



ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BÀI 8: HỆ GỢI Ý

Nội dung

1. Tổng quan về hệ gợi ý
2. Các phương pháp đánh giá
3. Lọc cộng tác dựa trên kNN
4. Lọc cộng tác dựa trên MF
5. NCF
6. Gợi ý theo phiên

1. Tổng quan về hệ gợi ý

Tại sao cần hệ gợi ý

- Người dùng bị quá tải thông tin trong môi trường web
- Nhà bán hàng cần đưa ra sản phẩm phù hợp để
 - Tăng doanh số bán hàng
 - Nâng cao chất lượng dịch vụ
- Xu hướng cá nhân hóa và số hóa là tất yếu

Hệ gợi ý vs hệ tìm kiếm

- Hệ tìm kiếm: Người dùng thể hiện mong muốn thông qua câu truy vấn
- Hệ gợi ý: Người dùng chưa biết mình muốn gì

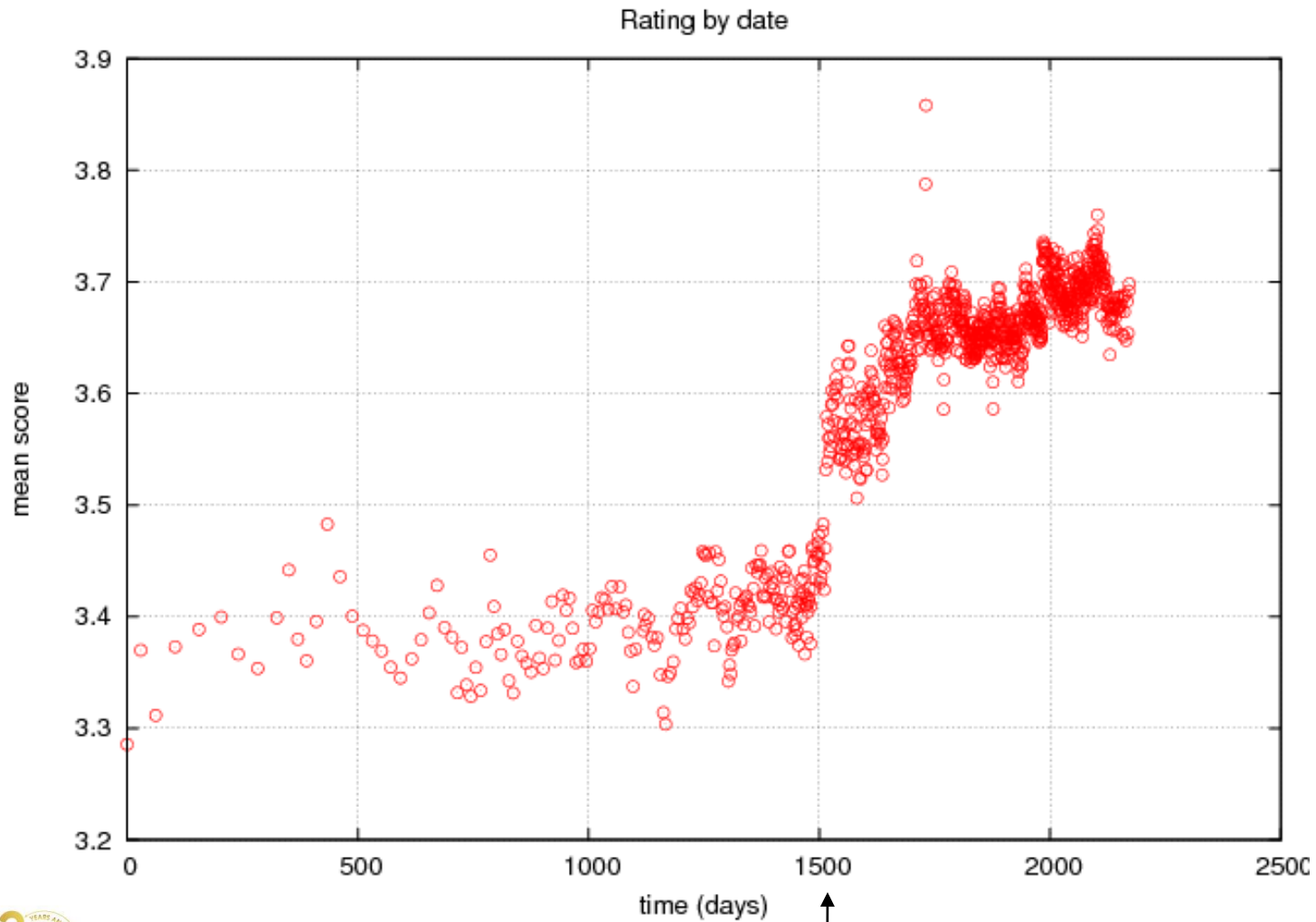
Lĩnh vực ứng dụng

- Thương mại điện tử
- Giải trí trực tuyến
- Tin tức trực tuyến
- Forum, mạng xã hội
- Nghiên cứu khoa học
- Hẹn hò trực tuyến

Lĩnh vực ứng dụng (tiếp)

- *Amazon:*
 - Gợi ý sản phẩm
 - Tăng hơn 30% doanh thu
- *Netflix:*
 - Gợi ý phim, chương trình TV
 - Mang về \$1B mỗi năm
- *Google News:*
 - Gợi ý tin tức
 - Tăng gần 40% lưu lượng truy cập

