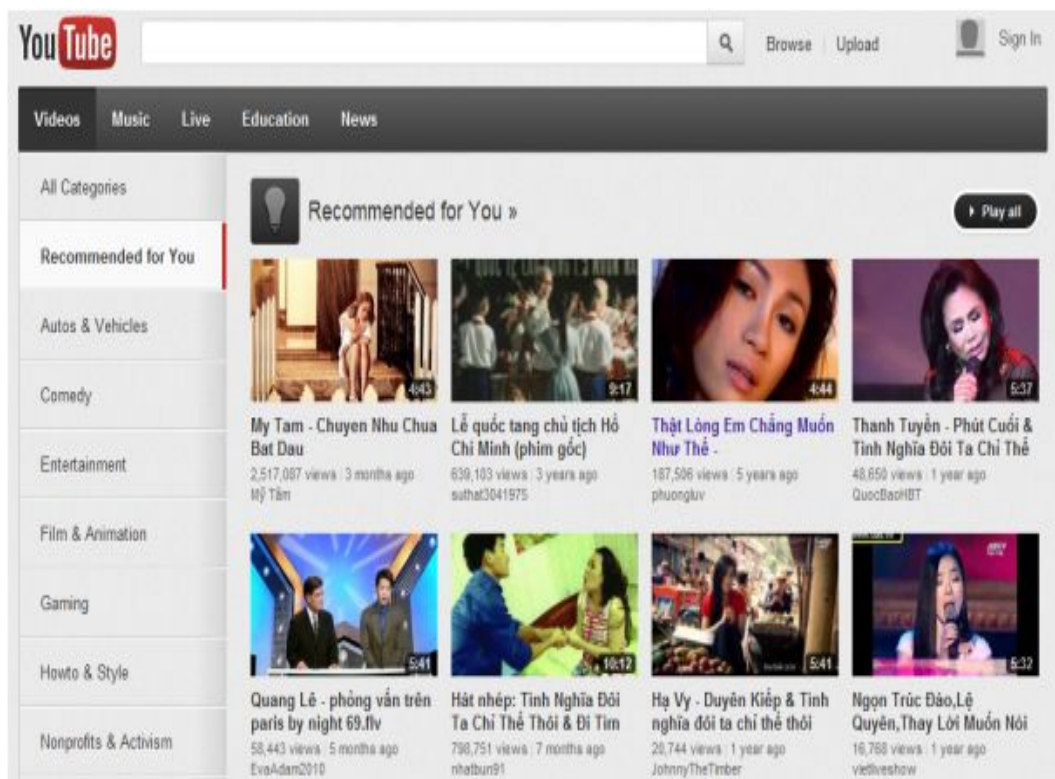
Popular Music on Last.fm
Popular | Hyped | Popular in Viet Nam

Coldplay
250,689,786 plays (4,279,722 listeners)
Similar to: Keane, Snow Patrol, The Killers, The Fray, Kings of Leon

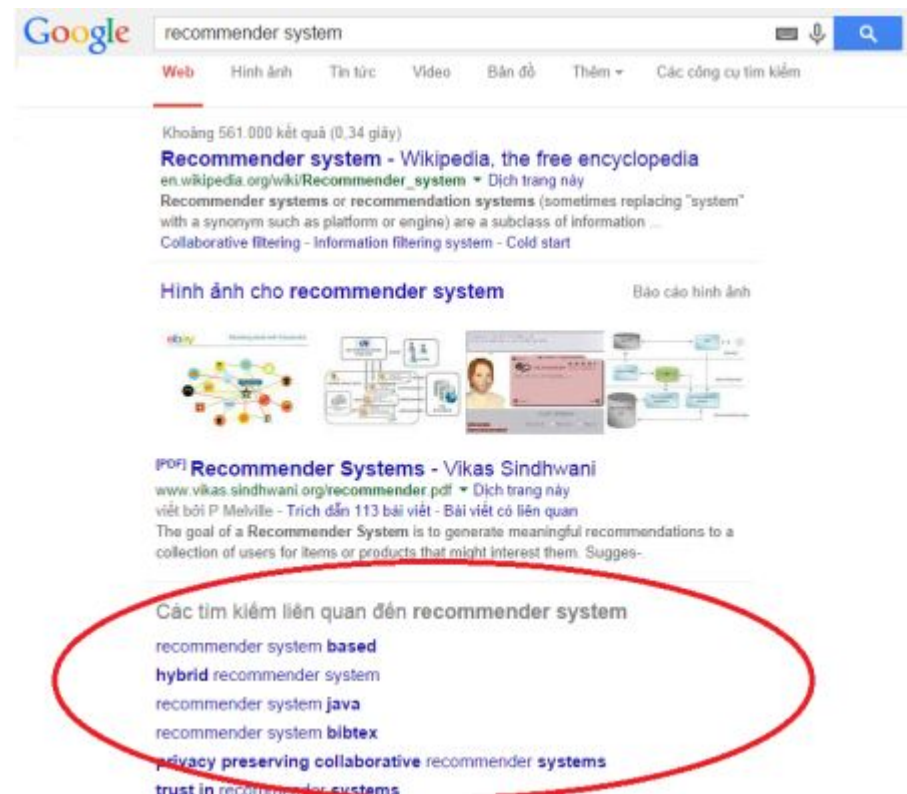
Radiohead
346,835,419 plays (3,865,320 listeners)

Giới thiệu về RS

- Gợi ý giải trí



- Gợi ý từ khóa



Giới thiệu về RS

- Gợi ý Bought together

"Bought together":

Keyboard FRU# 42T3671
by aCompatible
★★★★★ 7 customer reviews

Price: \$46.99 & FREE Shipping. Details

Only 19 left in stock.
Sold by Replacing and Fulfilled by Amazon. Gift-wrap available.

Frequently Bought Together

Price for both: **\$91.89**

☒ This item: Eathtek Brand new OEM IBM/ Lenovo Thinkpad X200, X201 Tablet Keyboard FRU# 42T3671 \$46.99

☒ Replacement for Lenovo IBM Thinkpad T420, X220 Series Laptop Keyboard US Layout \$44.90

Customers Who Bought This Item Also Bought

- Gợi ý theo Tag

QUAD MÀNG TỐT 27 NGƯỜI
Asus Zenfone 6
5.490.000đ

Sony Xperia C3
6.990.000đ
Khuyến mãi từ giá đến 990.000đ

Samsung Galaxy Tab 3 Lite
3G/8GB (T111)
4.190.000đ
Khuyến mãi từ giá đến 250.000đ

Samsung Galaxy Tab 4 7.0
SM-T231
5.990.000đ
Khuyến mãi từ giá đến 150.000đ

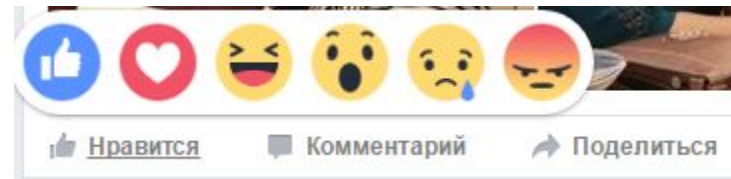
Samsung Galaxy Tab S
HTC Desire 516 Xperia Z2 Nokia XL
Nokia X Lumia 930 Zenfone 5 Nokia X2
Galaxy Note 4 Zenfone 4 iPhone 5S
iPhone 6 HTC Desire 616 Lumia 530
HTC One E8 Blackberry Z10 iOS 8 Sim số đẹp
Android L

Giới thiệu

- Gợi ý khác
 - Gợi ý theo bình luận (comments)
 - Gợi ý theo sản phẩm mới (new item)
 - Gợi ý theo số lần xem (views)
 - ...

