@@bbbbbbbbbbbbbbb

A green and orange logo

Description automatically generated

**TẬP ĐOÀN VIỄN THÔNG QUÂN ĐỘI**

**TÀI LIỆU MÔ TẢ**

**GIẢI PHÁP KHẢO SÁT HỆ THỐNG GỢI Ý VIDEO**

Hà Nội, 2023



# Mục lục

[Mục lục 2](#_Toc154570514)

[Danh mục hình ảnh 3](#_Toc154570515)

[Danh mục bảng biểu 5](#_Toc154570516)

[Phần 1. Đánh giá tìm hiểu bài toán 6](#_Toc154570517)

[1.1. Mô tả bài toán 6](#_Toc154570518)

[1.2. Đề xuất giải pháp 6](#_Toc154570519)

[1.3. Các vấn đề cần giải quyết 7](#_Toc154570520)

[Phần 2. KHẢO SÁT TỔNG QUAN 10](#_Toc154570521)

[2.1. Khảo sát kiến trúc gợi ý video 10](#_Toc154570522)

[2.2. Khảo sát dữ liệu 11](#_Toc154570523)

[Phần 3. Khảo sát dữ liệu 14](#_Toc154570524)

[3.1. Tổng quan dữ liệu 14](#_Toc154570525)

[3.2. Phân tích dữ liệu video 15](#_Toc154570526)

[3.2.2. Phân tích dữ liệu về video phổ biến 17](#_Toc154570527)

[3.2.3. Phân tích dữ liệu về video ít tương tác 20](#_Toc154570528)

[3.3. Phân tích dữ liệu người dùng 21](#_Toc154570529)

[3.3.2. Phân tích nhóm người dùng nhiều tương tác 22](#_Toc154570530)

[3.3.3. Phân tích nhóm người dùng ít tương tác 24](#_Toc154570531)

[3.4. Đánh giá hiện trạng dữ liệu 24](#_Toc154570532)

[Phần 4. KHẢO SÁT MÔ HÌNH 25](#_Toc154570533)

[4.1. Mô hình RecVAE 25](#_Toc154570534)

[4.1.1. Cơ sở lý thuyết 25](#_Toc154570535)

[4.1.2. Quy mô dữ liệu 29](#_Toc154570536)

[4.1.3. Kết quả mô hình 29](#_Toc154570537)

[4.2. Mô hình Bert4Rec 30](#_Toc154570538)

[4.2.1. Cơ sở lý thuyết 30](#_Toc154570539)

[4.2.2. Quy mô dữ liệu 32](#_Toc154570540)

[4.2.3. Kết quả mô hình 33](#_Toc154570541)

[Phần 5. Đề xuất cải tiến 34](#_Toc154570542)

# Danh mục hình ảnh

[Hình 3.1. Phân phối lượt xem video trong tháng 11 14](#_Toc154570498)

[Hình 3.2. Phân bố lượng người dùng xem video trên 4 tuần 15](#_Toc154570499)

[Hình 3.3. Dữ liệu mẫu của các đặc trưng video 16](#_Toc154570500)

[Hình 3.4. Phân phối lượt xem video với các chủ đề sub-category 17](#_Toc154570501)

[Hình 3.5. Top những video phổ biến trong tháng 11 17](#_Toc154570502)

[Hình 3.6. Phân bố top 50 video được xem nhiều nhất trong tuần 1 18](#_Toc154570503)

[Hình 3.7. Phân bố top 50 video được xem nhiều nhất trong tuần 2 18](#_Toc154570504)

[Hình 3.8. Phân bố top 50 video được xem nhiều nhất trong tuần 3 19](#_Toc154570505)

[Hình 3.9. Phân bố top 50 video được xem nhiều nhất trong tuần 4 19](#_Toc154570506)

[Hình 3.10. Phân bố top 50 người dùng xem video nhiều nhất trong tháng 11. 21](#_Toc154570507)

[Hình 3.11. Phân phối người dùng theo tương tác trong tuần 1 22](#_Toc154570508)

[Hình 3.12. Phân phối người dùng theo tương tác trong tuần 2 22](#_Toc154570509)

[Hình 3.13. Phân phối người dùng theo tương tác trong tuần 3 23](#_Toc154570510)

[Hình 3.14. Phân phối người dùng theo tương tác trong tuần 4 23](#_Toc154570511)

[Hình 4.1. Kiến trúc của RecVAE 27](#_Toc154570512)

[Hình 4.2. Kiến trúc của mô hình BERT4REC 31](#_Toc154570513)

# Danh mục bảng biểu

[Bảng 1.1. Tóm tắt các vấn đề cần giải quyết 6](#_Toc154570469)

[Bảng 2.1. Bảng mô tả dữ liệu thông tin video 10](#_Toc154570470)

[Bảng 2.2. Bảng tổng hợp dữ liệu hành vi phim theo ngày 11](#_Toc154570471)

[Bảng 2.3. Bảng tổng hợp dữ liệu hành vi phim của người dùng theo tháng 11](#_Toc154570472)

[Bảng 2.4. Bảng tổng hợp dữ liệu cho thông tin của người dùng xem phim 12](#_Toc154570473)

# Đánh giá tìm hiểu bài toán

## Mô tả bài toán

**Bài toán**: Dựa trên cơ sở của những khảo sát đánh giá của công việc tháng 10. Chúng tôi nhận thấy rằng hệ thống đang gặp phải vấn đề như:

* Mô hình đang thực hiện gợi ý cho cả hai phim bộ và phim lẻ. Hành vi của người dùng xem phim bộ và phim lẻ là khác nhau nên chúng tôi đề xuất việc chia dữ thành tập dữ liệu cho phim bộ và phim lẻ riêng để gợi ý.
* Mô hình đang bị vấn đề về popularity bias.
* Chưa có kịch bản ensemble rõ ràng.

