

Hệ tư vấn

(Recommender Systems)

Phan Xuân Hiếu

Khoa Công nghệ Thông tin
Trường ĐH Công nghệ (UET), ĐHQG Hà Nội (VNU)
hieupx@vnu.edu.vn

(last updated: 26–10–2016)

Nội dung

- ① Giới thiệu hệ tư vấn
- ② Các cách tiếp cận xây dựng hệ tư vấn
- ③ Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- ④ Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- ⑤ Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Nội dung

- 1 Giới thiệu hệ tư vấn
- 2 Các cách tiếp cận xây dựng hệ tư vấn
- 3 Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- 4 Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- 5 Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Hệ tư vấn là gì

Hệ tư vấn (recommender system) là một dạng công cụ lọc thông tin (information filtering) cho phép suy diễn và dự đoán các sản phẩm, dịch vụ, nội dung mà người dùng (có thể) quan tâm dựa trên những thông tin thu thập được về người dùng, về các sản phẩm, dịch vụ, về các hoạt động, tương tác cũng như đánh giá của người dùng đối với các sản phẩm, dịch vụ trong quá khứ.

- Tên gọi tiếng Anh: recommender system, recommendation system.
- Tên gọi tiếng Việt: hệ tư vấn, hệ khuyến nghị, hệ gợi ý.

Hệ tư vấn góp phần vào thành công của Amazon

<http://fortune.com/.../amazons-recommendation-secret/>

Amazon's recommendation secret

by JP Mangalindan

@jpmanga

JULY 30, 2012, 11:09 AM EST



Much is made of what the likes of Facebook, Google and Apple know about users. Truth is, Amazon may know more. And the massive retailer proves it every day.



Judging by Amazon's success, the recommendation system works. The company reported a 29% sales increase to \$12.83 billion during its second fiscal quarter, up from \$9.9 billion during the same time last year. A lot of that growth arguably has to do with the way Amazon has integrated recommendations into nearly every part of the purchasing process from product discovery to checkout. Go to Amazon.com and you'll find multiple panes of product suggestions; navigate to a particular product page and you'll see areas plugging items "Frequently Bought Together" or other items customers also bought. The company remains tight-lipped about how effective recommendations are. ("Our mission is to delight our customers by allowing them to serendipitously discover great products," an Amazon spokesperson told *Fortune*. "We believe this happens every single day and that's our biggest metric of success.")

Hệ tư vấn góp phần vào thành công của Amazon (2)



Frequently Bought Together



+



+



Price for all three: \$46.96

Add all three to Cart

Add all three to Wish List

Show availability and shipping details

- This item: The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful ... by Eric Ries Hardcover \$14.64
- Zero to One: Notes on Startups, or How to Build the Future by Peter Thiel Hardcover \$16.20
- The Hard Thing About Hard Things: Building a Business When There Are No Easy Answers by Ben Horowitz Hardcover \$16.12

Frequently Bought Together



Price for all three: \$48.78

Add all three to Cart

Add all three to Wish List

Show availability and shipping details

- This item: Golden Rose Matte Lipstick Set of 6 (SET1) \$24.99
- NYX Cosmetics Long Lasting Slim Lip Liner Pencils 6 Colors \$15.99
- Italia Eyeliners Set of 12 \$7.80

Hệ tư vấn góp phần vào thành công của Netflix

The Science Behind the Netflix Algorithms That Decide What You'll Watch Next



Star Trek, the first non-Trek title that Netflix is likely to suggest to you is the *Impossible* series (the one with the cool Lalo Schifrin soundtrack). Streaming *Who* is likely to net you the supernatural TV drama *Being Human* (the UK *From Dusk Till Dawn* and *300* and say hello to a new row on your homepage: Violent Action & Adventure. Trying to understand the invisible array of over your Netflix suggestions has long been a favorite sport, but what's in that galaxy of big data, those billions and billions of ratings stars? Turns out it's Netflix engineers working behind the scenes at their Silicon Valley HQ. The fact that 75 percent of viewer activity is driven by recommendation. This allows a profile feature enabling family members to demarcate their preferences and viewing histories. In March the company shipped its 4 billionth DVD, but in the first quarter of 2015 alone, it streamed more than 4 billion hours. We spoke with Netflix's recommendation dynamos—Carlos Gomez-Uribe, VP of product innovation and personalization algorithms (right), and Xavier Amatriain, engineering director—about how they control what you watch.

Nội dung

- 1 Giới thiệu hệ tư vấn
- 2 Các cách tiếp cận xây dựng hệ tư vấn
- 3 Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- 4 Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- 5 Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Các cách tiếp cận xây dựng hệ tư vấn

- Tiếp cận dựa trên luật hay tri thức (rule/knowledge-based).
 - ▶ Tư vấn dựa trên các luật được định nghĩa trước (predefined rules) hoặc các luật được trích rút từ dữ liệu (learned rules).
- Tiếp cận dựa trên lọc cộng tác (collaborative filtering - CF).
 - ▶ Tư vấn cho một người dựa trên sở thích, đánh giá của những người tương tự (similar users).
 - ▶ Sở thích (preferences) của người dùng có thể được thể hiện trực tiếp, rõ ràng (explicit) hoặc gián tiếp (implicit).
- Tiếp cận dựa trên nội dung (content-based).
 - ▶ Tư vấn dựa trên tính chất, nội dung của các sản phẩm, dịch vụ cũng như hồ sơ (profile) của người dùng.
- Tiếp cận dựa trên sự kết hợp giữa CF và content-based (hybrid).
 - ▶ Tư vấn bằng cách kết hợp ưu điểm của lọc cộng tác và tư vấn dựa trên nội dung.