Mục đích của RS

- Dự vào “Sở thích” của người dùng trong quá khứ, để dự đoán “Sở thích” trong tương lai và thực hiện gợi ý cho người dùng
- Hệ thống gợi ý tùy thuộc vào feedback của người dùng:
 - Xếp hạng * đến *****
 - Thích hoặc không thích
 - Số lần kích chuột
 - Thời gian quan sát sản phẩm



Dữ liệu truyền thống trong RS

•

		<i>Items</i>					
		<i>1</i>	<i>2</i>	<i>...</i>	<i>i</i>	<i>...</i>	<i>m</i>
<i>Users</i>	<i>1</i>	5	3		1	2	
	<i>2</i>		2				4
	<i>:</i>			5			
	<i>u</i>	3	4	?	2	1	
	<i>:</i>					4	
	<i>n</i>			3	2		

- $\hat{r} : U \times I \rightarrow R$
- \hat{r}_{ui} : xếp hạng của người dùng u cho sản phẩm i
- Dự đoán các sản phẩm chưa được xếp hạng (các ô trống)
- **Sắp xếp theo thứ tự**, để gợi ý cho người dùng

Mô hình hóa bài toán

- U : ID người dùng, I – ID sản phẩm, R giá trị đánh giá (rating)
- Tập dữ liệu: $D: U \times I \times R$
- Tập dữ liệu huấn luyện: $D^{Train} \subseteq D$
- Tập dữ liệu thử: $D^{Test} \subseteq D$
- Bài toán: cho D^{Train} , tìm $\hat{r} : U \times I \rightarrow R$ (giá trị dự đoán):
 $\varepsilon(\hat{r}, r)$ thỏa mãn điều kiện cho trước với $(u, i, r) \in D^{Test}$
 $r: U \times I \rightarrow R$
- ε là RMSE (root mean squared error) thì $\varepsilon(\hat{r}, r)$ cần phải tối thiểu.

$$RMSE = \sqrt{\frac{\sum_{(u,i,r) \in D^{test}} (r - \hat{r}_{(u,i)})^2}{|D^{test}|}}$$

Dữ liệu ví dụ

Training data

user	Item	rating
1	21	1
1	213	5
2	345	4
2	123	4
2	768	3
3	76	5
4	45	4
5	568	1
5	342	2
5	234	2
6	76	5
6	56	4

Test data

user	Item	rating
1	62	?
1	96	?
2	7	?
2	3	?
3	47	?
3	15	?
4	41	?
4	28	?
5	93	?
5	74	?
6	69	?
6	83	?

Các kỹ thuật cơ bản

- Gợi ý không cá nhân hóa (Non-Personalized Recommendation)
- Gợi ý cá nhân hóa cho người dùng
 - Lọc cộng tác
 - Lọc nội dung
 - Kết hợp

Các kỹ thuật cơ bản

- **Gợi ý không cá nhân hóa:** gợi ý sản phẩm
 - Được mua, được xem, được bình luận,... “**Nhiều nhất**”
 - Mới nhất,
 - Cùng tác giả, cùng nhà sản xuất, cùng thể loại
 - Được mua, được chọn cùng nhau (sử dụng luật kết hợp)



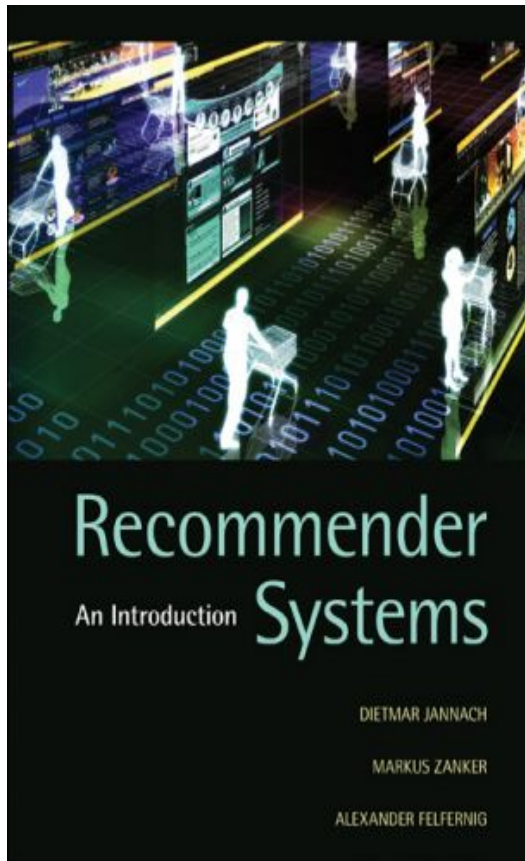
Các kỹ thuật cơ bản

- **Gợi ý theo cá nhân hóa cho người dùng:**
- **Lọc cộng tác**
- Cộng tác = sử dụng dữ liệu của người khác
 - Kỹ thuật “Láng giềng” (Neighborhood-based hay Memory-based)
 - Cơ sở người dùng: dựa vào dữ liệu quá khứ của người dùng tương tự
 - Cơ sở sản phẩm: dựa vào dữ liệu quá khứ của sản phẩm tương tự
 - Kỹ thuật dựa vào mô hình (model based)
 - **Matrix factorization**

Các kỹ thuật cơ bản

- **Gợi ý theo cá nhân hóa cho người dùng:**
- **Lọc nội dung**
 - Kỹ thuật dựa vào hồ sơ (profiles) người dùng
 - Kỹ thuật dựa vào sản phẩm có thuộc tính tương tự đã được người dùng xếp hạng trong quá khứ

Tài liệu



Educational Recommender Systems and Technologies: Practices and Challenges

Olga C. Santos (Spanish National University for Distance Education (UNED), Spain) and Jesus G. Boticario (Spanish National University for Distance Education (UNED), Spain)

Release Date: December, 2011. Copyright © 2012. 362 pages.

Select a Format: Hardcover \$175.00

[Add to Cart](#) [Quick Add](#)

In Stock. Have it as soon as Sep. 11 with express shipping!

DOI: 10.4018/978-1-61350-489-5, ISBN13: 9781613504895, ISBN10: 1613504896, EISBN13: 9781613504901

[Cite Book](#) [Favorite](#) [Send](#) [Like](#) [Tweet](#) [Share](#)

[Description](#) [Table of Contents](#) [Reviews and Testimonials](#) [Topics Covered](#) [Preface](#) [Author\(s\)/Editor\(s\) Bio](#) [Editorial Board](#)

[Access on Platform](#)

More Information

[Request examination copy](#)

[Brochure](#)

Recommend

[Send to a librarian](#)

[Send to a colleague](#)

Available In

Description

Recommender systems have shown to be successful in many domains where information overload exists. This success has motivated research on how to deploy recommender systems in educational scenarios to facilitate access to a wide spectrum of information. Tackling open issues in their deployment is gaining importance as lifelong learning becomes a necessity of the current knowledge-based society. Although Educational Recommender Systems (ERS) share the same key objectives as recommenders for e-commerce applications, there are some particularities that should be considered before directly applying existing solutions from those applications.

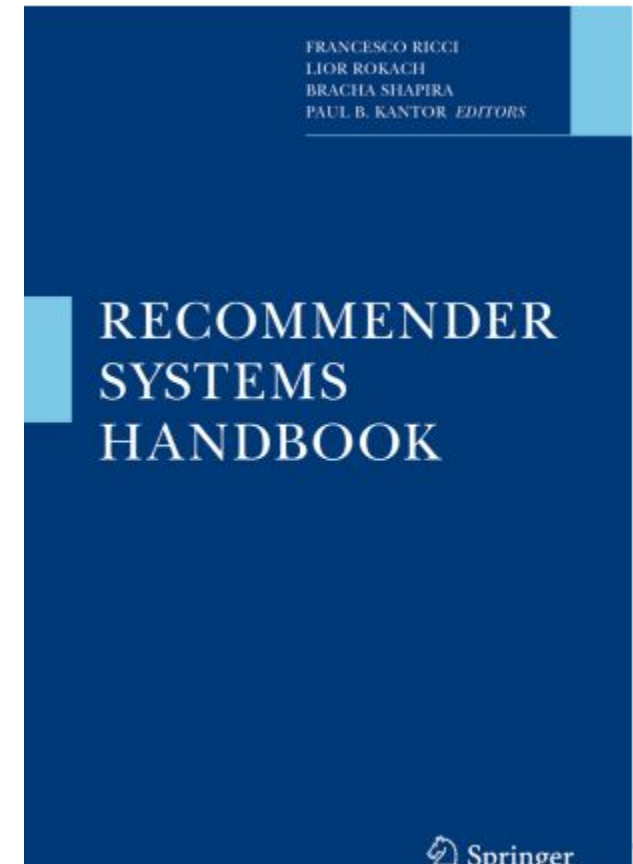
Chapter 6

Factorization Techniques for Predicting Student Performance (pages 129-153)

