

# Big Data

Sinh viên thực hiện:

Đàm Văn Tài

20122371

Phạm Văn Cường

20111231

# Hệ khuyến nghị

Hệ khuyến nghị là những hệ thống, công cụ, kĩ thuật được thiết kế để hướng đến mục đích hướng người dùng đến những đối tượng quan tâm, yêu thích khi lượng thông tin vượt quá khả năng xử lý của người dùng.

Hệ khuyến nghị có thể được xem là một giải pháp hỗ trợ tìm kiếm thông minh bằng cách hiểu sở thích của người dùng.



# Một số phương pháp

1. Hệ thống lọc dựa trên nội dung (Content-based)
2. Hệ thống lọc cộng tác (Collaborative Filtering)
3. Hệ thống kết hợp
4. Thuật toán dựa trên bộ nhớ
5. Thuật toán dựa trên mô hình



# Lọc dựa trên nội dung

Khuyến nghị các mặt hàng giống về nội dung với các mặt hàng khác được người mua đánh giá tốt

Đánh giá này dựa trên độ tương tự về nội dung (bao gồm tất cả các thuộc tính) giữa các mặt hàng



# Lọc cộng tác

Khuyến nghị các mặt hàng bằng việc so sánh độ tương đồng của hai người mua dựa vào bảng đánh giá các sản phẩm mà họ mua từ đó tìm ra tập người mua tương đồng với nhau. Sau đó sẽ đánh giá các mặt hàng mà người mua chưa đánh giá (đây gọi là lọc cộng tác dựa trên user-user). Hệ thống lọc cộng tác có thể chia ra làm hai phương pháp là dựa trên bộ nhớ (memory-based) và dựa trên mô hình.



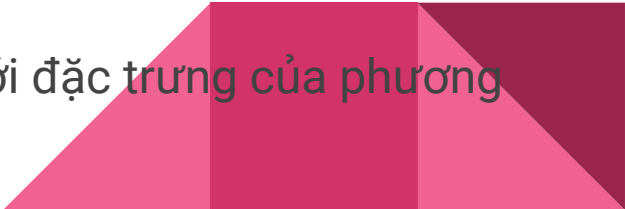
# Hệ thống kết hợp

Kết hợp giữa 2 phương pháp trên:

- Có thể sử dụng tách rời cả hai phương pháp lọc cộng tác (collaborative filtering) và dựa trên nội dung (content based), sau kết quả cuối cùng chúng ta kết hợp lại chọn những mặt hàng tốt nhất.

- Sử dụng phương pháp lọc cộng tác, kết hợp với các đặc trưng của nội dung.  
VD: trong dự đoán đánh giá phim, từ đặc điểm tuổi người xem, ta có thể gợi ý theo tuổi...

- Sử dụng phương pháp dựa trên nội dung, kết hợp với đặc trưng của phương pháp lọc cộng tác



# Thuật toán dựa trên bộ nhớ

- Sử dụng phương pháp thống kê để tìm tập người mua, được gọi là hàng xóm, tương đồng với người mua đang xét. Trong hệ khuyến nghị nó gồm:
  - + Lọc cộng tác dựa trên user-user
  - + Lọc cộng tác dựa trên item-item
- Để tìm được hàng xóm tương đồng thì bộ dữ liệu đầu vào, có mật độ lớn. VD trong bộ movielen dữ liệu đánh giá của mỗi người xem cho phim phải lớn



# Thuật toán dựa trên mô hình

- Sử dụng học máy, thuật toán khai phá dữ liệu mà tìm ra đặc trưng mẫu của dữ liệu và dự đoán đánh giá.
- Trong tập dữ liệu này, ta sử dụng SVD để phân tích ma trận đánh giá ban đầu về hai ma trận đặc trưng của user và item có số chiều thấp hơn.
- thường cho kết quả với độ chính xác cao.