Nội dung

- ① Giới thiệu hệ tư vấn
- ② Các cách tiếp cận xây dựng hệ tư vấn
- ③ Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- ④ Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- ⑤ Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Giới thiệu về lọc cộng tác

- Lọc cộng tác thực hiện tư vấn (gợi ý, khuyến nghị) các sản phẩm, dịch vụ, nội dung cho một người dùng nào đó dựa trên mối quan tâm, sở thích (preferences) của những người dùng tương tự đối với các sản phẩm, dịch vụ, nội dung đó.
- Lọc cộng tác được xem là một trong ba cách tiếp cận chính trong xây dựng các hệ thống tư vấn.
- Có nhiều kỹ thuật lọc cộng tác và được chia thành hai dạng chính:
 - ▶ Memory-based: lọc cộng tác dựa trên việc ghi nhớ toàn bộ dữ liệu.
 - ▶ Model-based: Lọc cộng tác dựa trên các mô hình phân lớp, hồi quy, dự đoán.

Amazon's collaborative filtering recommendation

Industry Report



Amazon.com Recommendations

Item-to-Item Collaborative Filtering

Greg Linden, Brent Smith, and Jeremy York • Amazon.com

	Book	CD	Movie	Game
User 1	Like	Dislike	Like	Like
User 2	Like	Dislike	Dislike	Dislike
User 3	Like	Like	Dislike	
User 4	Dislike		Like	
User 5	Like	Like	?	Dislike

	Book	CD	Movie	Game
User 1	Like	Dislike	Like	Like
User 2		Like	Dislike	Dislike
User 3	Like	Like	Dislike	
User 4	Dislike		Like	
User 5	Like	Like	?	Dislike

	Book	CD	Movie	Game
User 1	Like	Dislike	Like	Like
User 2		Like	Dislike	Dislike
User 3	Like	Like	Dislike	Dislike
User 4	Dislike		Like	
User 5	Like	Like	Like	Dislike

Giải thưởng Netflix - The Netflix Prize Challenge

Rank	Team	Best RMSE score	Improvement (%)
1	BellKor's Pragmatic Chaos	0.8556	10.07
2	Grand Prize Team	0.8571	9.91
3	Opera Solutions and Vandelay United	0.8573	9.89
4	Vandelay Industries!	0.8579	9.83
5	Pragmatic Theory	0.8582	9.80
6	BellKor in BigChaos	0.8590	9.71
7	Dace	0.8605	9.55
8	Opera Solutions	0.8611	9.49
9	BellKor	0.8612	9.48
10	BigChaos	0.8613	9.47

Nội dung

- ① Giới thiệu hệ tư vấn
- ② Các cách tiếp cận xây dựng hệ tư vấn
- ③ Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- ④ Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- ⑤ Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Ma trận đánh giá (user-item rating matrix)

Ma trận đánh giá:

- Tiếng Anh: rating matrix, user-item matrix, utility matrix.
- Tiếng Việt: ma trận đánh giá, ma trận người dùng – sản phẩm.

Ký hiệu:

- m người dùng $U = \{u_1, u_2, \dots, u_m\}$.
- n sản phẩm (items) $I = \{i_1, i_2, \dots, i_n\}$.
- Ma trận đánh giá: $\mathbf{R} = \{r_{u,i}\}_{m \times n}$ với $r_{u,i} \in \mathbb{R}$.

	I_1	I_2	I_3	I_4
U_1	4	?	5	5
U_2	4	2	1	
U_3	3		2	4
U_4	4	4		
U_5	2	1	3	5

Ví dụ về ma trận đánh giá

(a)

Alice: (like) Shrek, Snow White, (dislike) Superman

Bob: (like) Snow White, Superman, (dislike) spiderman

Chris: (like) spiderman, (dislike) Snow white

Tony: (like) Shrek, (dislike) Spiderman

(b)

	Shrek	Snow White	Spider-man	Super-man
Alice	Like	Like		Dislike
Bob		Like	Dislike	Like
Chris		Dislike	Like	
Tony	Like		Dislike	?

Xây dựng ma trận đánh giá

- Ma trận đánh giá tường minh (explicit rating matrix):
 - ▶ Người dùng đánh giá trực tiếp đối với các sản phẩm, dịch vụ, nội dung.
 - ▶ Thang điểm thường là:
 - ★ Nhị phân: like vs. dislike.
 - ★ Liên tục trong đoạn $[0, 1]$.
 - ★ Năm mức rời rạc: 1, 2, 3, 4, 5 (với 5 là mức tốt nhất).
- Ma trận đánh giá suy diễn (implicit rating matrix):
 - ▶ Suy diễn từ hành vi người dùng như:
 - ★ Duyệt web (browsing).
 - ★ Đọc (reading).
 - ★ Xem (watching)
 - ★ Chia sẻ (sharing).
 - ★ Mua (buying).
 - ▶ Ánh xạ các hành vi người dùng vào các mức điểm.

Nội dung

- ① Giới thiệu hệ tư vấn
- ② Các cách tiếp cận xây dựng hệ tư vấn
- ③ Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- ④ Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- ⑤ Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Các tính chất và thử thách của lọc cộng tác

- Dữ liệu thưa (data sparsity)
- Xuất phát nguội (cold start)
 - ▶ Vấn đề người dùng mới (new user problem)
 - ▶ Vấn đề sản phẩm mới (new item problem)
- Khả năng mở rộng (scalability)
- Vấn đề từ đồng nghĩa (synonymy)
- Gray sheep và Black sheep
- Shilling attacks

Dữ liệu thưa (data sparsity)

- Ma trận đánh giá (user-item/rating matrix) có thể rất thưa (sparse).
- Dữ liệu thưa ảnh hưởng rất nhiều đến hiệu quả tư vấn bởi rất khó tính toán sự tương tự giữa các người dùng (users) hoặc giữa các sản phẩm (items).
 - ▶ Hai sản phẩm có thể rất giống nhau nhưng có rất ít người cùng đánh giá (rating) đồng thời hai sản phẩm.
 - ▶ Hai người dùng có thể giống nhau về sở thích (preferences) nhưng chưa đánh giá cùng các sản phẩm.
- Dữ liệu thưa có thể dẫn đến độ phủ (coverage) của hệ tư vấn bị giảm.
- Giải pháp: áp dụng các kỹ thuật giảm số chiều (dimensionality reduction):
 - ▶ Singular value decomposition (SVD).
 - ▶ Principal component analysis (PCA).
 - ▶ Content–boosted CF.

