viettel
Theo cách của bạn

VHAC 2023 - Multi-domain Recommendation System
Next Content Recommendation
1st place solution

Team DHA

Ngô Hoàng Đăng - Trung tâm Không gian mạng Viettel
Nguyễn Huy Hoàng - Trung tâm Không gian mạng Viettel
Nguyễn Huy Anh - Trung tâm Không gian mạng Viettel

viettel cyberspace TRUNG TÂM KHÔNG GIAN MẠNG VIETTEL

viettel
cyberspace

PHẦN 1

XU HƯỚNG HỆ THỐNG ĐỀ XUẤT TRÊN THẾ GIỚI

1.1 Tổng quan về hệ thống đề xuất

Hệ thống đề xuất là một ứng dụng lọc thông tin được hình thành từ việc học các sở thích và tương tác của người dùng nhằm có thể đề xuất nội dung phù hợp tới người dùng.

Mục tiêu chính của hệ thống đề xuất là **cung cấp những gợi ý cá nhân hóa và tối ưu hóa trải nghiệm người dùng,** giúp họ **khám phá được nội dung, sản phẩm hoặc thông tin mới** mà họ có thể chưa biết hoặc chưa nhìn thấy.

viettel
cyberspace

1.2 Lợi ích của hệ thống đề xuất

NETFLIX

- Hệ thống đề xuất mang đến **80% lượt xem**
- Hệ thống đề xuất **tiết kiệm 1 tỷ đô la mỗi năm**

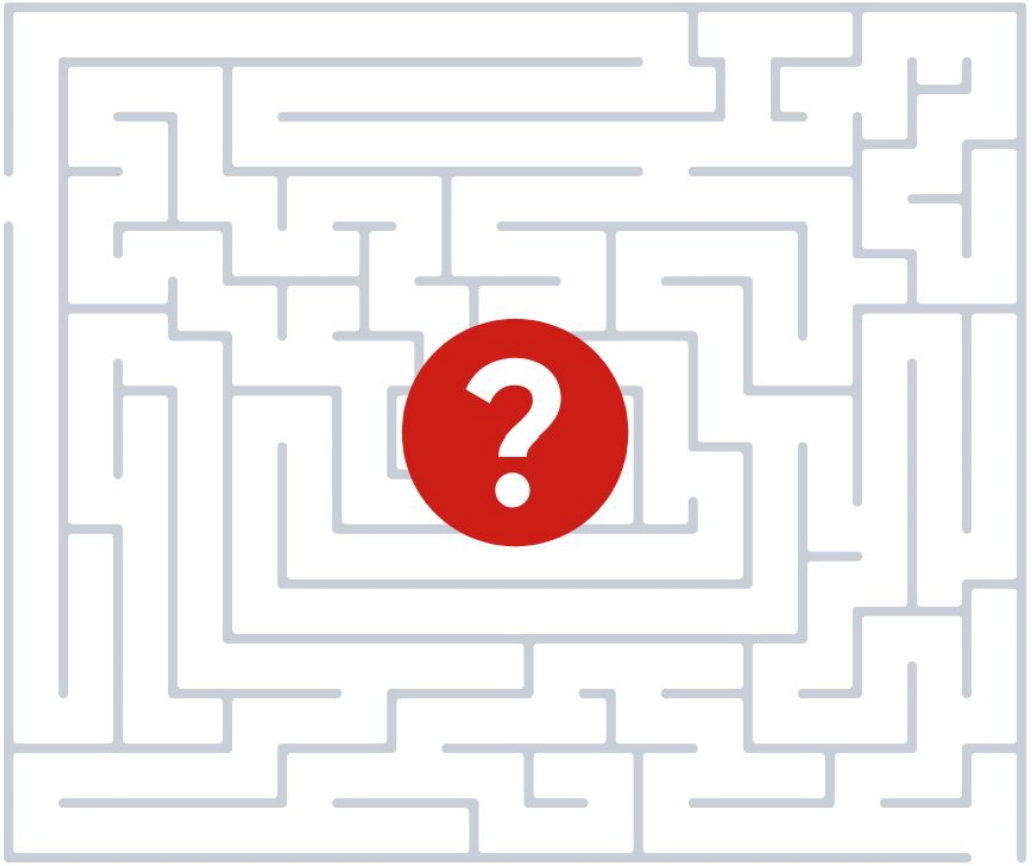
viettel
cyberspace

1.3 Các vấn đề liên quan đến hệ thống đề xuất

Vấn đề về chất lượng dữ liệu gặp phải khi vận hành hệ thống đề xuất hay gặp tại track Multidomain

