



Bài giảng môn học:

Học Máy (Machine Learning)

CHƯƠNG 5: HỆ THỐNG GỢI Ý (Recommender System)

Giảng viên: Đặng Văn Nam

Email: dangvannam@humg.edu.vn

Nội dung chương 5

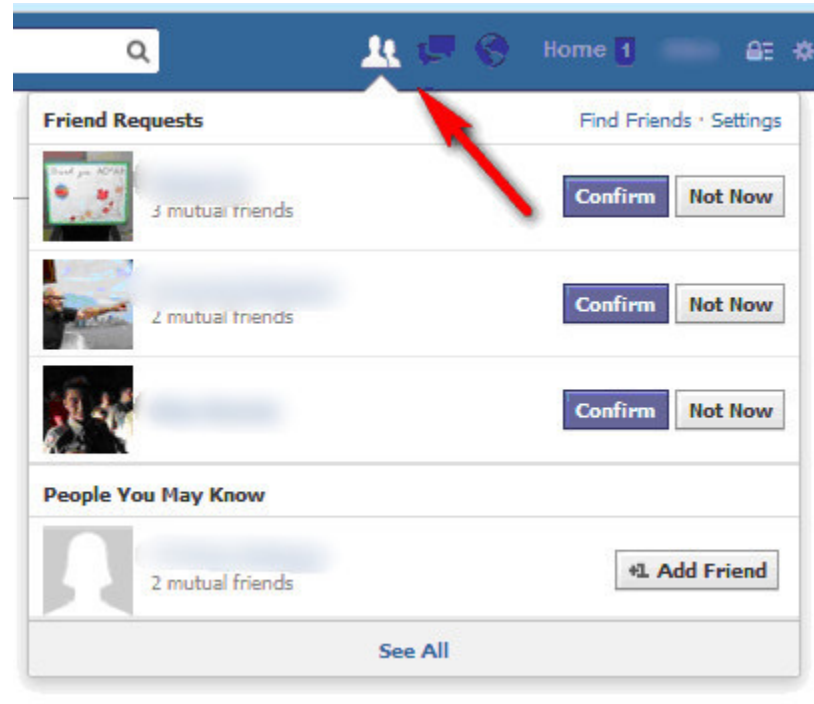
1. Giới thiệu chung
2. Phân loại hệ thống đề xuất
3. Các phương pháp tính toán độ tương đồng
4. Sơ đồ tổng quan và Thách thức
5. Ví dụ minh họa



1. Giới thiệu

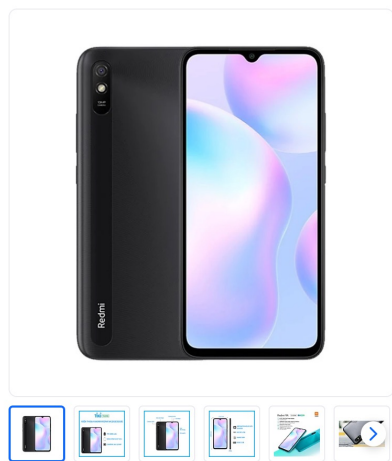
Giới thiệu

- Chúng ta sử dụng Facebook và mới chỉ kết bạn với một vài người. Tuy nhiên vài hôm sau, **Facebook đã tự gợi ý** cho chúng ta những người bạn khác nhau mà thậm chí ngay cả chúng ta cũng không biết họ???



Giới thiệu

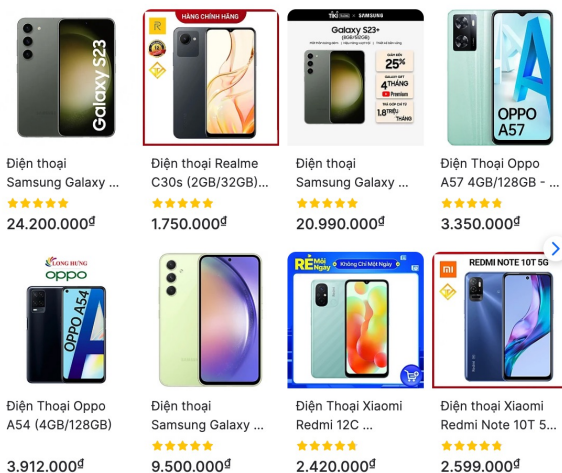
- Bạn đang dạo chơi trên một trang thương mại điện tử với mục đích ban đầu là tìm một chiếc váy.
- Khi click xem một chiếc váy thì hệ thống hiển thị một loạt những chiếc váy khác để gợi ý cho bạn???



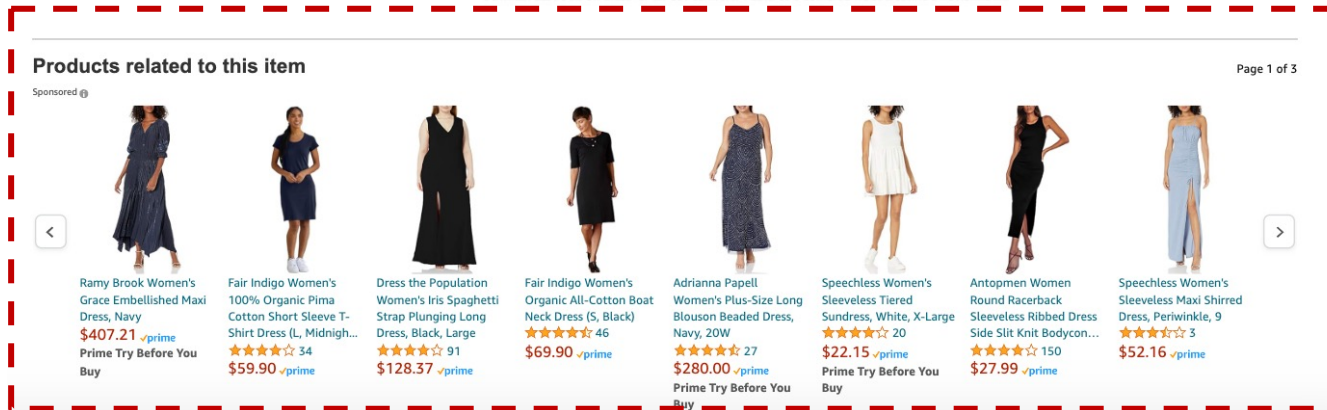
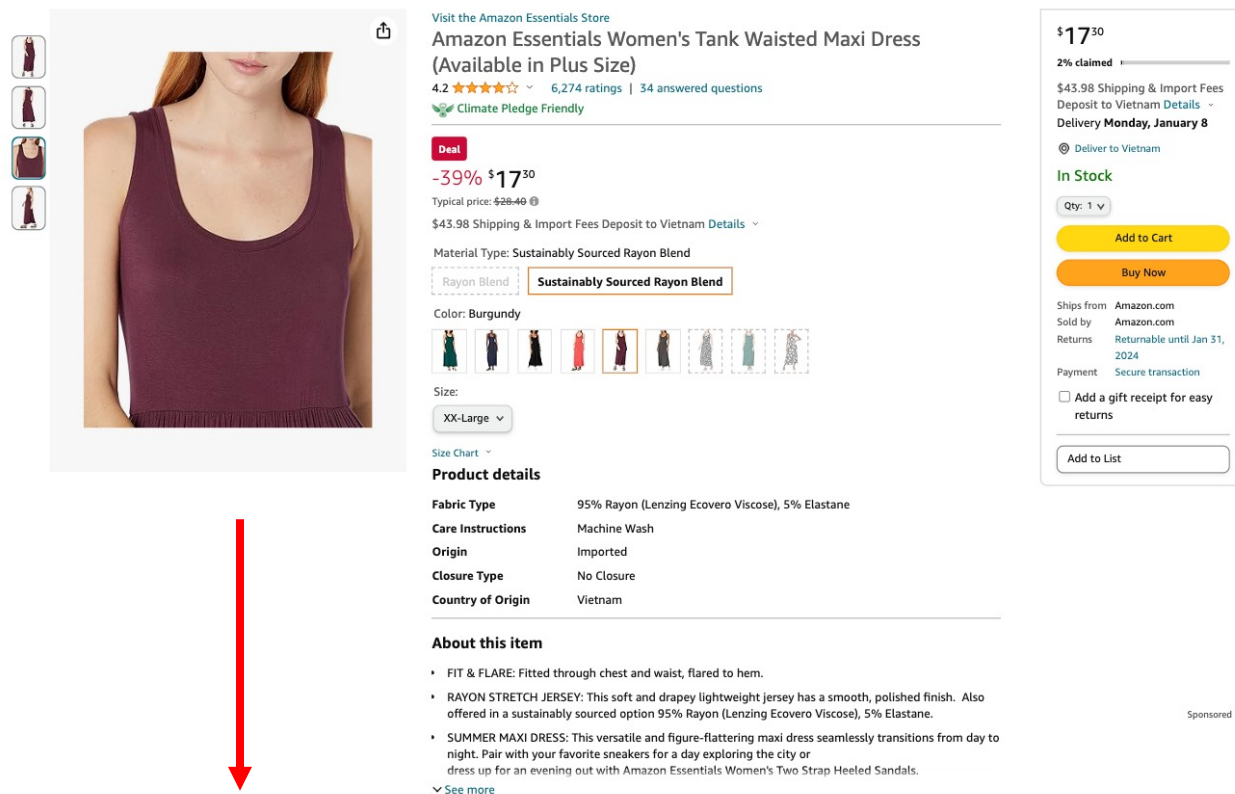
Đặc điểm nổi bật