Xuất phát nguội (cold start)

- Người dùng mới (new user problem):
 - ▶ Chưa tham gia đánh giá.
 - ▶ Chưa có dữ liệu lịch sử (xem, mua sắm ...).
- Sản phẩm mới (new item problem):
 - ▶ Chưa được ai (người dùng nào) đánh giá.
 - ▶ Chưa được xem, mua sắm ... bởi bất cứ người dùng nào.
- Giải pháp:
 - ▶ Tư vấn sản phẩm *hot*, phổ biến, hoặc ngẫu nhiên để tăng tính đa dạng.
 - ▶ Content–boosted CF: tích hợp thêm hồ sơ (profile) người dùng mới hoặc sử dụng thêm các đặc tính của sản phẩm.

Khả năng mở rộng (scalability)

- Khi ma trận rating lớn, tức số người dùng lẫn số sản phẩm lớn, thời gian tính toán tăng cao:
 - ▶ Tính toán trên ma trận.
 - ▶ Tính toán độ tương tự giữa các người dùng hoặc giữa các sản phẩm.
 - ▶ Tính toán và lựa chọn k người dùng láng giềng.
 - ▶ Khó đáp ứng tư vấn thời gian thực hoặc gần thời gian thực.
- Giải pháp:
 - ▶ Áp dụng các kỹ thuật giảm số chiều như SVD, PCA.
 - ▶ Item-based CF có khả năng mở rộng cao hơn so với user-based CF.

Vấn đề từ đồng nghĩa (synonymy)

- Vấn đề từ đồng nghĩa cũng gây cản trở cho việc tính toán độ tương tự.
- Ví dụ: *children movie* và *children film* có thể gây keyword-mismatch, làm ảnh hưởng đến việc tính toán độ tương tự.
- Giải pháp: có thể áp dụng các kỹ thuật phân tích ngữ nghĩa như LSI (Latent Semantic Indexing), mô hình chủ đề (Topic Models), hoặc Deep Learning để giải quyết vấn đề này.

Gray sheep và Black sheep

- Gray sheep:

- ▶ Có những người dùng có sở thích không tuân theo một nhóm người dùng nào cả.
- ▶ CF không thực sự hiệu quả trong những trường hợp này.
- ▶ Có thể kết hợp CF và content-based.

- Black sheep:

- Một số người đánh giá (rating) kỳ quặc và do đó không thể recommend chính xác cho những người này.

Shilling attacks

- Xảy ra khi có cạnh tranh không lành mạnh:
 - ▶ Đánh giá sản phẩm của mình rất cao.
 - ▶ Đánh giá sản phẩm của đối thủ rất thấp.
- Item-based CF ít bị ảnh hưởng bởi *shilling attacks* hơn so với user-based CF.
- Có thể phát hiện hiện tượng này ở bước tiền xử lý dữ liệu bằng phân tích phát hiện ngoại lệ.

Các cách tiếp cận và kỹ thuật lọc cộng tác (CF techniques)

CF categories	Representative techniques	Main advantages	Main shortcomings
Memory-based CF	* Neighbor-based CF (item-based/user-based CF algorithms with Pearson/vector cosine correlation) * Item-based/user-based top- <i>N</i> recommendations	* easy implementation * new data can be added easily and incrementally * need not consider the content of the items being recommended * scale well with co-rated items	* are dependent on human ratings * performance decrease when data are sparse * cannot recommend for new users and items * have limited scalability for large datasets
	* Bayesian belief nets CF * clustering CF	* better address the sparsity, scalability and other problems	* expensive model-building
Model-based CF	* MDP-based CF * latent semantic CF * sparse factor analysis * CF using dimensionality reduction techniques, for example, SVD, PCA	* improve prediction performance * give an intuitive rationale for recommendations	* have trade-off between prediction performance and scalability * lose useful information for dimensionality reduction techniques
Hybrid recommenders	* content-based CF recommender, for example, <i>Fab</i> * content-boosted CF * hybrid CF combining memory-based and model-based CF algorithms, for example, Personality Diagnosis	* overcome limitations of CF and content-based or other recommenders * improve prediction performance * overcome CF problems such as sparsity and gray sheep	* have increased complexity and expense for implementation * need external information that usually not available

Nội dung

- 1 Giới thiệu hệ tư vấn
- 2 Các cách tiếp cận xây dựng hệ tư vấn
- 3 Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- 4 Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- 5 Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Lọc cộng tác dựa trên ghi nhớ (memory-based CF)

- Memory-based CF sử dụng toàn bộ tập dữ liệu (ma trận đánh giá) để thực hiện dự đoán và tư vấn.
- Mỗi người dùng thuộc về ít nhất một nhóm những người dùng có chung sở thích/mối quan tâm.
- Người dùng cần được tư vấn, khuyến nghị tại một thời điểm nào đó được gọi là *active user*.
- Những người có sở thích tương tự với *active user* được gọi là những láng giềng (*neighbors*).
- Có hai cách tiếp cận:
 - User-based: dựa theo người dùng để dự đoán.
 - Item-based: dựa theo sản phẩm (items) để dự đoán.