Nguyen Thai-Nghe, Lucas Drumond, Tomáš Horváth, Artus Krohn-Grimberghe, Alexandros Nanopoulos, Lars Schmidt-Thieme

[Add to Cart](#) [Quick Add](#)

Recommendation systems are widely used in many areas, especially in e-commerce.



xây dựng hệ thống gợi ý

Hệ thống gợi ý hai chiều

- $\hat{r} : U \times I \rightarrow R$
- U – tập người dùng
- I – tập sản phẩm
- \hat{r} - hàm xác định độ đo của người dùng u với sản phẩm i

Hệ thống gợi ý hai chiều

- ❖ Người dùng u sẽ được giới thiệu sản phẩm I' , sao cho sản phẩm I' tương tự sản phẩm i .
- ❖ Người dùng u' được giới thiệu sản phẩm i , nếu sản phẩm i được đánh giá cao bởi người u , và người u và u' có cùng sở thích
- ❖ Kết hợp

Hệ thống gợi ý đa chiều

-

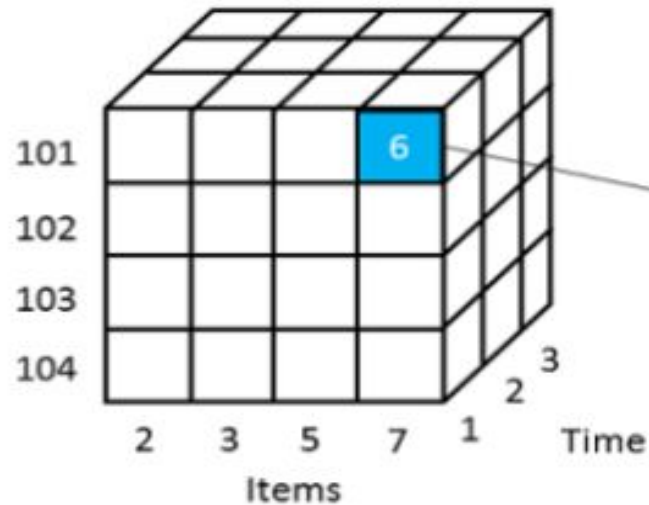
$$\hat{r} : U \times I \times C \rightarrow R$$

- U – tập người dùng
- I – tập sản phẩm
- C – tập ngữ cảnh
- \hat{r} - hàm xác định độ đo

Hệ thống gợi ý đa chiều

Id	Name	Age
101	John	25
102	Bob	18
103	Alice	27
104	Mary	24

Users



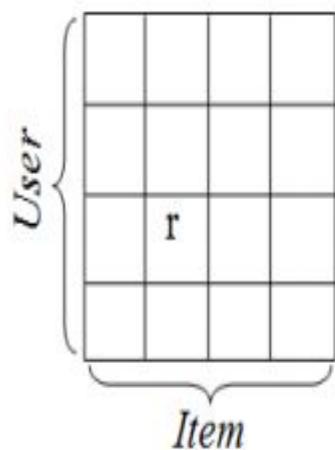
Id	Name	Cost
2	Item 2	10
3	Item 3	20
5	Item 5	15
7	Item 7	40

Id	Name
1	Weekday
2	Weekend
3	Holiday

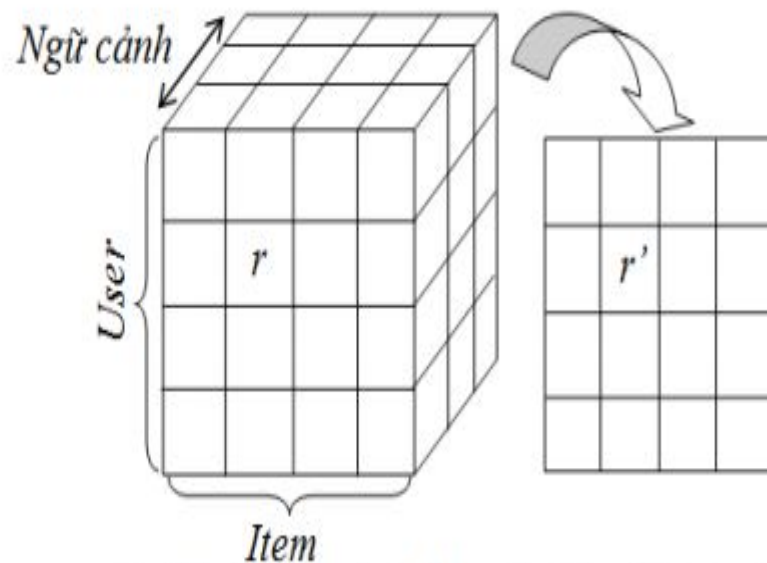
Hệ thống gợi ý

- Hướng tiếp cận
 - Đề xuất cải tiến phương pháp gợi ý đa chiều hiện có (khó khăn 😊)
 - Đề xuất can thiệp đơn giản -- > hệ thống mới
- Ý tưởng:
 - **Hệ thống gợi ý đa chiều ==> Hệ thống gợi ý 2 chiều**
 - Can thiệp xử lý đầu vào
 - Can thiệp xử lý đầu ra
 - Sử dụng phương pháp 2 chiều truyền thống.

Xử lý đầu vào



a. Dữ liệu trong hệ thống RS truyền thống



b. Xử lý ngữ cảnh đầu vào (Pre-filtering)

Xử lý đầu vào(tiếp)

user	item	time	Bạn Đồng hành	Thời tiết	rate
1	2	Cuối tuần	Bạn bè	Trời nắng	4
1	5	Cuối tuần	Một mình	Trời âm u	1
1	3	Lễ - tết	Gia đình	Trời trong xanh	5
2	2	Ngày trong tuần	Bạn bè	Trời nắng	2
2	1	Lễ - tết	Gia đình	Trời trong xanh	3
3	5	Lễ - tết	Gia đình	Trời trong xanh	4
3	4	Cuối tuần	Bạn bè	Trời nắng	3
4	3	Lễ - tết	Gia đình	Trời trong xanh	5

Xử lý đầu vào(tiếp)

user	item	time	Bạn Đồng hành	Thời tiết	rate
1	2	Cuối tuần	Bạn bè	Trời nắng	4
1	5	Cuối tuần	Một mình	Trời âm u	1
1	3	Lễ - tết	Gia đình	Trời trong xanh	5
2	2	Ngày trong tuần	Bạn bè	Trời nắng	2
2	1	Lễ - tết	Gia đình	Trời trong xanh	3
3	5	Lễ - tết	Gia đình	Trời trong xanh	4
3	4	Cuối tuần	Bạn bè	Trời nắng	3
4	3	Lễ - tết	Gia đình	Trời trong xanh	5

user	item	rate
1	3	5
2	1	3
3	5	4
4	3	5

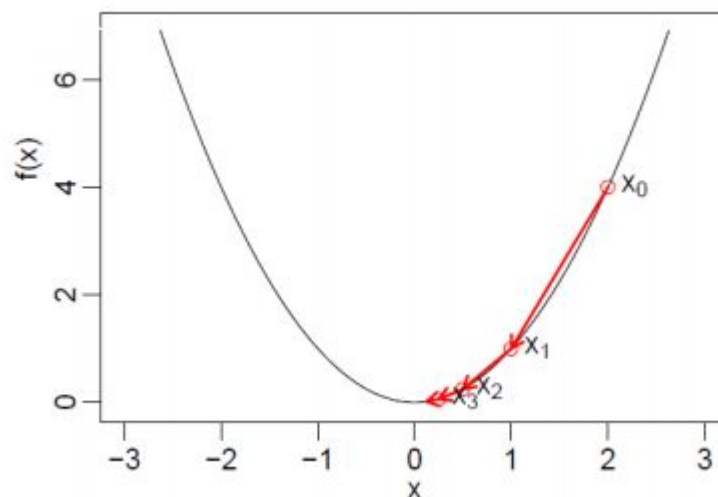
Xử lý đầu vào(tiếp)