Vấn đề phụ thuộc dữ liệu: Hệ thống đề xuất yêu cầu dữ liệu đầy đủ và chất lượng để có thể đưa ra gợi ý chính xác. Nếu dữ liệu đầu vào không đủ hoặc không đại diện cho sự đa dạng của người dùng, hệ thống có thể gặp khó khăn trong việc đưa ra gợi ý phù hợp, bao gồm:

1) **Cold-start problem:** Khi hệ thống đề xuất gặp người dùng mới sản phẩm mới hoặc hệ thống mới, không có đủ thông tin để xây dựng một mô hình gợi ý đáng tin cậy. Điều này có thể dẫn đến việc hiển thị gợi ý không chính xác hoặc không đủ phù hợp cho người dùng mới.
2) **Vấn đề mất cân bằng (bias):** Hệ thống đề xuất có thể chịu ảnh hưởng của các yếu tố định hướng hoặc thiên vị trong dữ liệu. Điều này có thể dẫn đến gợi ý bị mất cân bằng, bất công hoặc không công bằng cho một số nhóm người dùng.



viettel
cyberspace

1.4 Mục tiêu của track VHAC 2023 - Multi-domain Recommendation System

Thực tế: Hàng chục nghìn video, series film mới được đăng tải hàng ngày, việc xem hết các nội dung là **không thể**.

Bài toán: Xây dựng hệ thống khuyến nghị có khả năng dự đoán sản phẩm tương tác tiếp theo cho các phiên tương tác của người dùng với 2 domain là domain 1 (video) và domain 2 (series)

viettel
cyberspace

PHẦN
2

GIẢI PHÁP CỦA NHÓM

viettel
cyberspace

2.1 Phân tích khám phá dữ liệu

	Số sản phẩm	min published time	max published time
Domain 1 (video)	278,404	0	2023-07-31 23:35:37
Domain 2 (series)	5201	0	2023-09-23 06:00:00

Mẫu sample dữ liệu được cung cấp:
1_931450|2_18130|1_570697|2_17077|1_931483|2_15265|2_10402|1_320392|1_933439|2_18253|2_17977

	Số lượng chuỗi tương tác	Số sản phẩm tương tác thuộc domain 1	Số sản phẩm tương tác thuộc domain 2
Train	1,300,612	30783	2,993
Private test	147,735	40835	2,632

viettel
cyberspace

2.1 Phân tích khám phá dữ liệu

	Chuỗi tương tác trung bình		Chuỗi tương tác ngắn nhất		Chuỗi tương tác dài nhất	
	Domain 1	Domain 2	Domain 1	Domain 2	Domain 1	Domain 2
Train	9	9	0	0	485	344
Private test	21	11	0	0	247	184

	TOP 10 - DOMAIN 1	TOP 10 - DOMAIN 2
Train	930766, 931141, 925921, 929332, 697102, 869500, 932857, 934564, 891229, 704398	17881, 18199, 9040, 17077, 18196, 16207, 9103, 17684, 16438, 18193
Private test	982300, 784300, 988288, 990715, 1039835, 911879, 1043953, 1023391, 908417, 989515	19039, 17258, 19174, 20281, 9040, 18817, 19012, 11297, 19282, 19207

viettel
cyberspace

2.1 Phân tích khám phá dữ liệu

video_id: 930766 có mức độ trung tâm (centrality degree) rất cao (0,9) trong tập train trong khi với tập dữ liệu public test thì chỉ số mức độ trung tâm của video 930766 chỉ còn 0,3.
Trong tập public test, video_id có centrality degree cao nhất là video_id : 955807 (0.75). Khảo sát sâu hơn cho thấy video_id 930766 thì published_time là 14/01/2023, trong khi video_id: 955807 có published_time là 02/06/2023.
=> Có thể thấy tập dữ liệu bị lệch ~4 tháng và video_id 930766 có số lượng xem lớn vào 14/01/2023 trùng dịp với lễ tết. **Nhóm dự đoán là một video hài tết.**

=> Dữ liệu huấn luyện và các mô hình phức tạp về đậu sâu không áp dụng vào được trong **bài toán này vì data shift và hành vi người dùng…**

Do đó, nhóm thực hiện mô hình có tính chất tổng quan nhất.

viettel
cyberspace

2.2 Các thách thức của cuộc thi

Không có thông tin về đánh giá (ratings) giữa người dùng-sản phẩm: Khi thông tin đánh giá giữa người dùng - sản phẩm không có sẵn, các mô hình truyền thống như Matrix Factorization gặp khó khăn vì chúng dựa trên dữ liệu tương tác người dùng - sản phẩm để xây dựng hồ sơ người dùng hoặc sản phẩm.