- Màn hình tràn viền 6.53 inch, giọt nước, tạo không

Sản phẩm tương tự

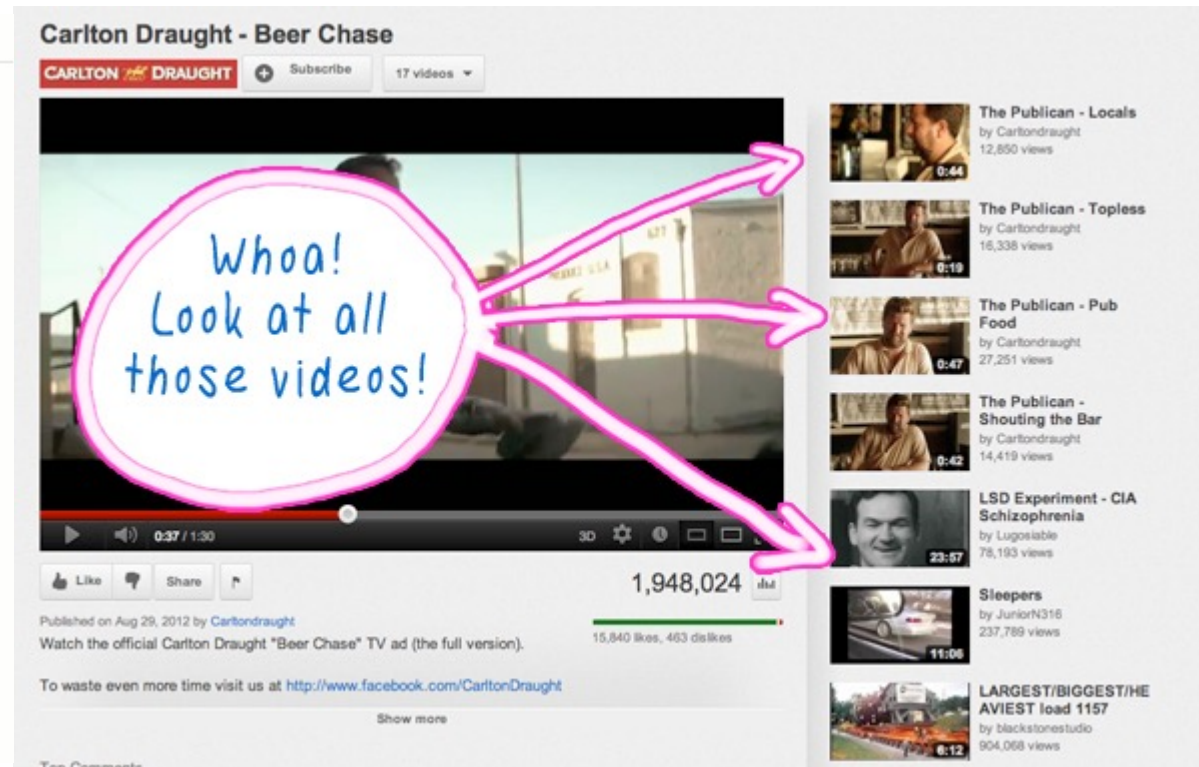
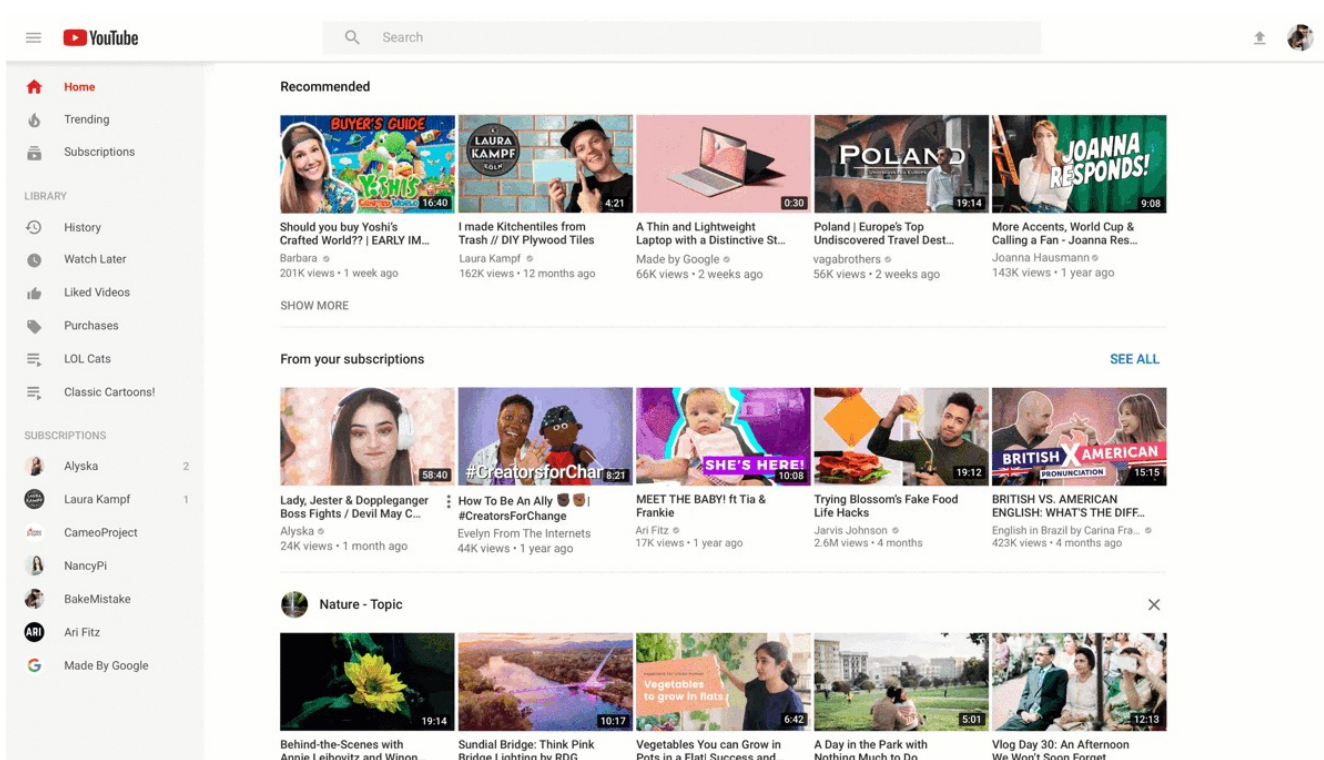


Back to results



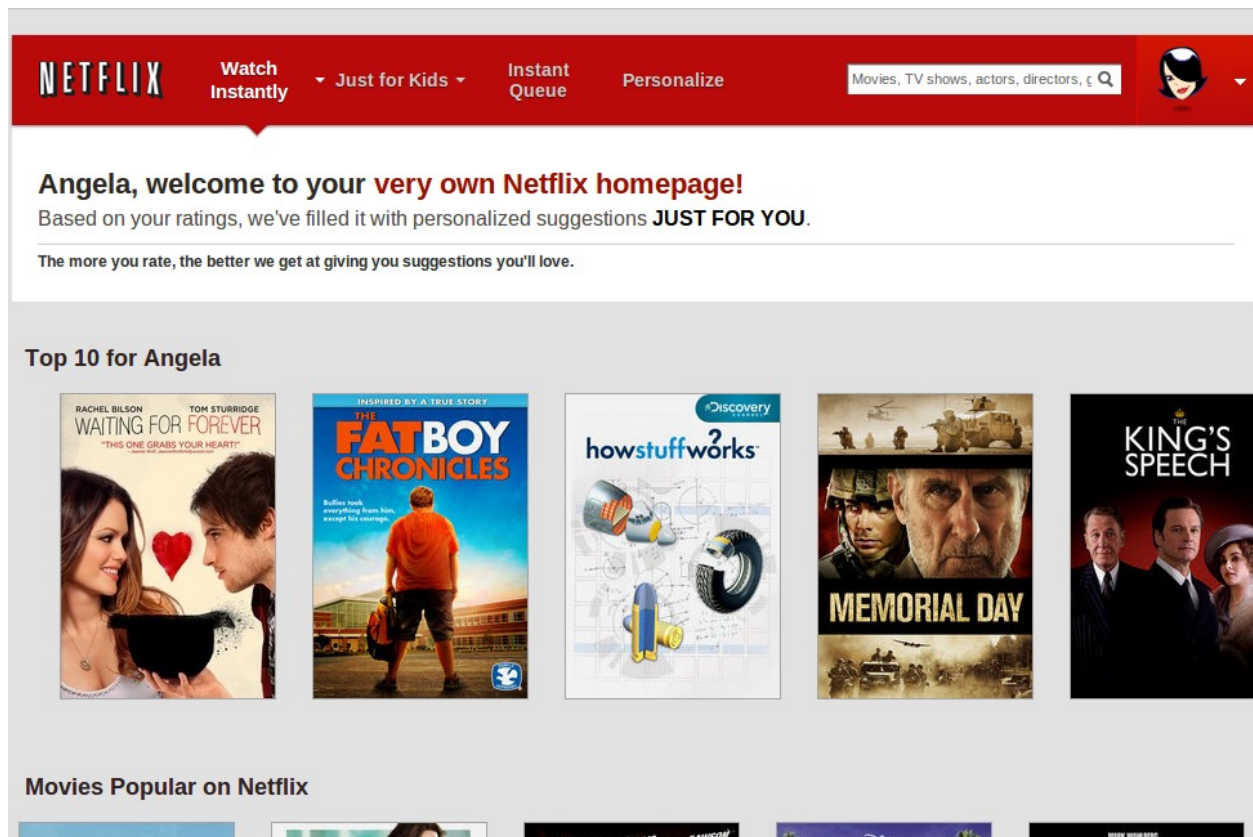
Giới thiệu

- Chúng ta mở **Youtube** và tìm kiếm một bản nhạc vàng, nằm xuống giường và thưởng thức âm nhạc, những bài hát tiếp theo sẽ được Youtube thân yêu tự động gợi ý và tất nhiên rất hiếm khi nó gợi ý một bản nhạc rock cho bạn trong khi bạn đang nghe nhạc vàng phải không nào ???



Giới thiệu

- Cuối tuần, bạn và người yêu lên NETFLIX để xem một bộ phim tình cảm sướt mướt của Hàn Quốc, xem xong bạn like, vote 5 sao và rồi một loạt phim có nội dung kiểu ấy, những bộ phim có diễn viên nữ chính, nam chính bạn đã xem ấy xuất hiện gợi ý để bạn xem tiếp.....



Giới thiệu



Hệ thống gợi ý (Recommender Systems)

- Hệ thống gợi ý (Recommender Systems) là một thành phần trong hệ thống thông tin. Mục đích của nó là **hỗ trợ người dùng tìm kiếm** được đúng thông tin cần thiết, dự đoán sở thích hay xếp hạng mà người dùng có thể dành cho một mục thông tin (item) nào đó mà họ chưa xem xét tới trong quá khứ.
- Các gợi ý được đưa ra là kết quả của việc tính toán dựa trên việc thu thập dữ liệu về người dùng như thói quen, sở thích, hành vi khi mua hàng, khi đưa ra các đánh giá cá nhân...
- Việc thực hiện tính toán được xây dựng trên các thuật toán Học máy, đưa ra các dự đoán tốt nhất về sản phẩm mà người dùng có thể thích, giúp gia tăng số lượng sản phẩm bán được.



Hệ thống gợi ý (Recommender Systems)



- Hệ thống gợi ý đang là một công cụ mạnh mẽ, được ứng dụng chủ yếu cho các mạng xã hội, giải trí trực tuyến, thương mại điện tử....

Lợi ích mang lại là gì?



Hệ thống gợi ý (Recommender Systems)

- Tăng doanh thu, tăng CRT (Click-through rate)...
- Tăng mức tín nhiệm và trung thành của khách hàng.
- Thêm hiểu biết về khách hàng.
- Khả năng đưa ra các dịch vụ các nhân hóa, hướng tới từng đối tượng khách hàng cụ thể.



Why Netflix thinks its personalized recommendation engine is worth \$1 billion per year



Nathan McAlone  
🕒 14.06.2016, 21:36 👁 228



FACEBOOK



LINKEDIN



TWITTER



EMAIL



PRINT

After a long refinement process, Netflix finally released its first “global” recommendation engine in December.