- Sử dụng các phương pháp gợi ý truyền thống cho tập dữ liệu

user	item	rate
1	3	5
2	1	3
3	5	4
4	3	5

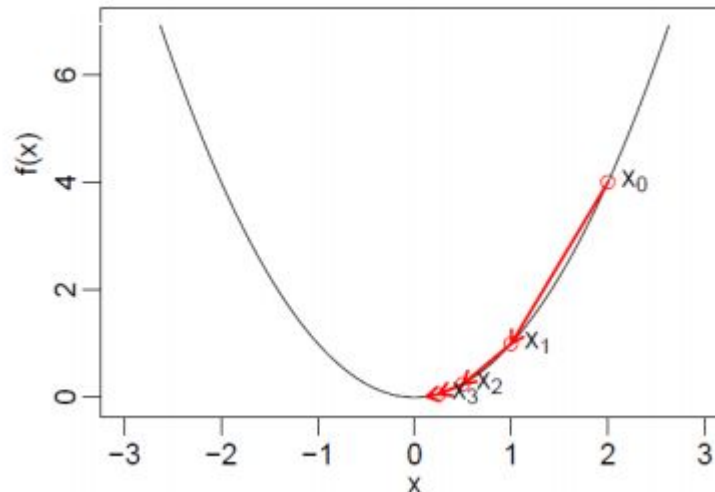
Kỹ thuật phân rã ma trận (matrix factorization)

- Cở sở lý luận (Tối ưu hóa bằng Gradient Descent)
- Cho $f: R^n \rightarrow R$, tìm x sao cho $f(x)$ nhỏ nhất
- Ý tưởng:
 - Từ ngẫu nhiên giá trị x_0 qua một bước cập nhật x_1 , có nghĩa là xây dựng $x_{n+1} : f(x_{n+1}) \leq f(x_n)$



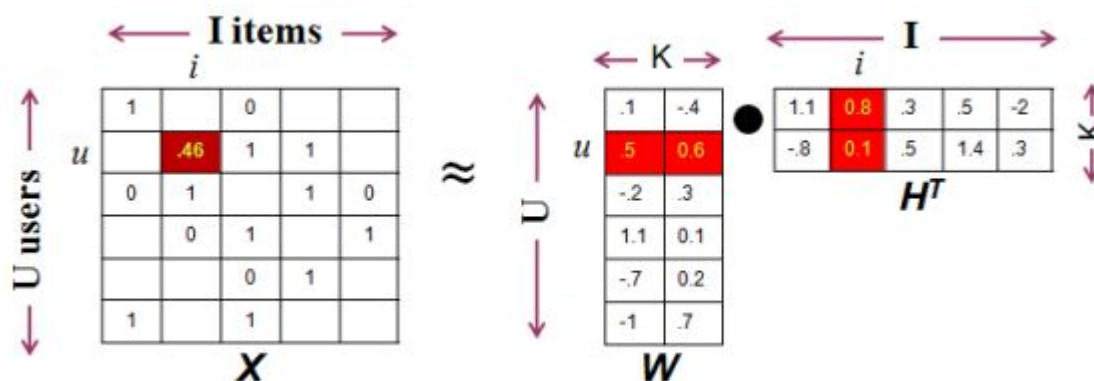
Kỹ thuật phân rã ma trận (matrix factorization)

- Cở sở lý luận (Tối ưu hóa bằng Gradient Descent)
- Chọn hướng để cập nhật: $-\frac{\partial f}{\partial x}(x_n)$
- $x_{n+1} = x_n - \beta \cdot \frac{\partial f}{\partial x}(x_n)$



Kỹ thuật phân rã ma trận (matrix factorization)

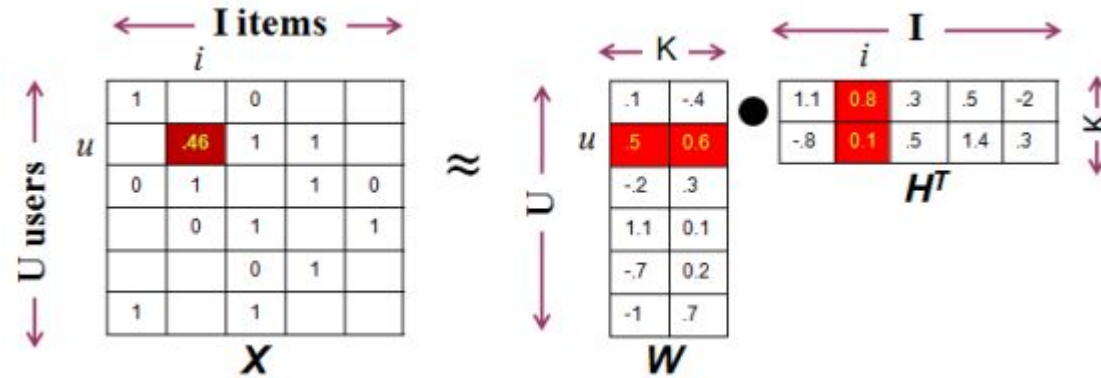
- Ý tưởng: Chia ma trận X thành W và H : $X \approx W \times H^T$
- W và H có thể xây dựng ma trận X càng chính xác càng tốt



- $W \in R^{U \times K}$, mỗi dòng (người dùng u) với K nhân tố
- $H \in R^{K \times I}$, mỗi dòng (sản phẩm i) với K nhân tố

Kỹ thuật phân rã ma trận

-



- Hàm dự đoán

$$\hat{r}_{ui} = \sum_{k=1}^K w_{uk} \times h_{ik}$$

Kỹ thuật phân rã ma trận

- Bản chất: xác định giá trị tham số W và H
- Hàm mục tiêu đạt min

$$O^{MF} = \sum_{u,i \in D^{train}} (r_{ui} - \hat{r}_{ui})^2 = \sum_{u,i \in D^{train}} \left(r_{ui} - \sum_{k=1}^K w_{uk} h_{ik} \right)^2$$

- Ý tưởng: Khởi tạo ngẫu nhiên giá trị của W và H , sau mỗi bước cập nhật giá trị, và kết thúc khi đạt giá trị min

Kỹ thuật MF

- Xác định tăng hay giảm W và H

$$\frac{\partial}{\partial w_{uk}} O^{MF} = -2(r_{ui} - \hat{r}_{ui})h_{ik}$$

$$\frac{\partial}{\partial h_{ik}} O^{MF} = -2(r_{ui} - \hat{r}_{ui})w_{uk}$$

- Cập nhật:

$$w_{uk}^{new} = w_{uk}^{old} - \beta \cdot \frac{\partial}{\partial w_{uk}} O^{MF} = w_{uk}^{old} + 2\beta \cdot (r_{ui} - \hat{r}_{ui})h_{ik}$$

$$h_{ik}^{new} = h_{ik}^{old} - \beta \cdot \frac{\partial}{\partial h_{ik}} O^{MF} = h_{ik}^{old} + 2\beta \cdot (r_{ui} - \hat{r}_{ui})w_{uk}$$

Kỹ thuật MF

$\lambda \in (0..1)$ và $\|\cdot\|_F$ là chuẩn Frobenius:

$$\|\mathbf{W}\|_F = \sqrt{\sum_{u=1}^{|U|} \sum_{k=1}^K |w_{uk}|^2}$$

- Ngăn ngừa học vẹt

$$O^{MF} = \sum_{u,i \in D^{train}} \left(r_{ui} - \sum_{k=1}^K w_{uk} h_{ik} \right)^2 + \lambda \cdot \left(\|W\|_F^2 + \|H\|_F^2 \right)$$

- Cập nhật đến khi chấp nhận hoặc số lần quy định trước

$$w_{uk}^{new} = w_{uk}^{old} + \beta \cdot \left(2(r_{ui} - \hat{r}_{ui}) h_{ik} - \lambda \cdot w_{uk}^{old} \right)$$

