

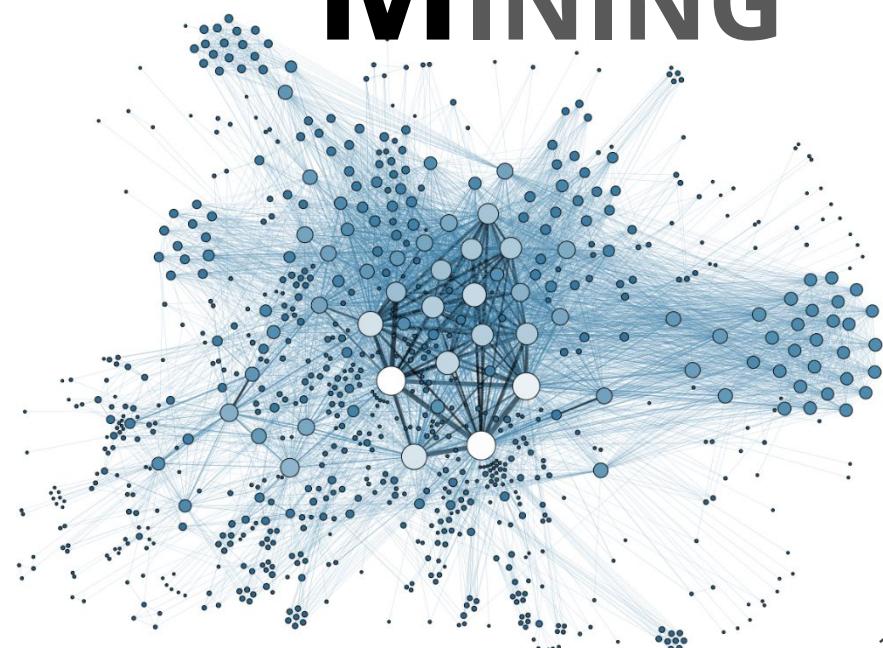
SOCIAL MEDIA MINING

Recommendation in Social Media

18TN

Nhóm:

Phạm Minh Khôi - 18120043
Nguyễn Hoàng Lân - 18120051



Sự khó khăn khi đưa ra quyết định

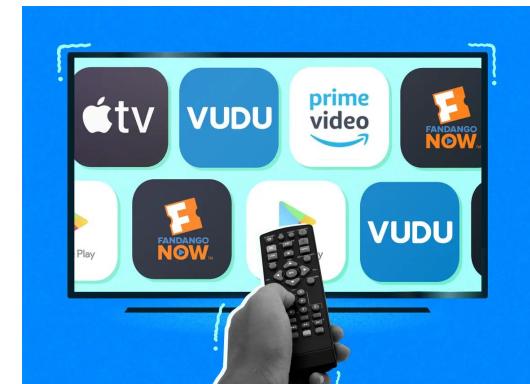
- Hằng ngày, chúng ta phải đưa ra quyết định đối với vô vàng các lựa chọn ở trên internet



Mua hàng online



Kết bạn



Tìm phim để xem

Sự khó khăn khi đưa ra quyết định

- Nên mua máy ảnh kỹ thuật số nào?
- Nên đi chơi dịp lễ ở đâu?
- Nên xem bộ phim nào
- Nên theo dõi ai trên Instagram?
- Nên đọc báo ở trang nào?
- Phim hay nhất cho gia đình của tôi?
- Có thể tìm xem thêm 2 hội nghị trước đây
 - SIGKDD2014, Recommendation in Social Media
 - RecSys2014, Personalized Location Recommendation

Khi nào khó khăn này xuất hiện?

- Khi có nhiều sự lựa chọn
- Khi không có lựa chọn nào thật sự nổi bật
- Khi chúng ta không thể cân nhắc hết mọi lựa chọn (**quá tải thông tin**)
- Khi ta không có đủ kiến thức và kinh nghiệm để chọn, hoặc
 - Do lười, nhưng vẫn muốn lựa chọn đúng
 - Không muốn chủ động quyết định

Mục tiêu của việc đề xuất:
Trả lời một danh sách các mục phù hợp với mong muốn của người dùng

Một số cách giải quyết thông thường

- Tham khảo bạn bè
- Tham khảo một nguồn tin cậy khác
- Thuê một nhóm các chuyên gia
- Tìm kiếm trên mạng
- Theo số đông:
 - Chọn các mục từ danh sách top-N
 - Chọn mua từ các best-seller trên Amazon

Có thể tự động hóa các công việc trên không?

- Sử dụng các thuật toán đề xuất
- Hay còn gọi là **các Hệ thống đề xuất**

Các hệ thống đề xuất - Một số ví dụ

Book recommendation in Amazon

Start reading *Networks: An Introduction* on your Kindle in under a minute.

Don't have a Kindle? Get your Kindle here, or download a FREE Kindle Reading App.

Get Your Copy For \$64.99
Receive a \$12.00 Amazon.com Gift Card for selling back this book. See other eligible items in our Books Trade-In Program. Restrictions Apply.

Used Price \$51.03
Buyback Price \$47.10
Price per Bookback \$3.93

More Buying Choices
36 used & new from \$51.03

Sell Your Copy
Sell on Amazon Listings start at \$0.99
Trade in Get a \$47.10 Amazon Gift Card

Share

Frequently Bought Together

Price For All Three: \$120.68

Add all three to Cart Add all three to Wish List

Show availability and shipping details

Customers Who Bought This Item Also Bought

Networks, Crowds, and Markets: Reasoning About a Highly Connected World by David Easley Hardcover \$41.47

Simply Complexity: A Clear Guide to Complexity Theory by Neil Johnson Paperback \$9.81

Networks, Crowds, and Markets: Reasoning About a... by David Easley Hardcover \$64.18

Dynamical Processes on Complex Networks by Alan L. Barabási Paperback \$44.52

Networks of the Brain by Olaf Sporns Paperback \$32.26

Networks, Crowds, and Markets: Reasoning About a... by David Easley Paperback \$41.47

Simply Complexity: A Clear Guide to Complexity Theory by Neil Johnson Paperback \$9.81

Social Network Analysis: Methods and Applications by Steven Wasserman Paperback \$44.52

Networks of the Brain by Olaf Sporns Paperback \$32.26

Page 1 of 20

Editorial Reviews

Video clip recommendation in YouTube

Arizona Wildfire Near Flagstaff at 10,000 Acres

fal2grace 1,023 videos Subscribe

Eagle Rock Wildfire Schultz Wildfire

ARIZONA

AP

510 views

Uploaded by fal2grace on Jun 22, 2010

AssociatedPress — Arizona authorities say a wildfire near Flagstaff has

15 likes, 0 dislikes

Uploader Comments: (fal2grace)

The "station fire" in California was in my home town. They evacuated my street further north in the hills. I spent a week just gazing at it with friends... powerful stuff that mother nature.

Like Add to Share

Suggestions

Schultz Fire - Flagstaff, AZ - June 20, 2010 by Bell0Virus 7,211 views

Flagstaff Father's Day Fire #2 - Schultz Wildfire by AssociatedPress 8,327 views

Winds Driving Fire in Ariz., Homes Threatened by AssociatedPress 1,491 views

Arizona wildfires rage on by NeverOnABC 141 views

Arizona wildfires third largest in state history by CBSNewsOnline 615 views

Arizona Governor Tours Growing Wildfire Near NM by AssociatedPress 18 views

Arizona wildfire barely contained by WMAR2News 59 views

Exercises in Chaos Theory

Product Recommendation in ebay

Recommendations for you

Dr Seuss's Second Beginner Book Collection \$21.00 See suggestions

Fox in Socks \$1.99 See suggestions

Ten Apples Up on Top! by Dr. Seuss \$1.95 Hardback \$3.50 See suggestions

Dr. Seuss's ABC by Dr. Seuss \$1.95 Hardback \$11.35 See suggestions

The Big Green Book of Beginner Books \$1.99 See suggestions

The Cat in the Hat \$0.99 See suggestions

Popular on ebay

10 x 3000mAh 18650 rechargeable batteries \$14.95 See suggestions

AAA 1800mAh 12V Rechargeable Battery X4 \$2.13 See suggestions

NEW 9V 900mAh Rechargeable Battery \$2.80 See suggestions

BP 9V Volt Ni-MH Rechargeable Battery \$1.99 See suggestions

8x AA 3000mAh Rechargeable Battery \$1.99 See suggestions

eBay stories eBay's hidden gem: eBay Radio

Concord, MA, Nov 20, 2011 Want to get the inside scoop on selling on ebay? Then join host Jon (Griff) Griffin on eBay Radio every Saturday morning at 8am ET for the eBay Stories. Read about how this show came into being. Continue reading →

See all stories

New eBay Go Together Easy to go to events with friends Check it out!

BUY Registration Buyer protection Bidder & buying help

SELL List selling Learn to sell Business sellers

EBAY COMPANIES eBay Classifieds Shopping Site Mail.com

ABOUT EBAY Company Info News Announcements

COMMUNITY Auction forums Answer center Discussion boards

Support Toys for Tots at the Give-a-Toy Store Visit now!

TOSHIBA Buy direct on ebay

Restaurant Recommendation in Yelp

Searching within Restaurant

www.yelp.com/?find_desc=8528&ll=&find_loc=Tempe%2C+AZ&n=1&attrs=ActiveDealfccff=restaurants&find_desc=restaurants&

yelp Real people. Real reviews.

Search for (e.g. taco, cheap dinner, Mac's) restaurants Near (Address, City, State or Zip) Tempe, AZ Search

Welcome About Me Write a Review Find Reviews Invite Friends Messaging Talk Events Member Search

OnStar FMV Now available for a special price! Click here to learn more

1 to 3 of 3 - Results per page 10

Concierge Too many options?

1. The Dhaba Categories Indian, Pakistani \$10 for \$20 Certificate

2. China Farm Chinese Buffet Categories Chinese, Buffet, Food Delivery Services \$8 for \$10 Certificate

3. Capriotti's Sandwich Shop Categories Sandwiches \$7 for \$15 Certificate

Less Map

Scotsdale Indian School Rd North East Village North Mesa Sherrill Heights McDowell Rd North Peoria Parkway Salt River Apache Blvd E. Broadway Rd. Gildland W. Southern Ave. Southern Garden Business Center E. Southern Ave. Sunbelt Galleria Cypres Southwest Firme Park Village

Got search feedback? Help us improve!

