BÁO CÁO GIẢI PHÁP CHO BÀI TOÁN NEWS 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. Đối với track VHAC 2023 - News Recommendation yêu cầu được đưa ra là xây dựng một hệ thống gợi ý cho Báo trí - Tin tức nhằm xác định nội dung có phù hợp với từng người đọc dựa trên thông tin người dùng, thuộc tính bài báo, và lịch sử tương tác giữa người dùng - bài báo.

Khởi động nguội (cold-start) và hành vi mới của người đọc là những thách thức quan trọng mà nhóm đối mặt tại track này. Một lượng lớn bài báo mới được xuất bản, đăng tải hàng ngày khiến việc áp dụng các phương pháp truyền thống gặp khó khăn. Ngoài ra, đôi khi người đọc có thể thực hiện hành động bất thường như đọc một thể loại mới mà trước đây họ chưa quan tâm.

1. **Nội dung và phương pháp:**

**2.1 Các hướng tiếp cận với bài toán đề xuất:**

Phương pháp phân rã ma trận là một trong những kỹ thuật phổ biến trong bài toán hệ thống đề xuất. Nó tách ma trận tiện ích thành hai ma trận thấp chiều hơn để khám phá mối quan hệ ẩn giữa người dùng và sản phẩm. Tuy nhiên, phương pháp này đối mặt với nhược điểm như khởi động nguội khi người dùng mới hoặc sản phẩm mới xuất hiện, vì khi này không có thông tin lịch sử tương tác giữa người dùng - sản phẩm. Sự thưa thớt của ma trận dữ liệu cũng là vấn đề, khiến cho việc xây dựng mô hình và đưa ra dự đoán trở nên khó khăn. Đồng thời, tính khả thi của phương pháp với tập dữ liệu lớn cũng là một thách thức do đòi hỏi khả năng tính toán và bộ nhớ lớn.

Một phương pháp nổi bật trong lĩnh vực đề xuất là Graph Neural Networks (GNN). GNN khám phá mối quan hệ giữa người dùng và sản phẩm thông qua việc xử lý đồ thị, tận dụng cấu trúc tương tác phức tạp. Tuy nhiên, như phương pháp phân rã ma trận, GNN cũng đối mặt với thách thức khi phải xử lý vấn đề khởi động nguội với người dùng hoặc sản phẩm mới. Ngoài ra, phương pháp này đòi hỏi lượng tính toán lớn do việc cập nhật đồ thị và tính toán trên các đỉnh và cạnh của đồ thị.

**2.2 Thách thức trong cuộc thi:**

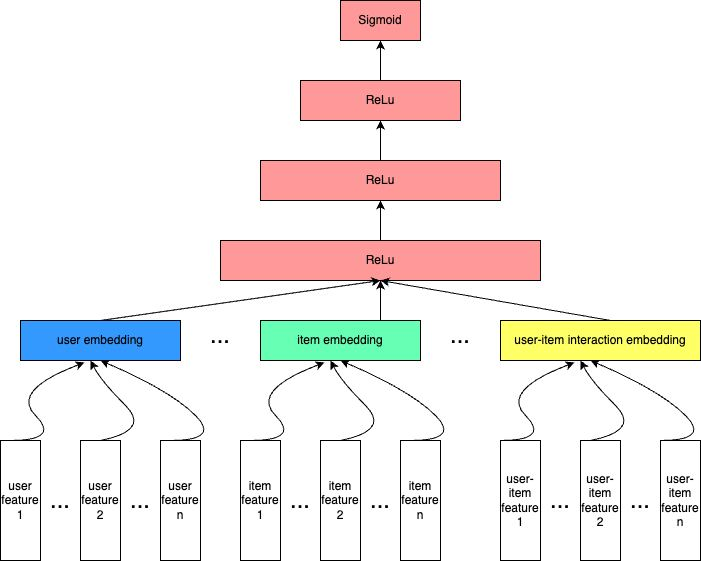
Khi áp dụng phương pháp phân rã ma trận nhóm nhận thấy phương pháp này chưa có khả năng giải quyết được các vấn đề dựa trên dữ liệu cung cấp bởi ban tổ chức, điển hình là vấn đề về khởi động nguội (cold start problem).