Đối với những vấn đề trên, trong nhiệm vụ thực hiện của tháng này, chúng tôi đề xuất thực hiện những bước cải tiến theo các kịch bản để nâng cấp mô hình và thực hiện việc ensemble cho dữ liệu của hệ thống.

## Đề xuất giải pháp

Với nguồn dữ liệu đầu vào lớn, cần đánh giá và tìm hiểu bài toàn và phân tích trực quan hoá các dữ liệu. Tập dữ liệu đầu vào có các đặc trưng khác nhau về hành vi và thông tin người dùng. Các hành vi sẽ được tổng hợp và trực quan hoá và kiểm tra đánh giá để đối xoát với các giả thiết được đề ra. Các bước sẽ được áp dụng trong bài toán như sau:

**Lập kế hoạch khảo sát:** Xác định mục tiêu của việc khảo sát. Mục tiêu có thể là để hiểu rõ hơn về hệ thống, để xác định các vấn đề và cơ hội, hoặc để đưa ra các khuyến nghị cho việc phát triển và triển khai hệ thống. Xác định phạm vi của việc khảo sát. Phạm vi có thể bao gồm các khía cạnh sau:

* Tổng quan của hệ thống
* Dữ liệu đầu vào
* Các kỹ thuật gợi ý
* Các chỉ số đánh giá

**Thu thập thông tin:** Sau khi đã lập kế hoạch khảo sát, bước tiếp theo là thu thập thông tin. Thông tin có thể được thu thập từ nhiều nguồn khác nhau, bao gồm:

* Dữ liệu thứ cấp: Tài liệu học thuật, bài báo nghiên cứu, báo cáo,...
* Dữ liệu sơ cấp: Khảo sát người dùng, phỏng vấn chuyên gia, thử nghiệm hệ thống,...

Khi thu thập dữ liệu thứ cấp, cần chú ý đến tính chính xác và cập nhật của thông tin. Khi thu thập dữ liệu sơ cấp, cần đảm bảo rằng thông tin được thu thập một cách khách quan và đáng tin cậy.

**Phân tích thông tin:** Một quá trình biến đổi thông tin thô thành thông tin có ý nghĩa. Phân tích thông tin giúp chúng tôi hiểu rõ hơn về hệ thống và đưa ra các kết luận và khuyến nghị chính xác. Phân tích thông tin giúp chúng tôi hiểu rõ hơn về thông tin thô. Chúng tôi có thể phân tích thông tin để tìm ra các xu hướng, các mối quan hệ, và các vấn đề. Phân tích thông tin cũng giúp chúng tôi xác định các cơ hội và các rủi ro. Phân tích thông tin là một quá trình quan trọng trong việc khảo sát hệ thống. Bằng cách phân tích thông tin một cách hiệu quả, chúng tôi có thể hiểu rõ hơn về hệ thống và đưa ra các quyết định sáng suốt hơn.

## Các vấn đề cần giải quyết

Trong giai đoạn đầu tiên, các vấn đề và công việc cần giải quyết được tóm tắt trong Bảng 1.1 như sau:

Tóm tắt các vấn đề cần giải quyết

| **STT** | **Công việc** |
| --- | --- |
| 1 | Khảo sát hệ thống gợi ý phim |
| 2 | Khảo sát hệ thống gợi ý phim |
| 3 | Đánh giá hệ thống gợi ý phim |
| 4 | Khảo sát kiến trúc luồng gợi ý phim |
| 5 | Đánh giá kiến trúc luồng gợi ý phim |
| 6 | Khảo sát quy mô dữ liệu |
| 7 | Đánh giá quy mô dữ liệu |
| 8 | Khảo sát các kỹ thuật của mô hình |
| 9 | Khảo sát phương pháp đánh giá hiệu quả của hệ thống |
| 10 | Đề xuất các phương pháp đánh giá hiệu quả |
| 11 | Khảo sát nguồn dữ liệu |
| 12 | Đề xuất thu thập thêm nguồn dữ liệu |
| 13 | Khảo sát luồng gợi ý real-time |
| 14 | Tổng hợp hiện trạng và các hạn chế của luồng gợi ý thời gian thực |
| 15 | Khảo sát dữ liệu |
| 16 | Tổng hợp dữ liệu người dùng |
| 17 | Khảo sát dữ liệu phim |
| 18 | Trực quan hóa và phân tích đặc trưng của người dùng |
| 19 | Tổng hợp các hành vi người dùng ít tương tác |
| 20 | Trực quan hoá và phân tích nhóm người dùng ít tương tác |
| 21 | Tổng hợp các hành vi người dùng nhiều tương tác |
| 22 | Trực quan hoá và phân tích nhóm người dùng nhiều tương tác |
| 23 | Trực quan hoá và phân tích các đặc trưng của phim bộ |
| 24 | Tổng hợp các phim bộ phổ biến |
| 25 | Trực quan hoá và phân tích các phim bộ phổ biến |
| 26 | Tổng hợp dữ liệu các phim ít tương tác |
| 27 | Trực quan hoá và phân tích các phim ít tương tác |
| 28 | Tổng hợp dữ liệu của phim lẻ |
| 29 | Trực quan hoá và phân tích các đặc trưng của phim lẻ |
| 30 | Trực quan hoá và phân tích các phim lẻ phổ biến |
| 31 | Trực quan hoá và phân tích hành vi người dùng với phim theo thời gian |
| 32 | Đánh giá hiện trạng dữ liệu |
| 33 | Tổng hợp hiện trạng và các hạn chế của dữ liệu |
| 34 | Khảo sát mô hình |
| 35 | Tổng hợp hiện trạng và các hạn chế của mô hình |
| 36 | Khảo sát mô hình RecVAE trong hệ thống |
| 37 | Kháo sát cấu hình của mô hình RecVAE |
| 38 | Khảo sát quy mô dữ liệu của mô hình RecVAE |
| 39 | Khảo sát các phương pháp đánh giá mô hình RecVAE |
| 40 | Đánh giá hiện trạng mô hình RecVAE |
| 41 | Đánh giá kết quả mô hình của RecVAE trên dữ liệu mẫu - TV360 - 1M |
| 42 | Khảo sát mô hình luật nghiệp vụ |
| 43 | Đề xuất cải tiến |
| 44 | Tổng hợp hiện trạng và các hạn chế của hệ thống |
| 45 | Đề xuất kế hoạch cải tiến hệ thống |
| 46 | Đề xuất kế hoạch cải tiến mô hình |
| 47 | Tổng hợp hiệu suất thực nghiệm tốt nhất của các models đề xuất |
| 48 | Phân tích và so sánh hiệu quả của các models đề xuất |

# KHẢO SÁT TỔNG QUAN

## Khảo sát kiến trúc gợi ý video

Hệ thống phim với 2 luồng gợi ý cho video là hệ thống cung cấp cho người dùng các gợi ý về video dựa trên dữ liệu thu thập được từ dữ liệu lịch sử của người dùng.

Hệ thống đang xử lý và chia thành các bài toán như: gợi ý video dựa trên trending, dựa trên nội dung đã xem và hành vi người dùng sử dụng các mô hình học sâu. Mô hình kết hợp các phương pháp.

Đầu vào của hệ thống:

* Lịch sử xem nội dung trên ứng dụng TV360
* Lịch sử tương tác app
* Thông tin demographic của người dùng

Đầu ra của giải pháp:

* 5 bộ 60 video gợi ý cá nhân hoá theo sở thích người dùng

**Xây dựng mô hình gợi ý video từ trending:** Thống kê các chỉ số tổng hợp người dùng, yêu thích đánh giá và trending trên nhiều chiều khác nhau như thời gian và độ tuổi người dùng.

Trong đó:

* N là số ngày đang xét
* videoScoreByDay là tổng điểm video theo từng ngày
* w\_k là giá trị trọng số (k càng nhỏ thì giá trị w\_k càng cao)

**Luồng gợi ý video:** Luồng gợi ý cho video được sử dụng để cung cấp cho người dùng các gợi ý ban đầu khi người dùng truy cập hệ thống. Luồng này dựa trên dữ liệu chung về các video phổ biến hoặc các video tương tự với các video mà người dùng đã xem trước đây.

Trong đó:

* lognum là số lượt xem của người dùng
* watch\_duration là thời gian xem của người dùng (được giới hạn từ 30 giây đến 1 ngày)

## Khảo sát dữ liệu

Tổng quan dữ liệu người dùng là một quá trình thu thập, sắp xếp và phân tích dữ liệu về người dùng để hiểu rõ hơn về họ. Dữ liệu người dùng có thể bao gồm thông tin về nhân khẩu học, hành vi, và sở thích của người dùng.

Dữ liệu của hệ thống gợi ý rất đa đạng và bao gồm nhiều thông tin khác nhau nhằm thực hiện và triển khai các bài toán khác nhau. Vậy nên, trong phạm vi của bài toán gợi ý phim cho người dùng. Chúng tôi sẽ sử dụng các nguồn dữ liệu từ các bảng dữ liệu sau để phục vụ cho việc khảo sát và phân tích dữ liệu:

Bảng mô tả dữ liệu thông tin video

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| f\_tv360\_video\_category | Thông tin thể loại video | video\_id | string | id video |
| video\_name | string | tên video |
| description | string | mô tả video |
| view\_count | int | số lượt xem |
| like\_count | int | số lượt thích |
| duration | int | thời lượng video |
| published\_time | string | thời gian xuất bản |
| is\_active | int | có phải nội dung hiện hành |
| sub\_category\_id | string | id thể loại con (1 video có thể nhận nhiều sub\_category) |
| sub\_category\_name | string | tên thể loại con (1 video có thể nhận nhiều sub\_category) |
| raw\_category\_id | string | id thể loại (dữ liệu gốc) |
| raw\_category\_name | string | tên thể loại (dữ liệu gốc) |
| category\_name | string | tên thể loại (nội suy) |
| category\_id | string | id thể loại (nội suy) |
| cluster\_id | string | id cụm |
| series\_name | string | tên chuỗi video |
| is\_series | int | có phải video chuỗi hay không |
| series\_id | string | id chuỗi video |
| episode\_number | int | số thứ tự tập |
| key | string | API kênh |
| value | string | API value |
| partition | string | partition |

Bảng tổng hợp dữ liệu hành vi phim theo ngày

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| f\_tv360\_log\_video\_daily | Thông tin tương tác video hàng ngày | user\_id | string | id người dùng |
| profile\_id | string | id tài khoản |
| content\_id | string | id nội dung |
| watch\_duration | bigint | thời lượng xem |
| lognum | bigint | số lần vào |
| pause\_count | bigint | số lần pause |
| seek\_count | bigint | số lần tua |
| change\_volumn\_count | bigint | số lần thay đổi âm lượng |
| partition | string | partition |

Bảng tổng hợp dữ liệu hành vi phim của người dùng theo tháng

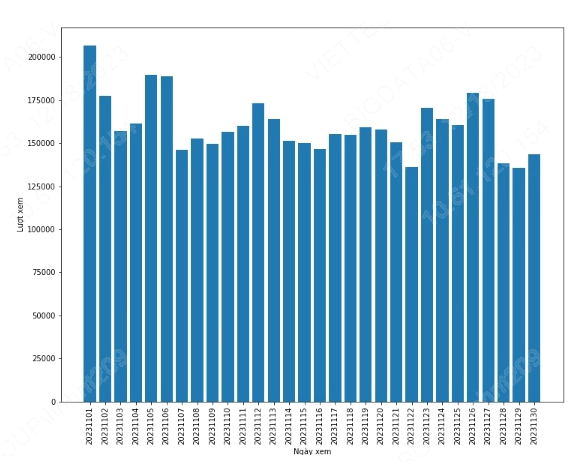
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| f\_tv360\_summary\_vod\_mon | Thông tin tương tác video hàng tháng | user\_id | string | id người dùng |
| profile\_id | string | id tài khoản |
| content\_id | string | id nội dung |
| watch\_duration | bigint | thời lượng xem |
| lognum | bigint | số lần vào |
| pause\_count | bigint | số lần pause |
| seek\_count | bigint | số lần tua |
| change\_volumn\_count | bigint | số lần thay đổi âm lượng |
| partition | string | partition (YYYYMM01) |

Bảng tổng hợp dữ liệu cho thông tin của người dùng xem phim

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| f\_tv360\_video\_user\_profile | Thông tin profile của người dùng xem video | profile\_id | string | id tài khoản |
| news\_politics | bigint | điểm thể loại |
| news\_news | bigint | điểm thể loại |
| news\_hot | bigint | điểm thể loại |
| news\_technology | bigint | điểm thể loại |
| news\_economy | bigint | điểm thể loại |
| news\_security | bigint | điểm thể loại |
| news\_world | bigint | điểm thể loại |
| sports\_general | bigint | điểm thể loại |
| sports\_football | bigint | điểm thể loại |
| sports\_golf | bigint | điểm thể loại |
| sports\_volleyball | bigint | điểm thể loại |
| sports\_tennis | bigint | điểm thể loại |
| sports\_basketball | bigint | điểm thể loại |
| kid\_animal | bigint | điểm thể loại |
| kid\_cartoon | bigint | điểm thể loại |
| kid\_education | bigint | điểm thể loại |
| kid\_music | bigint | điểm thể loại |
| kid\_toy | bigint | điểm thể loại |
| kid\_show | bigint | điểm thể loại |
| tvshow\_show | bigint | điểm thể loại |
| other\_discover | bigint | điểm thể loại |
| other\_other | bigint | điểm thể loại |
| health\_care | bigint | điểm thể loại |
| health\_beauty | bigint | điểm thể loại |
| gaming\_pubg | bigint | điểm thể loại |
| gaming\_aoe | bigint | điểm thể loại |
| gaming\_lol | bigint | điểm thể loại |
| gaming\_moba | bigint | điểm thể loại |
| gaming\_fun | bigint | điểm thể loại |
| music\_pop | bigint | điểm thể loại |
| music\_general | bigint | điểm thể loại |
| music\_bolero | bigint | điểm thể loại |
| entertainment\_other | bigint | điểm thể loại |
| entertainment\_tvshow | bigint | điểm thể loại |
| entertainment\_showbiz | bigint | điểm thể loại |
| education\_subject | bigint | điểm thể loại |
| education\_skills | bigint | điểm thể loại |
| education\_language | bigint | điểm thể loại |
| travel\_travel | bigint | điểm thể loại |
| travel\_food | bigint | điểm thể loại |
| comedy\_sitcom | bigint | điểm thể loại |
| partition | string | partition |

