BÁO CÁO GIẢI PHÁP CHO BÀI TOÁN MULTI-DOMAIN RECOMMENDATION – NEXT CONTENT RECOMMENDATION - DHA

|  |
| --- |
| Nguyễn Huy Hoàng1 , Ngô Hoàng Đăng2, Nguyễn Huy Anh3  *1 Trung tâm Không gian mạng Viettel, hoangnh69@viettel.com.vn*  *2 Trung tâm Không gian mạng Viettel, dangnh28@viettel.com.vn*  *3 Trung tâm Không gian mạng Viettel, anhnh151@viettel.com.vn* |

1. **Giới thiệu chung:**

Hệ thống gợi ý là một trong những ứng dụng quan trọng trong lĩnh vực khoa học dữ liệu và trí tuệ nhân tạo. Chúng được áp dụng rộng rãi trong các lĩnh vực như thương mại điện tử, mạng xã hội, đề xuất nội dung số, và đóng vai trò quan trọng trong cải thiện trải nghiệm khách hàng, tăng doanh số bán hàng. Mục tiêu của task VHAC 2023 - Multi-domain Recommendation System - Next Content Recommendation là 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).

1. **Nội dung và phương pháp:**

**2.1 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. Khi không có tương tác, không có thông tin để xây dựng hồ sơ này.

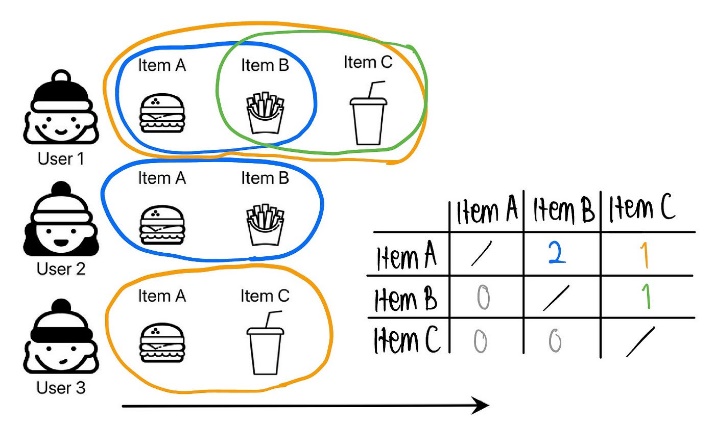
Dữ liệu được cung cấp dưới dạng tuần tự (sequence) nhưng không cung cấp thông tin về thời điểm tương tác: Khi nhận thấy dữ liệu tương tác được biểu diễn dưới dạng tuần tự, nhóm nghiêm cứu dự định áp dụng phương án sử dụng Mạng nơ-rơn đồ thị (Graph Neural Network) nhằm tối ưu đề xuất cho bài toán. Tuy nhiên, 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. Ví dụ: Trong tập dữ liệu train, video\_id: 930766 có mức độ trung tâm (centrality degree) rất cao (0,9) trong khi với tập dữ liệu public test thì chỉ số mức độ trung tâm của video 930766 chỉ còn 0,3. Nhóm nghiên cứu nhận thấy phân phối về mức độ trung tâm và mức độ thịnh hành giữa các video và series không tương đồng giữa tập dữ liệu train và test, điều này gây khó khăn lớn đối với phần lớn các mô hình khuyến khị hiện đại.

Thời điểm của dữ liệu giữa tập dữ liệu train và public không nhất quán: Sau khi đánh giá về mức độ trung tâm, nhóm nghiên cứu nhận ra thời gian trong tập train và thời gian diễn ra trong tập test không nhất quán. Ví dụ như video 930766 trong tập dữ liệu train có published\_time là ngày 14/01/2023, trong khi video  955807 trong tập dữ ilệu test có published\_time là ngày 02/06/2023. Do đó, nhóm nhận thấy có khoảng cách thời gian rất lớn trong hành vi dữ liệu giữa bộ dữ liệu huấn luyện với bộ dữ liệu cung cấp trong tập 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.

Từ những vấn đề trên về mặt dữ liệu, nhóm nghiên cứu nhận thấy việc áp dụng các mô hình truyền thống như Matrix Factorization là không khả thi và mô hình đề xuất tuần tự như Graph Neural Network không đem lại hiệu quả cao, điều này thể hiện kết quả thử nghiệm thực tế ở phần kết quả.