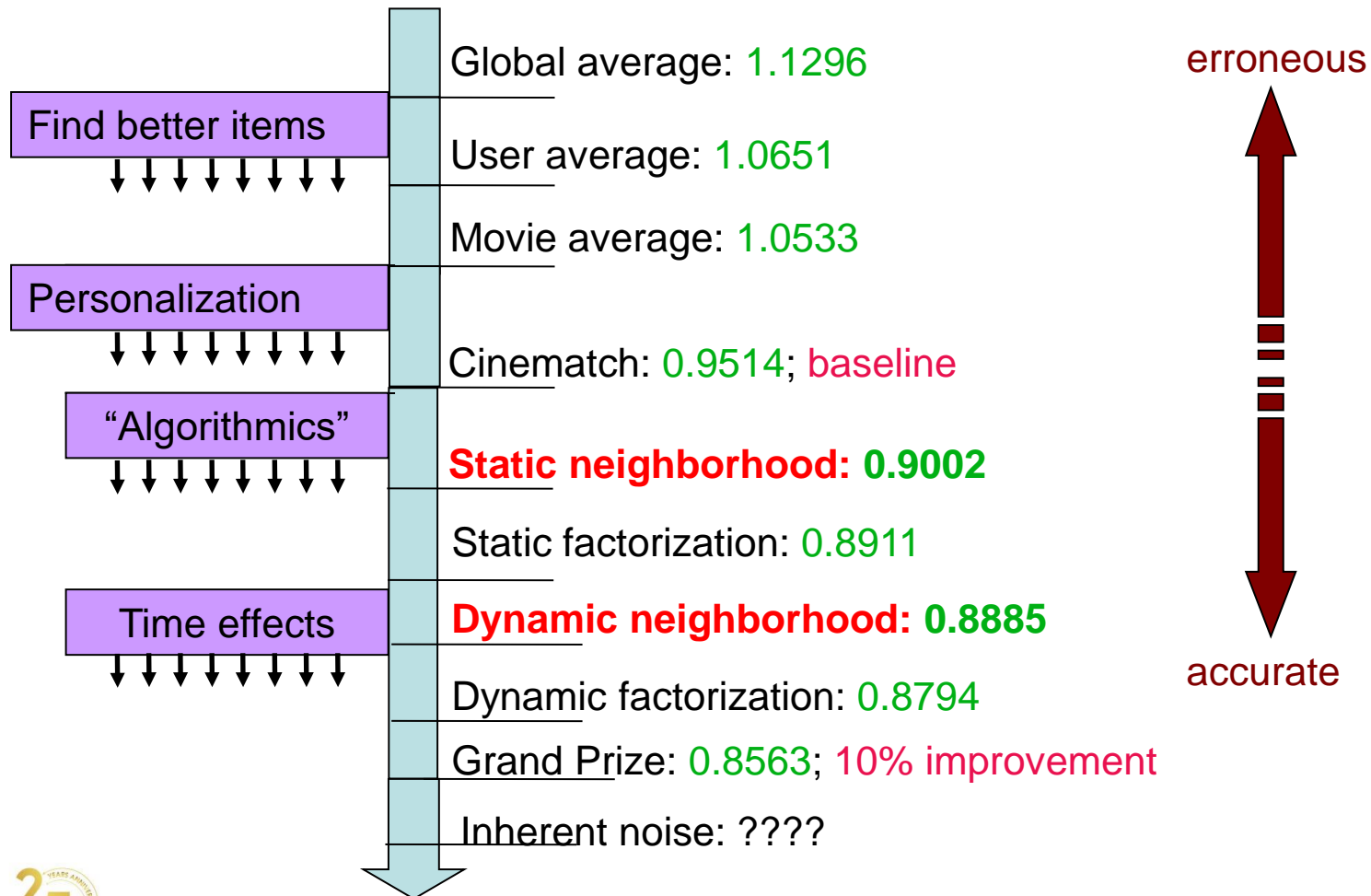
Lĩnh vực ứng dụng (tiếp)



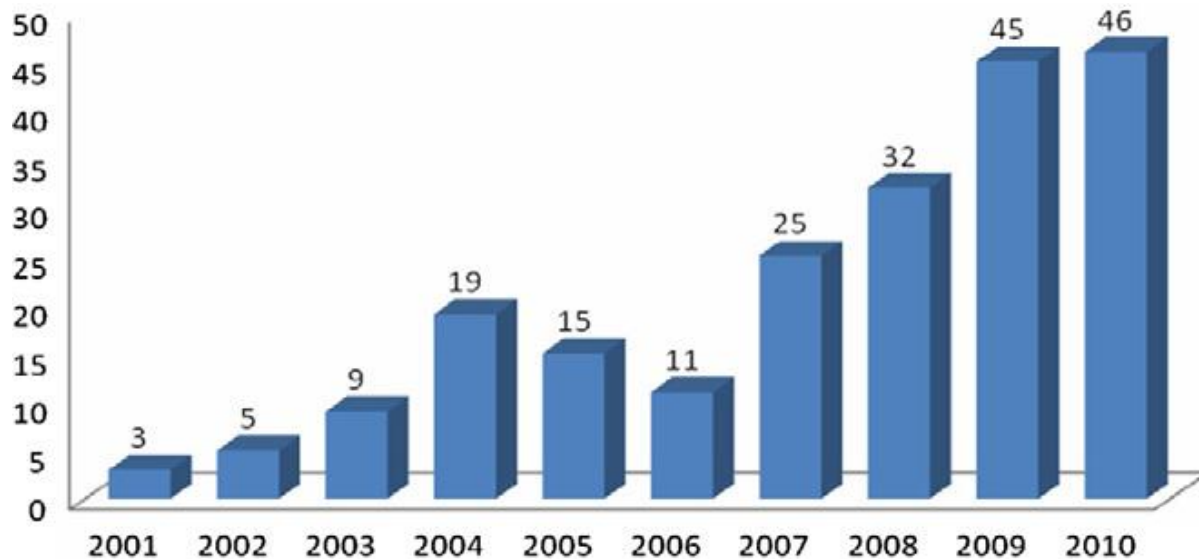
Các phương pháp gợi ý

- Gợi ý dựa trên nội dung:
 - Gợi ý dựa trên lịch sử giao dịch của người dùng
- Lọc cộng tác:
 - Gợi ý dựa trên người dùng có sở thích tương tự
- Gợi ý dựa trên phiên:
 - Gợi ý dựa trên chuỗi giao dịch
- Các phương pháp lai