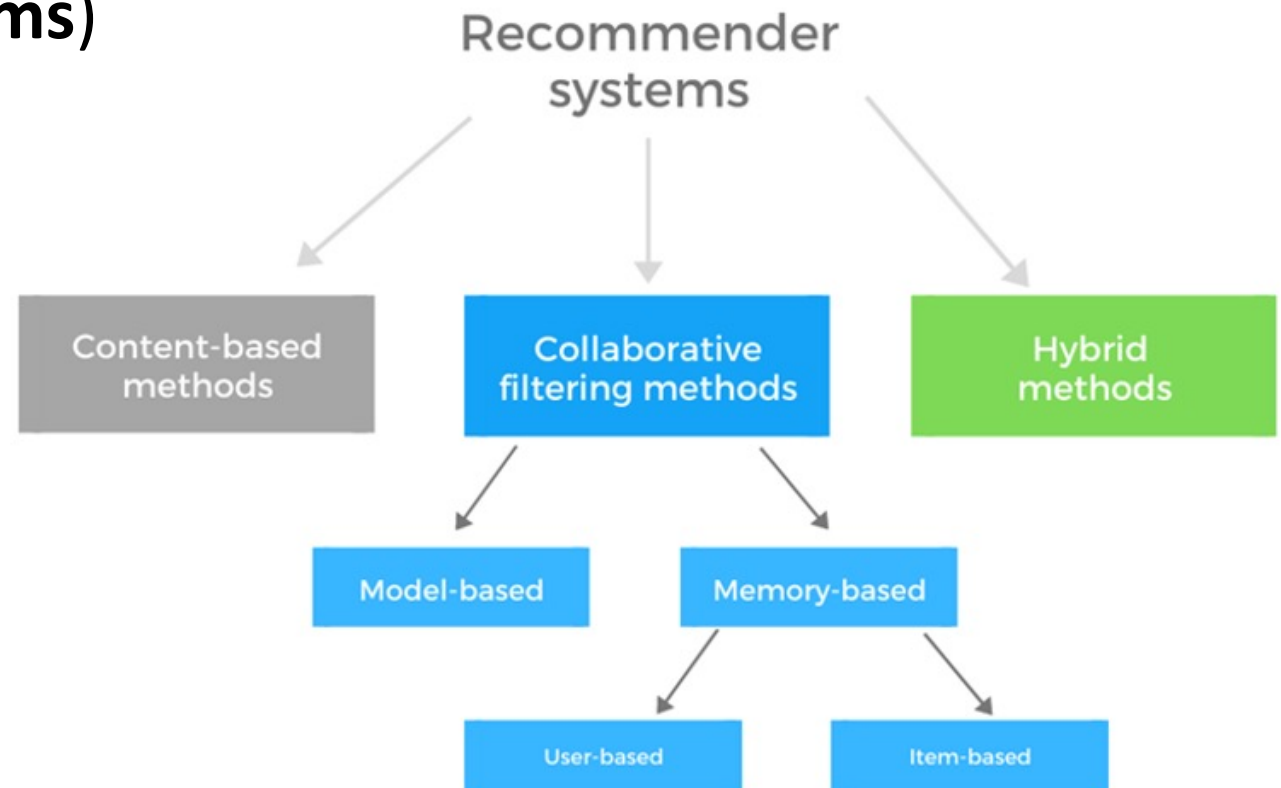
The engine takes dozens of algorithms into account and compares you with similar users in the more than 190 countries where Netflix's service is available.



2. Phân loại hệ thống gợi ý

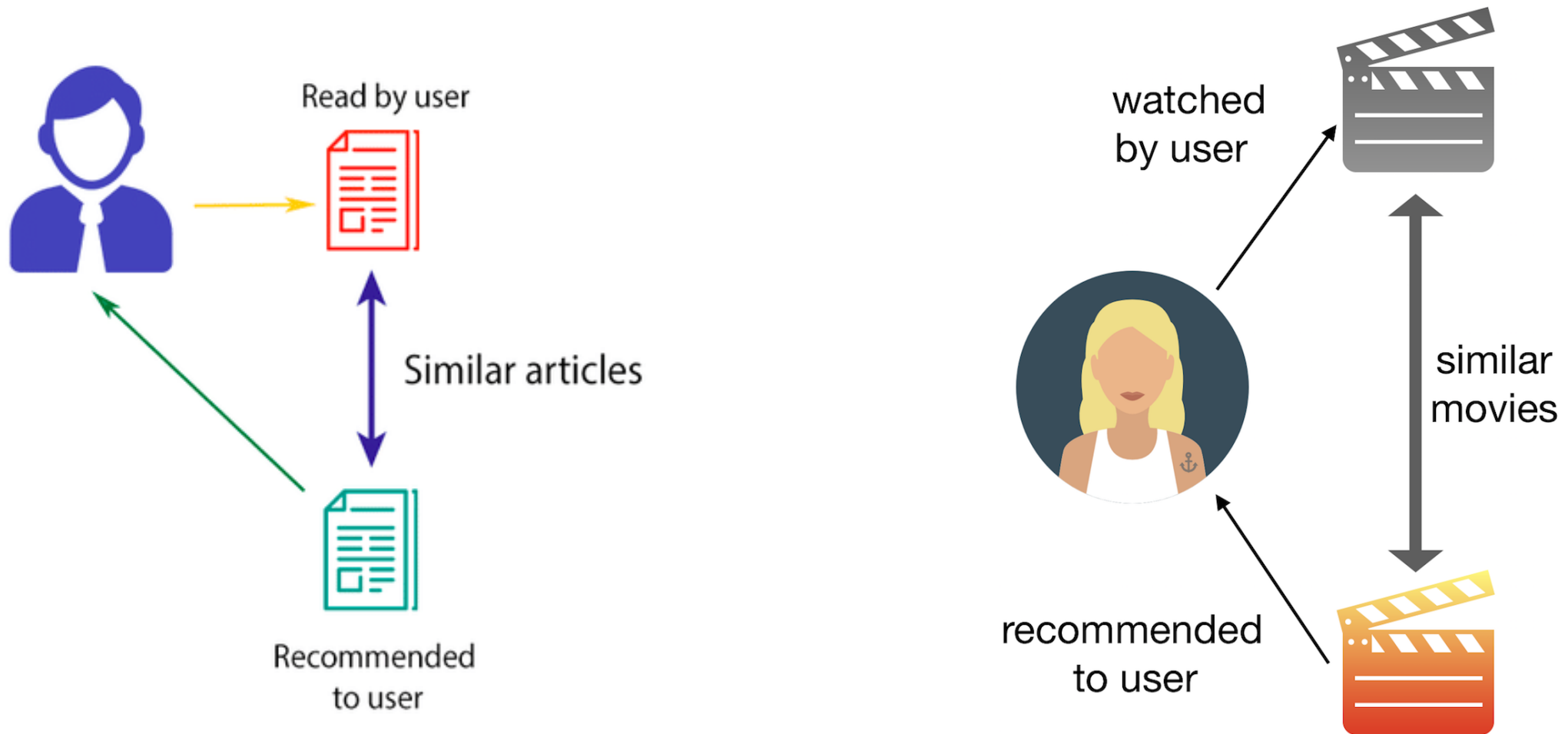
Phân loại các hệ thống gợi ý

1. Hệ thống gợi ý dựa trên nội dung (**Content based recommender systems**)
2. Hệ thống gợi ý dựa trên các user – Lọc cộng tác (**Collaborative filtering recommender systems**)
3. Hệ thống gợi ý lai (**Hybrid systems**)



Hệ thống gợi ý dựa trên nội dung

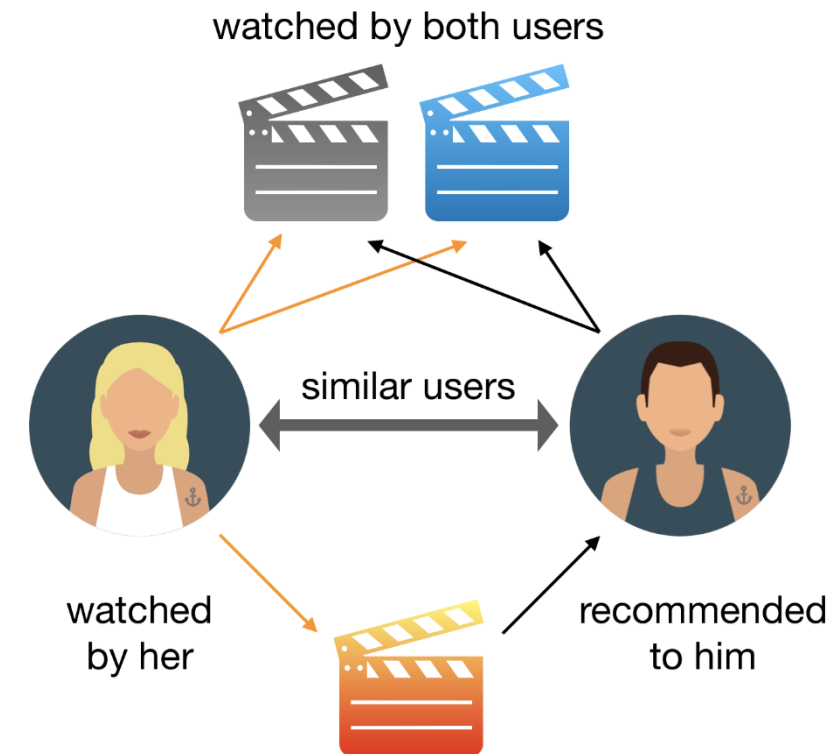
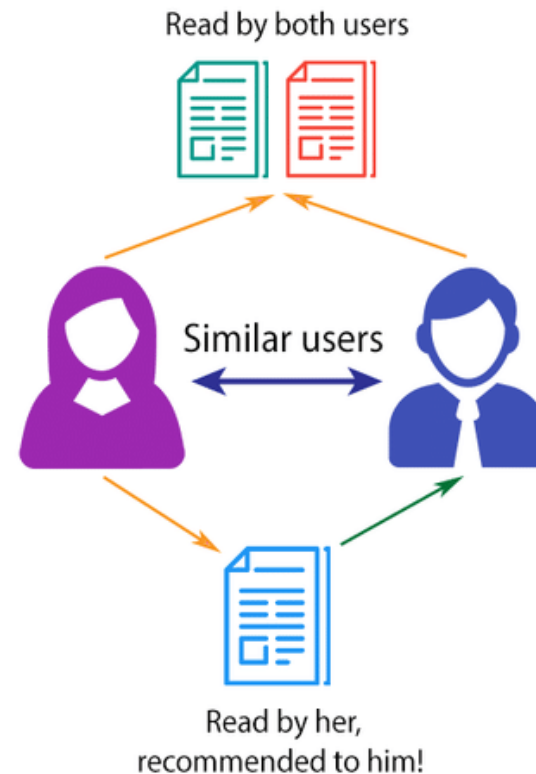
Hệ thống dựa trên nội dung, đặc điểm của mục tin, sản phẩm, video...hiện tại và sau đó gợi ý cho người dùng các mục tin, sản phẩm, video... tương tự.