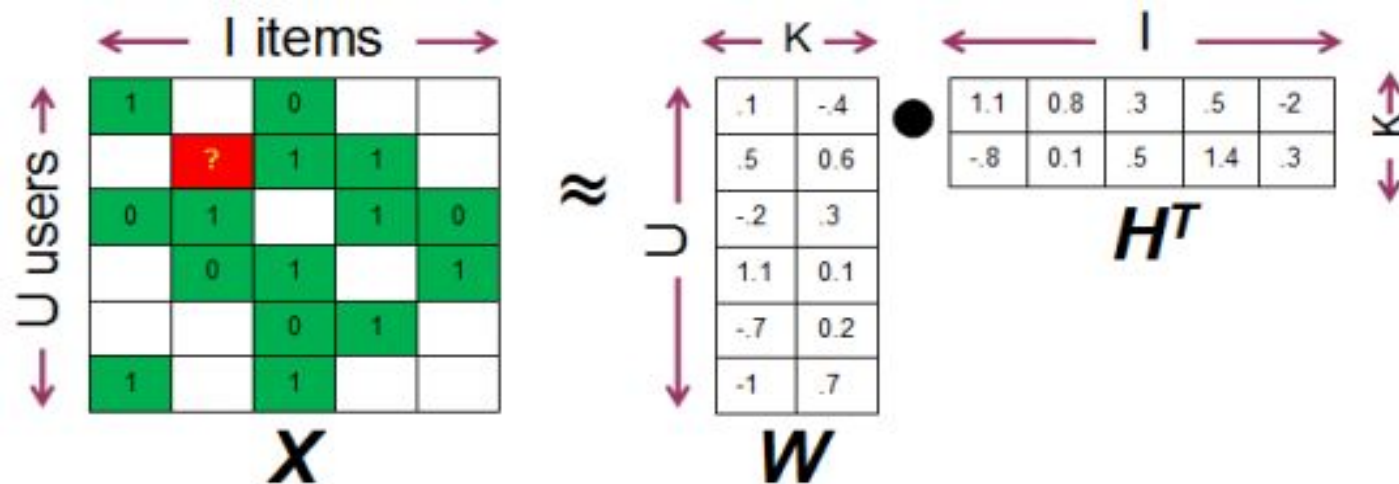
$$h_{ik}^{new} = h_{ik}^{old} + \beta \cdot \left(2(r_{ui} - \hat{r}_{ui}) w_{uk} - \lambda \cdot h_{ik}^{old} \right)$$

Kỹ thuật MF

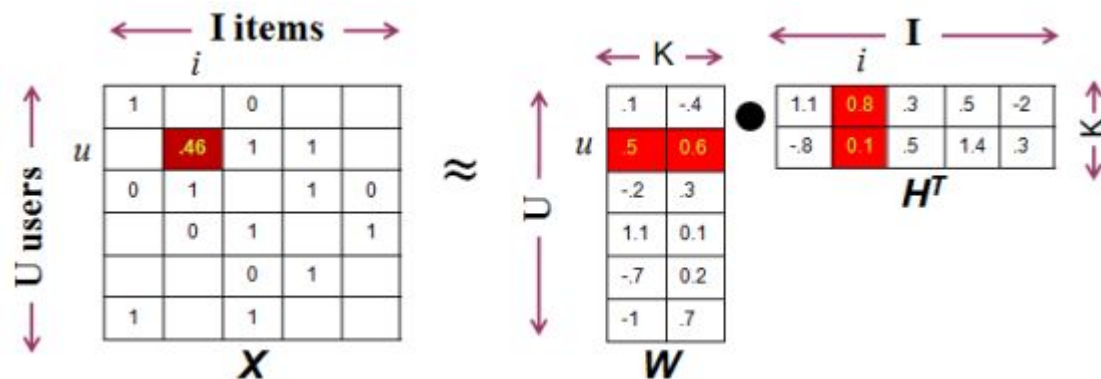
```
1: procedure MATRIXFACTORIZATION( $\mathcal{D}^{train}$ ,  $K$ ,  $\beta$ ,  $\lambda$ , stopping condition)
  // Let  $W[|U|][K]$  and  $H[|I|][K]$  be latent factors of users and items
2:    $W \leftarrow \mathcal{N}(0, \sigma^2)$ 
3:    $H \leftarrow \mathcal{N}(0, \sigma^2)$ 
4:   while (Stopping criterion is NOT met) do
5:     Draw randomly  $(u, i, r)$  from  $\mathcal{D}^{train}$ 
6:      $\hat{r} \leftarrow 0$ 
7:     for  $k \leftarrow 1, \dots, K$  do
8:        $\hat{r} \leftarrow \hat{r} + W[u][k] \cdot H[i][k]$ 
9:     end for
10:     $e_{ui} = r - \hat{r}$ 
11:    for  $k \leftarrow 1, \dots, K$  do
12:       $W[u][k] \leftarrow W[u][k] + \beta \cdot (e_{ui} \cdot H[i][k] - \lambda \cdot W[u][k])$ 
13:       $H[i][k] \leftarrow H[i][k] + \beta \cdot (e_{ui} \cdot W[u][k] - \lambda \cdot H[i][k])$ 
14:    end for
15:  end while
16:  return  $\{W, H\}$ 
17: end procedure
```

Kỹ thuật phân rã ma trận

- Sau khi có kết quả W và H
- Dự đoán thế nào?



Kỹ thuật phân rã ma trận



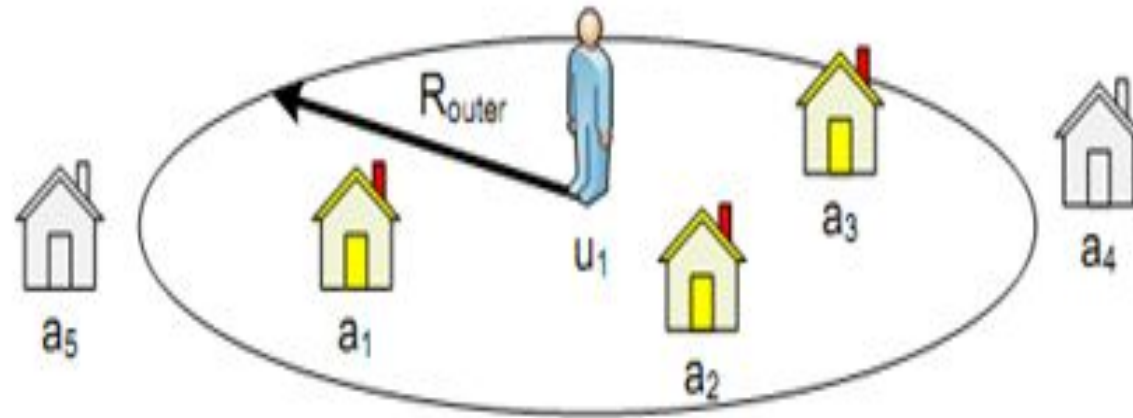
- Xếp hạng của người dùng u cho sản phẩm i được dự đoán bằng:

$$\hat{r}_{ui} = \mathbf{w} \cdot \mathbf{h}^T = \sum_{k=1}^K w_{uk} h_{ik}$$

- Ví dụ: dự đoán kết quả của người dùng 2 cho sản phẩm 2 là:

$$5 * 0.8 + 0.6 * 0.1 = 0.46$$

Xử lý đầu ra



Xử lý đầu ra (tiếp)

- Nếu không xử lý có kết quả xếp hạng theo thứ tự
- A – B – C – D,.....

STT	Tên địa điểm	Khoảng cách(km)	Dự đoán
1.	A	78.5	4.8
2.	B	14.3	4.6
3.	C	110.2	4.5
4.	D	33.4	4.1
5.	E	18.7	3.7
6.	F	2.4	3.5
7.	G	11.2	3.4
8.	H	45.3	3.3
9.	I	24.5	2.9
10.	J	62.1	2.6

Xử lý đầu ra (tiếp)

- $R < 20$ km, thì kết quả xếp hạng theo thứ tự
- B – E – F - G,.....