Cuộc thi Netflix

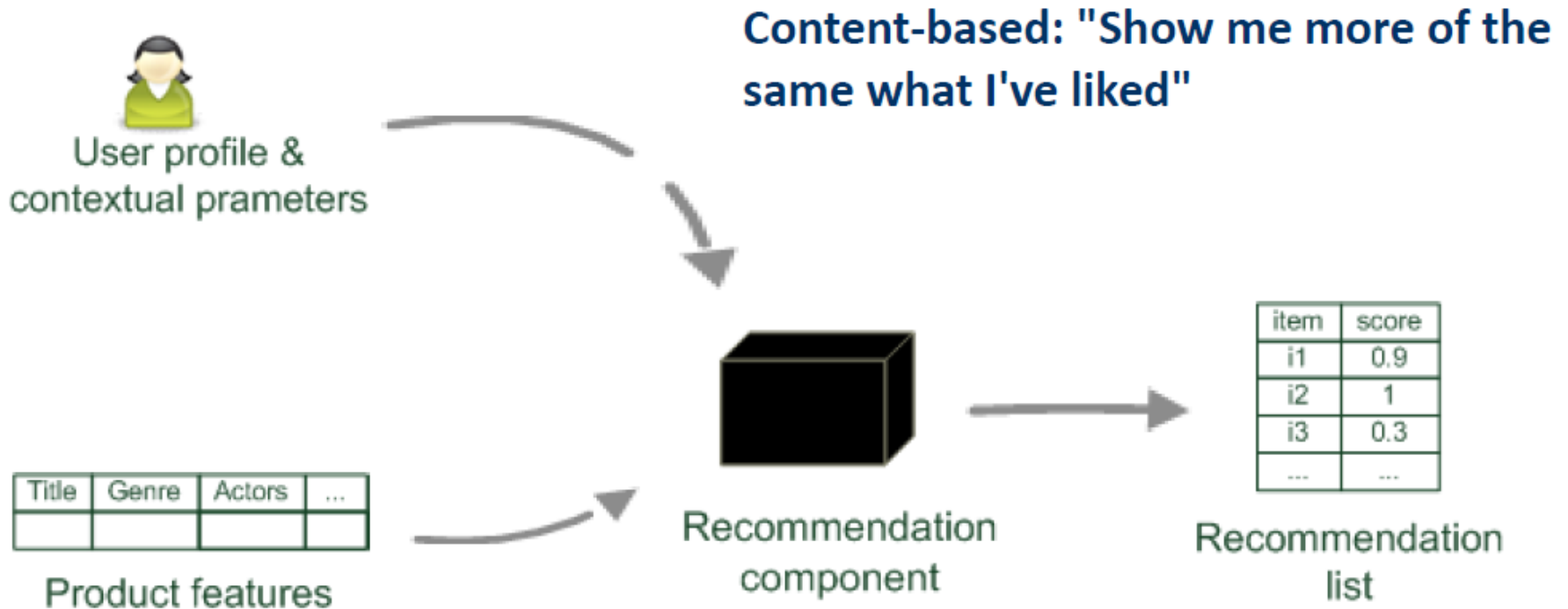


Cuộc thi Netflix (tiếp)

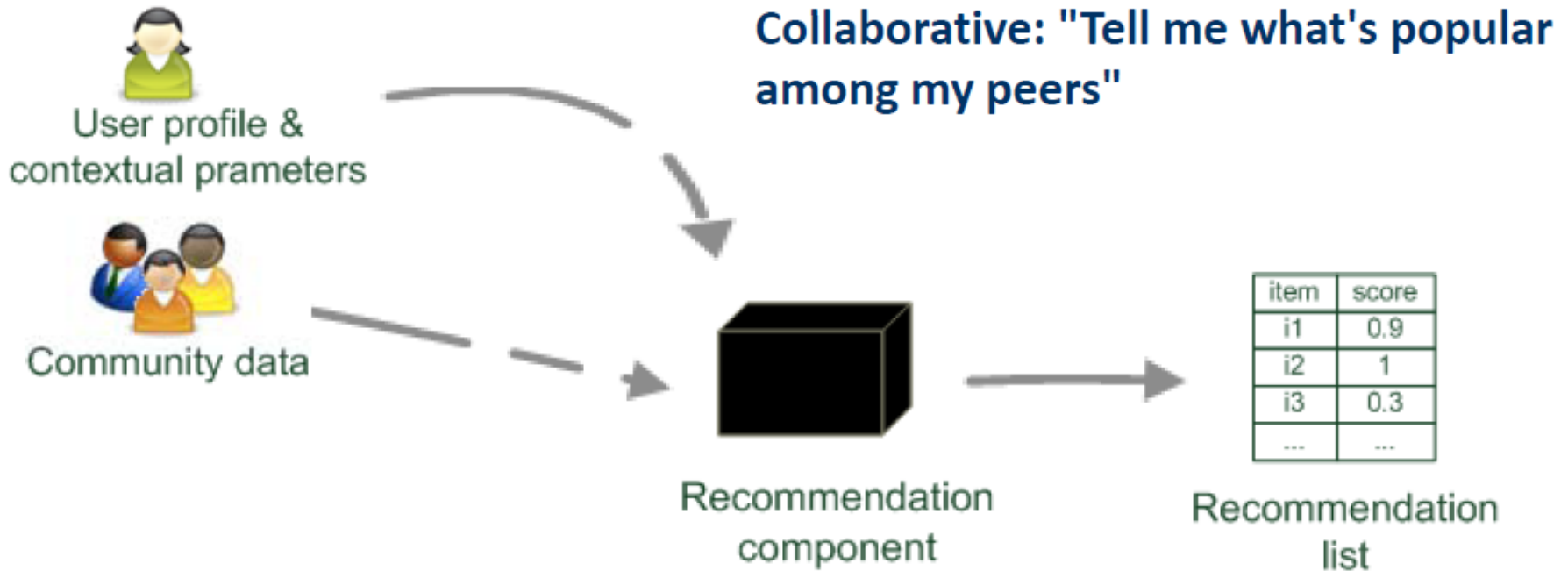


Number of papers on recsys by years

Gợi ý dựa trên nội dung



Lọc cộng tác



Những thách thức của hệ gợi ý

- Số giao dịch rất nhỏ so với số lượng người dùng và sản phẩm thực tế
- Không đủ thông tin về người dùng và sản phẩm mới
- Người dùng và sản phẩm thay đổi theo thời gian, theo mùa
- Thói quen tiêu dùng thay đổi theo thời gian, theo mùa
- Gợi ý theo thời gian thực

2. Các phương pháp đánh giá


- Cho
 - Tập người dùng U
 - Tập sản phẩm I
- DL gồm các giao dịch (u, i, r_{ui}, t)
 - u : người dùng $u \in U$
 - i : sản phẩm $i \in I$
 - r_{ui} : đánh giá của người dùng u đối với sản phẩm i
 - t : thời gian đánh giá

Các p_2 đánh giá (tiếp)