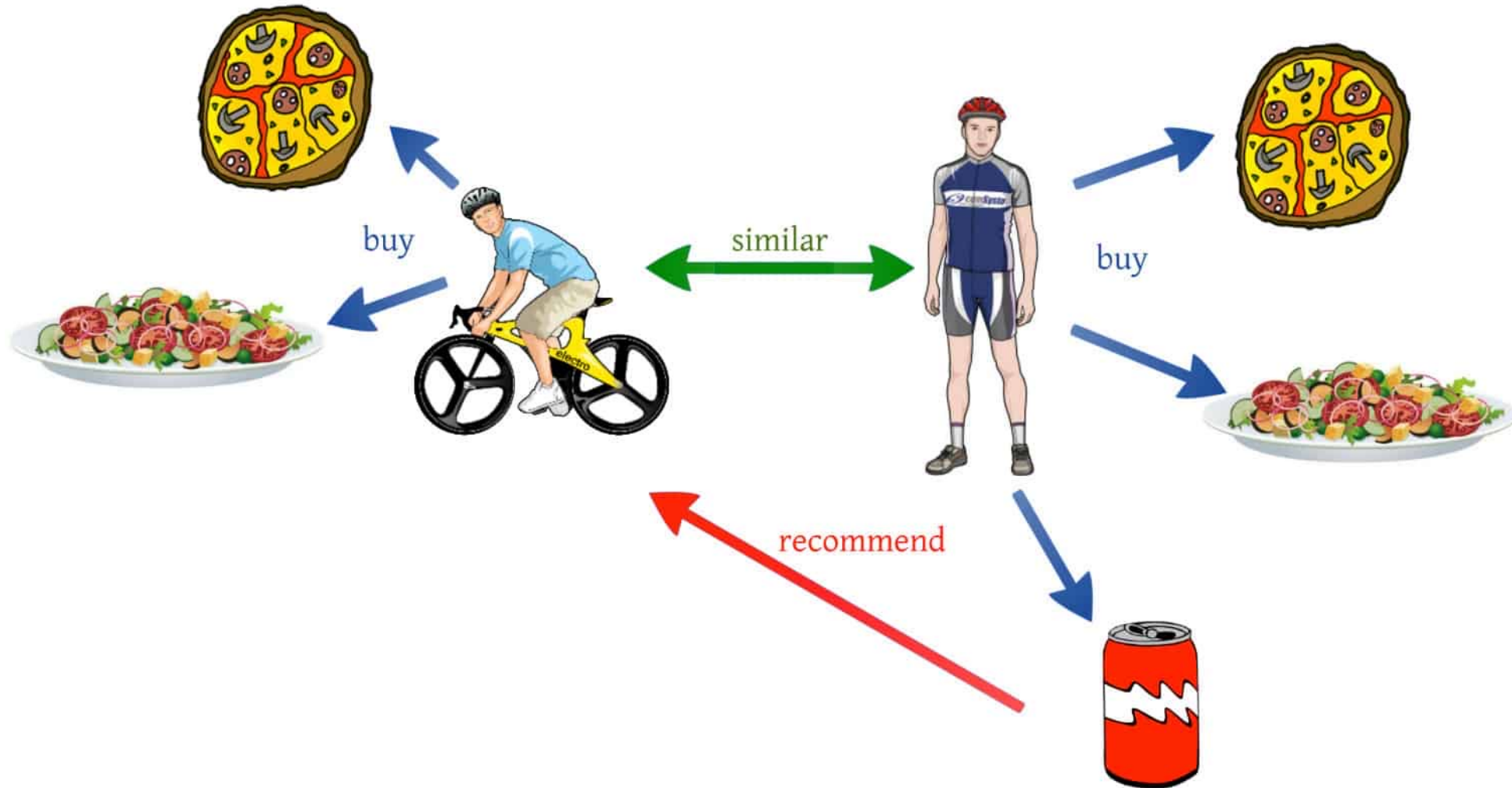
Hệ thống gợi ý dựa trên user – lọc cộng tác

Hệ thống sẽ phân tích các user có cùng đánh giá, cùng mua **mục tin** hiện tại. Sau đó tìm ra danh sách các **mục tin** khác cũng được đánh giá bởi các user này, xếp hạng và gợi ý cho người dùng. Tư tưởng của phương pháp này chính là dựa trên **sự tương đồng về sở thích giữa các người dùng** để đưa ra các gợi ý.

Về bản chất, Nó lọc trên những người có cùng sở thích, hay những người có cùng những hành vi tương tự, cùng bấm like, cho điểm đối với cùng một item.



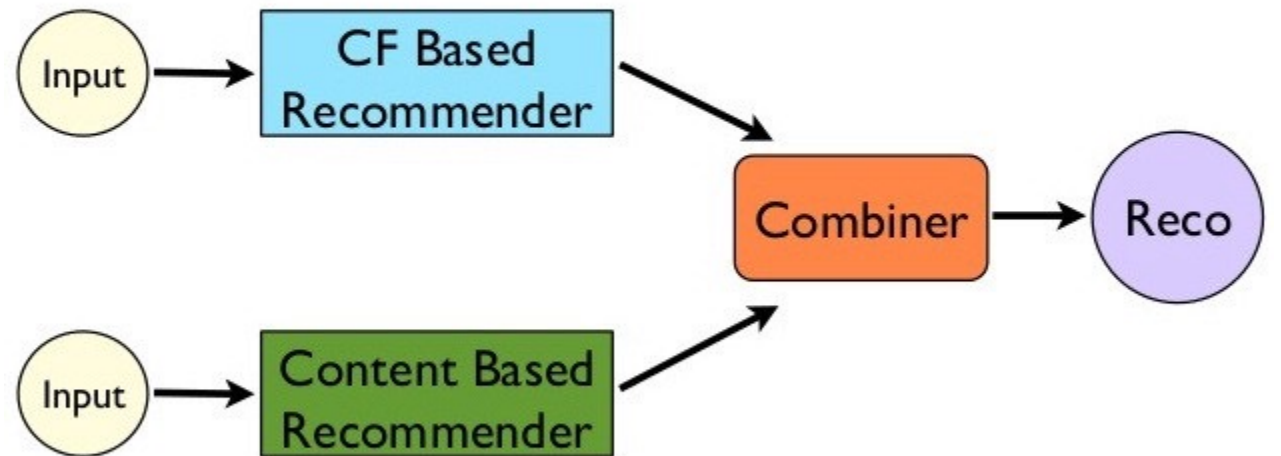
Hệ thống gợi ý dựa trên user – lọc cộng tác



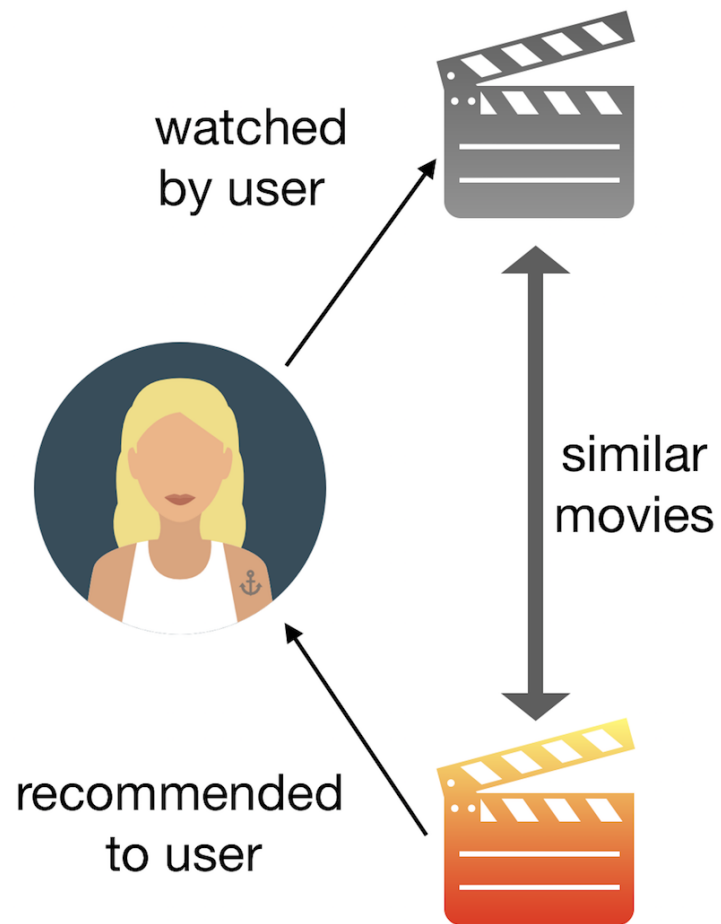
Hệ thống gợi ý lai

Kết hợp các ý tưởng của Content-based recommender và Collaborative filtering để xây dựng một hệ thống gợi ý.