Related Talk Topics

Các hệ thống đề xuất - Một số ví dụ



Amazon - Đề xuất các món hàng



Spotify - Đề xuất các bài nhạc



Facebook - Đề xuất kết bạn mới



Netflix - Đề xuất phim

Ý tưởng chính của các hệ thống đề xuất

Sử dụng dữ liệu quá khứ, chẳng hạn như sở thích của người dùng trong quá khứ để dự đoán cho tương lai

- Sở thích của người dùng thường không thay đổi, hoặc thay đổi chậm sau một thời gian.
 - Chúng ta dự đoán bằng cách theo dõi sở thích của họ hoặc của một nhóm người trong quá khứ
- Thuật toán đề xuất nhận vào một tập hợp người dùng U , một tập hợp các mục I , và học một hàm f sao cho:

$$f : U \times I \rightarrow \mathbb{R}$$

Ý tưởng chính của các hệ thống đề xuất

- Nói cách khác, hệ thống đề xuất sẽ **dự đoán mức độ đánh giá** của người dùng đối với các mục, sau đó gợi ý các mục có độ tương đồng cao nhất cho họ.

John	5	1	3	5
Tom	?	?	?	2
Alice	4	?	3	?

Ma trận Người dùng-Mục

Đề xuất so với Tìm kiếm

- Một cách khác để tìm kiếm câu trả lời là sử dụng **hệ thống Tìm kiếm**
- Hệ thống Tìm kiếm đưa ra kết quả tương ứng với câu truy vấn của người dùng
- Các kết quả này thường cung cấp một danh sách liên quan đến từ khóa tìm kiếm
- Xem xét câu truy vấn "**phim ngôn tình hay nhất năm 2014**"
 - Kết quả trả ra từ hệ thống Tìm kiếm cho một đứa trẻ 8 tuổi sẽ giống với cho một người lớn

Hệ thống tìm kiếm không thể tùy chỉnh phù hợp với người dùng

Các thử thách đối với hệ thống Đề xuất

- **Vấn đề khởi động nguội**

- Khi một người dùng mới truy cập vào hệ thống, họ chưa từng mua một sản phẩm nào do đó họ không có "lịch sử".
- Khó khăn trong việc dự đoán một người thích gì khi họ truy cập hệ thống lần đầu.
- Ví dụ: Netflix yêu cầu người dùng chọn một số phim yêu thích khi mới đăng ký tài khoản

- **Sự thưa thớt của dữ liệu**

- Do người dùng thường bỏ qua các bước đánh giá, hoặc chỉ đánh giá số ít sản phẩm → thiếu dữ liệu
- Khác với vấn đề khởi động nguội, vấn đề này xảy ra với toàn bộ hệ thống chứ không với một cá nhân nhất định

Các thử thách đối với hệ thống Đề xuất

- **Dễ bị tấn công**

- **Push Attack**: tạo nhiều tài khoản giả mạo để đánh giá sản phẩm khác cao → đẩy sản phẩm lên để được đề xuất cho nhiều người dùng hơn.
 - **Nuke attack**: tấn công DDoS, khiến cả hệ thống mất ổn định

- **Tính bảo mật**

- Người dùng thường không để lại các đánh giá do lo ngại thông tin nhạy cảm có thể bị lộ.
 - Hệ thống cần có các biện pháp để bảo mật thông tin khách hàng.

- **Tính giải thích được**

- Đa số các hệ thống đề xuất không thể giải thích được tại sao lại đề xuất các mục đó

Cuộc thi về Hệ thống Đề xuất - The Netflix Prize



The Netflix Prize tổ chức năm 2006-2009

Cuộc thi về Hệ thống Đề xuất - The Netflix Prize

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

Netflix is canceling its second \$1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine. Friday's announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to [...]

Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

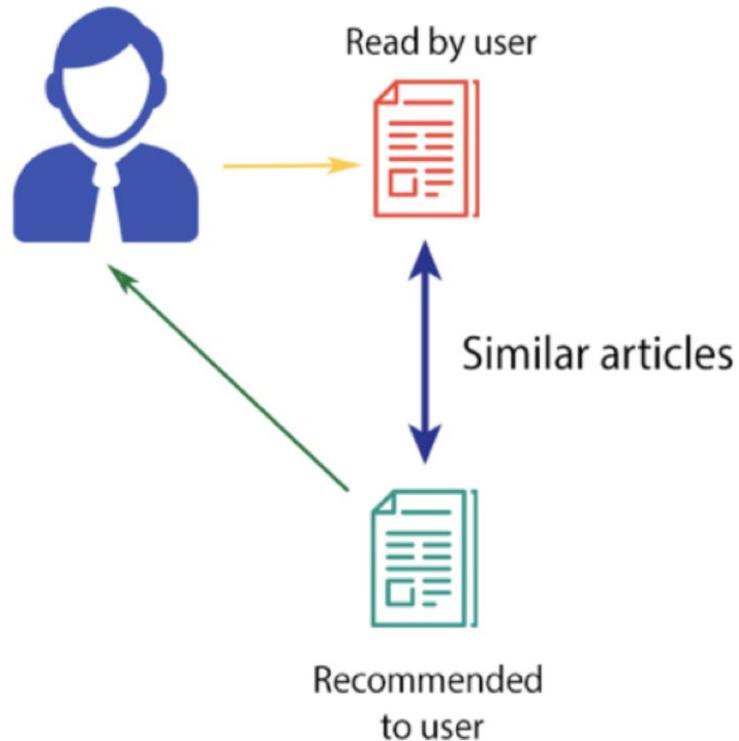
We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Các thuật toán Đề xuất cổ điển

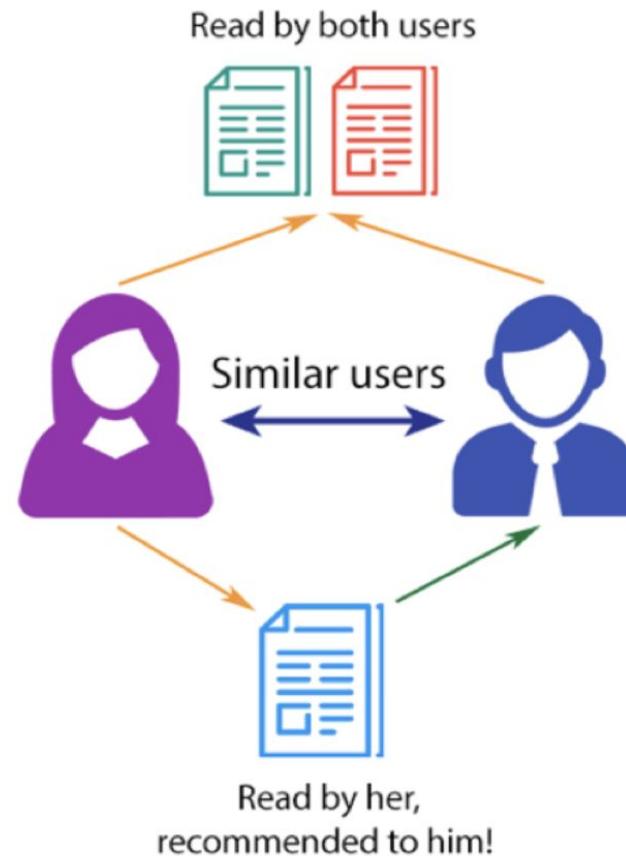
- Các thuật dựa trên nội dung (Content-based)
- Lọc cộng tác (Collaborative filtering)

Các thuật toán đề xuất cổ điển

CONTENT-BASED FILTERING



COLLABORATIVE FILTERING



Phương pháp dựa trên nội dung

Giả định: các mục mà hệ thống đề xuất cho người dùng nên tương đồng với sở thích của họ

- Độ tương đồng của các mục được đề xuất đối với người dùng càng cao, thì càng chắc chắn rằng người dùng sẽ thấy mục đó thú vị

Mục tiêu: tìm sự tương đồng giữa các người dùng và giữa các mục là cốt lõi của loại hệ thống Đề xuất này

Đề xuất dựa trên nội dung: Ví dụ

Cơ
sở
dữ
liệu

Title	Genre	Author	Type	Price	Keywords
<i>The Night of the Gun</i>	Memoir	David Carr	Paperback	29.90	press and journalism, drug addiction, personal memoirs, New York
<i>The Lace Reader</i>	Fiction, Mystery	Brunonia Barry	Hardcover	49.90	American contemporary fiction, detective, historical
<i>Into the Fire</i>	Romance, Suspense	Suzanne Brockmann	Hardcover	45.90	American fiction, murder, neo-Nazism
...					

Thông
tin
người
dùng

Title	Genre	Author	Type	Price	Keywords
...	Fiction, Suspense	Brunonia Barry, Ken Follett	Paperback	25.65	detective, murder, New York

Đề xuất dựa trên nội dung

1. Tạo thông tin mô tả của các mục đề xuất
2. Tạo hồ sơ của người dùng lưu thông tin những mục mà người đó thích
3. So sánh các mục với hồ sơ của người dùng để xác định nên đề xuất những gì

Hồ sơ người dùng thông thường sẽ được tạo và cập nhật tự động theo những đánh giá của họ cho với các mục.

Đề xuất dựa trên nội dung: Ví dụ

The screenshot shows the 'Edit Favorites' section of the Amazon website. It includes a 'Categories' section with checked boxes for Books, Biographies & Memoirs, Business & Investing, and Computers & Internet. There are also sections for 'Your Books Favorites' and 'Add to Your Favorites' with various other category options.

Hồ sơ người dùng

The screenshot shows the 'Recommended For You' section for Books on the Amazon website. It highlights recommendations based on the user's previous purchases. Two books are shown: 'The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture' by John Battelle and 'Writing Successful Science Proposals' by Andrew J. Friedland, Carol L Folt.

Mục được đề xuất bởi
hệ thống

Đề xuất dựa trên nội dung

- Hồ sơ người dùng và mô tả của mục sẽ được vecto hóa thành vecto có k đặc trưng (ví dụ sử dụng **TF-IDF**)
- Sau khi vecto hóa, ta so sánh độ tương đồng giữa 2 vecto này bằng độ tương đồng cosin

$$I_j = (i_{j,1}, i_{j,2}, \dots, i_{j,k}) \quad U_i = (u_{i,1}, u_{i,2}, \dots, u_{i,k})$$

$$\text{sim}(U_i, I_j) = \cos(U_i, I_j) = \frac{\sum_{l=1}^k u_{i,l} i_{j,l}}{\sqrt{\sum_{l=1}^k u_{i,l}^2} \sqrt{\sum_{l=1}^k i_{j,l}^2}}$$

Ta đề xuất top những mục tương đồng nhất cho người dùng

Đề xuất dựa trên nội dung: Mã giả

Algorithm 1: Phương pháp đề xuất dựa trên nội dung

Yêu cầu đầu vào: Thông tin hồ sơ của người dùng i , mô tả của mục $j \in \{1,2,\dots,n\}$, k từ khóa, r số lượng đề xuất.