- r_{ui}
 - Theo thang đo 5 bậc (1, 2, 3, 4, 5)
 - Theo thang đo nhị phân (0, 1)
- DL được chia làm các tập train/test
- Hệ gợi ý được huấn luyện trên tập train
- Trên tập test, hệ gợi ý dự đoán đánh giá p_{ui} của người dùng u với sản phẩm i

train/test

	i_1	i_2	i_3	i_4
u_1	-	5	3	-
u_2	4	-	2	3
u_3	4	1	-	5



	i_1	i_2	i_3	i_4
u_1	-	5		-
u_2		-	2	3
u_3	4	1	-	



	i_1	i_2	i_3	i_4
u_1	-		3	-
u_2	4	-		
u_3			-	5

Các độ đo đánh giá

- (N)MAE
- RMSE
- Xếp hạng:
 - Precision/Recall/F-score

MAE

$$MAE = \frac{\sum_{ui} |p_{ui} - r_{ui}|}{n}$$

- p_{ui} : Dự đoán của mô hình đối với đánh giá của người dùng u với sản phẩm i
- r_{ui} : Đánh giá của người dùng u đối với sản phẩm i
- n : Tổng số ví dụ trong tập test

NMAE

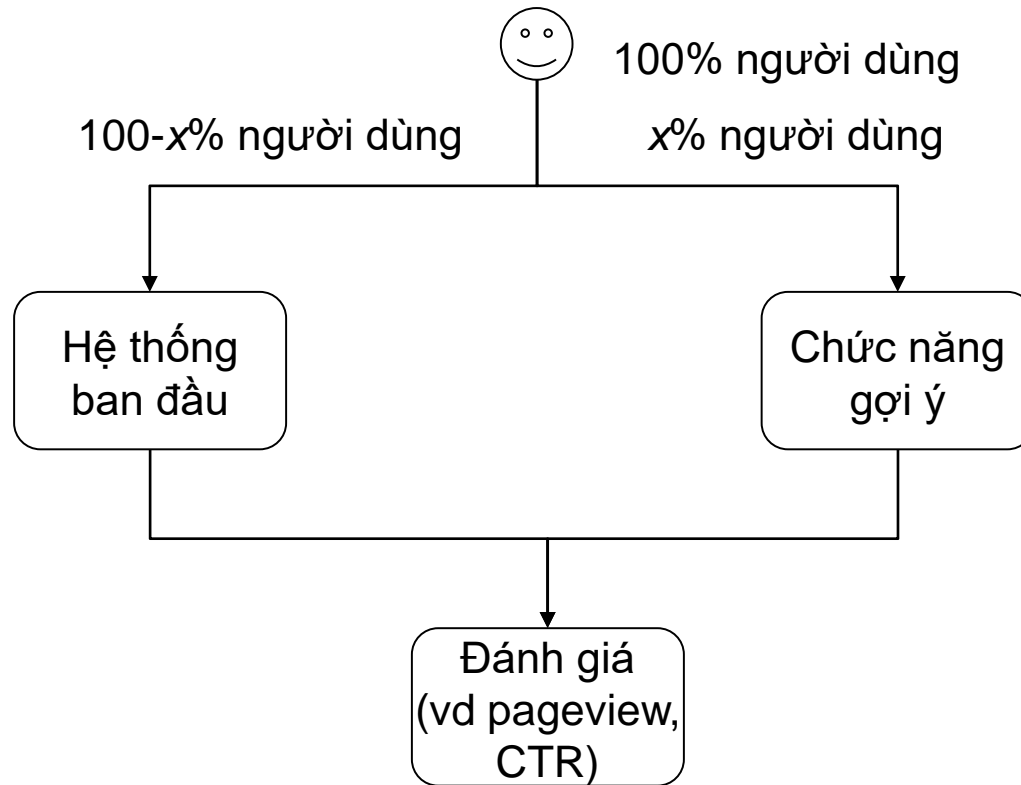
$$NMAE = \frac{MAE}{r_{max} - r_{min}}$$

- r_{max} : Giá trị dự đoán lớn nhất của người dùng
- r_{min} : Giá trị dự đoán bé nhất của người dùng

RMSE

$$RMSE = \sqrt{\frac{1}{n} \sum_{ui} (p_{ui} - r_{ui})^2}$$

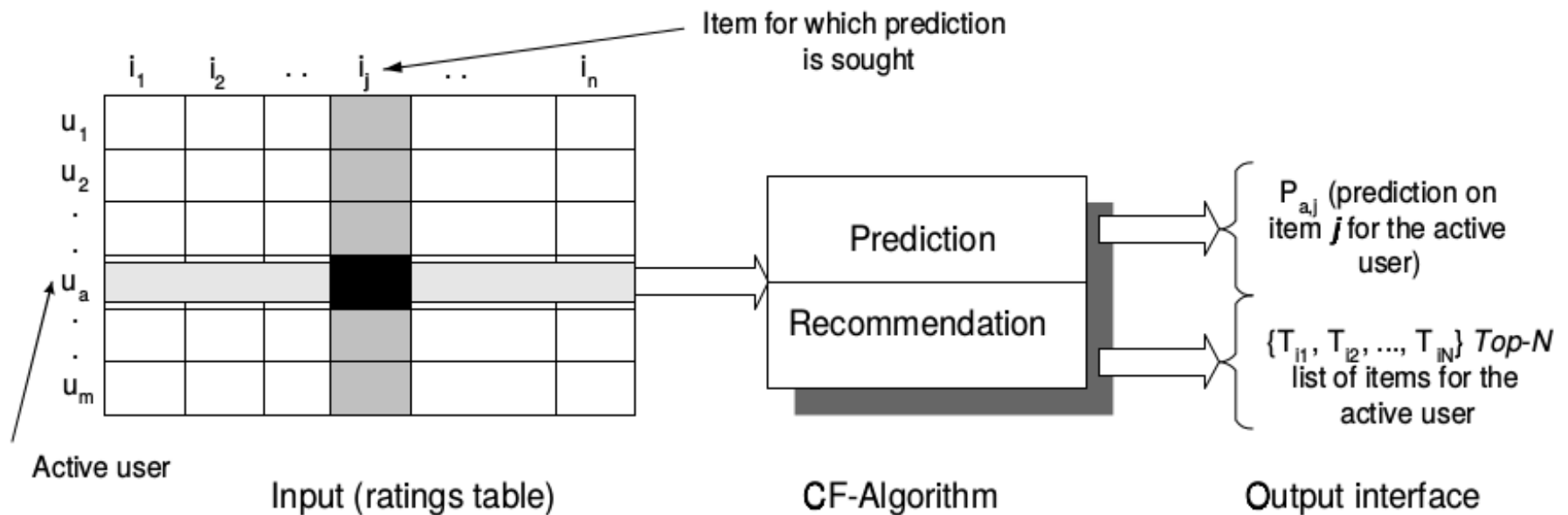
A/B testing



3. Lọc cộng tác dựa trên kNN

- Dựa trên ma trận tương tác người dùng sản phẩm
- Không có quá trình huấn luyện
- Gợi ý dựa trên người dùng
 - Tìm tập V bao gồm những người dùng tương tự với người dùng u
 - Tính toán đánh giá với sản phẩm i dựa trên đánh giá của những người dùng trong tập V đối với i