# Khảo sát dữ liệu

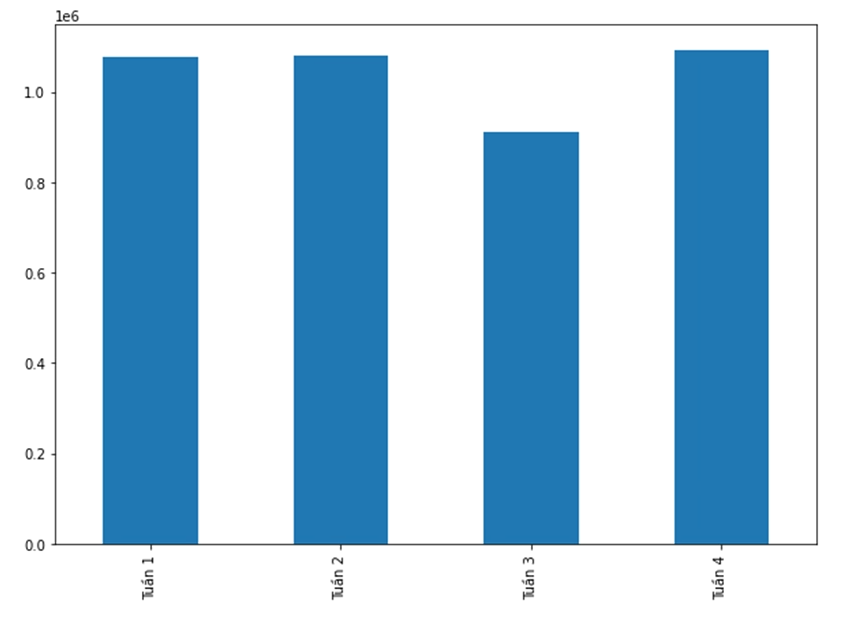
## Tổng quan dữ liệu



Phân phối lượt xem video trong tháng 11

Hình 3.1 cho thấy dữ liệu về lượt xem phim theo thời gian. Trục x thể hiện thời gian, trong khi trục y thể hiện số lượt xem. Dựa trên biểu đồ, dữ liệu xem video có lượt xem ổn định từ 130.000 lượt xem đến hơn 200.000 lượt xem.

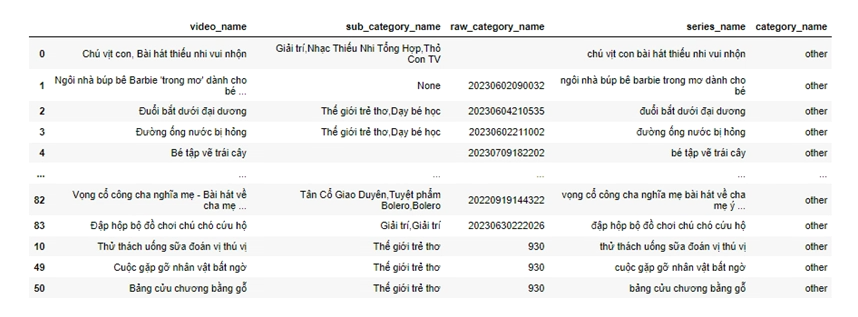
## Phân tích dữ liệu video



Phân bố lượng người dùng xem video trên 4 tuần

Phân phối lượng người dùng xem video trên 4 tuần được ánh xạ như biểu đồ 3.2. Trong 4 tuần, lượt người xem dao động từ hơn 800.000 lượt xem đến hơn 1.000.000 lượt xem. Tổng lượng tương tác của người dùng với video là 4.161.417 lượng xem. Trung bình mỗi tuần đạt hơn 1.000.000 lượt xem so với con số thực tế đạt mức kỳ vọng. Không có bất thường hay đột biến xảy ra trong các tuần của tháng 11.

Trong bảng dữ liệu video có nhiều trường thể hiện thể loại của video, trong đó có: raw\_category\_name, sub\_category\_name, category\_name, v.v… Để trực quan, chúng tôi đã hiển thị một vài dữ liệu mẫu được tham chiếu đến hình 3.3 dưới đây.

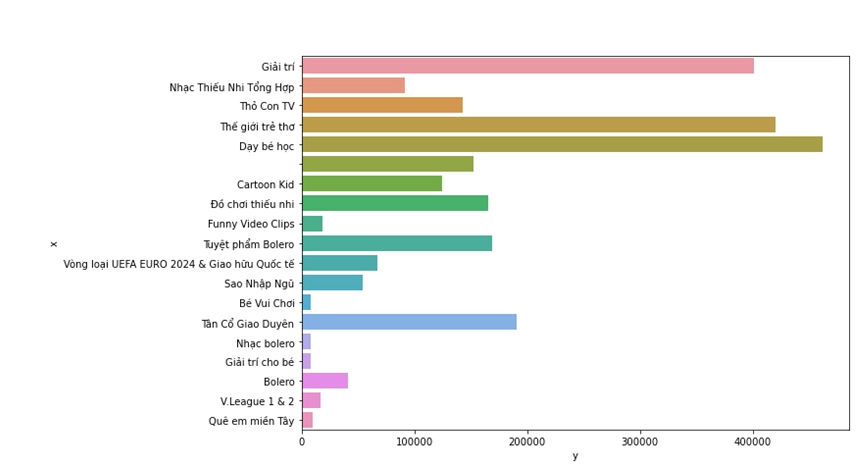


Dữ liệu mẫu của các đặc trưng video

Trong các đặc trưng mang giá trị biểu diễn thể loại video, các trường được thể hiện như:

* Sub\_category\_name: Thể hiện các thể loại nhỏ hơn và chi tiết hơn so với category\_name
* Category\_name: Thể hiện thể loại của video
* Raw\_category\_name: Biểu diễn thể loại phim nhưng mang dữ liệu hỗn loạn.

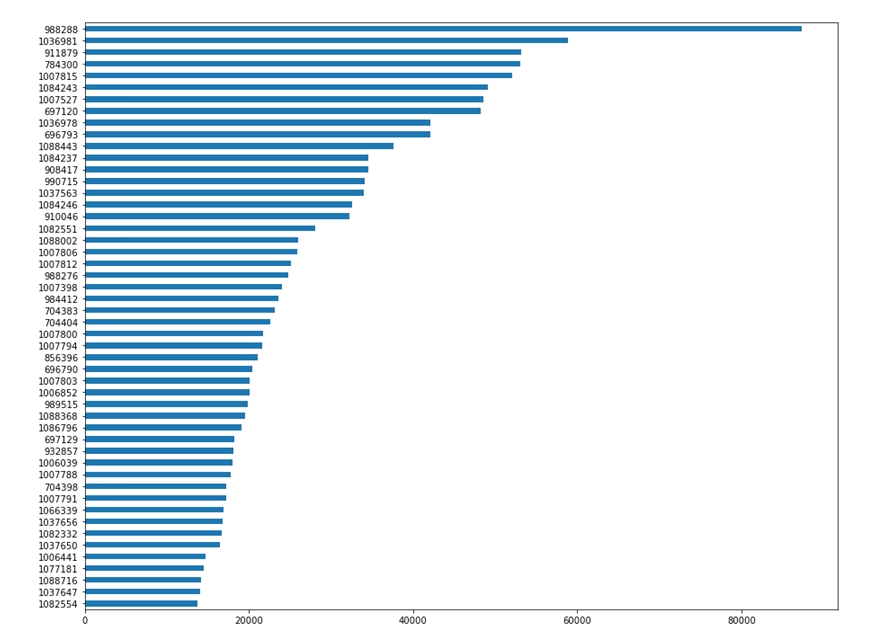
Vậy nên, chúng tôi trực quan hoá chiều thể loại phim theo trường *sub\_category\_name* để biểu diễn phân phối.



Phân phối lượt xem video với các chủ đề sub-category

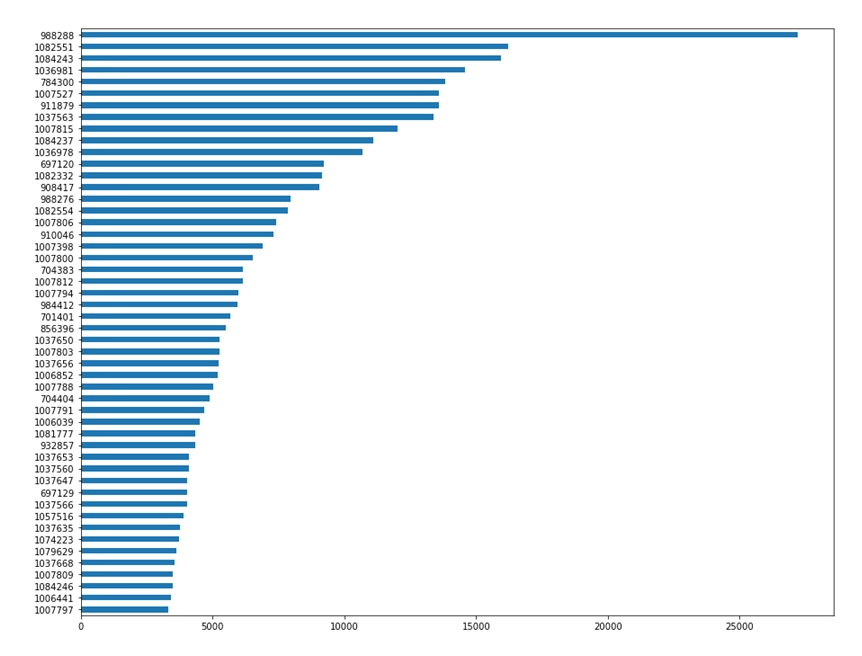
Đối với chủ đề sub-category của video, các thể loại được phổ biến thường liên quan đến trẻ em. Các thể loại được quan tâm nhiều thứ 2 là thể loại liên quan đến các dòng nhạc hoài cổ. Tiếp đến là các thể loại liên quan đến giải trí và thể loại. Tuy nhiên dữ liệu sub\_category\_name đang khá hỗn loạn và chưa được định nghĩa rõ ràng.