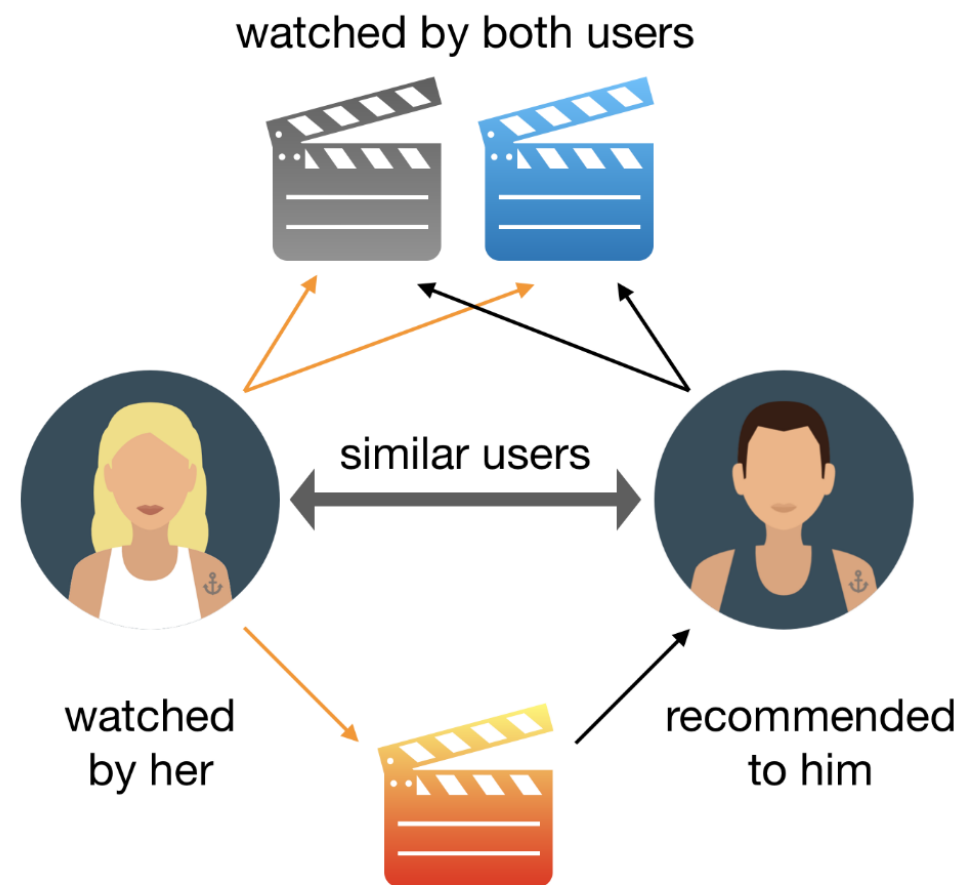
Hybrid Recommendations



So sánh



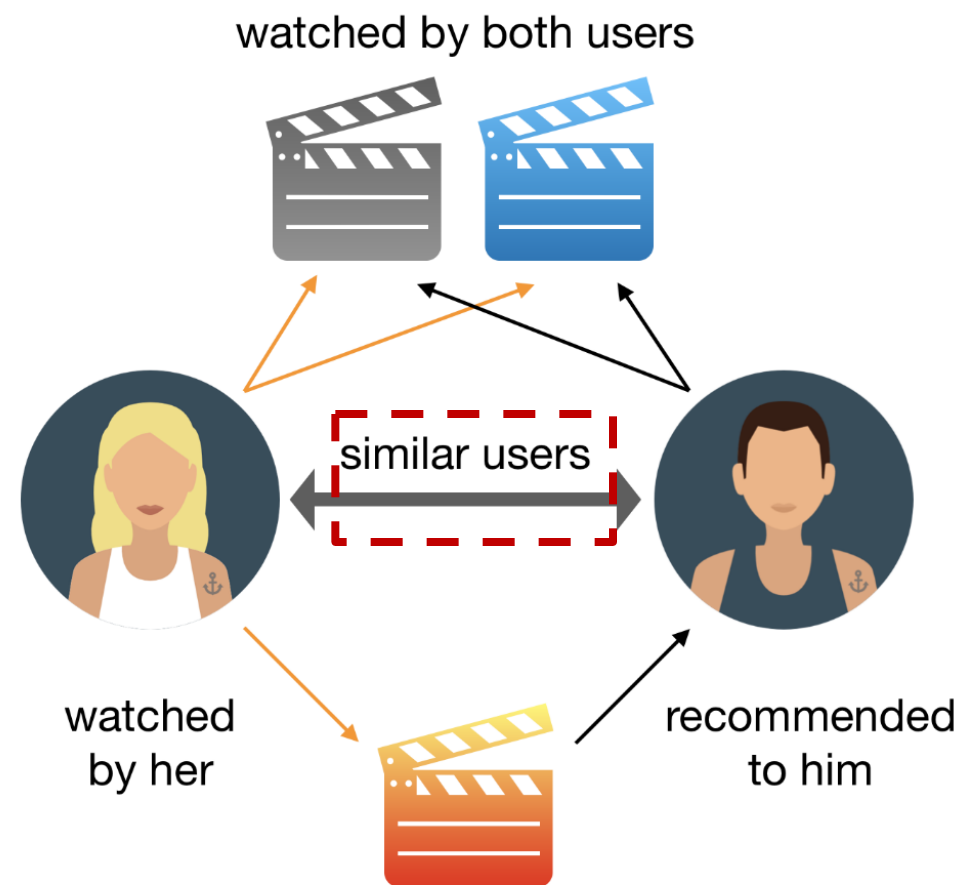
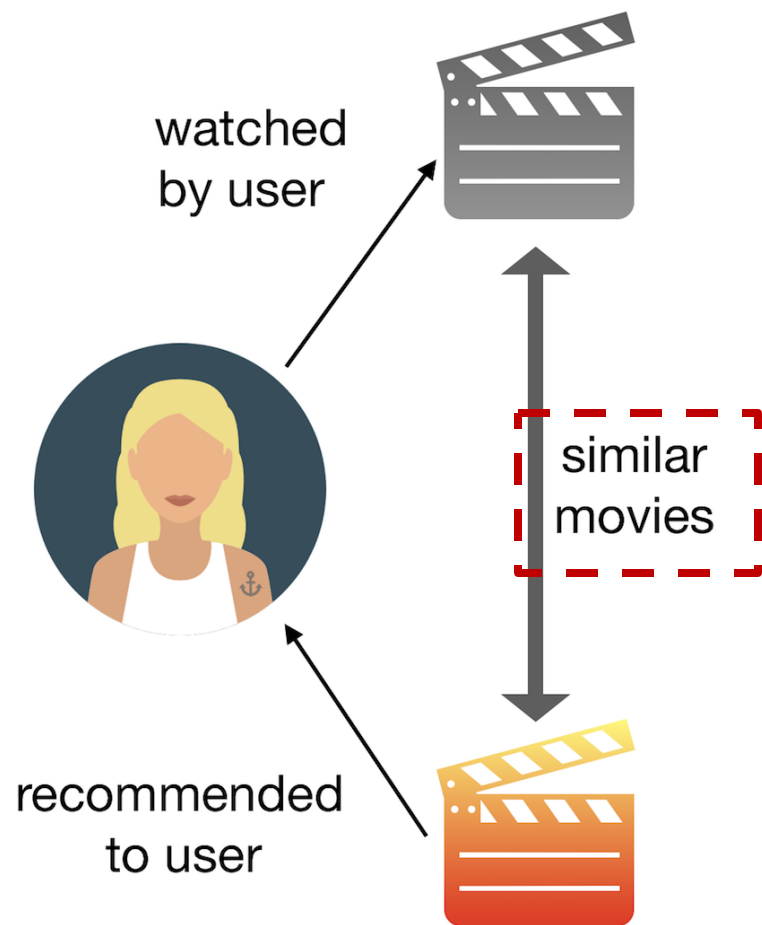
Hệ thống gợi ý dựa trên nội dung
(Content based recommender systems)



Hệ thống gợi ý dựa trên các user – Lọc cộng tác
(Collaborative filtering recommender systems)

3. Các phương pháp đánh giá độ tương đồng

Các phương pháp tính toán độ tương tự



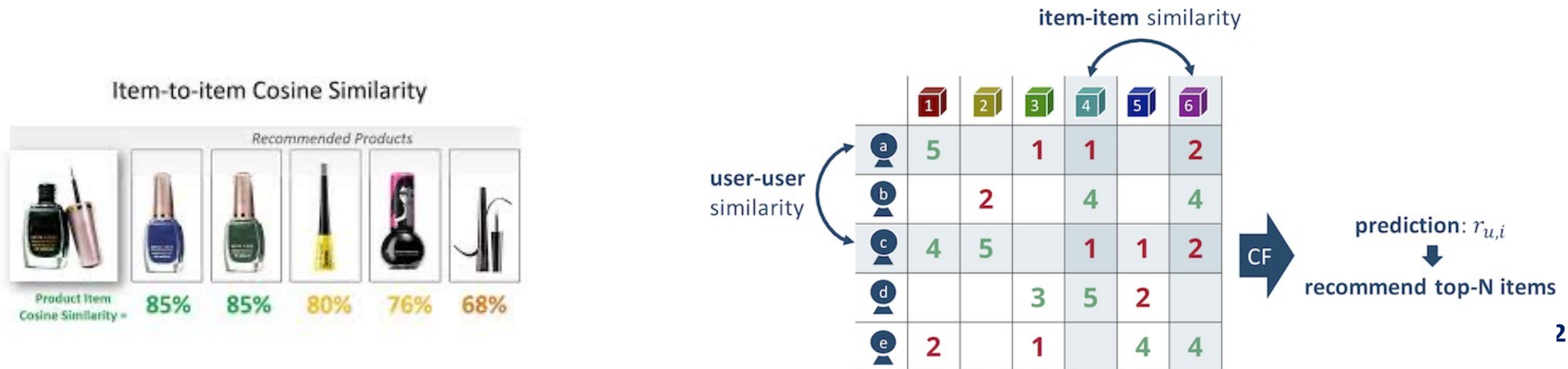
Rõ ràng chúng ta không thể xem xét **độ tương tự (Similar)** giữa các đối tượng bằng cảm tính được

Các phương pháp tính toán độ tương tự

Chúng ta cần một cơ sở toán học cụ thể để xác định **độ tương tự** và đại lượng này được gọi là **khoảng cách**:

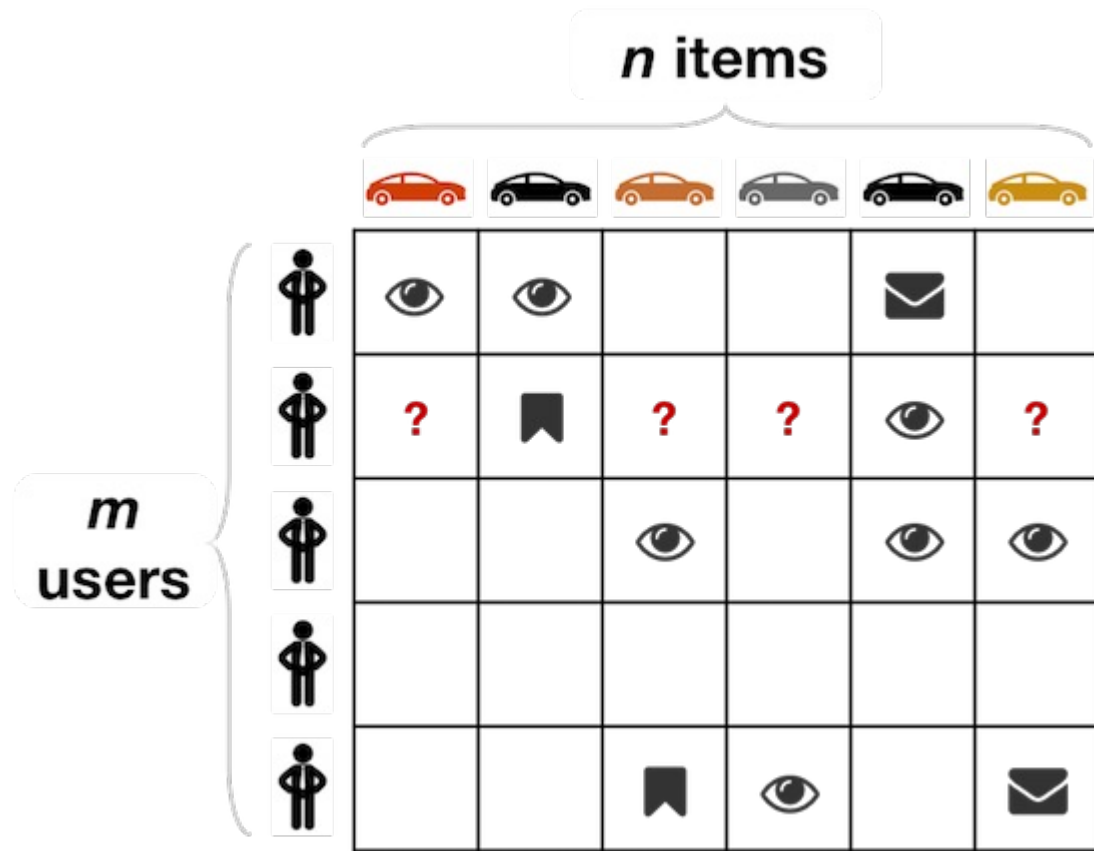
- Khoảng cách **càng nhỏ** => càng gần nhau => độ tương tự **càng lớn**
- Khoảng cách **càng lớn** => Càng xa nhau => độ tương tự **càng nhỏ**

Chúng ta có thể hiểu **độ đo tương tự** giống như **ngịch đảo của khoảng cách**. Sử dụng khoảng cách để tính toán đại lượng này.