1. Trả về r mục đề xuất
 2. $U_i = (u_1, u_2, \dots, u_k) =$ vectơ hồ sơ người dùng thứ i
 3. $\{I_j\}_{j=1}^n = \{(i_{j1}, u_{j2}, \dots, u_{jk}) =$ vectơ mô tả mục thứ $j\}_{j=1}^n$
 4. $s_{ij} = sim(U_i, I_j), 1 \leq j \leq n$
 5. Trả về top r mục với độ tương đồng cao nhất s_{ij}
-

- Ta chọn những mục có độ tương đồng cao nhất đối với người dùng j sau đó đề xuất chúng theo thứ tự

Lọc cộng tác

Lọc cộng tác: là quá trình lựa chọn thông tin hoặc khuôn mẫu sử dụng kĩ thuật liên quan đến việc cộng tác giữa nhiều đối tượng, góc nhìn, nguồn dữ liệu,...

Điểm lợi thế: chúng ta không nhất thiết phải có thêm thông tin chi tiết về người dùng hoặc mục

- Thuật toán hoạt động chỉ dựa trên đánh giá của người dùng hoặc lịch sử mua hàng

Ma trận Đánh giá: Ví dụ

Movies You've Rated

Based on your 745 movie ratings, this is the list of movies you've seen. As you discover movies on the website that you've seen, rate them and they will show up on this list. On this page, you may change the rating for any movie you've seen, and you may remove a movie from this list by clicking the 'Clear Rating' button.

TITLE	MPAA	GENRE	STAR RATING
Add 12 Angry Men (1957)	UR	Classics	Clear Rating
Add The 39 Steps (1935)	UR	Classics	Clear Rating
Add An American in Paris (1951)	UR	Classics	Clear Rating
Add The Andromeda Strain (1971)	G	Sci-Fi & Fantasy	Clear Rating
Add Apollo 13 (1995)	PG	Drama	Clear Rating
Add The Battle of Algiers (1965) La Battaglia di Algeri	UR	Foreign	Clear Rating
Add Being There (1979)	PG	Drama	Clear Rating
Add Big Deal on Madonna Street (1958) I soliti ignoti	UR	Foreign	Clear Rating
Add The Birds (1963)	PG-13	Thrillers	Clear Rating
Add Blade Runner (1982)	R	Sci-Fi & Fantasy	Clear Rating

Value	Graphic representation	Textual representation
5		Excellent
4		Very good
3		Good
2		Fair
1		Poor



Table 9.1: User-Item Matrix

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1

Ma trận Đánh giá

Người dùng sẽ đánh giá (xếp hạng) những mục (mà họ đã mua hoặc xem)

Đánh giá tường minh:

- Được nhập trực tiếp bởi người dùng
- Ví dụ, “Đánh giá trên thang điểm 0-10”



Đánh giá tiềm ẩn:

- Được suy ra từ hành vi của người dùng
- Ví dụ, những bài nhạc hoặc sách mà người dùng hay mở, lượng thời gian mà người dùng sử dụng một trang web

Lọc cộng tác

Các loại thuật toán Lọc cộng tác:

- **Dựa trên bộ nhớ:** Sự đề xuất được dựa trực tiếp trên những đánh giá trước đó được lưu dưới dạng ma trận Người dùng - mục
- **Dựa trên mô hình:** Giả sử rằng tồn tại một mô hình chi phối cách mà người dùng đánh giá các mục.
 - Mô hình này có thể được xấp xỉ và học
 - Mô hình này sau đó sẽ được dùng để dự đoán đánh giá của người dùng
 - **Ví dụ:** người dùng thường đánh giá thấp những phim có vốn đầu tư ít

Lọc cộng tác dựa trên bộ nhớ

Gồm hai phương pháp chính

Dựa trên người dùng

Những người dùng có những đánh giá **trước đó** tương tự nhau sẽ có khả năng đánh giá tiếp các mục tương tự trong tương lai

	I1	I2	I3	I4
U1	1	2	4	4
U2	1	2	4	?
U3	2	5	2	2
U4	5	2	3	3

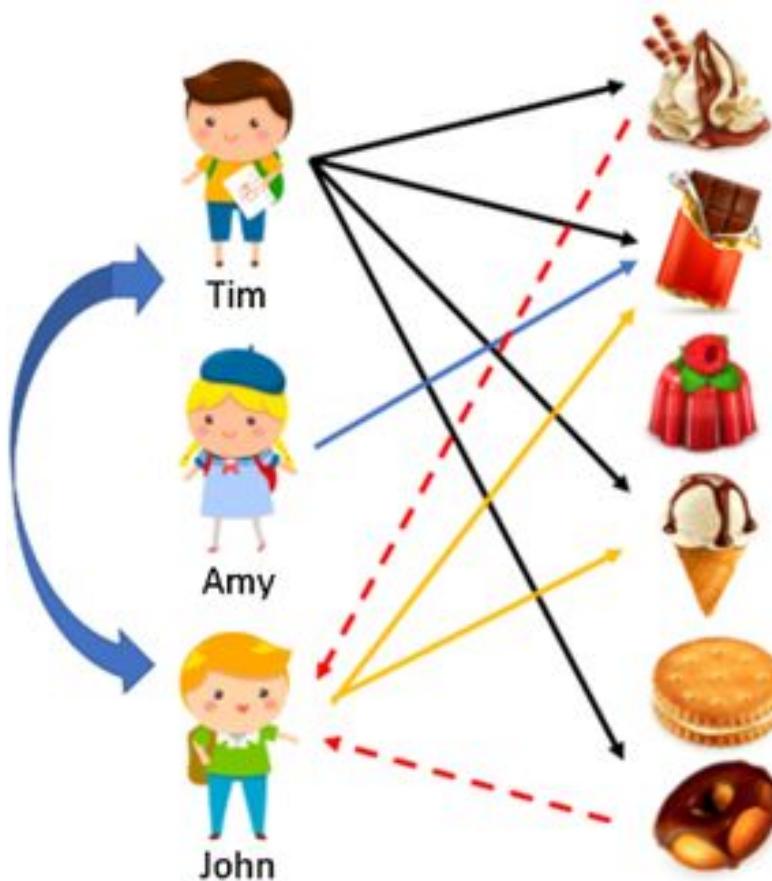
Dựa trên mục

Những mục **trước đó** đã nhận được đánh giá tương tự nhau có khả năng sẽ được đánh giá tương tự trong tương lai

	I1	I2	I3	I4
U1	1	2	4	4
U2	1	2	4	?
U3	2	5	2	2
U4	5	2	3	3

Lọc cộng tác dựa trên bộ nhớ

Dựa trên người dùng

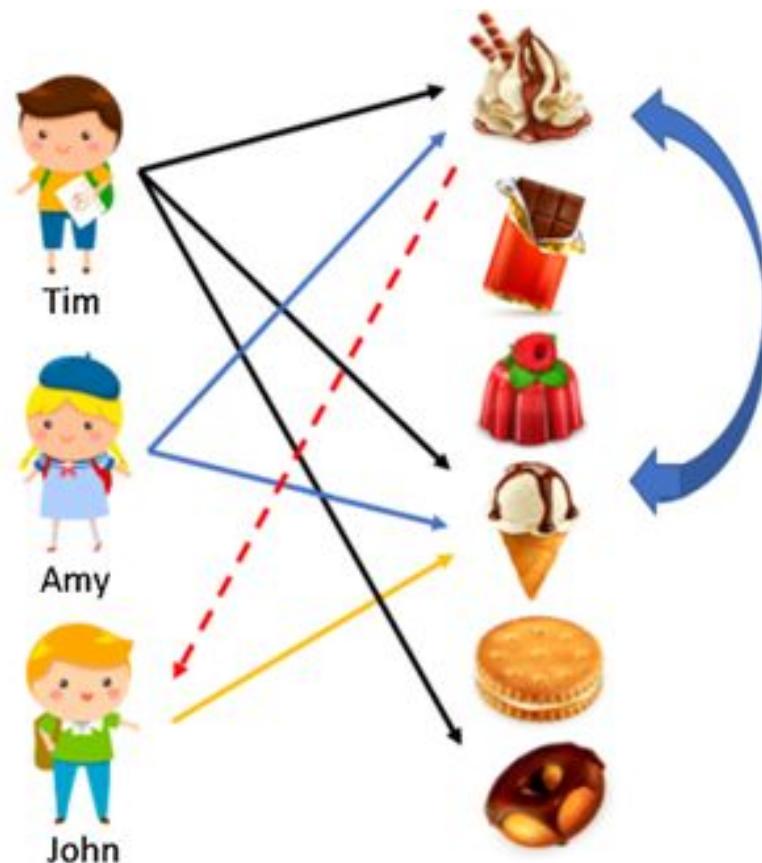


(a) User-based filtering

⇒ Những món V
thích sẽ được gợi ý
cho U

Lọc cộng tác dựa trên bộ nhớ

Dựa trên mục



⇒ Những người nào
thích món J, sẽ được
gợi ý cho món I

(b) Item-based filtering

Lọc cộng tác: Thuật toán chính

1. Đánh trọng số tất cả cặp người dùng / mục tương ứng với độ tương đồng của họ với mục
2. Chọn một tập láng giềng của người/mục cho việc đề xuất
3. Dự đoán đánh giá của người dùng cho mục sử dụng đánh giá của các láng giềng cho cùng mục đó
4. Đề xuất các mục với xếp hạng dự đoán cao nhất

Tính toán độ tương đồng giữa người dùng (hoặc mục)

Cosine Similarity

$$sim(U_u, U_v) = cos(U_u, U_v) = \frac{U_u \cdot U_v}{\|U_u\| \|U_v\|} = \frac{\sum_i r_{ui} r_{vi}}{\sqrt{\sum_i r_{ui}^2} \sqrt{\sum_i r_{vi}^2}}$$

Pearson Correlation Coefficient

$$sim(U_u, U_v) = \frac{\sum_i (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_i (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_i (r_{vi} - \bar{r}_v)^2}}$$

Lọc cộng tác dựa trên người dùng

Cập nhật đánh giá:

Điểm đánh giá trung bình
của người dùng u

$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in N(u)} sim(u,v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in N(u)} sim(u,v)}$$