Gợi ý dựa trên người dùng



Độ tương đồng người dùng

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\sum_{i \in C} (r_{ui} - \bar{r}_{\mathbf{u}})(r_{vi} - \bar{r}_{\mathbf{v}})}{\sqrt{\sum_{i \in C} (r_{ui} - \bar{r}_{\mathbf{u}})^2} \sqrt{\sum_{i \in C} (r_{vi} - \bar{r}_{\mathbf{v}})^2}}$$

- C : Tập những sản phẩm mà người dùng u và v đều đánh giá
- r_u : Đánh giá trung bình của người dùng u (chỉ tính trên các sản phẩm mà người dùng u đánh giá)

Dự đoán đánh giá

$$p_{ui} = \bar{r}_{\mathbf{u}} + \frac{\sum_{\mathbf{v} \in V} \text{sim}(\mathbf{u}, \mathbf{v})(r_{vi} - \bar{r}_{\mathbf{v}})}{\sum_{\mathbf{v} \in V} |\text{sim}(\mathbf{u}, \mathbf{v})|}$$

- V : top k người dùng tương tự với người dùng u

VD ($k = 2$)

	i_1	i_2	i_3	i_4
u_1	5	4	4	1
u_2	2	1		
u_3	5	4	4	?
u_4		1	2	5

$$r_1 = 14/4$$

$$r_2 = 3/2$$

$$r_3 = 13/3$$

$$r_4 = 8/3$$

$$\text{sim}(u_3, u_1) = \sim 0.492$$

$$\text{sim}(u_3, u_2) = \sim 0.948$$

$$\text{sim}(u_3, u_4) = \sim 0.919$$

$$p(u_3, i_4) = \sim 6.67$$

Nhược điểm

- Thường xuyên phải cập nhật lại vector người dùng khi người dùng có giao dịch mới
- Phải tính toán trên toàn bộ tập người dùng

Gợi ý dựa trên sản phẩm

- Biểu diễn sản phẩm dựa trên ma trận tương tác người dùng - sản phẩm
- Phù hợp với hệ thống có số sản phẩm \ll số người dùng
- Ít phải cập nhật lại véc-tơ sản phẩm
- Có thể tính toán trước độ tương tự sản phẩm - sản phẩm

Biểu diễn sản phẩm

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

Item-item similarity is computed by looking into co-rated items only. In case of items i and j the similarity s_{ij} is computed by looking into them. Note: each of these co-rated pairs are obtained from different users, in this example they come from users $1, u$ and $m-1$.

Độ tương đồng sản phẩm

$$\text{sim}(i, j) = \frac{\sum_{\mathbf{u} \in U} (r_{ui} - \bar{r}_{\mathbf{u}})(r_{uj} - \bar{r}_{\mathbf{u}})}{\sqrt{\sum_{\mathbf{u} \in U} (r_{ui} - \bar{r}_{\mathbf{u}})^2} \sqrt{\sum_{\mathbf{u} \in U} (r_{uj} - \bar{r}_{\mathbf{u}})^2}}$$

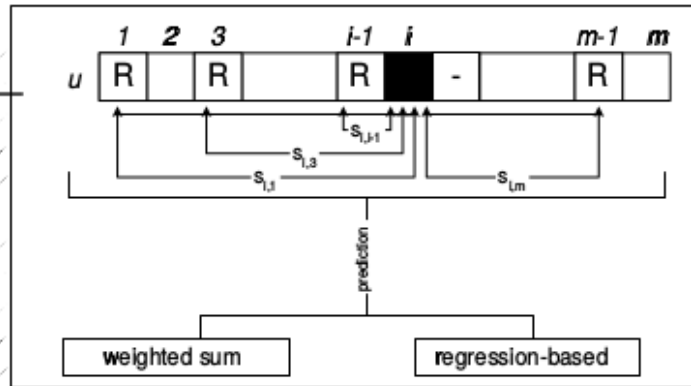
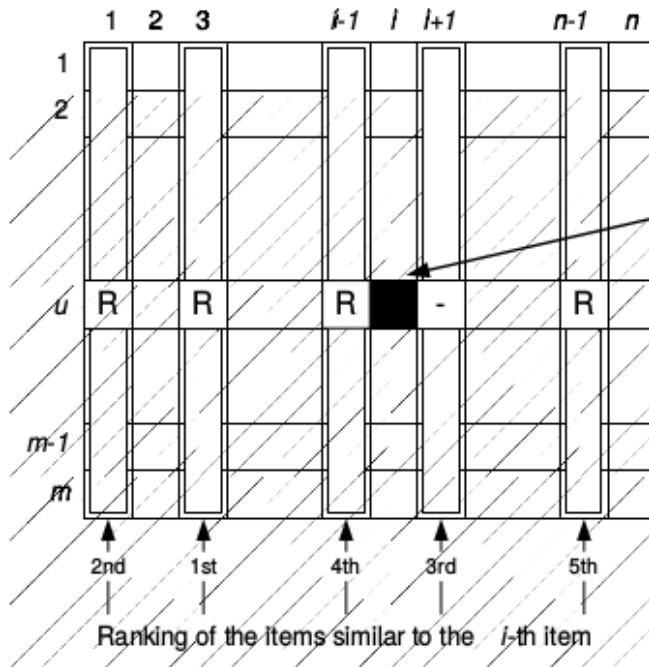
- U : tập người dùng cùng đánh giá sản phẩm i và sản phẩm j

Dự đoán đánh giá

$$p(\mathbf{u}, i) = \frac{\sum_{j \in J} r_{\mathbf{u}, j} \times \text{sim}(i, j)}{\sum_{j \in J} \text{sim}(i, j)},$$