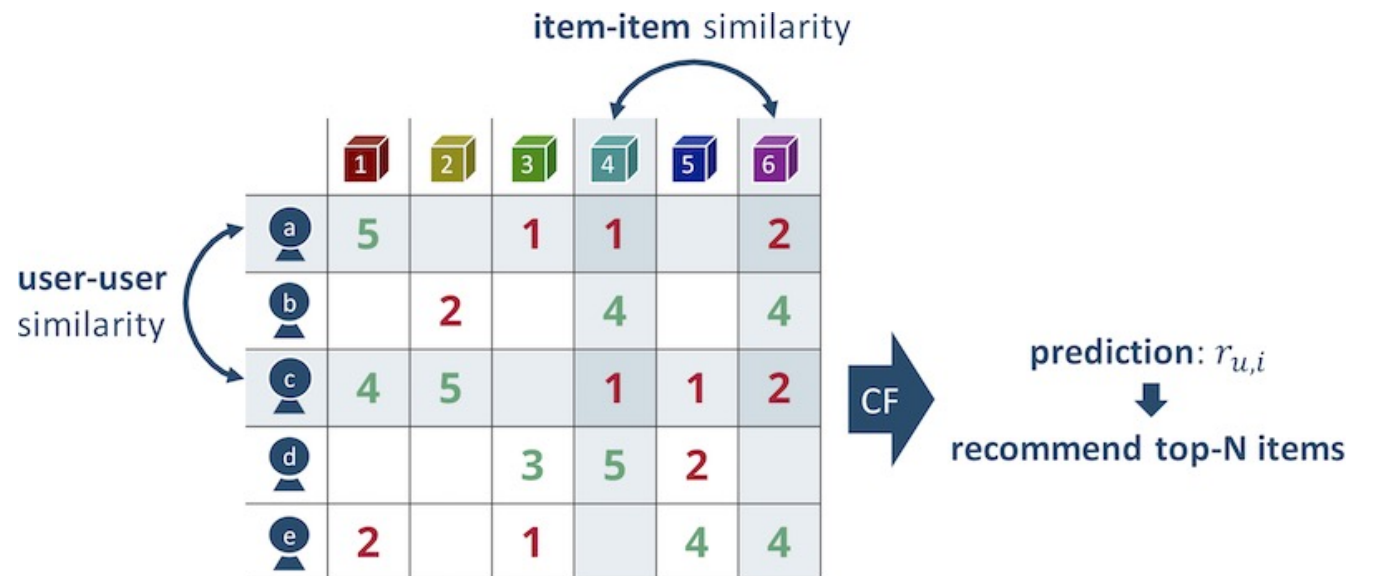


Các phương pháp tính toán độ tương tự

Khoảng cách được tính toán dựa trên ma trận *users - items*



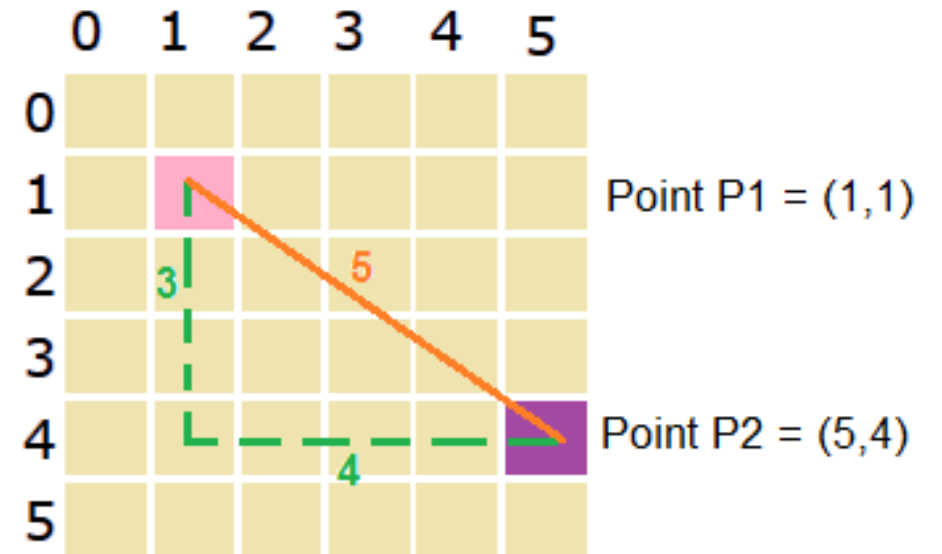
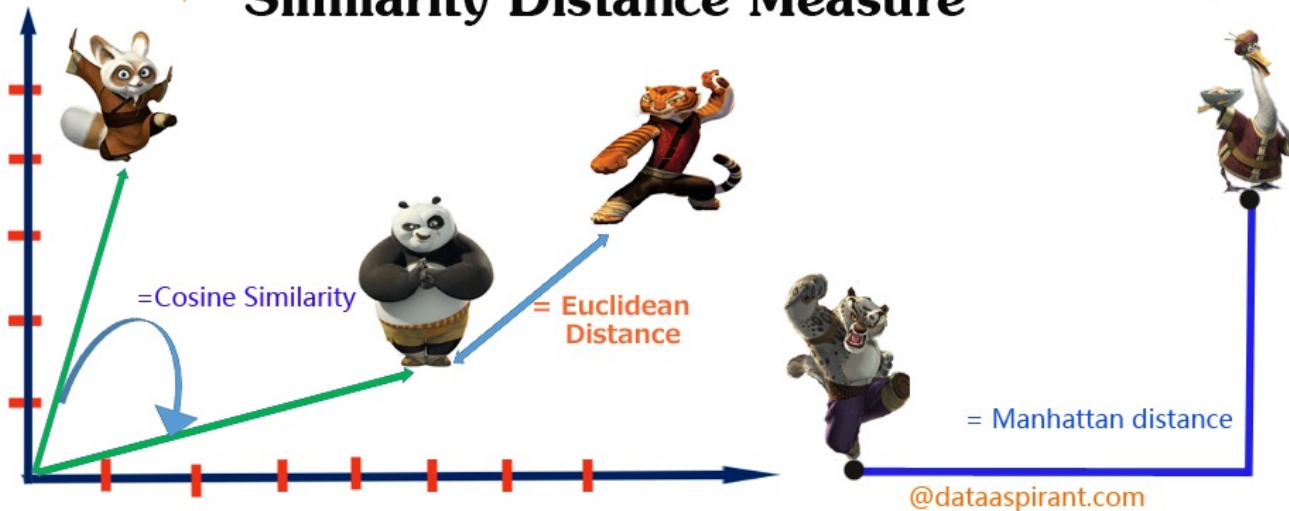
A		✓	✗	✓	✓
B			✓	✗	✗
C		✓	✓	✗	
D		✗		✓	
E		✓	✓	?	✗



Các phương pháp tính toán độ tương tự

Một số phương pháp tính khoảng cách:

Similarity Distance Measure

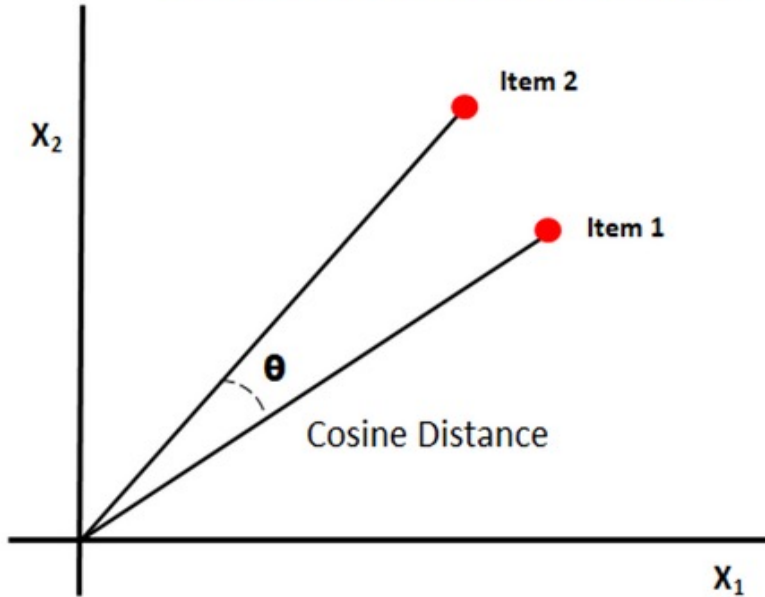


$$\text{Euclidean distance} = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$\text{Manhattan distance} = |5-1| + |4-1| = 7$$

Các phương pháp tính toán độ tương tự

Cosine Distance/Similarity



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

• Ví dụ:

• Điểm A(1, 1, 1, 1, 0, 1, 1, 2)

• Điểm B(1, 0, 0, 1, 1, 0, 1, 0)

$$\begin{aligned} \mathbf{A} \cdot \mathbf{B} &= \sum_{i=1}^n A_i B_i \\ &= (1 * 1) + (1 * 0) + (1 * 0) + (1 * 1) + (0 * 1) + (1 * 0) + (1 * 1) + (2 * 0) \\ &= 3 \end{aligned}$$