Điểm đánh giá
của người dùng
u cho mục i

Điểm đánh giá trung
bình của người dùng v

Điểm đánh giá của người
dùng v cho mục i

Lọc cộng tác dựa trên người dùng: Ví dụ

	Lion King	Aladdin	Mulan	Anastasia
John	3	0	3	3
Joe	5	4	0	2
Jill	1	2	4	2
Jane	3	?	1	0
Jorge	2	2	0	1

Cần dự đoán đánh giá của Jane cho phim Aladdin

1- Tính trung bình các đánh giá

$$\bar{r}_{John} = \frac{3 + 3 + 0 + 3}{4} = 2.25$$

$$\bar{r}_{Joe} = \frac{5 + 4 + 0 + 2}{4} = 2.75$$

$$\bar{r}_{Jill} = \frac{1 + 2 + 4 + 2}{4} = 2.25$$

$$\bar{r}_{Jane} = \frac{3 + 1 + 0}{3} = 1.33$$

$$\bar{r}_{Jorge} = \frac{2 + 2 + 0 + 1}{4} = 1.25$$

2- Tính độ tương đồng giữa Jane với những người còn lại

$$sim(Jane, John) = \frac{3 \times 3 + 1 \times 3 + 0 \times 3}{\sqrt{10} \sqrt{27}} = 0.73$$

$$sim(Jane, Joe) = \frac{3 \times 5 + 1 \times 0 + 0 \times 2}{\sqrt{10} \sqrt{29}} = 0.88$$

$$sim(Jane, Jill) = \frac{3 \times 1 + 1 \times 4 + 0 \times 2}{\sqrt{10} \sqrt{21}} = 0.48$$

$$sim(Jane, Jorge) = \frac{3 \times 2 + 1 \times 0 + 0 \times 1}{\sqrt{10} \sqrt{5}} = 0.84$$

Lọc cộng tác dựa trên người dùng: Ví dụ

3- Tính toán đánh giá của Jane cho phim Aladdin, xét số láng giềng = 2

$$\begin{aligned} r_{Jane,Aladdin} &= \bar{r}_{Jane} + \frac{sim(Jane, Joe)(r_{Joe,Aladdin} - \bar{r}_{Joe})}{sim(Jane, Joe) + sim(Jane, Jorge)} \\ &\quad + \frac{sim(Jane, Jorge)(r_{Jorge,Aladdin} - \bar{r}_{Jorge})}{sim(Jane, Joe) + sim(Jane, Jorge)} \\ &= 1.33 + \frac{0.88(4 - 2.75) + 0.84(2 - 1.25)}{0.88 + 0.84} = 2.33 \end{aligned}$$

Lọc cộng tác dựa trên mục

- Tính toán độ tương đồng giữa các mục, sau đó dự đoán đánh giá của mục mới dựa trên những mục tương đồng nhất

$$r_{u,i} = \bar{r}_i + \frac{\sum_{j \in N(i)} sim(i,j)(r_{u,j} - \bar{r}_j)}{\sum_{j \in N(i)} sim(i,j)}$$

Điểm đánh giá trung
bình của mục i

i và j là 2 mục tương tự
nhau

Lọc cộng tác dựa trên người mục: Ví dụ

1- Tính toán đánh giá trung bình của các phim

$$\bar{r}_{Lion\ King} = \frac{3 + 5 + 1 + 3 + 2}{5} = 2.8$$

$$\bar{r}_{Aladdin} = \frac{0 + 4 + 2 + 2}{4} = 2.$$

$$\bar{r}_{Mulan} = \frac{3 + 0 + 4 + 1 + 0}{5} = 1.6$$

$$\bar{r}_{Anastasia} = \frac{3 + 2 + 2 + 0 + 1}{5} = 1.6$$

2- Tính toán độ tương đồng giữa các mục

$$sim(Aladdin, Lion\ King) = \frac{0 \times 3 + 4 \times 5 + 2 \times 1 + 2 \times 2}{\sqrt{24} \sqrt{39}} = 0.84$$

$$sim(Aladdin, Mulan) = \frac{0 \times 3 + 4 \times 0 + 2 \times 4 + 2 \times 0}{\sqrt{24} \sqrt{25}} = 0.32$$

$$sim(Aladdin, Anastasia) = \frac{0 \times 3 + 4 \times 2 + 2 \times 2 + 2 \times 1}{\sqrt{24} \sqrt{18}} = 0.67$$

3- Tính toán đánh giá của Jane cho phim Aladdin, xét số lảng giềng = 2

$$\begin{aligned} r_{Jane, Aladdin} &= \bar{r}_{Aladdin} + \frac{sim(Aladdin, Lion\ King)(r_{Jane, Lion\ King} - \bar{r}_{Lion\ King})}{sim(Aladdin, Lion\ King) + sim(Aladdin, Anastasia)} \\ &\quad + \frac{sim(Aladdin, Anastasia)(r_{Jane, Anastasia} - \bar{r}_{Anastasia})}{sim(Aladdin, Lion\ King) + sim(Aladdin, Anastasia)} \\ &= 2 + \frac{0.84(3 - 2.8) + 0.67(0 - 1.6)}{0.84 + 0.67} = 1.40 \end{aligned}$$

Lọc cộng tác dựa trên bộ nhớ

- Lọc cộng tác dựa trên người dùng
 - Hệ thống tìm những người dùng tương tự với người dùng hiện tại (các láng giềng) và sử dụng sở thích của họ để đề xuất
- Phương pháp dựa trên người dùng không thông dụng bằng phương pháp dựa trên mục
 - **Vì sao?** Với một lượng lớn người dùng, chỉ cần một thay đổi nhỏ trong dữ liệu người dùng có thể thay đổi nhóm các người dùng tương tự

Lọc cộng tác dựa trên mô hình

- **Phương pháp dựa trên bộ nhớ**
 - Dự đoán đánh giá còn thiếu dựa trên độ tương đồng giữa người dùng và mục
- **Phương pháp dựa trên mô hình**
 - Giả sử rằng có một mô hình điều khiển đánh giá của người dùng
 - Ta tìm một mô hình sử dụng để dự đoán đánh giá.
 - Một phương pháp dựa trên mô hình phổ biến là phân rã SVD

Singular Value Decomposition (SVD)

- SVD là một kỹ thuật đại số tuyến tính, cho trước một ma trận thực $X \in \mathbb{R}^{m \times n}$, $m \geq n$, phân tích nó thành ba ma trận:

$$X = U\Sigma V^T$$

- Ma trận $U \in \mathbb{R}^{m \times m}$ và $V \in \mathbb{R}^{n \times n}$ là các ma trận trực giao và ma trận $\Sigma \in \mathbb{R}^{m \times n}$ là ma trận đường chéo
- Tích của những ma trận này bằng với ma trận ban đầu

Không xảy ra mất mát thông tin!

Xấp xỉ hạng thấp của ma trận

- Một xấp xỉ ma trận hạng thấp của ma trận $X \in \mathbb{R}^{m \times n}$ là một ma trận $C \in \mathbb{R}^{m \times n}$
- Ma trận C xấp xỉ X , và hạng của C (số cột độc lập tuyến tính tối đa) là hằng số $k \ll \min(m, n)$
$$Rank(C) = k$$
- Xấp xỉ ma trận hạng thấp tốt nhất là khi C cực tiểu hàm
$$\|X - C\|_F$$

$$\|X\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n X_{ij}^2}$$

- Xấp xỉ hạng thấp có thể giúp khử nhiễu bằng cách xem ma trận không phải ngẫu nhiên và có cấu trúc đơn giản
- SVD có thể dùng để tính xấp xỉ hạng thấp của ma trận

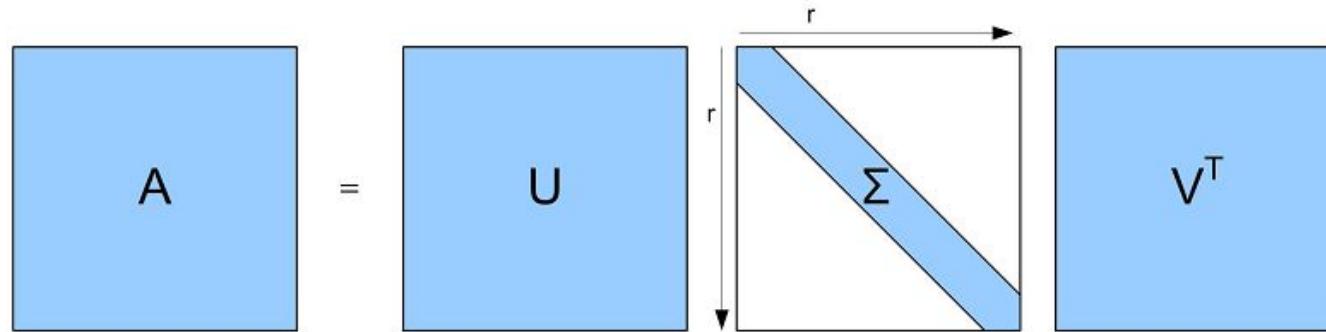
Xấp xỉ hạng thấp của ma trận bằng SVD

1. Tạo Σ_k từ Σ bằng cách chỉ giữ k phần tử đầu tiên trên đường chéo. Khi đó $\Sigma_k \in \mathbb{R}^{k \times k}$
2. Giữ lại k cột đầu của U và gọi đó là $U_k \in \mathbb{R}^{m \times k}$, và giữ lại k dòng đầu của V^T và gọi đó là $V_k^T \in \mathbb{R}^{k \times n}$
3. Gọi $X_k = U_k \Sigma_k V_k^T, X_k \in \mathbb{R}^{m \times n}$