J : top k sản phẩm tương tự với sản phẩm i

Ví dụ



4. Lọc cộng tác dựa trên MF

- Dựa trên ngữ cảnh tổng thể của tương tác người dùng - sản phẩm
- Biểu diễn người dùng, sản phẩm theo các khía cạnh ẩn
- Sử dụng kỹ thuật MF (phân rã ma trận)
- Dựa trên véc-tơ người dùng, sản phẩm để đưa ra dự đoán

SVD

- Ma trận người dùng - phim \mathbf{R}
- Phân rã \mathbf{R} thành ma trận người dùng - khía cạnh \mathbf{U} và ma trận phim - khía cạnh \mathbf{M}
- Mỗi người dùng được biểu diễn bởi một véc-tơ K chiều, trong đó K là số khía cạnh ẩn
- Mỗi phim được biểu diễn bởi một véc-tơ K chiều

$$\mathbf{R} \approx \mathbf{U}^T \mathbf{M} . \quad r_{ij} \approx \mathbf{u}_i^T \mathbf{m}_j = \sum_{k=1}^K u_{ki} \times m_{kj} ,$$

Dự đoán đánh giá

$$p_{ij} = \sum_{k=1}^K u_{ki} \times m_{kj}.$$

- p_{ij} : Dự đoán đánh giá của người dùng i với phim j
- u_i : Véc-tơ người dùng i
- m_j : Véc-tơ phim j

Hàm lỗi: e_{ij}^2

$$e_{ij} = r_{ij} - p_{ij}.$$

Học mô hình

- U, M là các tham số cần học
- R là tập DL huấn luyện
- Hàm bình phương lỗi: $E_{ij} = e_{ij}^2 = (p_{ij} - r_{ij})^2$
- Kỹ thuật học: Suy giảm gradient

Suy giảm gradient

Đạo hàm của hàm lỗi theo u_{ki}

$$\frac{\partial (e_{ij})^2}{\partial u_{ki}} = 2e_{ij} \frac{\partial e_{ij}}{\partial u_{ki}}.$$

$$\frac{\partial e_{ij}}{\partial u_{ki}} = -\frac{\partial p_{ij}}{\partial u_{ki}}.$$

$$\frac{\partial (e_{ij})^2}{\partial u_{ki}} = 2e_{ij} (-m_{kj}) = -2(r_{ij} - p_{ij})m_{kj}.$$

Suy giảm gradient (tiếp)

Đạo hàm của hàm lỗi theo m_{kj}

$$\frac{\partial (e_{ij})^2}{\partial m_{kj}} = 2e_{ij}(-u_{ki}) = -2(r_{ij} - p_{ij})u_{ki}.$$

$$u_{ki}^{t+1} = u_{ki}^t - \gamma \frac{\partial (e_{ij})^2}{\partial u_{ki}} = u_{ki}^t + 2\gamma(r_{ij} - p_{ij})m_{kj}^t.$$

$$u_{ki}^{t+1} = u_{ki}^t + 2\gamma(r_{ij} - p_{ij})m_{kj}^t,$$

$$m_{kj}^{t+1} = m_{kj}^t + 2\gamma(r_{ij} - p_{ij})u_{ki}^t.$$

Suy giảm gradient (tiếp)

Cập nhật u_{ki} theo đạo hàm:

$$u_{ki}^{t+1} = u_{ki}^t + \gamma(2(r_{ij} - p_{ij})m_{kj}^t - \lambda u_{ki}^t),$$

Cập nhật m_{kj} theo đạo hàm:

$$m_{kj}^{t+1} = m_{kj}^t + \gamma(2(r_{ij} - p_{ij})u_{ki}^t - \lambda m_{kj}^t).$$

λ : Giá trị chuẩn hóa

γ : Tốc độ cập nhật

5. NCF (Neural CF)

- Biểu diễn người dùng, sản phẩm theo một mức khía cạnh ẩn có thể không thể hiện được hết tính chất phức tạp của tương tác người dùng - sản phẩm
- Mạng nơ-ron tiến nhiều tầng cho phép tự động học ra các mức khía cạnh ẩn từ đơn giản đến trừu tượng
- MF là trường hợp đơn giản nhất của mạng nơ-ron

Giới hạn của MF

- MF bảo toàn tính chất tương tự của véc-tơ người dùng
- Giả sử độ tương tự của người dùng được tính theo hệ số Jaccard
- Véc-tơ người dùng và véc-tơ sản phẩm cùng được biểu diễn trên không gian khóa cạnh ẩn K chiều

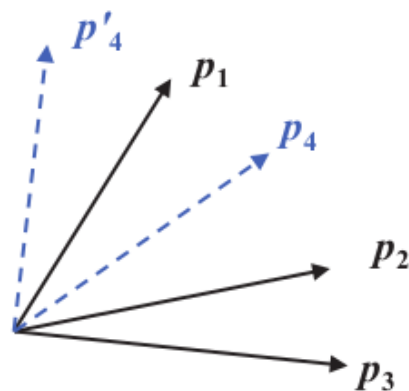
Giới hạn của MF (tiếp)

- $s_{23} (0.66) > s_{12} (0.5) > s_{13} (0.4)$
- $s_{41} (0.6) > s_{43} (0.4) > s_{42} (0.2)$