Kịch bản lọc cộng tác dựa trên ghi nhớ

Gợi ý cho người dùng a (active user):

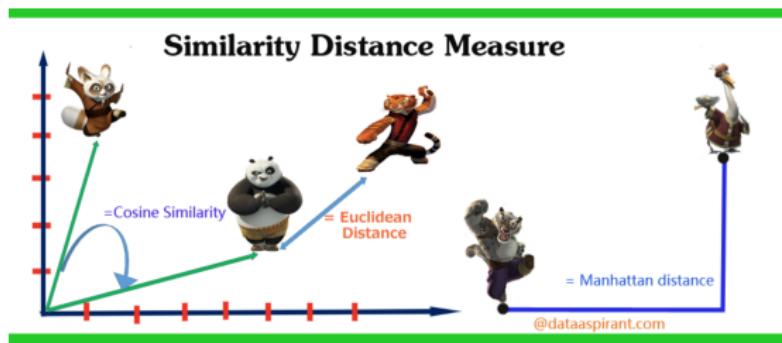
- Bước 1 - similarity computation: tính toán độ tương tự giữa các người dùng ($w_{u,v}$) đối với user-based và giữa các sản phẩm ($w_{i,j}$) đối với item-based.
- Bước 2 - prediction: ước lượng hay dự đoán giá trị rating của active user đối với một sản phẩm nào đó dựa trên thông tin ở bước 1.

Gợi ý top- N (top- N recommendation):

- Xác định k người dùng tương tự nhất với active user và gọi là k người dùng láng giềng (neighbors).
- Tổng hợp từ k người dùng láng giềng top- N sản phẩm (items) phổ biến nhất để tư vấn cho active user.

Tính toán mức độ tương tự (similarity computation)

- Là một bước quan trọng trong lọc cộng tác dựa trên bộ nhớ (memory-based CF).
- Đối với item-based CF: tính toán độ tương tự $w_{i,j}$ giữa hai item i và j dựa trên những người đã cùng đánh giá (rate) hai items này.
- Đối với user-based CF: tính toán độ tương tự $w_{u,v}$ giữa hai người dùng u và v dựa trên những đánh giá của hai người dùng này trên các items.



Vector cosine-based similarity

	I_1	I_2	I_3	I_4
U_1	4	?	5	5
U_2	4	2	1	
U_3	3		2	4
U_4	4	4		
U_5	2	1	3	5

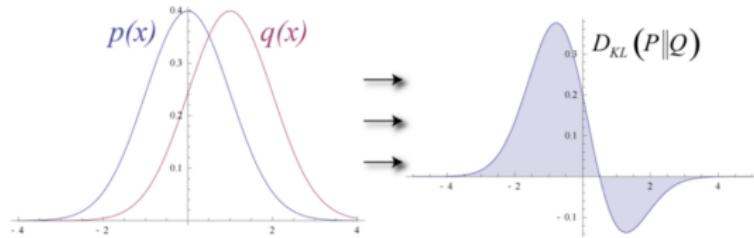
$$w_{i,j} = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\| \times \|\vec{j}\|} \quad (1)$$

- \vec{i} và \vec{j} là hai véc tơ trong ma trận rating của hai sản phẩm i và j .
- Ở ma trận trên: $\vec{l}_1 = (4, 4, 3, 4, 2)$ và $\vec{l}_2 = (?, 2, ?, 4, 1)$.
- Khi tính toán $w_{i,j}$: $\vec{l}_1 = (4, 4, 2)$ và $\vec{l}_2 = (2, 4, 1)$.

Khoảng cách Kullback–Leiber

Ký hiệu $p(x)$ và $q(x)$ là hai phân bố xác suất:

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (2)$$



$$\text{KL-similarity} = \frac{D(p||q) + D(q||p)}{2} \quad (3)$$

- Với user-based: hai phân bố là hai hàng trong ma trận đánh giá \mathbf{R} .
- Với item-based: hai phân bố là hai cột trong ma trận đánh giá \mathbf{R} .
- Các phân bố cần được chuẩn hoá trước khi tính $D(p||q)$ và $D(q||p)$.

Tương quan Pearson - user-based CF

	I_1	I_2	I_3	I_4
U_1	4	?	5	5
U_2	4	2	1	
U_3	3		2	4
U_4	4	4		
U_5	2	1	3	5

$$w_{u,v} = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I} (r_{v,i} - \bar{r}_v)^2}} \quad (4)$$

- I : tập các items cả hai người dùng u và v cùng đánh giá.
- \bar{r}_u và \bar{r}_v là rating trung bình của u và v trên các sản phẩm trong I .
- Với ma trận ví dụ trên, $w_{1,5} = 0.756$.

Tương quan Pearson - item-based CF

	1	2	...	i	j	...	m-1	m
1				R	?			
2				R		R		
:								
l				R		R		
:								
n-1				?		R		
n				R		R		

$$w_{i,j} = \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U} (r_{u,i} - \bar{r}_i)^2} \sqrt{\sum_{u \in U} (r_{u,j} - \bar{r}_j)^2}} \quad (5)$$

- U : tập các users cùng đánh giá cả hai sản phẩm i và j .
- $r_{u,i}$: rating u đối với i , tương tự cho $r_{u,j}$.
- \bar{r}_i , \bar{r}_j : trung bình rating của các người dùng trong U đối với i và j .