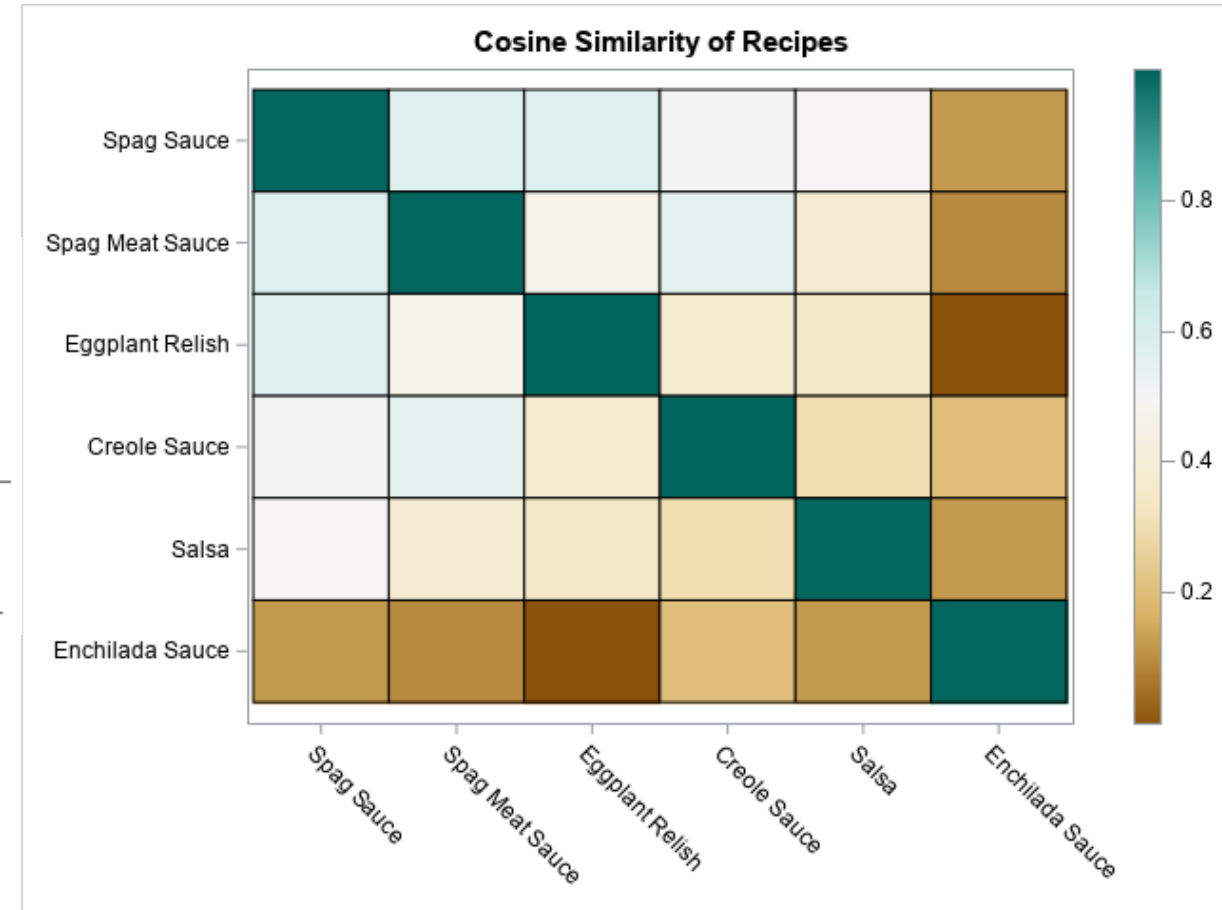
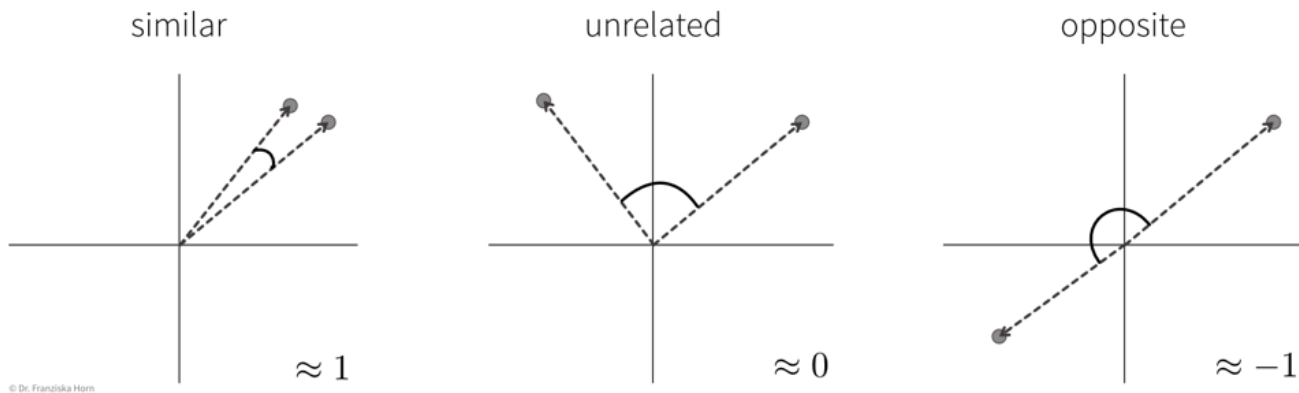
$$\sqrt{\sum_{i=1}^n A_i^2} = \sqrt{1 + 1 + 1 + 1 + 0 + 1 + 1 + 4} = \sqrt{10}$$

$$\sqrt{\sum_{i=1}^n B_i^2} = \sqrt{1 + 0 + 0 + 1 + 1 + 0 + 1 + 0} = \sqrt{4}$$

$$\text{cosine similarity} = \cos\theta = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{3}{\sqrt{10} * \sqrt{4}} = 0.4743$$

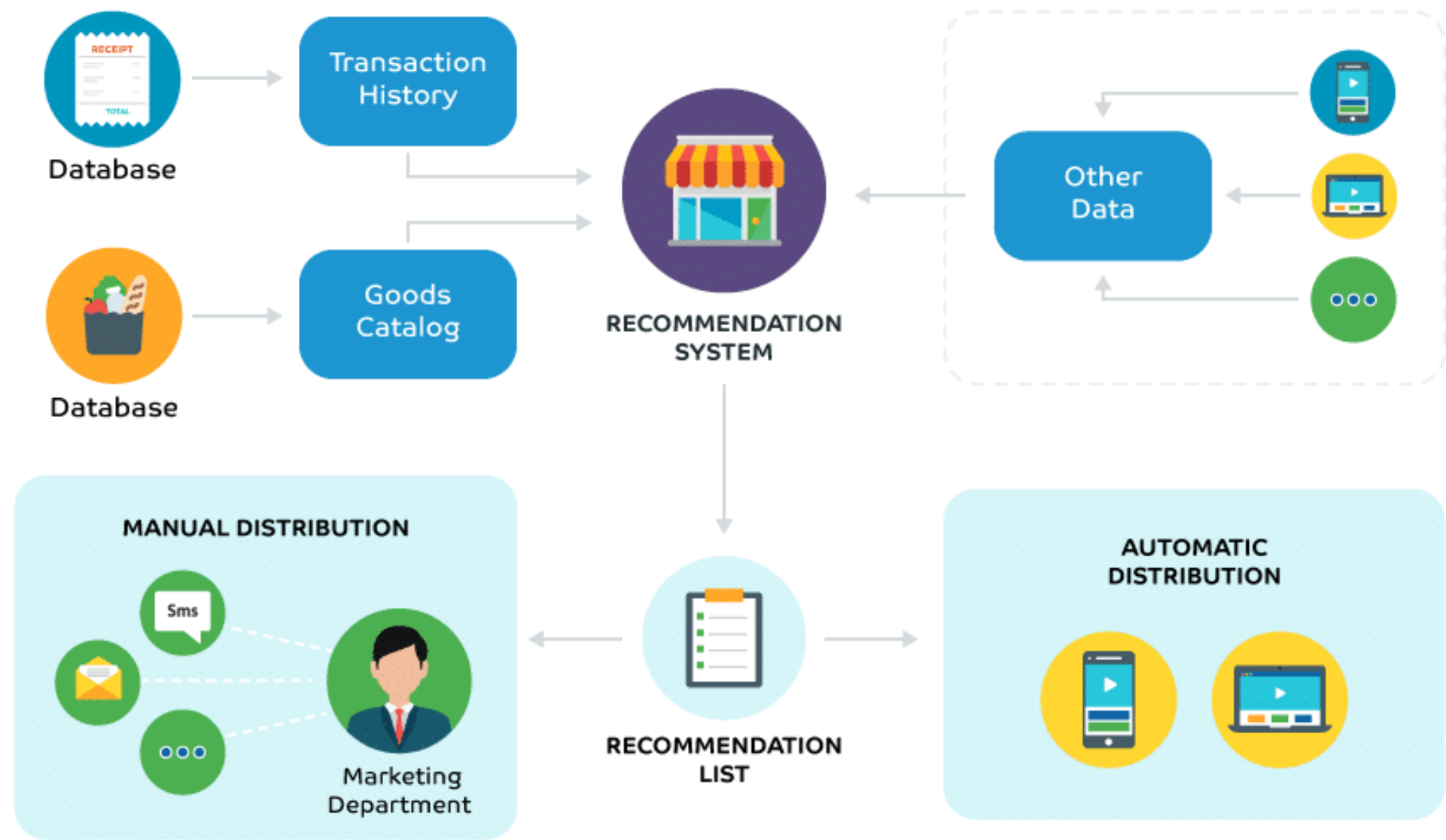
Các phương pháp tính toán độ tương tự

Độ tương tự cosine:



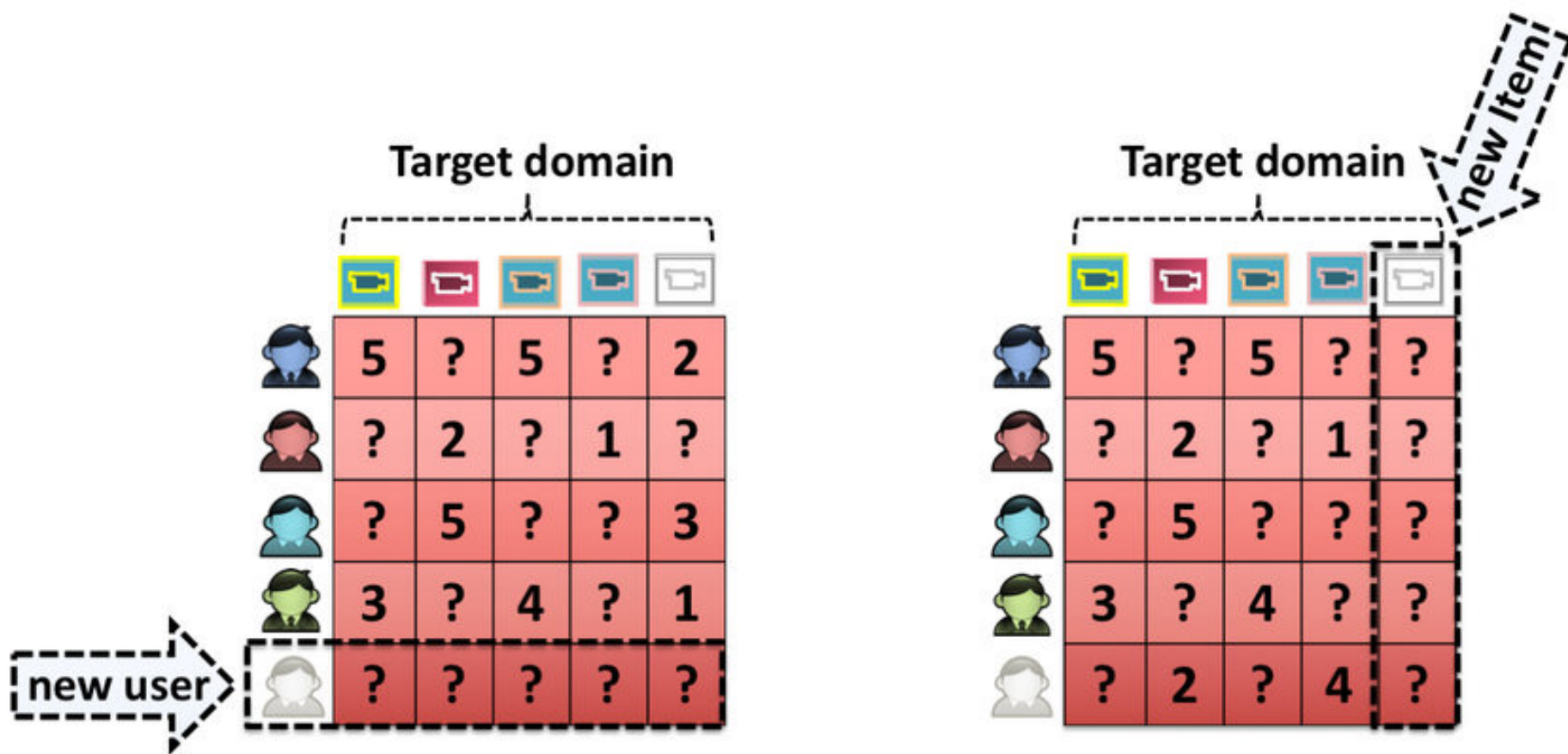
4. Sơ đồ tổng quan và Thách thức khi xây dựng hệ thống gợi ý