Dữ liệu được cung cấp dưới dạng tuần tự (sequence) nhưng không cung cấp thông tin về thời điểm tương tác: Khi khảo sát dữ liệu nhóm nhận thấy phân bố dữ liệu rất khác nhau trong tập dữ liệu train và dữ liệu public-test. Ngoài ra, việc nối các tương tác tạo thành chuỗi không đảm bảo việc lấy trong một phiên, có thể được lấy ra từ nhiều phiên.

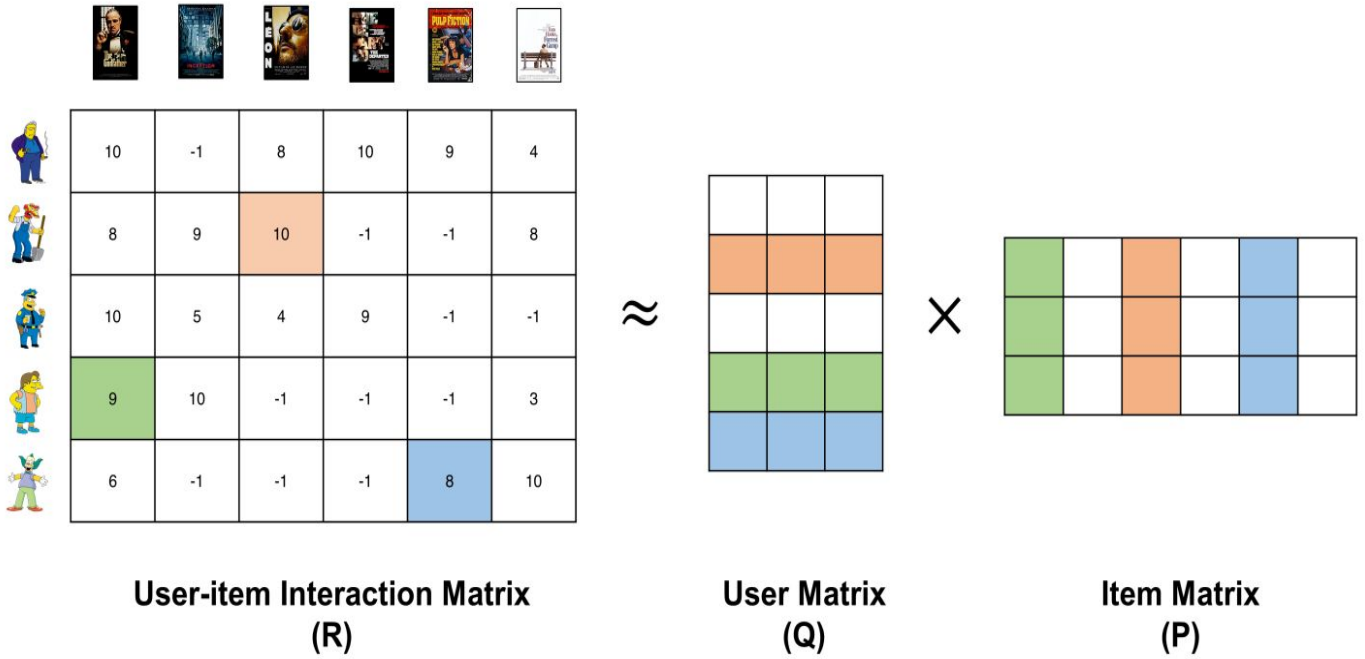
viettel
cyberspace

2.2 Các hướng tiếp cận với bài toán đề xuất

2.2.1 Hướng tiếp cận sử dụng phương pháp phân rã ma trận (Matrix factorization)

Không là phương pháp khả thi cho bài toán:

- No user-item interaction feedback (clicked, ratings, …)



viettel
cyberspace

2.2 Các hướng tiếp cận với bài toán đề xuất

2.2.2 Hướng tiếp cận sử dụng mạng nơ-ron đồ thị (Graph neural network)

Qua thực nghiệm, mạng nơ-ron đồ thị (Graph Neural Network) không là phương pháp khả thi cho bài toán vì:

- Phân bố dữ liệu rất khác nhau trong tập dữ liệu train và dữ liệu public-test, ~50% dữ liệu sản phẩm ở tập test không xuất hiện trong tập train.