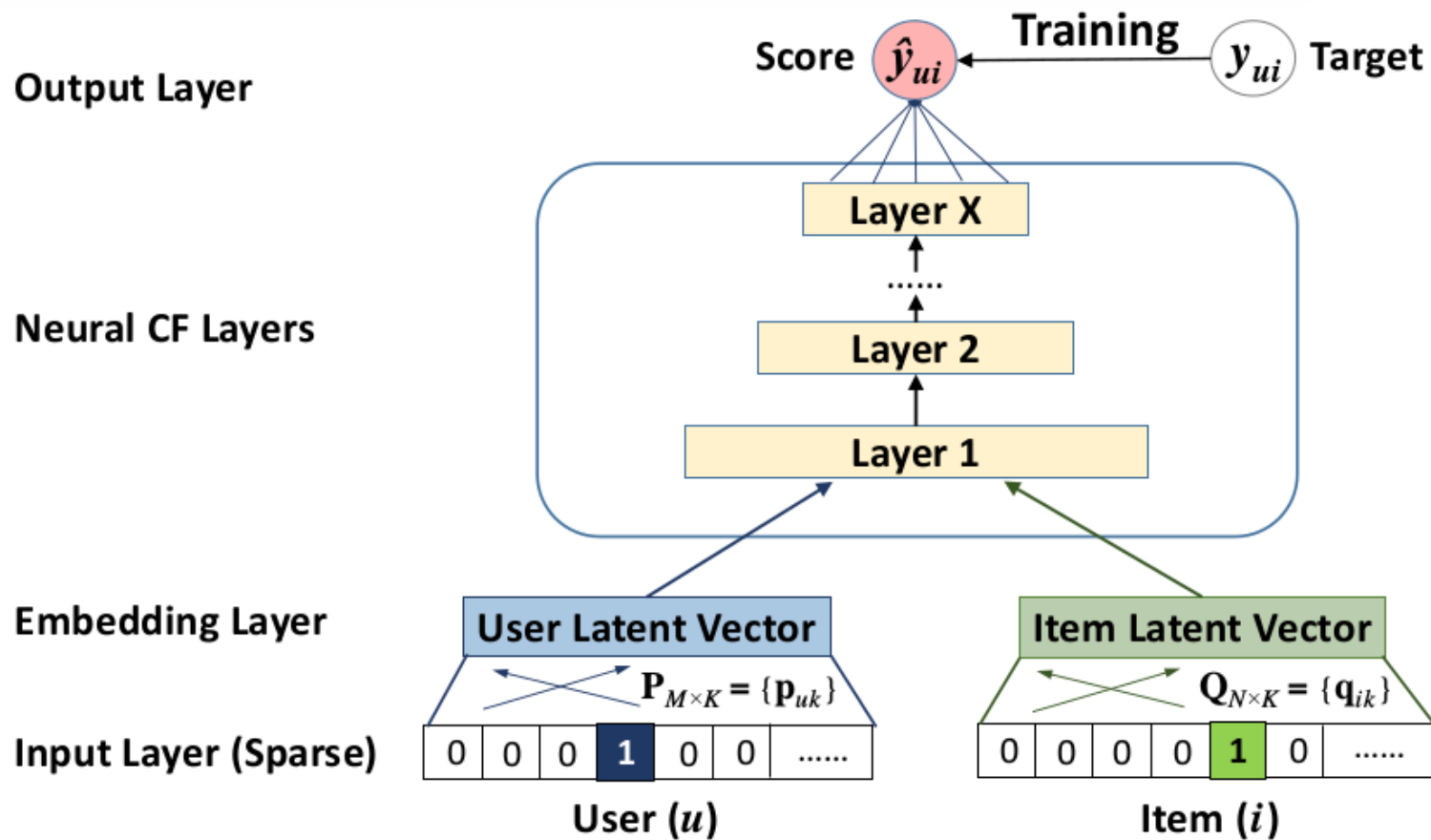
	i_1	i_2	i_3	i_4	i_5
u_1	1	1	1	0	1
u_2	0	1	1	0	0
u_3	0	1	1	1	0
u_4	1	0	1	1	1

← items →

↑ users



Kiến trúc NCF



Tầng đầu vào

- Người dùng được biểu diễn bởi véc-tơ one-hot gồm M chiều
 - M là số người dùng
 - Mỗi người dùng u có một giá trị 1 tương ứng, các giá trị còn lại đều bằng 0
- Sản phẩm được biểu diễn bởi véc-tơ one-hot gồm N chiều, trong đó N là số sản phẩm

Tầng nhúng

- Biểu diễn người dùng và sản phẩm độc lập
- $M \times K$ trọng số liên kết để biểu diễn người dùng
- $N \times K$ trọng số liên kết để biểu diễn sản phẩm
- Mô hình thử nghiệm: $K = 16$

MLP

- Biểu diễn người dùng và sản phẩm được ghép nối để đưa vào một mạng MLP nhiều tầng nhằm học được các tương tác phức tạp của người dùng và sản phẩm
- Mô hình thử nghiệm:
 - Hàm kích hoạt *ReLU*
 - 3 tầng với kích thước giảm dần $32 \rightarrow 16 \rightarrow 8$

Hàm lỗi

- Phản hồi ẩn:
 - 1: Người dùng tương tác với sản phẩm
 - 0: Người dùng không tương tác với sản phẩm
- Bài toán phân loại hai lớp có yếu tố xác suất
- Sử dụng hàm lỗi binary cross entropy

$$- \sum_{(u,i) \in \mathcal{Y} \cup \mathcal{Y}^-} y_{ui} \log \hat{y}_{ui} + (1 - y_{ui}) \log(1 - \hat{y}_{ui}).$$

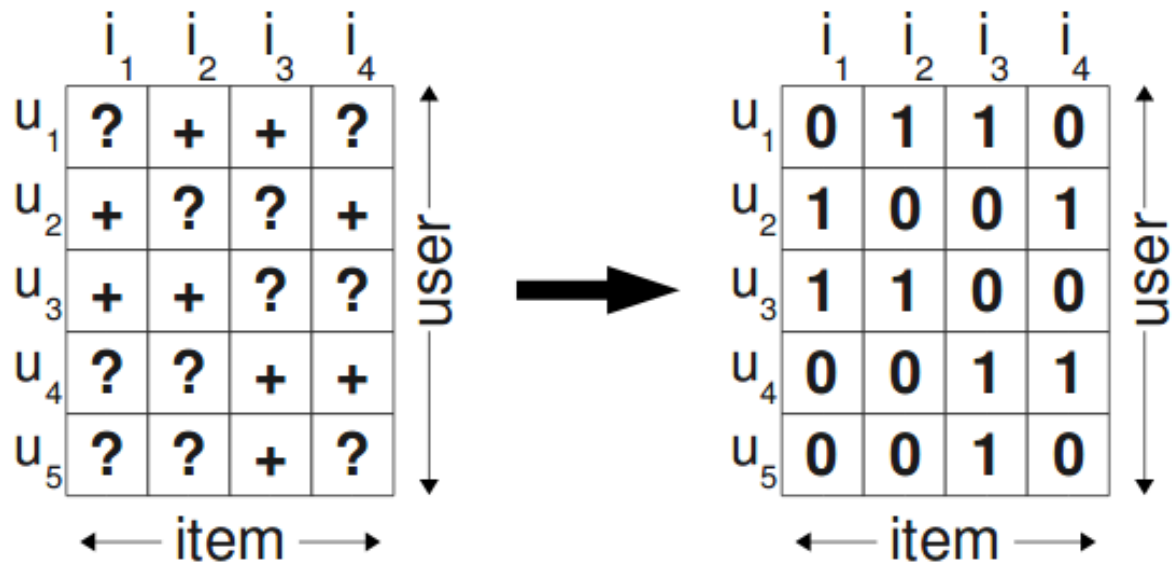
6. Gợi ý theo phiên

- Trong nhiều trường hợp, khó có thể định danh người dùng và thu thập đánh giá. VD:
 - Các website thương mại nhỏ
 - Các trang tin tức
- Kỹ thuật gợi ý theo phiên
 - Không yêu cầu định danh người dùng
 - Mỗi phiên giao dịch, bao gồm thứ tự giao dịch được sử dụng làm dữ liệu huấn luyện mô hình