Dự đoán và tư vấn - Weight Sum of Others' Ratings

Dự đoán mức độ rating của *active user* a đối với một sản phẩm i nào đó, ký hiệu là $P_{a,i}$:

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_u) \times w_{a,u}}{\sum_{u \in U} |w_{a,u}|} \quad (6)$$

- \bar{r}_a và \bar{r}_u là rating trung bình của a và u trên các sản phẩm.
- $w_{a,u}$ là mức độ tương tự giữa hai người dùng a và u .
- U là tập tất cả người dùng (trừ a) đã đánh giá sản phẩm i .

Dự đoán và tư vấn - Weight Sum of Others' Ratings (2)

	I_1	I_2	I_3	I_4
U_1	4	?	5	5
U_2	4	2	1	
U_3	3		2	4
U_4	4	4		
U_5	2	1	3	5

$$P_{1,2} = \bar{r}_1 + \frac{\sum_u (r_{u,2} - \bar{r}_u) \times w_{1,u}}{\sum_u |w_{1,u}|} \quad (7)$$

$$= \bar{r}_1 + \frac{(r_{2,2} - \bar{r}_2)w_{1,2} + (r_{4,2} - \bar{r}_4)w_{1,4} + (r_{5,2} - \bar{r}_5)w_{1,5}}{|w_{1,2}| + |w_{1,4}| + |w_{1,5}|} \quad (8)$$

$$= 4.67 + \frac{(2 - 2.5)(-1) + (4 - 4)0 + (1 - 3.33)0.756}{1 + 0 + 0.756} \quad (9)$$

$$= 3.95 \quad (10)$$

Dự đoán và tư vấn - Simple Weighted Average

Với tư vấn dựa trên sản phẩm (item-based), giá trị dự đoán rating của một người dùng u trên sản phẩm i , ký hiệu $P_{u,i}$, được tính như sau:

$$P_{u,i} = \frac{\sum_{j \in J} r_{u,j} w_{i,j}}{\sum_{j \in J} |w_{i,j}|} \quad (11)$$

Trong đó:

- J là tập tất cả các sản phẩm (ngoài i) mà người dùng u đã đánh giá.
- $w_{i,j}$ là mức độ tương tự giữa hai sản phẩm i và j .
- $r_{u,j}$ là rating của người dùng u đối với sản phẩm j .

Gợi ý top- N (Top- N recommendations)

- Tư vấn, gợi ý N sản phẩm có thể được quan tâm nhất hay liên quan nhất đến một người dùng nào đó.
- Gợi ý top- N có thể được thực hiện theo hai cách:
 - ▶ Theo người dùng (user-based).
 - ▶ Theo sản phẩm (item-based).
- Việc tư vấn dựa vào các tính toán trên ma trận đánh giá R .

Gợi ý top- N theo người dùng (user-based)

- Gọi a là *active user*, \mathbf{R} là ma trận đánh giá.
- Tìm U_a là tập k người dùng tương tự nhất với a .
 - ▶ Dùng độ đo tương quan Pearson hoặc cosine.
- Gọi C là tập tất cả các sản phẩm mà các người dùng trong U_a đã mua (hoặc đánh giá) mà a chưa từng mua hay đánh giá.
- Xếp hạng các sản phẩm trong C giảm dần theo số người dùng (trong U_a) mua hoặc đánh giá cao.
- Lấy top- N sản phẩm từ C theo thứ tự xếp hạng trên để tư vấn hay gợi ý cho a .

Gợi ý top- N theo sản phẩm (item-based)

- Gọi a là *active user*, \mathbf{R} là ma trận đánh giá.
- Gọi I_a là tập các sản phẩm mà a đã mua hoặc đánh giá cao.
- Với mỗi sản phẩm i trong I_a , xác định k sản phẩm tương tự nhất với i , ký hiệu I_i^k .
- Gọi C là tập các sản phẩm bằng cách lấy hợp các I_i^k (với $i \in I_a$).
- Loại bỏ khỏi C các sản phẩm có trong I_a .
- Tính độ tương tự giữa các sản phẩm trong C với tập I_a .
 - ▶ Có thể tính độ tương tự bằng việc tính tổng độ tương tự giữa mỗi sản phẩm trong C và mỗi sản phẩm trong I_a .
 - ▶ Có thể tính bằng cách lấy trung bình độ tương tự giữa các sản phẩm trong C và mỗi sản phẩm trong I_a .
- Xếp hạng C giảm dần theo mức độ tương tự nói trên.
- Lấy top- N sản phẩm từ C theo thứ tự giảm dần của độ tương tự, sau đó tư vấn cho người dùng a .

Nội dung

- 1 Giới thiệu hệ tư vấn
- 2 Các cách tiếp cận xây dựng hệ tư vấn
- 3 Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - **Lọc cộng tác dựa trên mô hình (model-based CF)**
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- 4 Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- 5 Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Lọc cộng tác dựa trên mô hình (model-based CF)

- Thực hiện tư vấn dựa trên các mô hình học máy.
- Các mô hình này được xây dựng từ dữ liệu huấn luyện.
- Các phương pháp để xây dựng mô hình lọc cộng tác thường dùng:
 - ▶ Bayesian models.
 - ▶ Clustering models.
 - ▶ Dependency networks.
- Các mô hình tư vấn dựa trên học có giám sát có thể là
 - ▶ dạng mô hình phân lớp (classification), hoặc
 - ▶ dạng mô hình hồi quy (regression).