STT	Tên địa điểm	Khoảng cách(km)	Dự đoán
1.	A	78.5	4.8
2.	B	14.3	4.6
3.	C	110.2	4.5
4.	D	33.4	4.1
5.	E	18.7	3.7
6.	F	2.4	3.5
7.	G	11.2	3.4
8.	H	45.3	3.3
9.	I	24.5	2.9
10.	J	62.1	2.6

Mô hình đề xuất

```
1: procedure ContextAware-MF ( $D^{\text{Train}}$ , Iter, K,  $\beta$ ,  $\lambda$ )  
//  $W[[U]][K]$  và  $H[[I]][K]$  là 2 tham số cần tìm  
2:  $W := N(0, \sigma^2)$  //khởi tạo giá trị theo phân phối chuẩn  
3:  $H := N(0, \sigma^2)$  //khởi tạo giá trị theo phân phối chuẩn  
4:  $D^{\text{TrainC}} = \text{Pre-filtering}(D^{\text{Train}})$   
5: for (iter:=1; iter <= Iter *  $|D^{\text{TrainC}}|$ ; iter++)  
6:   Chọn ngẫu nhiên một dòng (u, i,  $r_{ui}$ ) từ  $D^{\text{TrainC}}$   
7:    $\hat{r}_{ui} := 0$   
8:   for (k:=1; k<=K; k++)  
9:      $\hat{r}_{ui} := \hat{r}_{ui} + W[u][k] * H[i][k]$   
10:  end for  
11:   $e_{ui} = r_{ui} - \hat{r}_{ui}$   
12:  for (k:=1; k<=K; k++)  
13:     $W[u][k] := W[u][k] + \beta * (e_{ui} * H[i][k] - \lambda * W[u][k])$   
14:     $H[i][k] := H[i][k] + \beta * (e_{ui} * W[u][k] - \lambda * H[i][k])$   
15:  end for  
16:  Break nếu đã hội tụ  
17: end for  
18: return {W, H}  
19: Post-filtering(Tập kết quả được dự đoán dùng W, H)  
20: end procedure
```

- Các tham số:
 - Iter - số lần lặp,
 - K – số nhân tố tiềm ẩn,
 - B – tốc độ học,
 - λ – hệ số chính tắc hóa
- Tìm kiếm theo phương pháp siêu tham số

Gợi ý minh họa ứng dụng

- Hệ thống gợi ý địa điểm du lịch tại tp Đà Nẵng
 - Thành viên: họ tên, ngày sinh, giới tính, tên đăng nhập, mật khẩu,...vv
 - Địa điểm du lịch: tên, địa chỉ, hình đại diện, nội dung, chủ đề, lịch sử truy cập,...vv
 - Ngữ cảnh: bạn đồng hành, thời gian, thời tiết, vị trí, khoảng cách, tốc độ mạng,...vv.
 - Đánh giá:

Cải tiến kỹ thuật MF

(Biased matrix factorization)

- Cơ sở kỹ thuật phân rã ma trận

- Giá trị trung bình toàn cục :

$$\mu = \frac{\sum_{(u,i,r) \in \mathcal{D}^{train}} r}{|\mathcal{D}^{train}|}$$

- Độ lệch của người dùng u:

$$b_u = \frac{\sum_{(u',i,r) \in \mathcal{D}^{train} | u'=u} (r - \mu)}{|\{(u',i,r) \in \mathcal{D}^{train} | u' = u\}|}$$

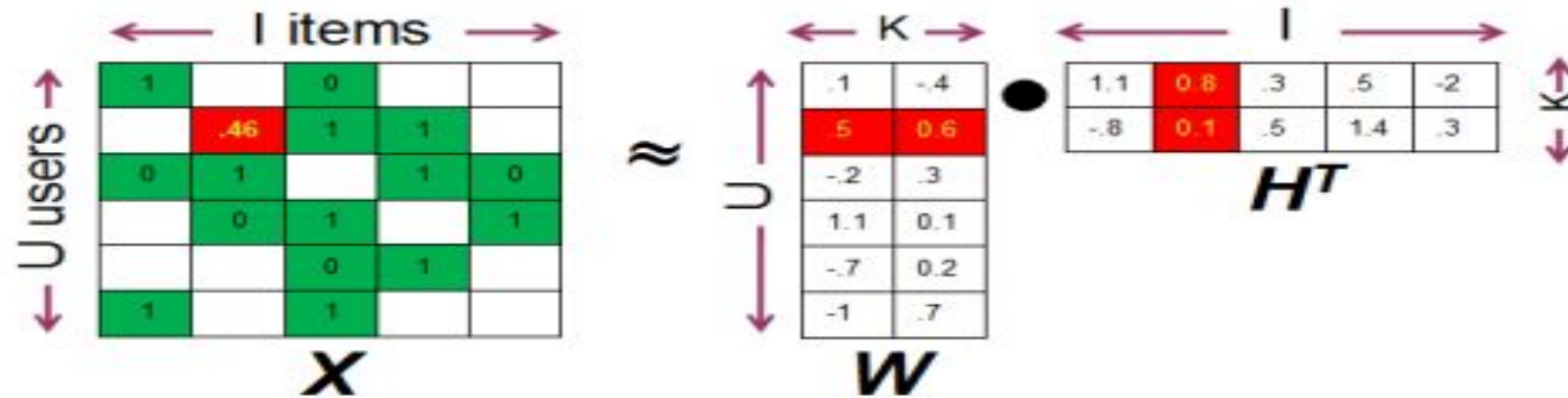
- Độ lệch của sản phẩm i:

$$b_i = \frac{\sum_{(u,i',r) \in \mathcal{D}^{train} | i'=i} (r - \mu)}{|\{(u,i',r) \in \mathcal{D}^{train} | i' = i\}|}$$

Training data			Test data		
user	Item	rating	user	Item	rating
1	21	1	1	82	?
1	213	5	1	96	?
2	345	4	2	7	?
2	123	4	2	3	?
2	768	3	3	47	?
3	76	5	3	15	?
4	45	4	4	41	?
5	568	1	4	28	?
5	342	2	5	93	?
5	234	2	5	74	?
6	76	5	6	69	?
6	56	4	6	83	?

Cải tiến kỹ thuật MF

(Biased matrix factorization)



Hàm dự đoán

$$\hat{r}_{ui} = \mu + b_u + b_i + \sum_{k=1}^K w_{uk} h_{ik}$$

Cải tiến MF

1. Procedure: ResultPrediction_BMF(D^{train} , K , β , λ , stopping condition)

Let $s \in S$ be a student, $i \in I$ a item, $p \in P$ a score

Let $W[S][K]$ and $H[I][K]$ be latent factors of students and tasks

Let $b_s[S]$ and $b_i[I]$ be students-bias and task-bias

$$2. \quad \mu \leftarrow \frac{\sum_{p \in D^{train}} p}{|D^{train}|}$$

3. for each student s do

$$4. \quad b_s[s] \leftarrow \frac{\sum_i (p_{si} - \mu)}{|D_s^{train}|}$$

5. end for

6. for each task i do

$$7. \quad b_i[i] \leftarrow \frac{\sum_u (p_{ui} - \mu)}{|D_i^{train}|}$$

8. end for

$$9. \quad W \leftarrow N(0, \sigma^2)$$

$$10. \quad H \leftarrow N(0, \sigma^2)$$

11. while (Stopping criterion is NOT met) do

12. Draw randomly (s, i, p_{si}) from D^{train}

$$13. \quad \hat{p}_{si} \leftarrow \mu + b_s[s] + b_i[i] + \sum_k^K (W[s][k] * H[i][k])$$

$$14. \quad e_{si} = p_{si} - \hat{p}_{si}$$

$$15. \quad \mu \leftarrow \mu + \beta * e_{si}$$

$$16. \quad b_s[s] \leftarrow b_s[s] + \beta * (e_{si} - \lambda * b_s[s])$$

$$17. \quad b_i[i] \leftarrow b_i[i] + \beta * (e_{si} - \lambda * b_i[i])$$

18. for $k \leftarrow 1, \dots, K$ do

$$19. \quad W[s][k] \leftarrow W[s][k] + \beta * (2e_{si} * H[i][k] - \lambda * W[s][k])$$

$$20. \quad H[i][k] \leftarrow H[i][k] + \beta * (2e_{si} * W[s][k] - \lambda * H[i][k])$$