Đối với dữ liệu được cung cấp, khi đánh giá dữ liệu của tương tác người dùng với bài báo trên khung phiên tự quy định là 15 phút một phiên, nhóm nhận thấy không có nhiều tương tác đọc bài theo báo tuần tự - chuỗi tuần tự các bài đọc báo ngắn với độ dài tuần tự trung bình của các bài báo được đọc của một người dùng nhỏ hơn 2, hoặc độ trung tâm của bài báo (centrality degree) không thể hiện rõ, điều đó cũng thể hiện hành vi dữ liệu người đọc báo là có xu hướng xem các bài báo mới. Từ những khảo sát trên, nhóm nhận định với bài toán này không áp dụng phương pháp đồ thị vào trong đề xuất.

**2.3 Hướng tiếp cận của nhóm với bài toán:**

**2.3.1. Sử dụng mô hình đề xuất Two-Tower (TT)**

Quan sát dữ liệu được cung cấp của cuộc thi ngoài dữ liệu lịch sử tương tác giữa người dùng - bài báo, còn có dữ liệu cá nhân về thông tin người dùng và thông tin bài báo. Giai đoạn đầu, nhóm thử nghiệm mô hình Two-Tower (kiến trúc mô hình được minh họa ở hình bên dưới), là một kiểu mô hình học sâu được sử dụng rộng rãi gần đây trong bài toán đề xuất, và nó có một số ưu điểm so với phương pháp sử dụng thuần tương tác người dùng như sau:



Tích hợp thông tin người dùng/ sản phẩm (Side-information): Thông tin chi tiết về người dùng/ sản phẩm được tích hợp vào trong mô hình, do đó hệ thống có thể giải quyết vấn đề về cold-start của các bài báo mới.

Khả năng học đặc trưng phức tạp: TT sử dụng mạng nơron cho cả người dùng và sản phẩm để học các đặc trưng, thông tin ẩn, ngữ cảnh ẩn trên tương tác người dùng và bài báo. Trong khi matrix factorization (MF) dựa vào phân tách ma trận tiện ích thành ma trận con thấp chiều, TT có khả năng học các biểu diễn phức tạp hơn.

**2.3.2 Hướng tiếp cận sử dụng lớp mô hình tree-based**

Trong quá tình xây dựng giải pháp, nhận thấy bài toán hoàn toàn có thể đưa về dạng bảng và kết quả sẽ rất khả quan nếu có thể xây dựng các đặc trưng (features) tốt dựa trên lịch sử tương tác giữa người dùng - bài báo, đặc trưng của người dùng, đặc trưng của bài báo được cung cấp. Nhóm thử nghiệm sử dụng thêm mô hình XGBoost (XGB). Lớp mô hình tree-based được sử dụng rộng rãi trong thực tế với bài toán xây dựng hệ gợi ý vì một số ưu điểm quan trọng sau:

**Hạn chế các vấn đề về mất cân bằng dữ liệu:** Trong bài toán recommendation, dữ liệu thường mất cân bằng vì một số sản phẩm được tương tác nhiều hơn so với các sản phẩm khác. Thuật toán tree-based có khả năng xử lý mất cân bằng dữ liệu bằng cách sử dụng trọng số hoặc kỹ thuật oversampling / undersampling;

**Tích hợp thông tin sản phẩm:** Lớp mô hình tree-based có thể tích hợp thông tin chi tiết về sản phẩm, bao gồm các đặc trưng sản phẩm (miêu tả sản phẩm, danh mục sản phẩm, …) để cải thiện khả năng gợi ý;

**Dễ dàng sử dụng và triển khai:** Có sẵn nhiều thư viện và framework hỗ trợ các giải thuật tree-based như XGBoost, LightGBM và CatBoost giúp đơn giản hóa việc triển khai và sử dụng.

**2.3.3 Hướng tiếp cận sử dụng Ensemble learning**

Sử dụng ensemble learning để kết hợp giữa hai lớp mô hình deep learning và gradient boosted decision trees (GBDT) là một chiến lược mạnh mẽ để tận dụng ưu điểm của cả hai phương pháp.