### Phân tích dữ liệu về video phổ biến

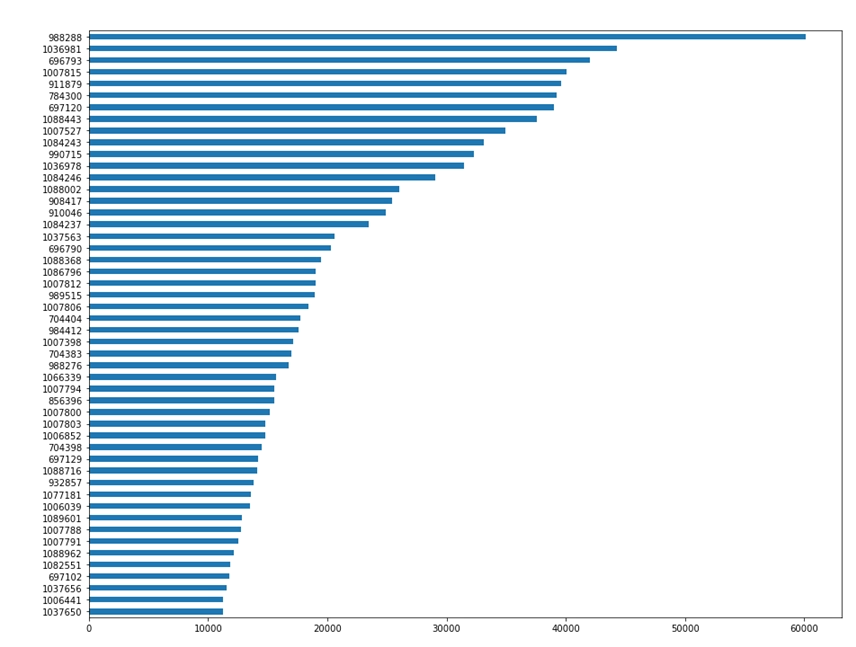


Top những video phổ biến trong tháng 11

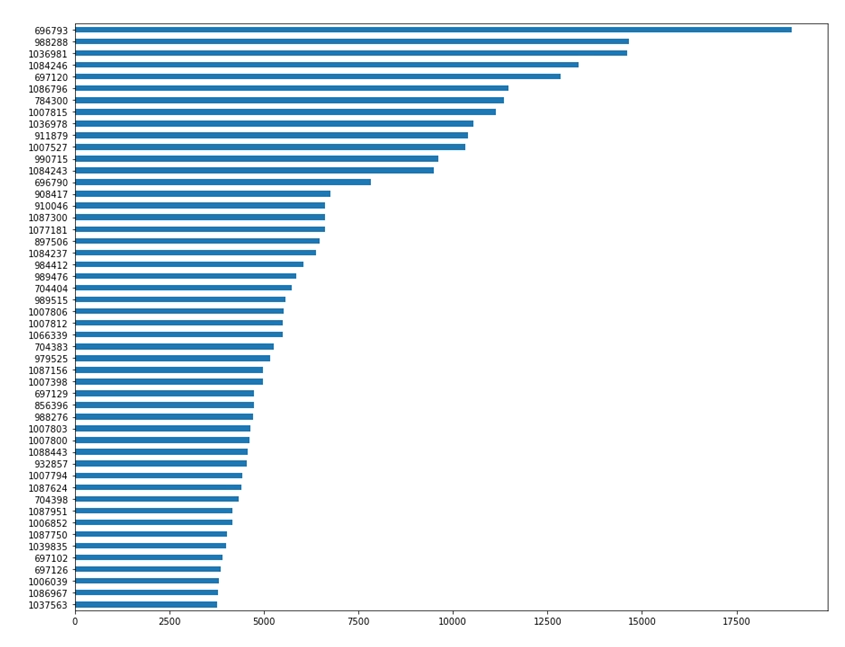
Dựa trên biểu đồ hình 3.4, phân bố video phổ biến trong tháng 11, có lượt xem từ 15.000 đến hơn 80.000 lượt xem. Trong xếp hạng top 50 video được xem nhiều nhất trong tháng 11. Với đặc trưng của video: Video thường có thời lượng ngắn, thường là vài phút hoặc thậm chí chỉ vài giây. Mục tiêu là truyền đạt thông điệp một cách nhanh chóng để giữ sự chú ý của người xem.