21. end for

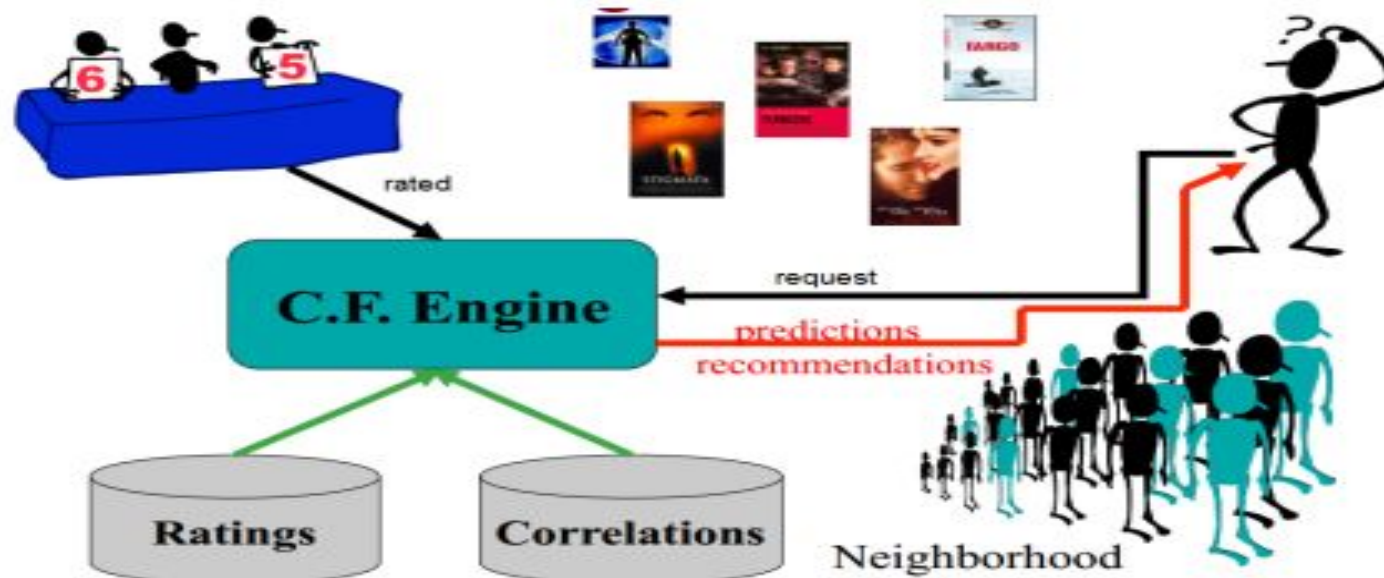
22. end while

23. return $\{W, H, b_s, b_i, \mu\}$

24. end procedure.

Kỹ thuật láng giềng

- Ý tưởng của Học cộng tác: “người tương tự” có thể thích “sản phẩm tương tự” hoặc ngược lại
- Xác định “mối tương quan” giữa các người dùng và các sản phẩm



Picture from <https://class.coursera.org/recsys-001/lecture>

Kỹ thuật láng giềng (cơ sở người dùng)

- Sử dụng “sự tương tự người dùng”
- Đo đo tương đồng của 2 người dùng
 - Cosine

$$sim_{cosine}(u, u') = \frac{\sum_{i \in I_{uu'}} r_{ui} \cdot r_{u'i}}{\sqrt{\sum_{i \in I_{uu'}} r_{ui}^2} \sqrt{\sum_{i \in I_{uu'}} r_{u'i}^2}}$$

- Pearson

$$sim_{pearson}(u, u') = \frac{\sum_{i \in I_{uu'}} (r_{ui} - \bar{r}_u)(r_{u'i} - \bar{r}_{u'})}{\sqrt{\sum_{i \in I_{uu'}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{uu'}} (r_{u'i} - \bar{r}_{u'})^2}}$$

Kỹ thuật mô hình láng giềng (tiếp)

- Hàm dự đoán
 - Tổng

$$\hat{r}_{ui} = \frac{\sum_{u' \in K_u} \text{sim}(u, u') \cdot r_{u'i}}{\sum_{u' \in K_u} |\text{sim}(u, u')|}$$

- Độ lệch

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{u' \in K_u} \text{sim}(u, u') \cdot (r_{u'i} - \bar{r}_{u'})}{\sum_{u' \in K_u} |\text{sim}(u, u')|}$$

Recommendation tasks: Example

Rating prediction from explicit feedback

- How would Steve rate the Titanic movie?

	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Joe	1	4	5		3
Ann	5	1		5	2
Mary	4	1	2	5	
Steve	?	3	4		4

Item recommendation from implicit feedback

- Which movie(s) Steve would like to see/buy?

	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Joe	1	1	1		1
Ann	1	1		1	1
Mary	1	1	1	1	
Steve	?	1	1	?	1

Thank Tomas Horváth for this example!

User similarity: Example

Cosine similarity:

$sim(u, u')$	Joe	Ann	Mary	Steve
Joe	1.0	0.283	0.372	0.962
Ann	—	1.0	0.915	0.232
Mary	—	—	1.0	0.254
Steve	—	—	—	1.0

Pearson similarity:

$sim(u, u')$	Joe	Ann	Mary	Steve
Joe	1.0	-0.716	-0.762	-0.005
Ann	—	1.0	0.972	0.565
Mary	—	—	1.0	0.6
Steve	—	—	—	1.0

Prediction using 2 most similar users: Example

rating prediction using 2 most similar users:

$$\blacktriangleright U_{Titanic} = \{Joe, Ann, Mary\},$$

$$K_{Steve,2}^{Titanic} = \{Mary, Ann\}$$

$$\blacktriangleright \bar{r}_{Steve} = \frac{11}{3} = 3.67 \quad \bar{r}_{Mary} = \frac{12}{4} = 3 \quad \bar{r}_{Ann} = \frac{13}{4} = 3.25$$

Using Pearson sim:

$$\blacktriangleright \hat{r}_{ST} = \bar{r}_S + \frac{sim(S,M) \cdot (r_{MT} - \bar{r}_M) + sim(S,A) \cdot (r_{AT} - \bar{r}_A)}{|sim(S,M)| + |sim(S,A)|} =$$
$$3.67 + \frac{0.6 \cdot (4 - 3) + 0.565 \cdot (5 - 3.25)}{0.6 + 0.565} = 1.36$$

Mô hình láng giềng(tiếp)

```
1: procedure USERKNN-CF ( $\bar{r}_u, r, D^{train}$ )
2: for  $u=1$  to  $N$  do
3:   Tính Sim_uu'
4: end for
5: Sort Sim_uu'
6: for  $k=1$  to  $K$  do
7:    $K_u \leftarrow k$ 
8: end for
9: for  $i = 1$  to  $M$  do
10:  Tính  $\widehat{r}_{ui}$ 
11: end for
12: end procedure
```

- Ưu điểm
 - Tính toán đơn giản
 - Có độ chính xác cao
- Nhược điểm
 - Vấn đề người dùng mới
 - Vấn đề sản phẩm mới
- Cách khắc phục
 - Kết hợp lọc cộng tác và lọc dựa trên một số thuộc tính của người dùng
 - Bổ sung thông tin về sản phẩm mới “NEW”

Kỹ thuật lán giềng (cơ sở sản phẩm)



Kỹ thuật láng giềng (tiếp)

- Đo đo tương đồng
 - Cosine

$$sim_{cosine}(i, i') = \frac{\sum_{u \in U_{ii'}} r_{ui} r_{ui'}}{\sqrt{\sum_{u \in U_{ii'}} r_{ui}^2} \cdot \sqrt{\sum_{u \in U_{ii'}} r_{ui'}^2}}$$

- Pearson

$$sim_{pearson}(i, i') = \frac{\sum_{u \in U_{ii'}} (r_{ui} - \bar{r}_i)(r_{ui'} - \bar{r}_{i'})}{\sqrt{\sum_{u \in U_{ii'}} (r_{ui} - \bar{r}_i)^2} \cdot \sqrt{\sum_{u \in U_{ii'}} (r_{ui'} - \bar{r}_{i'})^2}}$$

Kỹ thuật láng giềng (tiếp)

- Hàm dự đoán
 - Tổng

$$\hat{r}_{ui} = \frac{\sum_{i' \in K_i} \text{sim}(i, i') \cdot r_{ui'}}{\sum_{i' \in K_i} |\text{sim}(i, i')|}$$

- Độ lệch

$$\hat{r}_{ui} = \bar{r}_i + \frac{\sum_{i' \in K_i} \text{sim}(i, i') \cdot (r_{ui'} - \bar{r}_{i'})}{\sum_{i' \in K_i} |\text{sim}(i, i')|}$$

Recommendation tasks: Example

Rating prediction from explicit feedback

- How would Steve rate the Titanic movie?