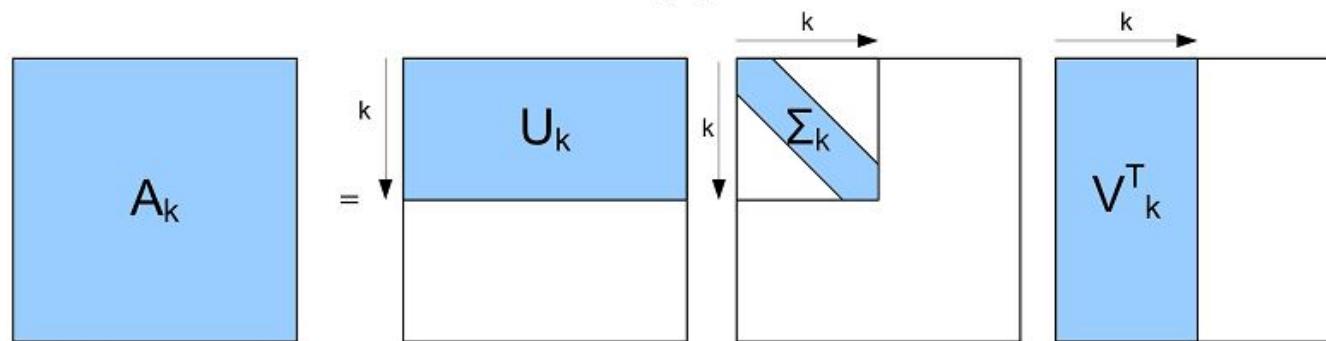
Định lý 2. (Xấp xỉ ma trận bậc thấp Eckart-Young-Mirsky). Gọi X là ma trận và C là xấp xỉ cấp thấp nhất của X ; nếu $\|X - C\|_F$ là cực tiểu và $\text{hạng}(C) = k$, thì $C = X_k$

X_k là xấp xỉ hạng thấp tốt nhất của ma trận X

Lọc cộng tác dựa trên mô hình: Ví dụ



(a)



(b)

Xếp xỉ hạng k của ma trận

Lọc cộng tác dựa trên mô hình: Ví dụ

Table 9.2: An User-Item Matrix

	Lion King	Aladdin	Mulan
John	3	0	3
Joe	5	4	0
Jill	1	2	4
Jorge	2	2	0

$$U = \begin{bmatrix} -0.4151 & -0.4754 & -0.7679 & 0.1093 \\ -0.7437 & 0.5278 & 0.0169 & -0.4099 \\ -0.4110 & -0.6626 & 0.6207 & -0.0820 \\ -0.3251 & 0.2373 & 0.1572 & 0.9018 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 8.0265 & 0 & 0 \\ 0 & 4.3886 & 0 \\ 0 & 0 & 2.0777 \\ 0 & 0 & 0 \end{bmatrix}$$

$$V^T = \begin{bmatrix} -0.7506 & -0.5540 & -0.3600 \\ 0.2335 & 0.2872 & -0.9290 \\ -0.6181 & 0.7814 & 0.0863 \end{bmatrix}$$

Xét một xấp xỉ hạng 2 ($k = 2$), ta cắt bỏ các dòng và cột tương ứng từ 3 ma trận

$$U_k = \begin{bmatrix} -0.4151 & -0.4754 \\ -0.7437 & 0.5278 \\ -0.4110 & -0.6626 \\ -0.3251 & 0.2373 \end{bmatrix}$$

$$\Sigma_k = \begin{bmatrix} 8.0265 & 0 \\ 0 & 4.3886 \end{bmatrix}$$

$$V_k^T = \begin{bmatrix} -0.7506 & -0.5540 & -0.3600 \\ 0.2335 & 0.2872 & -0.9290 \end{bmatrix}$$

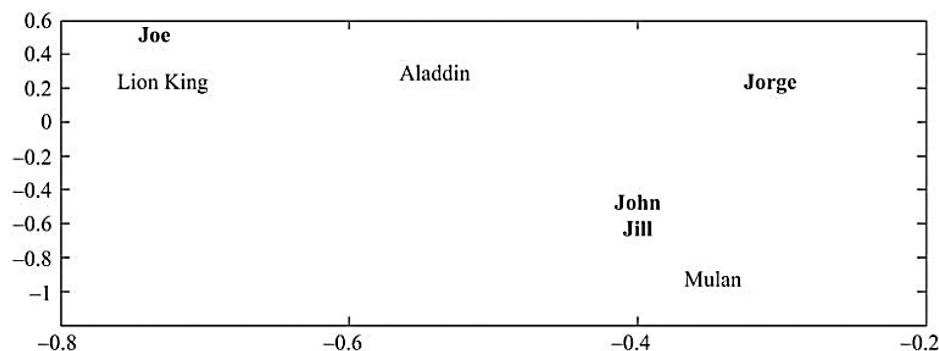


Figure 9.1: Users and Items in the 2-D Space.

Đề xuất cho một nhóm/ cộng đồng

Đề xuất cho các nhóm

- Tìm kiếm nội dung nào được yêu thích chung giữa một nhóm các cá nhân
- Ví dụ:
 - Một bộ phim cho nhóm bạn xem chung
 - Một địa điểm du lịch cho gia đình
 - Một nhà hàng tốt cho các đồng nghiệp ăn trưa
 - Một bài nhạc để có thể mở nơi công cộng

Nhiệm vụ của các hệ thống đề xuất cho nhóm

- Tìm kiếm sở thích chung
- Tạo ra các đề xuất
- Giải thích được các đề xuất
- Giúp nhóm người có thể nhất trí với nhau

Các chiến lược tổng hợp

Cực đại hóa mức hài lòng trung bình

- Tính đánh giá chung bình của nhóm người cho mục i, và chọn giá trị lớn nhất

$$R_i = \frac{1}{n} \sum_{u \in G} r_{u,i}$$

Ít mâu thuẫn nhất

- Tối thiểu hóa sự không hài lòng của nhóm người (giá trị lớn nhất của cực tiểu)

$$R_i = \min_{u \in G} r_{u,i}$$

Hài lòng nhiều nhất

- Lấy giá trị lớn nhất giữa những đánh giá lớn nhất của mọi người

$$R_i = \max_{u \in G} r_{u,i}$$

Đề xuất cho một nhóm: Ví dụ

Table 9.3: User-Item Matrix

	Soda	Water	Tea	Coffee
John	1	3	1	1
Joe	4	3	1	2
Jill	2	2	4	2
Jorge	1	1	3	5
Juan	3	3	4	5

Giả sử John, Jill, Juan là những người bạn của nhau

Cực đại hóa mức hài lòng trung bình

$$R_{Soda} = \frac{1 + 2 + 3}{3} = 2.$$

$$R_{Water} = \frac{3 + 2 + 3}{3} = 2.66$$

$$R_{Tea} = \frac{1 + 4 + 4}{3} = 3.$$

$$R_{Coffee} = \frac{1 + 2 + 5}{3} = 2.66$$

Ít mâu thuẫn nhất

$$R_{Soda} = \min\{1, 2, 3\} = 1$$

$$R_{Water} = \min\{3, 2, 3\} = 2$$

$$R_{Tea} = \min\{1, 4, 4\} = 1$$

$$R_{Coffee} = \min\{1, 2, 5\} = 1$$

Hài lòng nhiều nhất

$$R_{Soda} = \max\{1, 2, 3\} = 3$$

$$R_{Water} = \max\{3, 2, 3\} = 3$$

$$R_{Tea} = \max\{1, 4, 4\} = 4$$

$$R_{Coffee} = \max\{1, 2, 5\} = 5$$

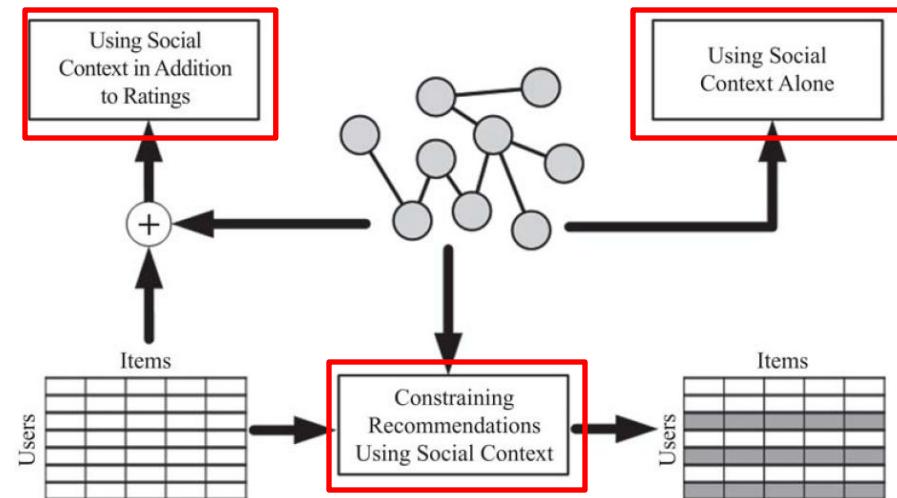
Đề xuất sử dụng ngữ cảnh xã hội

- Đề xuất chỉ sử dụng ngữ cảnh xã hội
- Mở rộng các phương pháp cổ điển bằng với ngữ cảnh xã hội
- Đề xuất bị ràng buộc bởi ngữ cảnh xã hội

Thông tin có sẵn trong ngữ cảnh xã hội

- Trên phương tiện truyền thông xã hội, ngoài đánh giá xếp hạng của sản phẩm, còn có thông tin bổ sung
 - VD: mạng lưới quan hệ giữa cá nhân và bạn bè của họ,...
- Thông tin này có thể được sử dụng để cải thiện các đề xuất

- Giả sử rằng bạn bè có tác động đến xếp hạng của cá nhân đó.
- Tác động này có thể là do tình trạng không ổn định hoặc bị can thiệp.



I. Chỉ sử dụng riêng thông tin ngữ cảnh xã hội

- Hãy xem xét một mạng bạn bè mà không có ma trận đánh giá mục người dùng nào được cung cấp.
- Trong mạng này, chúng ta vẫn có thể đề xuất người dùng trong mạng với các người dùng khác.
- Phương pháp kinh điển Link Prediction, sử dụng để đề xuất bạn bè trong mạng xã hội sẽ được trình bày ở chương 10.

Ví dụ

Trong mạng xã hội, các cá nhân thường hình thành bộ ba bạn bè.

- Bộ ba cá thể a, b, c với ba cạnh $e(a, b)$, $e(b, c)$, $e(c, a)$.
 - Một bộ ba bị thiếu một trong ba cạnh là một bộ ba mở.
- > Tìm bộ ba mở để đề xuất bạn bè.