Ưu và nhược điểm của lọc cộng tác dựa trên mô hình

CF categories	Representative techniques	Main advantages	Main shortcomings
Memory-based CF	<ul style="list-style-type: none">* Neighbor-based CF (item-based/user-based CF algorithms with Pearson/vector cosine correlation)* Item-based/user-based top-N recommendations	<ul style="list-style-type: none">* easy implementation* new data can be added easily and incrementally* need not consider the content of the items being recommended* scale well with co-rated items	<ul style="list-style-type: none">* are dependent on human ratings* performance decrease when data are sparse* cannot recommend for new users and items* have limited scalability for large datasets
Model-based CF	<ul style="list-style-type: none">* Bayesian belief nets CF* clustering CF* MDP-based CF* latent semantic CF* sparse factor analysis* CF using dimensionality reduction techniques, for example, SVD, PCA	<ul style="list-style-type: none">* better address the sparsity, scalability and other problems* improve prediction performance* give an intuitive rationale for recommendations	<ul style="list-style-type: none">* expensive model-building* have trade-off between prediction performance and scalability* lose useful information for dimensionality reduction techniques
Hybrid recommenders	<ul style="list-style-type: none">* content-based CF recommender, for example, Fab* content-boosted CF* hybrid CF combining memory-based and model-based CF algorithms, for example, Personality Diagnosis	<ul style="list-style-type: none">* overcome limitations of CF and content-based or other recommenders* improve prediction performance* overcome CF problems such as sparsity and gray sheep	<ul style="list-style-type: none">* have increased complexity and expense for implementation* need external information that usually not available

Xây dựng hệ tư vấn với Naive Bayes

- Áp dụng khi các giá trị đánh giá của người dùng là rời rạc và hữu hạn.
- Dựa vào đề tư vấn về bài toán phân lớp.
 - ▶ Mỗi mức đánh giá của người dùng tương ứng với một lớp (class).
 - ▶ Ví dụ: năm mức đánh giá tương ứng với năm lớp {1, 2, 3, 4, 5}.
- Dựa trên giả thiết độc lập (independence assumption): *các thuộc tính độc lập với nhau nếu biết nhãn lớp.*

Xây dựng hệ tư vấn với Naive Bayes (2)

- Ký hiệu $C = \{c_1, c_2, \dots, c_k\}$ là k lớp tương ứng với k mức đánh giá của người dùng.
- Việc tư vấn sẽ dựa trên dự đoán nhãn lớp theo công thức sau:

$$c^* = \operatorname{argmax}_{j \in C} p(c_j) \prod_h P(X_h = x_h | c_j) \quad (12)$$

- Để tránh vấn đề zero probability, chúng ta dùng Laplace Estimator để tính $P(C = c | X = x)$:

$$P(X = x | Y = y) = \frac{\#(X = x, Y = y) + 1}{\#(Y = y) + |X|} \quad (13)$$

- Ví dụ:

- ▶ $P(X = 0 | Y = 1) = 0/2$ sẽ thành $(0 + 1)/(2 + 2) = 1/4$.
- ▶ $P(X = 1 | Y = 1) = 2/2$ sẽ thành $(2 + 1)/(2 + 2) = 3/4$.

Xây dựng hệ tư vấn với tiếp cận phân cụm (clustering)

- Mỗi cụm là một tập các đối tượng tương tự nhau (theo một tiêu chí nào đó).
- Đo độ tương tự giữa các đối tượng theo các độ đo như Cosine, Minkowski, tương quan Pearson, ... Ví dụ với hai đối tượng $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ và $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$, độ đo khoảng cách Minkowski có công thức:

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^n |x_i - y_i|^q \right)^{\frac{1}{q}} \quad (14)$$

- ▶ $q = 1$: độ đo Manhattan.
- ▶ $q = 2$: độ đo Euclid.
- Nhiều cách tiếp cận phân cụm: phân cụm phân cấp, phân cụm phân hoạch, phân cụm dựa trên mật độ, phân cụm dựa trên đồ thị ...
- Có thể áp dụng memory-based CF trên kết quả phân cụm để thực hiện tư vấn cho từng cụm.

Xây dựng hệ tư vấn với SVD (singular value decomposition)

$$\begin{pmatrix} X \\ \begin{matrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{matrix} \\ m \times n \end{pmatrix} = \begin{pmatrix} U \\ \begin{matrix} u_{11} & \dots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{matrix} \\ m \times r \end{pmatrix} \begin{pmatrix} S \\ \begin{matrix} s_{11} & 0 & \dots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{matrix} \\ r \times r \end{pmatrix} \begin{pmatrix} V^T \\ \begin{matrix} v_{11} & \dots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{matrix} \\ r \times n \end{pmatrix}$$

Phân tích SVD trên ma trận đánh giá

SVD

$$\left(\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{array} \right) = \left(\begin{array}{ccc} & U & \\ & \left(\begin{array}{ccc} u_{11} & \dots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{array} \right) & \\ & m \times r & \end{array} \right) \left(\begin{array}{ccc} S & & \\ \left(\begin{array}{ccc} s_{11} & 0 & \dots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{array} \right) & & \\ r \times r & & \end{array} \right) \left(\begin{array}{ccc} V^T & & \\ \left(\begin{array}{ccc} v_{11} & \dots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{array} \right) & & \\ r \times n & & \end{array} \right)$$

$$\left(\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{array} \right)_{m \times n} = \boxed{\left(\begin{array}{ccc} & U & \\ u_{11} & \dots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{array} \right)_{m \times r}}$$

Giảm số chiều với SVD

$$\left(\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{array} \right)_{m \times n} = \left(\begin{array}{ccc} U & & \\ u_{11} & \dots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{array} \right)_{m \times d} \left(\begin{array}{ccc} S & & \\ s_{11} & 0 & \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{array} \right)_{d \times d} \left(\begin{array}{cc} V^T & \\ v_{11} & \dots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{array} \right)_{d \times n}$$

$m \times d$ $d \times d$ $d \times n$

Nội dung

- ① Giới thiệu hệ tư vấn
- ② Các cách tiếp cận xây dựng hệ tư vấn
- ③ Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- ④ Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- ⑤ Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Đánh giá hiệu quả (evaluation metrics) cho lọc cộng tác

Có rất nhiều cách tiếp cận và độ đo hiệu quả khác nhau:

- Mean Absolute Error (MAE).
- Normalized Mean Absolute Error (NMAE).
- Root Mean Squared Error (RMSE).
- ROC Sensitivity.