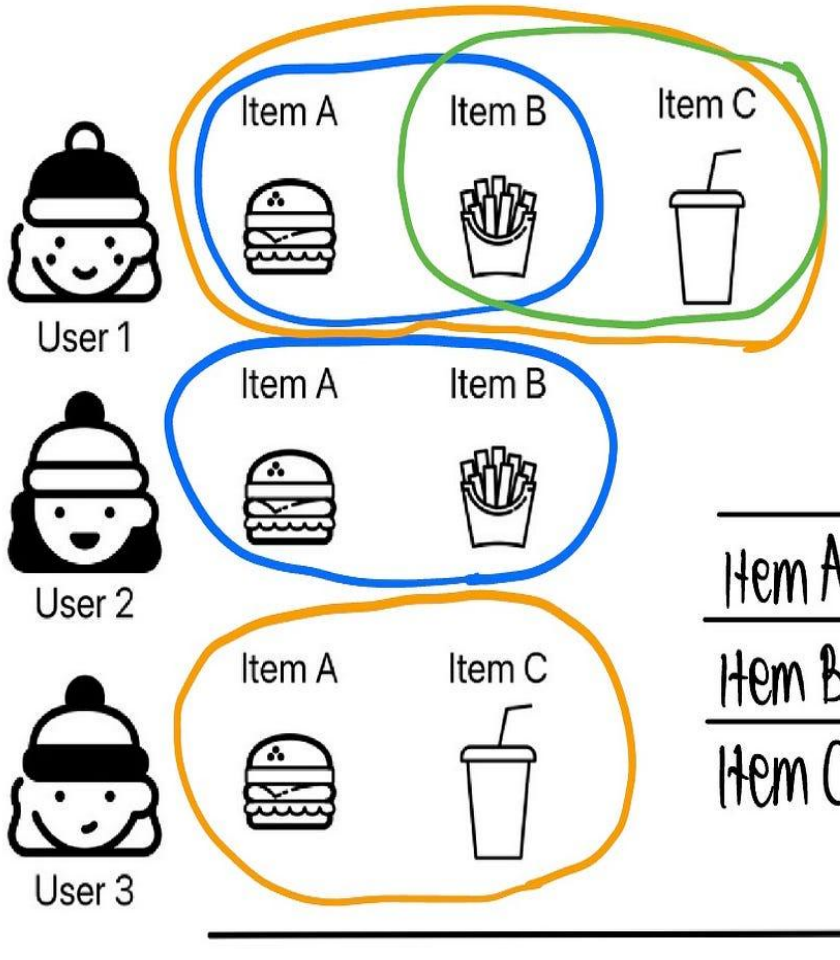
- Các tương tác tạo thành chuỗi không đảm bảo việc lấy trong một phiên.

viettel
cyberspace

2.3 Hướng tiếp cận của nhóm

Hướng tiếp cận sử dụng covisitation matrix (bảng tương tác)

- Khả năng hoạt động trong trường hợp thiếu dữ liệu xếp hạng (rating)
- Sự đơn giản và hiệu quả tính toán
- **Thích hợp với dữ liệu dạng sequence:** covisitation matrix giúp đo lường mối quan hệ giữa các sản phẩm trong các chuỗi và tạo ra gợi ý cho sản phẩm tiếp theo.



	Item A	Item B	Item C
Item A	/	2	1
Item B	0	/	1
Item C	0	0	/

viettel
cyberspace

PHẦN
3

KẾT QUẢ ĐẠT ĐƯỢC

Kết quả thực nghiệm

Phương pháp	Public test		Private test	
	Recal@50 Domain 1	Recall@50 Domain 2	Recall@50 Domain 1	Recall@50 Domain 2
Graph neural network	0.027	0.192	Không áp dụng	Không áp dụng
Covisitation matrix	0.132	0.229	Không áp dụng	Không áp dụng
User watch history Covisitation matrix Domain popularity Items that new users usually watch	0.191	0.316	0.178	0.288

viettel
cyberspace

PHẦN
4

KẾT LUẬN

viettel
cyberspace

Đánh giá của nhóm

Nhờ vào việc thực hiện mô hình trực tiếp cho bài toán, nhóm nghiên cứu hiểu thêm được các hạn chế của các hệ thống đề xuất SOTA, từ đó có khả năng linh hoạt trong việc lựa chọn các giải pháp phù hợp với dữ liệu đặc biệt là các log dữ liệu đặc biệt.

Ngoài ra, nhóm nghiên cứu cũng có thêm các kinh nghiệm quý giá trong việc phát hiện các lỗi bất thường của dữ liệu trong hệ thống đề xuất từ đó phòng tránh các vấn đề trên trong quá trình vận hành huấn luyện các hệ thống đề xuất trong Tập đoàn.