II. Mở rộng các phương pháp cổ điển

- Sử dụng thông tin ngữ cảnh xã hội vào ma trận xếp hạng mục người dùng để cải thiện đề xuất
- Thông tin xã hội được bổ sung
 - Chúng ta giả định rằng bạn bè đều đánh giá các mục tương tự nhau.

$$R_{ij} = U_i^T V_j$$

$$R = U^T V$$



$$\min_{U, V} \frac{1}{2} \|R - U^T V\|_F^2$$



$$\min_{U, V} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m I_{ij} (R_{ij} - U_i^T V_j)^2$$

Tối ưu cho các mục đã
được đánh giá

Sở thích của người
dùng $U_i \in \mathbb{R}^{k \times 1}$ $R \in \mathbb{R}^{n \times m}$, $U \in \mathbb{R}^{k \times n}$, $V \in \mathbb{R}^{k \times m}$

Mục sản phẩm $V_j \in \mathbb{R}^{k \times 1}$ n: số lượng người dùng
m: số lượng sản phẩm

Mô hình hóa thông tin xã hội trong đề xuất

- Độ tương tự cosine: biểu thị sở thích của người dùng i gần với sở thích của tất cả bạn bè của anh ta $j \in F(i)$

$$\sum_{i=1}^n \sum_{j \in F(i)} sim(i, j) \|U_i - U_j\|_F^2$$

- trong đó $sim(i, j)$ biểu thị sự giống nhau giữa người dùng i và j

Công thức tổng quát:

$$\min_{U,V} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m I_{ij} (R_{ij} - U_i^T V_j)^2 + \beta \sum_{i=1}^n \sum_{j \in F(i)} sim(i, j) \|U_i - U_j\|_F^2$$

$$+ \frac{\lambda_1}{2} \|U\|_F^2 + \frac{\lambda_2}{2} \|V\|_F^2$$

**Đại lượng regularization
kiểm soát độ thưa thớt**

3. Đề xuất ràng buộc bởi ngữ cảnh xã hội

- Trong đề xuất cổ điển,
 - Để ước lượng đánh giá, chúng ta xác định những người dùng hoặc mặt hàng tương tự.
 - Bất cứ người dùng nào đồng với cá nhân đều có ảnh hưởng đến dự đoán đánh giá cho cá nhân đó

Ví dụ trong lọc cộng tác dựa trên người dùng, chúng ta xác định vùng lân cận của hầu hết các cá nhân giống nhau.

Chúng ta có thể lấy phép giao của vùng lân cận này với tập hợp bạn bè của người đó để lấy đề xuất từ những người bạn đủ giống nhau.

$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in N(u) \cap F(u)} sim(u, v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in N(u) \cap F(u)} sim(u, v)}$$

Đề xuất ràng buộc bởi ngữ cảnh xã hội

- Chúng ta có thể giới hạn nhóm các cá nhân có thể ảnh hưởng đó bằng nhóm “bạn” của người dùng.
 - $S(i)$ là nhóm k người bạn tương đồng nhất đối với cá nhân i.

$$r_{u,i} = \bar{r}_u + \frac{\sum_{v \in S(u)} sim(u, v)(r_{v,i} - \bar{r}_v)}{\sum_{v \in S(u)} sim(u, v)}$$

Ví dụ

	John	Joe	Jill	Jane	Jorge
John	0	1	0	0	1
Joe	1	0	1	0	0
Jill	0	1	0	1	1
Jane	0	0	1	0	0
Jorge	1	0	1	0	0

	Lion King	Aladdin	Mulan	Anastasia
John	4	3	2	2
Joe	5	2	1	5
Jill	2	5	?	0
Jane	1	3	4	3
Jorge	3	1	1	2

Đánh giá trung bình

$$\bar{r}_{John} = \frac{4 + 3 + 2 + 2}{4} = 2.75.$$

$$\bar{r}_{Joe} = \frac{5 + 2 + 1 + 5}{4} = 3.25.$$

$$\bar{r}_{Jill} = \frac{2 + 5 + 0}{3} = 2.33.$$

$$\bar{r}_{Jane} = \frac{1 + 3 + 4 + 3}{4} = 2.75.$$

$$\bar{r}_{Jorge} = \frac{3 + 1 + 1 + 2}{4} = 1.75.$$

$$sim(Jill, John) = \frac{2 \times 4 + 5 \times 3 + 0 \times 2}{\sqrt{29} \sqrt{29}} = 0.79$$

$$sim(Jill, Joe) = \frac{2 \times 5 + 5 \times 2 + 0 \times 5}{\sqrt{29} \sqrt{54}} = 0.50$$

$$sim(Jill, Jane) = \frac{2 \times 1 + 5 \times 3 + 0 \times 3}{\sqrt{29} \sqrt{19}} = 0.72$$

$$sim(Jill, Jorge) = \frac{2 \times 3 + 5 \times 1 + 0 \times 2}{\sqrt{29} \sqrt{14}} = 0.54$$

$$\begin{aligned}
 r_{Jill, Mulan} &= \bar{r}_{Jill} + \frac{sim(Jill, Jane)(r_{Jane, Mulan} - \bar{r}_{Jane})}{sim(Jill, Jane) + sim(Jill, Jorge)} \\
 &\quad + \frac{sim(Jill, Jorge)(r_{Jorge, Mulan} - \bar{r}_{Jorge})}{sim(Jill, Jane) + sim(Jill, Jorge)} \\
 &= 2.33 + \frac{0.72(4 - 2.75) + 0.54(1 - 1.75)}{0.72 + 0.54} = 2.72
 \end{aligned}$$

Điểm tương đồng

Đánh giá hệ thống đề xuất

Đánh giá hệ thống đề xuất có nhiều khó khăn

- Các thuật toán khác nhau có thể tốt hơn hoặc tệ hơn trên các tập dữ liệu (ứng dụng) khác nhau
 - Nhiều thuật toán được thiết kế đặc biệt cho các tập dữ liệu có nhiều người dùng hơn các mục hoặc ngược lại.
 - Sự khác biệt tồn tại đối với mật độ xếp hạng, thang xếp hạng và các thuộc tính khác của bộ dữ liệu
- Các mục tiêu để thực hiện đánh giá có thể khác nhau
 - Công việc đánh giá (evaluation) ban đầu tập trung đặc biệt vào "độ chính xác" của các thuật toán trong việc "dự đoán" xếp hạng (ratings).
 - Các thuộc tính khác với độ chính xác cũng có ảnh hưởng quan trọng đến sự hài lòng và hiệu suất của người dùng.
- Có một thách thức đáng kể trong việc quyết định nên sử dụng kết hợp các độ đo nào trong so sánh đánh giá.

Đánh giá hệ thống đề xuất

- Có vô số thuật toán được đề xuất, nhưng
 - Cái nào là tốt nhất trong một miền ứng dụng nhất định?
 - Các yếu tố thành công của các thuật toán khác nhau là gì?
 - Phân tích so sánh dựa trên một tiêu chí tối ưu?
- RA (Rank Accuracy) có hiệu quả đối với các tiêu chí cụ thể như độ chính xác, sự hài lòng của người dùng, thời gian phản hồi, v.v. không?
- Khách hàng có thích / mua các mặt hàng được giới thiệu không?
- Khách hàng có mua hay không mua các mặt hàng không?
- Họ có hài lòng với một đề xuất sau khi mua hàng không?

Làm thế nào để đánh giá các đề xuất

- Kết quả ứng dụng
 - Quảng cáo giảm giá
 - Tỷ lệ nhấp / bấm
 - Số lượng sản phẩm đã mua, sản phẩm không hoàn lại
- Các biện pháp đo đạc
 - Sự thỏa mãn của người tiêu dùng
- Metrics
 - Để dự đoán trước những điều trên

Accuracy Metrics

- **Độ dự đoán chính xác - Predictive Accuracy**
 - *Đánh giá dự đoán của hệ thống để xuất gần như thế nào so với đánh giá thực tế người dùng?*
- **Độ chính xác phân loại - Classification Accuracy**
 - Tỷ lệ mà hệ thống đề xuất đưa ra quyết định chính xác (correct) so với quyết định không chính xác (incorrect) về việc một mặt hàng có tốt hay không.
 - Do đó, các chỉ số phân loại thích hợp cho các nhiệm vụ như “Tìm mặt hàng tốt” khi người dùng có lựa chọn nhị phân (tốt/không tốt).
- **Xếp hạng độ chính xác - Rank Accuracy**

I. Độ dự đoán chính xác - Các chỉ số đo lường tỷ lệ lỗi

Sai số tuyệt đối trung bình (MAE), tính toán sự khác biệt tuyệt đối trung bình giữa xếp hạng được dự đoán và xếp hạng thực

$$MAE = \frac{\sum_{ij} |\hat{r}_{ij} - r_{ij}|}{n}$$

Chuẩn hóa MAE (NMAE), bằng cách chia cho phạm vi xếp hạng.

$$NMAE = \frac{MAE}{r_{max} - r_{min}}$$

Chi phí bình phương trung bình gốc, giống MAE nhưng nhẫn mạnh hơn về độ lệch

$$RMSE = \sqrt{\frac{1}{n} \sum_{i,j} (\hat{r}_{ij} - r_{ij})^2}$$

Ví dụ về đánh giá

<i>Item</i>	<i>Predicted Rating</i>	<i>True Rating</i>
1	1	3
2	2	5
3	3	3
4	4	2
5	4	1

$$MAE = \frac{|1 - 3| + |2 - 5| + |3 - 3| + |4 - 2| + |4 - 1|}{5} = 2$$

$$NMAE = \frac{MAE}{5 - 1} = 0.5$$

$$\begin{aligned} RMSE &= \sqrt{\frac{(1 - 3)^2 + (2 - 5)^2 + (3 - 3)^2 + (4 - 2)^2 + (4 - 1)^2}{5}} \\ &= 2.28 \end{aligned}$$

II. Độ chính xác phân loại: Precision và Recall

Độ dự đoán tích cực - Precision: thước đo độ chính xác, xác định phần nhỏ của các mục có liên quan được truy xuất trong số tất cả các mục được truy xuất

$$P = \frac{N_{rs}}{N_s}$$

Độ nhạy Recall: thước đo mức độ đầy đủ, xác định phần nhỏ của các mục có liên quan được lấy ra từ tất cả các mục có liên quan để đề xuất

$$R = \frac{N_{rs}}{N_r}$$

$$F = \frac{2PR}{P + R}.$$

	Selected	Not Selected	Total
Relevant	N_{rs}	N_m	N_r
Irrelevant	N_{is}	N_{in}	N_i
Total	N_s	N_n	N

Ví dụ đánh giá mức độ liên quan của đề xuất

	<i>Selected</i>	<i>Not Selected</i>	<i>Total</i>
<i>Relevant</i>	9	15	24
<i>Irrelevant</i>	3	13	16
<i>Total</i>	12	28	40

$$P = \frac{9}{12} = 0.75$$

$$R = \frac{9}{24} = 0.375$$

$$F = \frac{2 \times 0.75 \times 0.375}{0.75 + 0.375} = 0.5$$

III. Đánh giá xếp hạng các đề xuất

Tương quan của thứ hạng Spearman:

$$\rho = 1 - \frac{6 \sum_{i=1}^n (x_i - y_i)^2}{n^3 - n}$$

x_i : thứ hạng được dự đoán cho mục i.