Sơ đồ tổng quan hệ thống



Thách thức

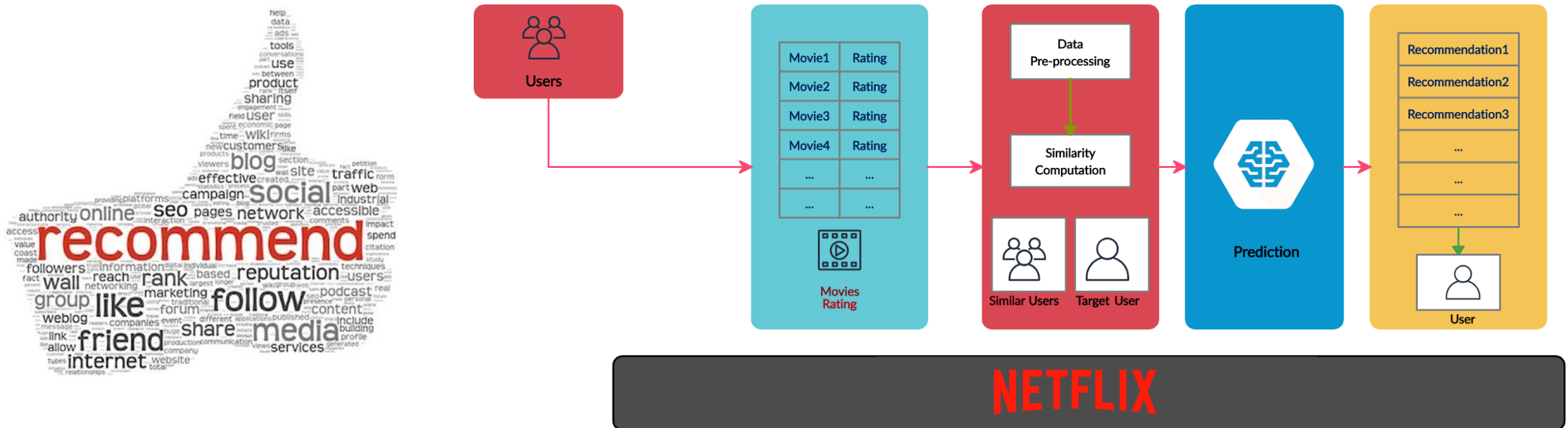
Một trong những thách thức với hệ thống gợi ý đó là vấn đề người dùng mới (new user) – item mới (new item), còn gọi là “cold-start problem”



5. Ví dụ demo hệ thống gợi ý

Giới thiệu bài toán

Xây dựng hệ thống gợi ý phim!



Giới thiệu bài toán

Dựa vào dữ liệu của trên 12 182 bộ films, xây dựng hệ thống đề xuất đưa ra danh sách 15 bộ film liên quan. Có hai loại Recommender system được xây dựng trong project này:

- Simple Recommender
- Content-Based Recommender

Các file dữ liệu sử dụng bao gồm:

Data_Movies.csv:File này chứa thông tin tổng hợp của ~ 12 000 bộ film, mỗi bộ film có 24 thuộc tính khác nhau, một số thuộc tính chính bao gồm:

1. **adult:** Bộ phim dành cho người lớn hay không. Dữ liệu boolean (True - False)
2. **original_language:** Ngôn ngữ ban đầu; dữ liệu categorical
3. **genres:** Thể loại phim
4. **original_title:** Tiêu đề của phim, dữ liệu text
5. **overview:** Tóm tắt nội dung của phim; Dữ liệu text
6. **release_date:** Ngày phát hành films
7. **vote_average:** Tỷ lệ vote trung bình [0-10]
8. **vote_count:** Số lượt vote





Thực hành

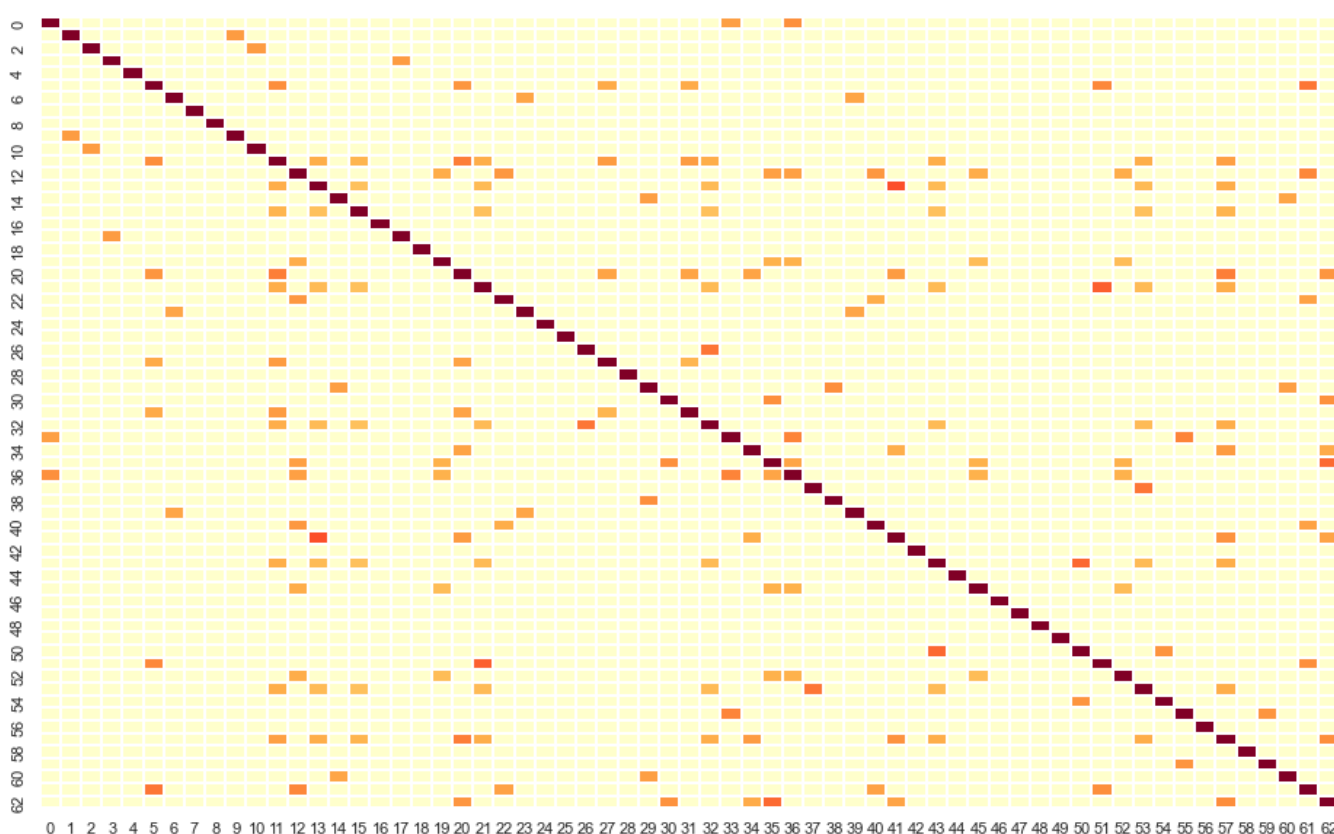
Thực hành

Sinh viên đọc file dữ liệu Data_VN_2021.xlsx, lưu trữ thông tin của 63 tỉnh, thành phố của Việt Nam

	Tỉnh/Thành phố ▾	Diện tích (km ²) ▾	Dân số (ngư ^o) ▾	Vĩ độ ▾	Kinh độ ▾
1					
2	Nghệ An	16493.7	3417809	18.6733	105.6922
3	Gia Lai	15510.8	1566882	13.9833	108
4	Sơn La	14123.5	1286068	21.327	103.9141
5	Đắk Lắk	13030.5	1897710	12.6667	108.05
6	Thanh Hóa	11114.7	3690022	19.8075	105.7764
7	Quảng Nam	10574.7	1510960	15.8733	108.3327
8	Lâm Đồng	9783.2	1319952	11.9359	108.4429
9	Kon Tum	9674.2	565685	14.3417	107.9792
10	Điện Biên	9541	623295	21.3833	103.0169
11	Lai Châu	9068.8	480588	22.3991	103.4393
12	Lạng Sơn	8310.2	791872	21.8478	106.7578
13	Quảng Bình	8065.3	905895	17.4833	106.6
14	Hà Giang	7929.5	883388	22.8233	104.9836
15	Bình Thuận	7812.8	1243977	10.9375	108.1583
16	Yên Bái	6887.7	838181	21.7	104.8667
17	Bình Phước	6877	1020839	11.6504	106.6
18	Cao Bằng	6700.3	535098	22.6731	106.25
19	Đắk Nông	6509.3	652766	12.0042	107.6907
20	Lào Cai	6364	756083	22.4194	103.995
21	Kiên Giang	6348.8	1730117	10.2289	103.9572
22	Quảng Ninh	6177.7	1358490	20.95	107.0833

Thực hành

1. Sử dụng phương pháp trích chọn đặc trưng TF-IDF để vector hóa dữ liệu tên các tỉnh/thành phố
2. Sử dụng độ đo cosine để tính toán độ tương đồng giữa các tên từ vector TF-IDF, trực quan hóa kết quả
3. Xây dựng hàm trả về tên 5 tỉnh/thành phố gần nhất với tên một tỉnh/thành phố đưa vào?



```
1 get_recommend_city('Quảng Ninh')
```

Tỉnh/Thành phố

11	Quảng Bình
57	Ninh Bình
5	Quảng Nam
62	Bắc Ninh
41	Ninh Thuận



Thank you!