# Đánh giá

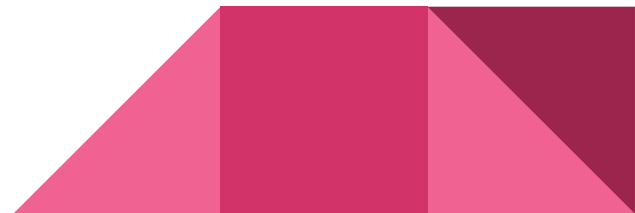
1. Đánh giá trung bình tuyệt đối của lỗi:

$$MAE = \frac{1}{n} \sum_{i=1}^n |rating_{actual_i} - rating_{predicted_i}|$$

Trong đó: n là tổng số lượt đánh giá (rating)

Ratingactual là đánh giá của người dùng trong thực tế

Ratingpredict là đánh giá của hệ thống dự đoán



# Đánh giá

## 2. Căn bậc 2 trung bình của bình phương lỗi

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (rating_{actual_i} - rating_{predicted_i})^2}$$



# Đánh giá chéo

Đánh giá chéo được hiểu như là đánh giá xoay vòng, ta chia dữ liệu thành các  $n$  bộ sau chọn  $n-1$  bộ để làm tập dữ liệu đào tạo và tập còn lại để đánh giá, cứ làm như vậy cho đến khi mọi tập đều được chọn làm đánh giá. Kết quả cuối cùng sẽ là giá trị trung bình của mỗi lần chọn.



# Áp dụng

Lọc cộng tác dựa trên item-item:

1. Dựa trên độ tương đồng Cosin

$$similarities(m, n) = cosine(\vec{m}, \vec{n}) = \frac{\vec{m} * \vec{n}}{\|\vec{m}\|_2 * \|\vec{n}\|_2}$$

Do tỉ lệ đánh giá của mỗi người khác nhau nên ta có thể giải quyết bằng cách sau:

$$similarity(m, n) = \frac{\sum_{u \in U} (R_{u,m} - \bar{R}_u)(R_{u,n} - \bar{R}_u)}{\sqrt{(R_{u,m} - \bar{R}_u)^2} \sqrt{(R_{u,n} - \bar{R}_u)^2}}$$

# Áp dụng

	1	2		<i>i</i>	<i>j</i>	<i>n</i>
1				R	R	
2				-	R	
				-	-	
				.	.	
<i>u</i>				.	.	
				.	.	
<i>m-2</i>				R	R	
<i>m-1</i>				R	-	
<i>m</i>				R	R	

Item-item similarity is computed by looking into co-rated items only. In case of items *i* and *j* the similarity is computed by calculating similarity between ratings in rows 1, *m-2* and *m*.

# Áp dụng

Lọc cộng tác dựa trên user-user:

	1	2	3			n-1	n
1							
2							
$i$	R		R			-	R
$j$	R		R			R	R
$m$							

User-item similarity is computed by looking into co-rated items only. In case of users  $i$  and  $j$  the similarity is computed by calculating similarity between ratings in columns 1, 3 and  $n$ .

# Áp dụng

## 2. Dựa trên hệ số tương quan Pearson

$$\text{similarity}(m, n) = \frac{\sum_{u \in U} (R_{u,m} - \bar{R}_m)(R_{u,n} - \bar{R}_n)}{\sqrt{(R_{u,m} - \bar{R}_m)^2} \sqrt{(R_{u,n} - \bar{R}_n)^2}}$$



# Áp dụng

## 3. Tính toán dự đoán

$$P_{u,m} = \frac{(\sum_N (s_{m,k} * R_{u,k}))}{\sum_N (|s_{m,k}|)}$$

Tròn đó:

$P_{u,m}$  là dự đoán đánh giá của người dùng  $u$  với mặt hàng  $m$

$s_{m,k}$  là độ tương tự của mặt hàng  $m$  và mặt hàng  $k$

$N$  là tập các mặt hàng tương tự với  $m$





# Áp dụng

Đánh giá:

```
print("MAE trung binh cua loc cong tac tren user: ", np.mean(MAE_user))
print("MAE trung binh cua loc cong tac tren item: ", np.mean(MAE_item))
print("RMSE trung binh cua loc cong tac tren user: ", np.mean(RMSE_user))
print("RMSE trung binh cua loc cong tac tren item: ", np.mean(RMSE_item))

print("RMSE trung binh cua loc con tac tren user tinh do tuong dong bang correlation: ", np.mean(RMSE_user_
_correlation))
```

```
('MAE trung binh cua loc cong tac tren user: ', 9.8185771729523132)
('MAE trung binh cua loc cong tac tren item: ', 12.10000866588871)
('RMSE trung binh cua loc cong tac tren user: ', 3.1333664762378204)
('RMSE trung binh cua loc cong tac tren item: ', 3.478443082625057)
('RMSE trung binh cua loc con tac tren user tinh do tuong dong bang correlation: ', 3.1283312456013621)
```