y_i : thứ hạng thực của mục i theo quan điểm của người dùng.

n: tổng số mục

III. Đánh giá xếp hạng các đề xuất

Thước đo tương quan thứ hạng khác: Kendall's Tau

Cặp mục (i, j) là tương đồng (concordant) nếu thứ hạng của chúng x_i, y_i và x_j, y_j theo thứ tự:

$$x_i > x_j, \quad y_i > y_j \quad \text{or} \quad x_i < x_j, \quad y_i < y_j$$

Một cặp mục sẽ bất đồng (discordant) nếu thứ hạng tương ứng của chúng không theo thứ tự:

$$x_i > x_j, \quad y_i < y_j \quad \text{or} \quad x_i < x_j, \quad y_i > y_j$$

III. Đánh giá xếp hạng các đề xuất

Kendall's Tau tính toán sự khác biệt giữa các cặp mục ở 2 trạng thái, được chuẩn hóa bằng tổng số cặp mục $\binom{n}{2}$. Ví dụ có 4 mục trong danh sách xếp hạng, thì sẽ tạo thành $4 \times 3 / 2 = 6$ cặp mục:

$$\tau = \frac{c - d}{\binom{n}{2}}$$

c: tổng số cặp mục tương đồng (concordant)

d: tổng số cặp mục bất đồng (discordant)

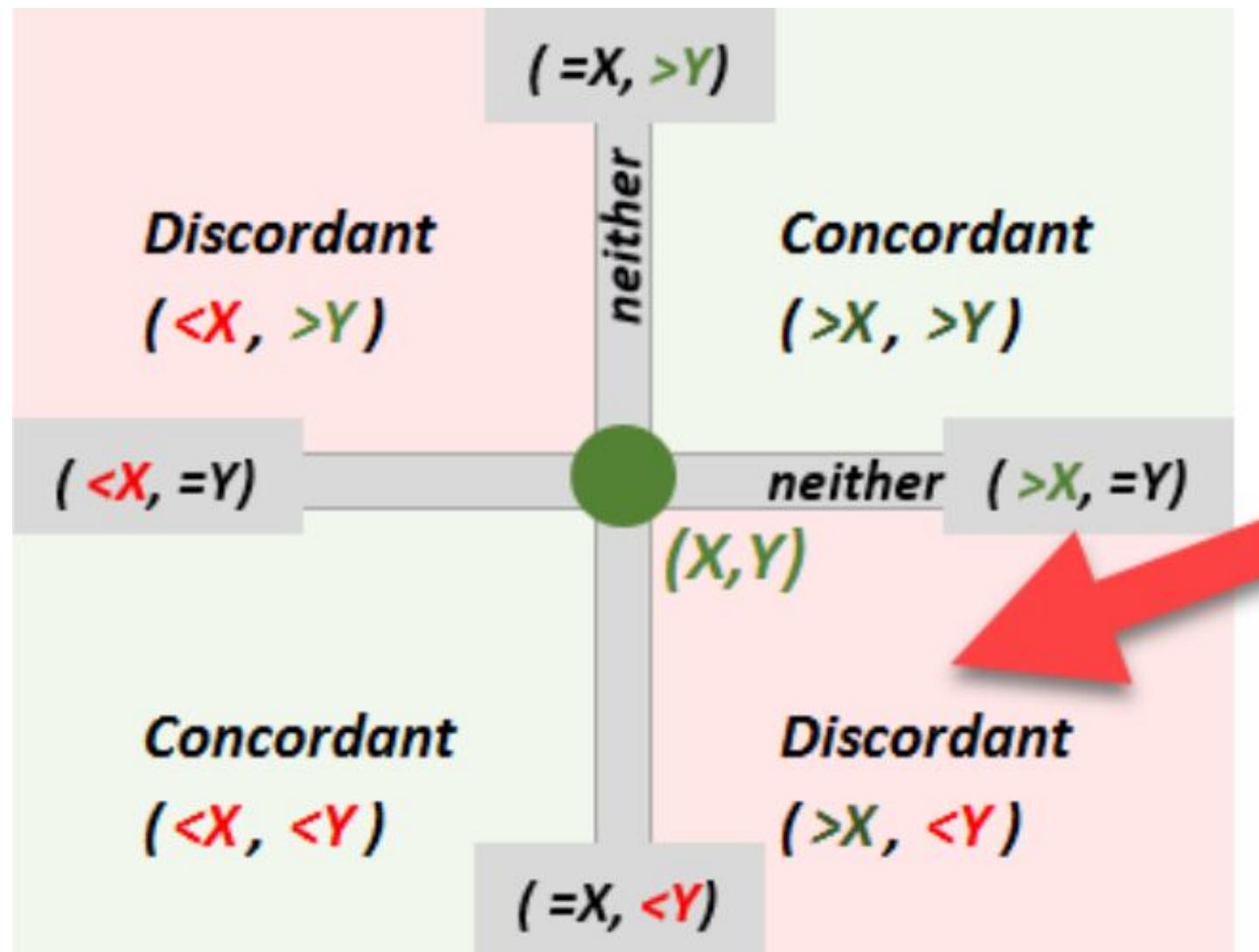
-> Kendall's Tau nhận giá trị trong khoảng [-1, 1]

Ví dụ

Ví dụ: Cho tập hợp bốn mục $I = \{i_1, i_2, i_3, i_4\}$ có thứ hạng được dự đoán và thực tế như sau:

	Predicted Rank	True Rank
i_1	1	1
i_2	2	4
i_3	3	2
i_4	4	3

Ví dụ



Ví dụ

Trạng thái tương đồng (concordant) và bất đồng (discordant):

(i_1, i_2)	: concordant
(i_1, i_3)	: concordant
(i_1, i_4)	: concordant
(i_2, i_3)	: discordant
(i_2, i_4)	: discordant
(i_3, i_4)	: concordant

Kendall's Tau nhận giá trị:

$$\tau = \frac{4 - 2}{6} = 0.33$$

Những tính chất khác

- Tính bao phủ
 - Đo lường miền của các mục trong hệ thống mà hệ thống có thể hình thành các dự đoán hoặc đưa ra các đề xuất
 - Tính mới lạ và độc đáo
 - Giúp người dùng tìm thấy một mặt hàng thú vị đáng ngạc nhiên mà họ có thể chưa phát hiện ra
 - Độ tự tin
 - Giúp RS chắc chắn rằng đề xuất của nó là chính xác đến mức nào?
 - Tính đa dạng
 - Độ rủi ro
- Tính chắc chắn
- Tính riêng tư
 - Tính thích nghi
 - Tính mở rộng

Đề xuất nội dung trên phương tiện truyền thông xã hội

Đề xuất video

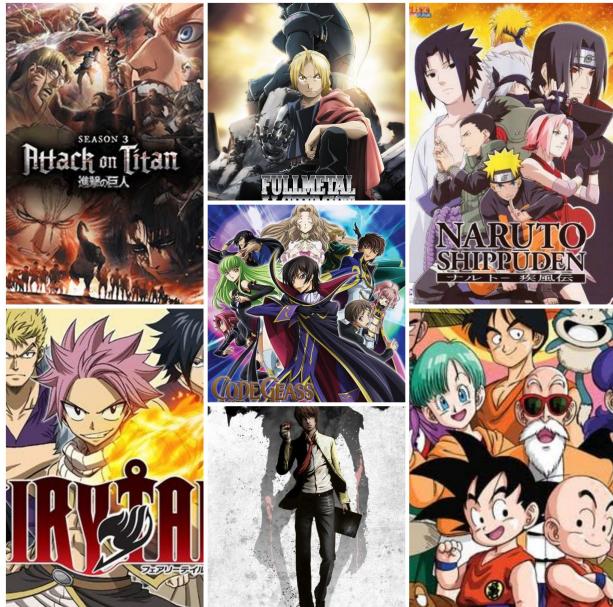
- Video có liên quan được định nghĩa là video mà người dùng có thể sẽ xem sau khi đã xem video
- Các cách tiếp cận với đề xuất video:
 - Quy tắc học kết hợp - Association rule mining
 - co-visitation

Đề xuất video: Association Rule Mining

Hệ thống tính toán xác suất xem v_j sau khi người dùng xem v_i và đề xuất top-N các video được xếp hạng cao

$$Rank(v_i, v_j) = P(v_j|v_i)$$

Association Rule Minining



1. Demon Slayer, Your Name, Doraemon

2. Pokemon, Conan, Your Name, Dragon Ball

3. Conan, Doraemon, Death Note

4. Demon Slayer, Death Note, Conan, Pokemon

Doraemon, X?

Association Rule

Support

Confidence

Đề xuất video: Co-visitation

Điểm Co-visitation là một con số hiển thị trong một khoảng thời gian nhất định, tần suất hai video cùng xem trong các phiên

$$Co-visiting(v_i, v_j) = \frac{c_{ij}}{f(v_i, v_j)}$$

c_{ij} là số lượt xem đồng thời video v_i và v_j . Với $f(v_i, v_j)$ là một yếu tố chuẩn hóa hóa liên quan đến mức độ phổ biến của hai video, ...

Đề xuất gắn thẻ

- Đề xuất gắn thẻ là quy trình đề xuất các thẻ thích hợp được người dùng áp dụng cho mỗi chủ thích mục cụ thể
- Tiếp cận:
 - Đề xuất các thẻ phổ biến nhất
 - Lọc cộng tác - CF
 - Đề xuất thẻ dựa trên nội dung
 - Dựa trên đồ thị

Đề xuất gắn thẻ: Cách tiếp cận dựa trên mức độ phổ biến và CF

Đề xuất các thẻ phổ biến nhất

- Các thẻ phổ biến đã được chỉ định cho mục target.
- Các thẻ thường xuyên được người dùng sử dụng trước đây
- Các thẻ cùng xảy ra với các thẻ đã được chỉ định.

Lọc cộng tác

- Sử dụng các phương pháp tiếp cận dựa trên mục hoặc dựa trên người dùng.
- Sử dụng phương pháp lai bằng cách đề xuất các thẻ do những người dùng giống nhau cung cấp cho các mục giống nhau.

Đề xuất gắn thẻ: Phương pháp tiếp cận dựa trên nội dung và đồ thị

Đề xuất dựa trên nội dung

- Thực hiện bằng cách đề xuất các từ khóa từ văn bản hoặc thẻ được liên kết của mục có mức độ đồng xuất hiện cao nhất với các từ khóa quan trọng.