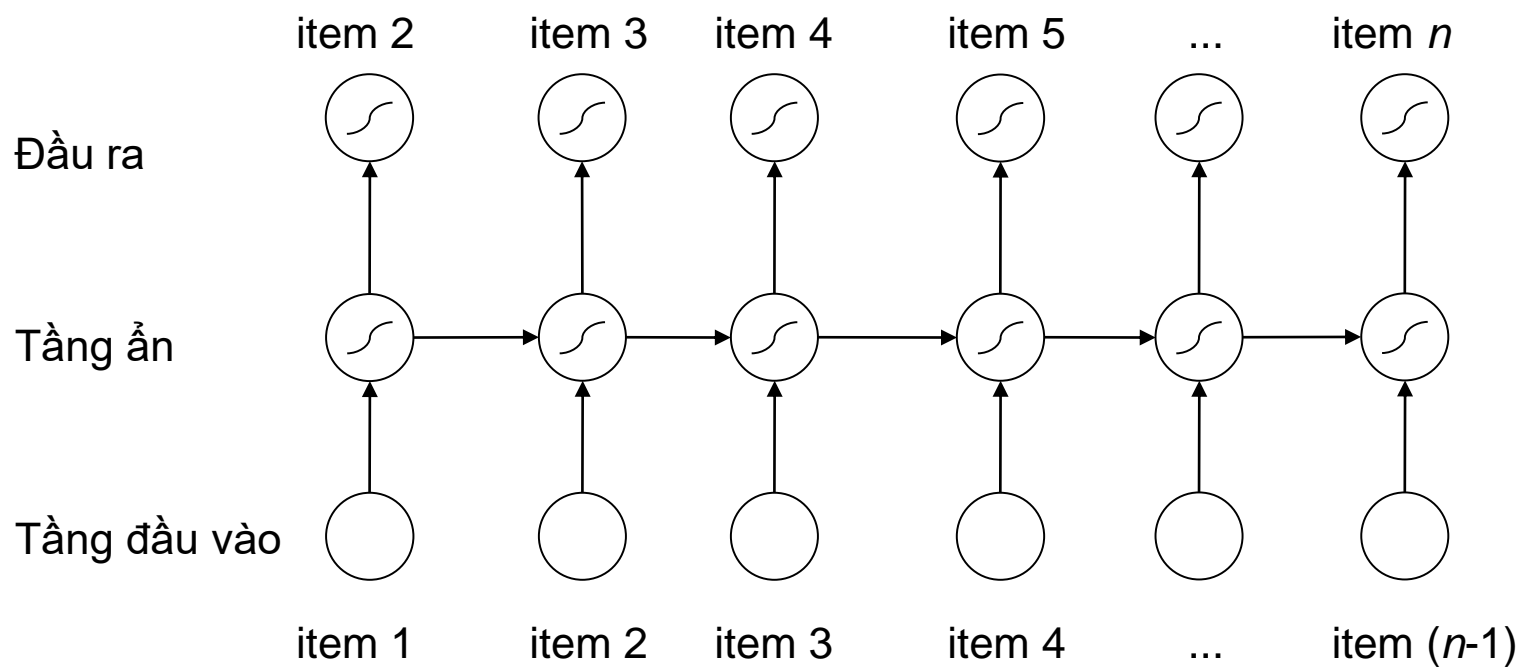
Phát biểu bài toán

- Đầu vào: Chuỗi các sản phẩm i_1, i_2, \dots, i_{n-1}
- Đầu ra: Gợi ý sản phẩm tiếp theo
- Đưa ra danh sách các sản phẩm có xác suất lớn nhất $\Pr(i_n | i_1, i_2, \dots, i_{n-1})$
- Phù hợp với phản hồi ẩn

Phản hồi ẩn



Kiến trúc RNN



Tầng đầu vào

- Biểu diễn mỗi sản phẩm tại một thời điểm t dưới dạng one-hot
 - Giá trị 1 ứng với vị trí của sản phẩm trong tập từ vựng, các giá trị khác = 0
 - V – từ vựng: Số sản phẩm khác nhau

Tầng nhúng

- Biến đổi biểu diễn one-hot thành biểu diễn K chiều
 - K là số nơ-ron của tầng nhúng
 - Số trọng số liên kết giữa tầng đầu vào và tầng nhúng $V \times K$

Tầng hồi quy

- Tầng hồi quy lưu trữ các thông tin quá khứ thông qua các liên kết hồi quy
- Nhiều tầng hồi quy có thể được “chồng” (stack) lên nhau để học được các đặc trưng trừu tượng mức cao
- Thay vì nhân cơ bản, có thể dùng các nhân LSTM hoặc GRU

Tầng MLP

- Đầu ra của tầng hồi quy được dùng làm vào đầu vào của tầng MLP để sinh ra dự đoán
- Tầng MLP có thể bao gồm các tầng ẩn để học được hàm phi tuyến
- Tầng đầu ra của MLP có V nơ-ron ứng với V sản phẩm

Hàm mất mát

- Hàm mất mát thứ hạng theo cặp

$$L = -\frac{1}{N} \sum_{j=1}^N \log(\sigma(\hat{r}_i - \hat{r}_j))$$

- N : số mẫu negative
- \hat{r}_i : Điểm của sản phẩm mong muốn positive i
- \hat{r}_j : Điểm của sản phẩm không mong muốn negative j
- Giả thiết phản hồi ẩn: Sản phẩm i được lựa chọn nên mức độ ưu tiên của i cao hơn các sản phẩm j còn lại



25 YEARS ANNIVERSARY
SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

**Thank you for
your attentions!**