Mean Absolute Error và Normalized Mean Absolute Error

- MAE và NMAE được sử dụng rất rộng rãi trong việc đánh giá hiệu quả lọc cộng tác.
- Công thức tính:

$$MAE = \frac{\sum_{\{u,i\}} |p_{u,i} - r_{u,i}|}{n} \quad (15)$$

- ▶ n là tổng số các đánh giá của tất cả các người dùng.
 - ▶ $r_{u,i}$ là đánh giá thực sự của người dùng u trên sản phẩm i .
 - ▶ $p_{u,i}$ là giá trị đánh giá dự đoán của người dùng u đối với sản phẩm i .
- NMAE được chuẩn hoá theo độ lớn của các giá trị đánh giá:

$$NMAE = \frac{MAE}{r_{\max} - r_{\min}} \quad (16)$$

- ▶ r_{\max} và r_{\min} là giá trị đánh giá lớn nhất và nhỏ nhất trong ma trận.

Root Mean Squared Error (RMSE)

- RMSE trở nên phổ biến vì được sử dụng trong giải thưởng Netflix.
- Công thức tính:

$$RMSE = \sqrt{\frac{1}{n} \sum_{\{u,i\}} (p_{u,i} - r_{u,i})^2} \quad (17)$$

- ▶ n là tổng số các đánh giá của tất cả các người dùng.
- ▶ $r_{u,i}$ là đánh giá thực sự của người dùng u trên sản phẩm i .
- ▶ $p_{u,i}$ là giá trị đánh giá dự đoán của người dùng u đối với sản phẩm i .
- RMSE khuếch đại sai số tuyệt đối giữa giá trị đánh giá thực sự và giá trị đánh giá dự đoán.

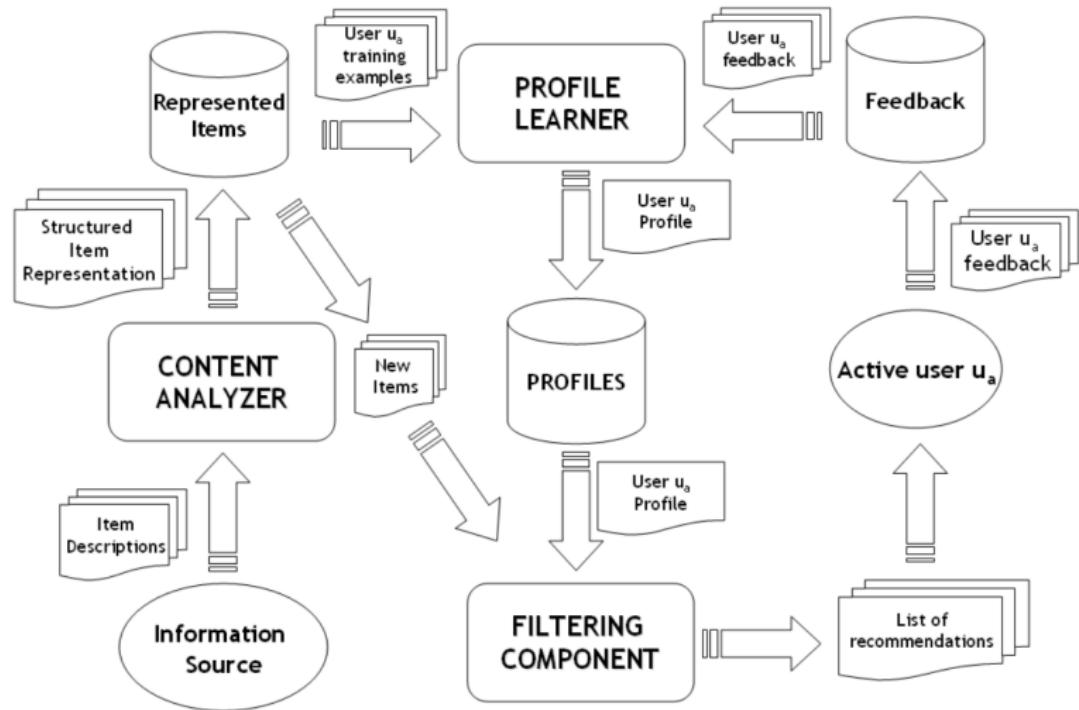
Nội dung

- 1 Giới thiệu hệ tư vấn
- 2 Các cách tiếp cận xây dựng hệ tư vấn
- 3 Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- 4 Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- 5 Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Tư vấn dựa trên nội dung (content-based)

- Tư vấn cho một người nào đó các sản phẩm tương tự với (các) sản phẩm mà người đó đã mua hoặc quan tâm trong quá khứ.
- Đối sánh thuộc tính (attributes) của hồ sơ người dùng (user profile) với thuộc tính của các sản phẩm để quyết định nội dung tư vấn.
- Cần lưu trữ hồ sơ người dùng và cập nhật hồ sơ người dùng dựa trên những hành vi người dùng thực hiện trong quá khứ.