Dựa trên đồ thị

- Thuật toán Folk Rank là một ví dụ.
- Ý tưởng: một tài nguyên được gắn thẻ bởi những người dùng quan trọng sẽ quan trọng.

Đề xuất tin tức

- Một tin tức được đề xuất nên được người dùng quan tâm nếu nó mới, đa dạng và không quá giống nhau với những tin tức khác mà người dùng đã đọc gần đây.
- Hệ thống đề xuất thông thường có thể không được sử dụng cho tin tức vì lần truy cập gần nhất chính là một trong những yếu tố quan trọng nhất đối với một mẫu tin tức.

Đề xuất blog

- Đề xuất blog là nhiệm vụ tìm kiếm các blog có liên quan để trả lời một truy vấn
- Xếp hạng mức độ liên quan của blog khác với xếp hạng và truy xuất tài liệu với các nguyên nhân như:
 - Làm thế nào để giải quyết / quản lý các bài đăng trên blog
 - Làm thế nào để đưa ra các truy vấn đáng tin cậy vì các truy vấn của người dùng đại diện cho các mối quan tâm hiện tại về chủ đề này

Thuật toán đề xuất blog

- Giải pháp cho vấn đề đầu tiên là sử dụng hai mô hình tài liệu khác nhau:
 - Mô hình tài liệu lớn: dùng toàn bộ blog như một tài liệu
 - Mô hình tài liệu nhỏ: mỗi bài đăng trên blog như một tài liệu.
- Vấn đề thứ hai có thể được giải quyết bằng cách mở rộng truy vấn
 - Query Expansion (QE) là quá trình định dạng lại một truy vấn gốc để cải thiện hiệu suất truy xuất trong các hoạt động truy xuất thông tin.
 - Sử dụng wikipedia là một phương pháp thường được sử dụng để mở rộng truy vấn: truy vấn được gửi đến Wikipedia và các bài báo wiki được truy xuất hiện là truy vấn mới và sẽ được sử dụng để tìm kiếm trên blog.

Đề xuất nội dung truyền thông xã hội: Dựa trên thẻ

- Mọi người sử dụng thẻ để tóm tắt, ghi nhớ và sắp xếp thông tin.
- Thẻ là một công cụ mạnh mẽ để điều hướng xã hội, giúp mọi người chia sẻ và khám phá thông tin mới do các thành viên khác trong cộng đồng đóng góp.
- Các thẻ thúc đẩy điều hướng xã hội bằng từ vựng của chúng hoặc tập hợp các thẻ được các thành viên của cộng đồng sử dụng.
- Thay vì áp đặt từ vựng hoặc danh mục được kiểm soát, từ vựng của hệ thống gắn thẻ xuất hiện một cách tự nhiên từ các thẻ do các thành viên riêng lẻ chọn hoặc tạo.
- Hệ thống đề xuất dựa trên thẻ sử dụng các thẻ để đề xuất các mục / mặt hàng.

Đề xuất dựa trên thẻ

- Các thuật toán kết hợp thẻ với người giới thiệu cung cấp cả tính năng tự động hóa vốn có trong người giới thiệu và tính khả thi và khả năng hiểu khái niệm vốn có trong hệ thống gắn thẻ.

Các hướng tiếp cận hỗn hợp cho việc đề xuất

Mô hình Lai tạo tuần tự

- Phương pháp này sử dụng nhiều hơn một hệ thống đề xuất. Các hệ thống sẽ nối tiếp nhau
- Kết quả của hệ thống này sẽ là đầu vào của hệ thống kế tiếp.
- Những hệ thống thế hệ đầu tiên có thể tạo ra một danh sách các đề xuất để làm đầu vào cho hệ thống kế tiếp

Mô hình Lai tạo song song

- Cũng sử dụng nhiều hơn một thuật toán để xuất, nhưng sẽ hoạt động song song
- Quy trình lai tạo hóa kết hợp các kết quả của hệ thống để xuất và tạo ra đề xuất cuối cùng
- Các phương pháp khác nhau có thể được sử dụng để kết hợp kết quả là:
 - Mixed (hợp lại kết quả của các hệ thống)
 - Weighted (tổ hợp có trọng số của các kết quả),
 - Switching (sử dụng kết quả của 1 hệ thống cho 1 công việc cụ thể),
 - Majority voting (trả lời theo số đông)

Giải thích các đề xuất

Máy ảnh X là sự lựa chọn tốt nhất cho bạn vì ...

Sự tin tưởng

- Tạo sao có người lại không tin tưởng vào đề xuất
 - Tôi có thể tin vào người bán không
 - Hệ thống này hoạt động thế nào?
 - Hệ thống này có hiểu tôi đủ không?
 - Hệ thống này có hiểu các mục nó đề xuất không?
 - Hệ thống chắc chắn bao nhiêu phần?

Thử thách đối với sự tin tưởng

- Tại sao người dùng tin tưởng các đề xuất
- Khi nào người dùng nên tin chúng
- Các hướng tiếp cận
 - Đưa ra chỉ số tin cậy
 - Giải thích các đề xuất
 - Đưa ra dữ liệu và các quy trình
 - Minh chứng các dữ liệu, hồ sơ
 - Luôn đón nhận cơ hội để sửa chữa lỗi

Giải thích các hệ thống đề xuất

- Động lực
 - "Máy ảnh *Profishot* là bắt buộc mua cho bạn vì . . ."
 - Thật sự tại sao hệ thống đề xuất cần giải thích?
 - Câu trả lời liên quan đến hai bên: bên cung cấp và bên nhận đề xuất:
 - Người bán hàng sẽ quan tâm đến việc quảng cáo sản phẩm cụ thể
 - Người mua sẽ quan tâm đến việc đưa ra quyết định mua hàng đúng đắn
- Thông tin bổ sung cần để giải thích hệ thống sẽ phải theo một số mục tiêu nhất định

Mục tiêu của sự giải thích

- **Sự minh bạch**
 - Cung cấp thông tin để người dùng có thể hiểu được lý do được sử dụng để tạo ra một đề xuất cụ thể
- **Tính hợp lệ**
 - Giải thích có thể được tạo để cho phép người dùng kiểm tra tính hợp lệ của đề xuất
- **Độ tin cậy**
 - Các giải thích nhằm mục đích xây dựng lòng tin cho các đề xuất, làm giảm sự không chắc chắn về chất lượng của đề xuất
- **Tính dễ hiểu**
 - Giải thích về sự hiểu biết nhằm mục tiêu hỗ trợ người dùng bằng cách liên hệ các khái niệm đã biết của họ với các khái niệm được hệ thống đề xuất sử dụng
- **Giáo dục**
 - Kiến thức sâu sắc về chuyên ngành giúp khách hàng suy nghĩ lại về sở thích của họ và đánh giá ưu và nhược điểm của các lựa chọn khác nhau

Mục tiêu của sự giải thích

- Tính thuyết phục
 - Các giải thích thuyết phục cho các đề xuất nhằm mục đích thay đổi hành vi mua hàng của người dùng
- Tính hiệu quả
 - Người dùng nhận được sự hỗ trợ mà để đưa ra các quyết định chất lượng cao
- Tính năng suất
 - Hỗ trợ người dùng giảm thời gian ra quyết định
- Sự hài lòng
 - Các giải thích có thể cố gắng cải thiện sự hài lòng tổng thể.
- Sự liên quan
 - Thông tin bổ sung có thể cần cho hệ thống đề xuất đối thoại. Sự giải thích có thể được cung cấp để giải thích lý do tại sao hệ thống cần thông tin bổ sung từ người dùng

Các loại giải thích

- **Sự giải thích cho lảng giềng gần nhất**
 - Khách hàng khi mua sản phẩm X, thường cũng mua luôn sản phẩm Y, Z
 - Do đó sản phẩm Y được đề xuất cho bạn vì bạn đã đánh giá sản phẩm X
- **Giải thích dựa trên nội dung**
 - Mẩu truyện này thuộc về chủ đề X, Y cũng là chủ đề mà bạn thích
- **Giải thích dựa trên mạng xã hội**
 - Mọi người sử dụng mạng xã hội để nhò vào các mối quan hệ tin cậy để lọc thông tin.
 - Người bạn X của bạn đã viết bài blog đó
 - 50% người bạn của bạn đã Thích món hàng này (trong khi chỉ có 5 không thích)

Ví dụ

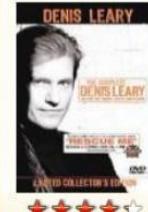
- **Sự tương đồng giữa các mục**
- **Sự tương đồng giữa người dùng**
- **Tags**
 - Các tag liên quan (cho các mục)
 - Các tag sơ thích (của người dùng)

Because you enjoyed:

Suicide Kings
Clerks

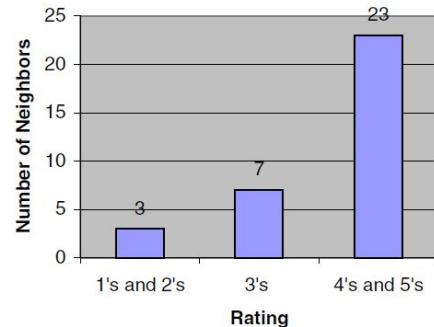
We think you'll enjoy:
[Denis Leary: The Complete Denis Leary](#)

Add



★★★☆☆ Not Interested

Your Neighbors' Ratings for this Movie



Your prediction is based on how MovieLens thinks you like these aspects of the film:

Relevance ↓		Your preference
██████	wes anderson	★★★★★
██████	deadpan	★★★★★
██████	quirky	★★★★★
██████	witty	★★★★★
██████	off-beat comedy	★★★★★
██████	notable soundtrack	★★★★★
██████	stylized	★★★★★

Các thuật toán đề xuất khác

- Hệ thống đề xuất đa tiêu chí
- Hệ thống đề xuất nhận thức rủi ro
- Hệ thống đề xuất hỗn hợp
- Hệ thống đề xuất trên di động
- Hệ thống đề xuất theo phiên
- Hệ thống đề xuất sử dụng học tăng cường

Các độ đo để đánh giá khác

- Độ đo Rank Aware top-N: MRR, NDCG
- Độ đo về tính đa dạng - Diversity: Intra-list Similarity, Lathia's Diversity
- Độ đo về mức độ phản hồi ngầm - Implicit Feedback: Mean Percentage Ranking, User-Centric Evaluation Frameworks

Cảm ơn thầy và các bạn đã lắng nghe!