**2.2 Hướng tiếp cận của nhóm**

Thay vì sử dụng các thuật toán phức tạp đòi hỏi lượng dữ liệu lớn và không có nhiều nhiễu cũng như dữ liệu tương tác giữa người dùng - sản phẩm có sẵn để xây dựng mô hình, nhóm sử dụng phương pháp xây dựng covisitation matrix (bảng tương tác) giữa các sản phẩm trong cùng một session. Covisitation matrix được sử dụng để đo lường mức độ tương tác giữa các cặp sản phẩm trong cùng một chuỗi tương tác. Phương pháp này là một công cụ quan trọng trong bài toán đề xuất dạng chuỗi tương tác, đặc biệt là trong việc xây dựng hệ thống gợi ý vào bài toán next content recommendation.



Hình 1: Minh họa cách xây dựng covisitation matrix từ 3 chuỗi tương tác đơn giản.

Cách tiếp cận bài toán sử dụng covisitation matrix đặc biệt có ý nghĩa đối với trường hợp thiếu dữ liệu đánh giá (rating) của người dùng, covisitation matrix xây dựng lại dữ liệu tương tác tương tác người dùng-sản phẩm mà không yêu cầu thông tin xếp hạng cụ thể và đặc biệt hữu ích trong các trường hợp của bài toán chỉ có dữ liệu theo dạng chuỗi xem không xác định thời điểm như trong bài thi. Cách tiếp cận sử dụng phương pháp này thích ứng dễ dàng với thay đổi trong dữ liệu và môi trường người dùng, giúp cải thiện khả năng tự động hóa quá trình đề xuất nội dung.

Thêm vào đó, việc tạo covisitation matrix là một quá trình đơn giản và hiệu quả trong tính toán, đặc biệt khi lượng tương tác người dùng-sản phẩm lớn trong chức năng đề xuất video tiếp theo vì quá trình tính toán covisitation scores giữa cặp sản phẩm thường không phức tạp mà vẫn đạt được hiệu quả cao.

1. **Kết quả thực hiện**

Kết quả của từng mô hình trên tập validation và private test được minh họa ở bảng dưới đây:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Public test** | | **Private test** | |
| **Chỉ số** | **Recall\_1** | **Recall\_2** | **Recall\_1** | **Recall\_2** |
| **GNN** | 0.027 | 0.192 | Không áp dụng | Không áp dụng |
| **CN** | 0.132 | 0.229 | 0.178 | 0.288 |

Kết quả trên tập public test cho thấy, mô hình lớn dựa trên chuỗi tương tác sequence như Graph neural network cho kết quả không tốt khi dữ liệu cung cấp chứa nhiều nhiễu và có yếu tố thay đổi lớn theo thời gian. Từ đó, nhóm quyết định không sử dụng mô hình GNN trong phase private test của cuộc thi mà chỉ tập trung tối ưu phương pháp sử dụng covisitation matrix.

1. **Đánh giá**

Với những vấn đề nêu ở trên, nhóm nghiên cứu nhận thấy phương án sử dụng covisitation matrix hiện đang phù hợp nhất với dữ liệu hiện tại. Do ảnh hưởng chênh lệch lớn về mặt thời gian giữa tập dữ liệu huấn luyện và tập dữ liệu test, kết quả đạt được trên tập dữ liệu test vẫn chưa thực sự ấn tượng đối với cả hai domain 1 và domain 2. Nhóm hi vọng trong thời gian tới sẽ có nhiều cách tiếp cận mới được đề xuất trong trường hợp dữ liệu có phân bố không đồng đều như trong cuộc thi.

1. **Giá trị đem lại**

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.

1. **Kết luận:**

Đối với hiện trạng dữ liệu hiện tại, nhóm nghiên cứu đã thực hiện các nghiên cứu có khả thi áp dụng bài toán đề xuất, tuy nhiên trong tương lai nhóm mong muốn được thực hiện bài toán với dữ liệu đầy đủ hơn nhằm khai thác được đầy đủ sức mạnh của các công nghệ đề xuất mới.