Kiến trúc chung của hệ tư vấn dựa trên nội dung



Các thành phần chính của hệ tư vấn theo nội dung

- Content Analyzer:
 - ▶ Phân tích và trích chọn thông tin, nội dung từ các sản phẩm.
 - ▶ Biểu diễn thông tin trích chọn được theo một cấu trúc nào đó.
 - ▶ Ví dụ: document → bag-of-word; image → object labels
- User Profiles:
 - ▶ Là cơ sở dữ liệu lưu trữ thông tin, sở thích của người dùng, ví dụ:
 - ▶ Thông tin nhân khẩu học (demographic): tuổi, giới tính, nơi sống ...
 - ▶ Thông tin sở thích (preferences/interests): sở thích đọc, xem, mua ...
 - ▶ Thông tin về tình trạng, ngữ cảnh: mang bầu? có con nhỏ? ...
 - ▶ Thông tin về ý định (intent): ý định mua/bán, ý định đi du lịch ...
- Profile Learner: cập nhật hồ sơ người dùng dựa trên hành vi của họ:
 - ▶ Các hành vi là một dạng thông tin phản hồi (feedback) của họ.
 - ▶ Các hành vi: đọc (read), xem (watch), thích (like), đánh giá (rate), bình luận (comment), chia sẻ (share), mua (purchase) ...
- Filtering Component: thực hiện tư vấn bằng cách
 - ▶ Đồi sánh những nội dung sản phẩm mới với hồ sơ người dùng.
 - ▶ Xếp hạng các sản phẩm theo mức độ tương đồng với người dùng.

Ưu và nhược điểm của tư vấn theo nội dung

• Ưu điểm:

- ▶ Tính độc lập giữa các người dùng (user independence).
- ▶ Tính dễ hiểu (transparency, white-box recommendation).
- ▶ Giải quyết được phần lớn vấn đề xuất phát nguội.
 - ★ Giải quyết được vấn đề sản phẩm mới (new item problem).
 - ★ Giải quyết được phần nào vấn đề người dùng mới (new user problem).

• Nhược điểm:

- ▶ Cần phải phân tích và trích chọn nội dung sản phẩm.
 - ★ Một số dạng sản phẩm rất khó trích chọn nội dung (âm thanh, hình ảnh, phim ảnh).
 - ★ Cần có tri thức miền (domain knowledge).
- ▶ Có thể tạo lối mòn (over-specification, serendipity).
 - ★ Thường tư vấn những dạng sản phẩm (quá) quen thuộc.
 - ★ Rất khó để tư vấn các sản phẩm mới mẻ (novel) hay thú vị bất thường (unexpected).
- ▶ Vẫn còn khó khăn khi tư vấn cho người dùng mới (new user problem).

Biểu diễn, phân tích nội dung và cập nhật hồ sơ

- Biểu diễn sản phẩm và hồ sơ người dùng:
 - ▶ Biểu diễn theo không gian véc tơ (vector space model).
 - ▶ Biểu diễn theo <attribute:value>.
 - ▶ Biểu diễn trên đồ thị: <node:property:value>.
- Phân tích nội dung:
 - ▶ Phân tích dữ liệu text/web.
 - ▶ Phân tích dữ liệu hình ảnh, video (vấn đề khó).
 - ▶ Các kỹ thuật phân tích tự động: phân cụm, phân tích chủ đề, ...
- Cập nhật hồ sơ người dùng (user profile update).

Nội dung

- 1 Giới thiệu hệ tư vấn
- 2 Các cách tiếp cận xây dựng hệ tư vấn
- 3 Xây dựng hệ tư vấn với lọc cộng tác (collaborative filtering - CF)
 - Ma trận đánh giá (rating matrix)
 - Các tính chất và thử thách của lọc cộng tác
 - Lọc cộng tác dựa trên ghi nhớ (memory-based CF)
 - Lọc cộng tác dựa trên mô hình (model-based CF)
 - Đánh giá hiệu quả các phương pháp lọc cộng tác
- 4 Xây dựng hệ tư vấn dựa trên nội dung (content-based)
- 5 Xây dựng hệ tư vấn với các phương pháp lai (hybrid approach)

Hệ tư vấn với phương pháp lai (hybrid)

- Kết hợp hai cách tiếp cận: lọc cộng tác và dựa trên nội dung.
- Cải thiện được những nhược điểm của cả hai cách tiếp cận.
- Có nhiều cách để kết hợp:
 - ▶ Xây dựng CF và content-based rời rạc, kết hợp kết quả đoán nhận của hai mô hình.
 - ▶ Tích hợp một số đặc tính về nội dung vào trong hệ CF.
 - ▶ Tích hợp một số đặc điểm CF vào hệ content-based.
 - ▶ Xây dựng một hệ tư vấn hợp nhất tích hợp được đặc trưng của cả hai cách tiếp cận.

Tài liệu tham khảo

-  M. D. Ekstrand, J. T. Riedl, and J. A. Konstan.
Collaborative Filtering Recommender Systems.
Foundations and Trends in Human–Computer Interaction, Vol.4, No.2,
pp.81–173, 2010.
-  X. Su and T. M. Khoshgoftaar.
A Survey of Collaborative Filtering Techniques.
Advances in Artificial Intelligence, Article ID.421425, 2009.
-  P. Lops, M. Gemmis, and G. Semeraro.
Content-based Recommender Systems: State of the Art and Trends,
Recommender Systems Handbook.
Springer, 2010.
-  F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor.
Recommender Systems Handbook.
Springer, 2011.

Tài liệu tham khảo (2)



M. Deshpande and G. Karypis.

Item-Based Top- N Recommendation Algorithms.

ACM Transactions on Information Systems, Vol.22, No.1, pp.143–177, 2004.



G. Linden, B. Smith, and J. York.

Amazon.com Recommendation: Item-to-Item Collaborative Filtering.
IEEE Internet Computing, Vol.7, No.1, pp.76–80, 2003.



G. Adomavicius and A. Tuzhilin.

Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions.

IEEE Transactions on Knowledge and Data Engineering, Vol.17, No.6, pp.735–749, 2005.