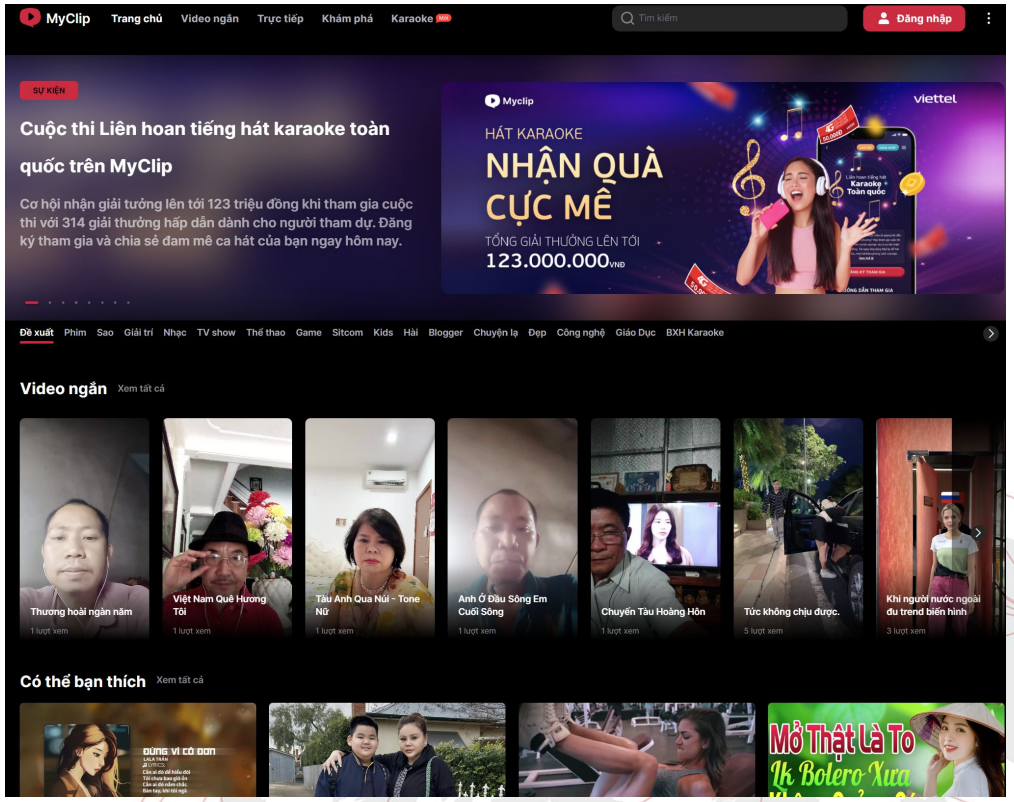
viettel
cyberspace

Mô hình hệ thống đề xuất thiết kế cho nền tảng Myclip

Hệ thống đề xuất được nghiên cứu và phát triển bởi đội ngũ kỹ sư VTCC nhằm giải quyết các yêu cầu phức tạp cho hệ thống Myclip:

Thách thức:

1) Hệ thống Myclip là nền tảng chia sẻ số lượng nội dung lớn với **3,5 triệu nội dung và 1,5 triệu khách hàng.**
2) Hệ thống Myclip có mức độ tương tác không đồng đều giữa các nội dung, có nội dung rất nhiều tương tác, nội dung ít tương tác…
3) Là hệ thống chia sẻ Video, nên thời gian phản hồi của hệ thống đề xuất rất thấp < 0.1ms.
4) Dữ liệu được lưu trữ với mục đích vận hành, nên chưa được chuẩn hóa theo các nhu cầu thiết kế hệ thống đề xuất…



Kết quả:

1) Việc áp dụng các mô hình kỹ thuật đề xuất và thiết kế hệ thống giúp đạt được **15-35% Click-through rate (tăng 8% so với hệ thống cũ)** với **0.058ms tốc độ phản hồi** cho toàn bộ tải hệ thống. Sẵn sàng mở rộng cho toàn bộ nhu cầu phát triển của nền tảng Myclip.
2) Hệ thống tự động chuẩn hóa, điều chỉnh dữ liệu cho các video.
3) Xây dựng cơ sở hành vi khách hàng, hành vi tương tác video, cơ sở lưu logs… nhằm giúp cho việc phân tích chuyên sâu về khách hàng, tạo tiền đề cho việc xây dựng các sản phẩm mới có tính cạnh tranh trên thị trường.