Phương pháp này tận dụng ensemble learning để kết hợp hai loại mô hình, deep learning và gradient boosted decision trees (GBDT), nhằm tạo ra hệ thống gợi ý mạnh mẽ và đáng tin cậy. Việc kết hợp những ưu điểm đặc trưng của cả deep learning và GBDT giúp mô hình học được đặc trưng phức tạp và xử lý mối quan hệ phi tuyến tính một cách hiệu quả. Sự kết hợp này mang lại hiệu suất dự đoán tốt hơn so với việc sử dụng một mô hình đơn lẻ, tạo ra một ensemble model có khả năng đưa ra dự đoán chính xác và ổn định trong nhiều ứng dụng thực tế

1. **Kết quả thực hiện:**

Quan sát dữ liệu tập private test là các tương tác giữa người dùng - sản phẩm trong ngày 12/04/2022, tức là ở ngày thứ t+1 so với dữ liệu ở trập training kéo dài từ ngày 05/04/2022 - 11/04/2022. Nhóm chia tập dữ liệu training thành 2 tập train và validation sao cho tập dữ liệu validation có cùng tính chất với tập dữ liệu private test, tức là cùng ở thời điểm t+1. Điều này đảm bảo sự tương quan giữa 2 tập validation và test.

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

|  |  |  |
| --- | --- | --- |
| **Model** | **Validation ROC-AUC score** | **Private test ROC-AUC score** |
| **MF** | 0.80732 | 0.80747 |
| **TT** | 0.98876 | 0.98890 |
| **XGB** | 0.98828 | 0.98855 |
| **TT+XGB** | **0.98946** | **0.98957** |

**Bảng 1.** Kết quả các mô hình học máy

1. **Đánh giá hiệu quả:**

Kết quả cho thấy cách tiếp cận của nhóm là hiệu quả trong việc đưa ra các đề xuất phù hợp với người dùng. Với việc kết hợp 2 mô hình: Two-tower và XGBoost, nhóm đã đạt được kết quả cao vượt trội so với cách tiếp cận cổ điển là phân rã ma trận. Khi tận dụng các đặc trưng của người dùng, sản phẩm, nhóm đã tăng được độ chính xác trong các trường hợp đề xuất khó như cold-start và hành vi tương tác mới của người dùng.

Ngoài ra, nhóm nhận thấy khả năng đề xuất của thuật toán có khả năng bao quát hóa tốt, điều này thể hiện kết quả không chênh lệch giữa tập dữ liệu validation và test mặc dù tập dữ liệu private test ở ngày t+1 so với ngày huấn luyện mô hình. Điều này có nghĩa, với 2 mô hình đề xuất ở trên, sản phẩm có thể triển khai trên thực tế mà không gặp phải vấn đề về overfitting dữ liệu.

1. **Lợi ích:**

Dựa trên các kết quả đã đạt được trong quá trình nghiên cứu và tham gia cuộc thi xây dựng hệ thống đề xuất cho track News Recommendation, nhóm nghiên cứu nhận thấy lợi ích của việc đưa hệ thống tương tự cho các sản phẩm báo đề xuất của tập đoàn như Tiin.vn nhằm nâng cao chất lượng, trải nghiệm khách hàng hay Viettel Family nhằm tăng tương tác của cán bộ nhân viên với các sự kiện, thông tin của Tập đoàn.

1. **Kết luận:**

Nhóm nghiên cứu đã thử nghiệm một số giải pháp đánh giá khả năng đọc của người đọc với các bài báo qua nhiều phương án. Các phương án đều cho các kết quả khả quan trong việc bao quát hóa bài toán. Để giải quyết các vấn đề như cold-start, click-noise, clickbait…, việc kết hợp giữa các mô hình đem lại các kết quả tốt nhất trong các giải pháp trên.

Tuy nhiên, điểm hạn chế trong sáng kiến là việc chưa tận dụng hết được phần tiêu đề bài viết do để sử dụng tối ưu tiêu đề cần phải sử dụng dữ liệu, các mô hình pretrain và thời gian cuộc thi có hạn. Do đó, trong tương lai nhóm nghiên cứu mong muốn có cơ hội kết hợp thêm các dữ liệu bài viết nhằm cải thiện hiệu quả của hệ thống đề xuất.

1. **Tài liệu tham khảo:**

Yi, X., Yang, J., Hong, L., Cheng, D. Z., Heldt, L., Kumthekar, A. A., … Chi, E. (Eds.). (2019). *Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations.*