	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Joe	1	4	5		3
Ann	5	1		5	2
Mary	4	1	2	5	
Steve	?	3	4		4

Item recommendation from implicit feedback

- Which movie(s) Steve would like to see/buy?

	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Joe	1	1	1		1
Ann	1	1		1	1
Mary	1	1	1	1	
Steve	?	1	1	?	1

Thank Tomas Horváth for this example!

User similarity: Example

Cosine similarity:

$sim(u, u')$	Joe	Ann	Mary	Steve
Joe	1.0	0.283	0.372	0.962
Ann	—	1.0	0.915	0.232
Mary	—	—	1.0	0.254
Steve	—	—	—	1.0

Pearson similarity:

$sim(u, u')$	Joe	Ann	Mary	Steve
Joe	1.0	-0.716	-0.762	-0.005
Ann	—	1.0	0.972	0.565
Mary	—	—	1.0	0.6
Steve	—	—	—	1.0

Item similarity: Example

Cosine similarity:

$sim(i, i')$	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Titanic	1.0	0.386	0.299	0.982	0.372
Pulp Fiction	—	1.0	0.975	0.272	0.929
Iron Man	—	—	1.0	0.211	0.858
Forrest Gump	—	—	—	1.0	263
The Mummy	—	—	—	—	1.0

Pearson similarity:

$sim(i, i')$	Titanic	Pulp Fiction	Iron Man	Forrest Gump	The Mummy
Titanic	1.0	-0.956	-0.815	NaN	-0.581
Pulp Fiction	—	1.0	0.948	NaN	0.621
Iron Man	—	—	1.0	NaN	0.243
Forrest Gump	—	—	—	1.0	NaN
The Mummy	—	—	—	—	1.0

NaN values are usually converted to zero, such cases should be rare in case of enough data

Prediction using 2 most similar items: Example

► $I_{\text{Steve}} = \{\text{Pulp Fiction}, \text{Iron Man}, \text{The Mummy}\}$

$$K_{\text{Titanic},2}^{\text{Steve}} = \{\text{Iron Man}, \text{The Mummy}\}$$

► $\bar{r}_T = \frac{10}{3} = 3.34, \quad \bar{r}_I = \frac{11}{3} = 3.67, \quad \bar{r}_M = \frac{9}{3} = 3$

Using Pearson sim:

►
$$\hat{r}_{ST} = \bar{r}_T + \frac{\text{sim}(T,I) \cdot (r_{SI} - \bar{r}_I) + \text{sim}(T,M) \cdot (r_{SM} - \bar{r}_M)}{|\text{sim}(T,I)| + |\text{sim}(T,M)|} =$$
$$3.34 + \frac{-0.815 \cdot (4 - 3.67) - 0.581 \cdot (4 - 3)}{0.815 + 0.581} = 2.73$$

Prediction using 2 most similar users: Example

rating prediction using 2 most similar users:

$$\blacktriangleright U_{Titanic} = \{Joe, Ann, Mary\},$$

$$K_{Steve,2}^{Titanic} = \{Mary, Ann\}$$

$$\blacktriangleright \bar{r}_{Steve} = \frac{11}{3} = 3.67 \quad \bar{r}_{Mary} = \frac{12}{4} = 3 \quad \bar{r}_{Ann} = \frac{13}{4} = 3.25$$

Using Pearson sim:

$$\blacktriangleright \hat{r}_{ST} = \bar{r}_S + \frac{sim(S,M) \cdot (r_{MT} - \bar{r}_M) + sim(S,A) \cdot (r_{AT} - \bar{r}_A)}{|sim(S,M)| + |sim(S,A)|} =$$
$$3.67 + \frac{0.6 \cdot (4 - 3) + 0.565 \cdot (5 - 3.25)}{0.6 + 0.565} = 1.36$$

Phương pháp dự đoán cơ sở

- Baseline : dung để kiểm tra
- Giải thuật đề xuất tốt hơn bao nhiêu
- Mục đích chính là kiểm tra chứ không phải so sánh
- Baseline thông dụng:
 - Trung bình toàn cục
 - Trung bình người dung
 - Trung bình sản phẩm
 - Phương pháp dự đoán cơ sở (baseline predictor)

Dự đoán toàn cục

- Hàm dự đoán

$$\hat{r}_{ui} = \mu = \frac{\sum_{(u,i,r) \in \mathcal{D}^{train}} r}{|\mathcal{D}^{train}|}$$

Training data

user	Item	rating
1	21	1
1	213	5
2	345	4
2	123	4
2	768	3
3	76	5
4	45	4
5	568	1
5	342	2
5	234	2
6	76	5
6	56	4

Test data

user	Item	rating
1	62	?
1	96	?
2	7	?
2	3	?
3	47	?
3	15	?
4	41	?
4	28	?
5	93	?
5	74	?
6	69	?
6	83	?

Trung bình người dùng

- Hàm dự đoán

$$\hat{r}_{ui} = \frac{\sum_{(u', i, r) \in \mathcal{D}^{train} | u' = u} r}{|\{(u', i, r) \in \mathcal{D}^{train} | u' = u\}|}$$

Training data

user	Item	rating
1	21	1
1	213	5
2	345	4
2	123	4
2	768	3
3	76	5
4	45	4
5	568	1
5	342	2
5	234	2
6	76	5
6	56	4

Test data

user	Item	rating
1	62	?
1	96	?
2	7	?
2	3	?
3	47	?
3	15	?
4	41	?
4	28	?
5	93	?
5	74	?
6	69	?
6	83	?

Trung bình sản phẩm

- Hàm dự đoán

$$\hat{r}_{ui} = \frac{\sum_{(u,i',r) \in \mathcal{D}^{train} | i'=i} r}{|\{(u,i',r) \in \mathcal{D}^{train} | i'=i\}|}$$

Training data

user	Item	rating
1	21	1
1	213	5
2	345	4
2	123	4
2	768	3
3	76	5
4	45	4
5	568	1
5	342	2
5	234	2
6	76	5
6	56	4

Test data

user	Item	rating
1	62	?
1	96	?
2	7	?
2	3	?
3	47	?
3	15	?
4	41	?
4	28	?
5	93	?
5	74	?
6	69	?
6	83	?

Phương pháp dự đoán cơ sở (baseline predictor)

- Hàm dự đoán

$$\hat{r}_{ui} = \mu + b_u + b_i$$

$$\mu = \frac{\sum_{(u,i,r) \in \mathcal{D}^{train}} r}{|\mathcal{D}^{train}|}$$

$$b_u = \frac{\sum_{(u',i,r) \in \mathcal{D}^{train} | u'=u} (r - \mu)}{|\{(u',i,r) \in \mathcal{D}^{train} | u' = u\}|}$$

$$b_i = \frac{\sum_{(u,i',r) \in \mathcal{D}^{train} | i'=i} (r - \mu)}{|\{(u,i',r) \in \mathcal{D}^{train}\} | i' = i|}$$