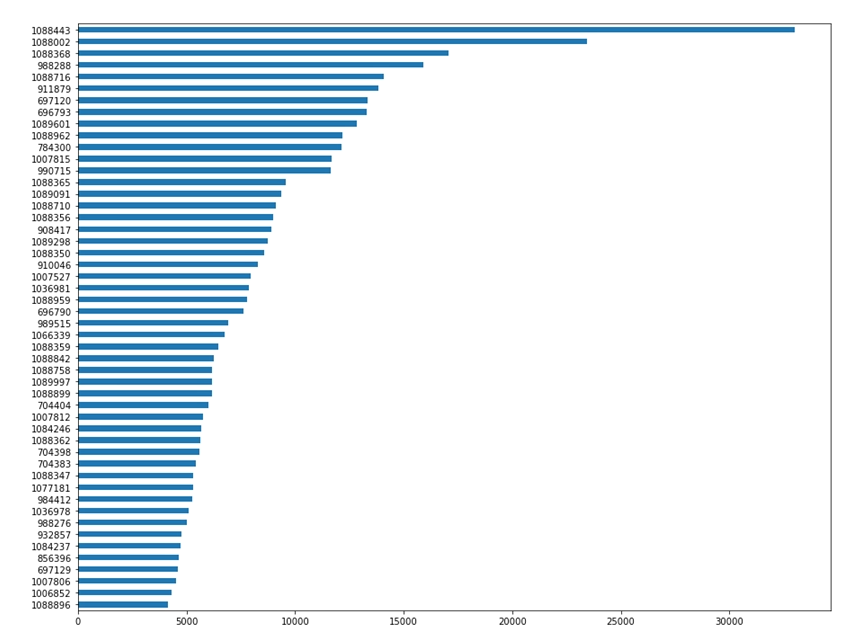
Phân bố top 50 video được xem nhiều nhất trong tuần 1



Phân bố top 50 video được xem nhiều nhất trong tuần 2



Phân bố top 50 video được xem nhiều nhất trong tuần 3



Phân bố top 50 video được xem nhiều nhất trong tuần 4

Trong 4 tuần của tháng 11, phân phối của 4 tuần qua các biểu đồ được duy trì ổn định. Đối với tuần 1, tuần 3 và tuần 4, lượt xem ở mức từ 5.000 – 30.000. Tại tuần thứ 2, lượt xem ở mức từ 10.000 – 60.000 cao hơn gấp đôi so với bình thường. Đối với tính ngắn và cô dọng của video, chúng tôi nhận định vài điểm như sau:

* Số lượng video thường được TV360 cập nhật thường xuyên thông qua các ngày, tuy nhiên, các video trending thường luôn đứng top đầu và duy trì trong nhiều tuần.
* Lượng video đứng top đầu thường liên quan về chủ đề trẻ em.
* Tại tuần 3, video mang mã id là 1088443 đạt gần 5.000 lượt xem nhưng đến tuần 4, video đã đứng đầu với lượt xem hơn 30.000 lượt xem.

Hạn chế về vấn đề kỹ thuật:

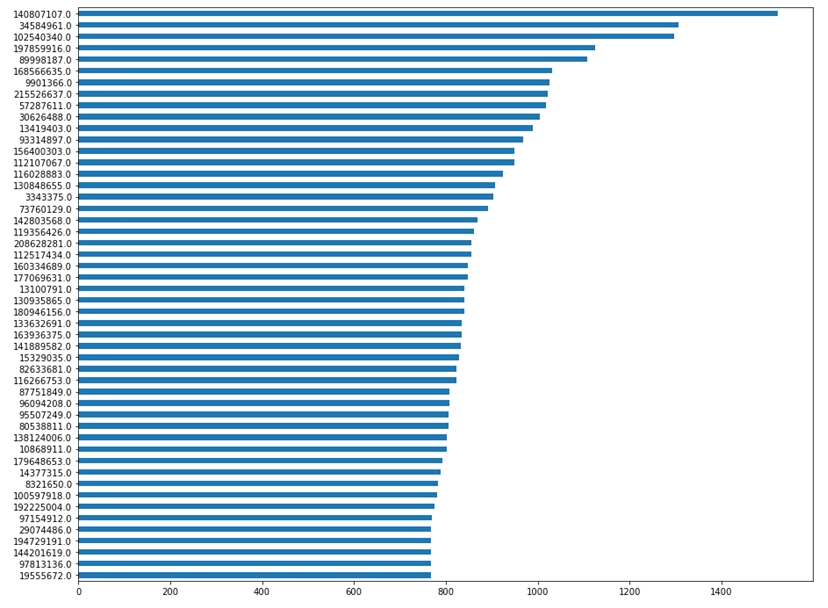
* Hiện tại, chúng tôi chưa tìm thấy được log lịch sử gợi ý phim của mô hình học sâu như RecVAE và Bert4Rec. Việc truy vết lại được log mô hình giúp chúng tôi kiểm tra được chất lượng mô hình và kiểm tra được các yếu tố khiến phim được xem và yêu thích nhiều hơn.

### Phân tích dữ liệu về video ít tương tác

Những video có ít tương tác trong hệ thống có lượng tương tác là 1. Đây là những video được gọi là cold-start video. Tuy nhiên, đối mặt với hệ thống lớn, có rất nhiều video được coi là cold-start item là việc bình thường. Để thúc đẩy cold-start item, việc cần làm là đánh giá nội dung và phân loại nội dung theo video hoặc thumball để gợi ý những video có mức độ tương tự với những video có gần sở thích người dùng. Ví dụ, video từ youtube, đã đề cử những video có lượt tương tác hạn chế nhưng nội dung gần gũi với những nội dung mà người dùng theo dõi.

Nếu chỉ dựa trên mình trường văn bản để đánh giá xếp loại và gợi ý rất khó để có thể làm active được những item mang tính cold-start.

## Phân tích dữ liệu người dùng

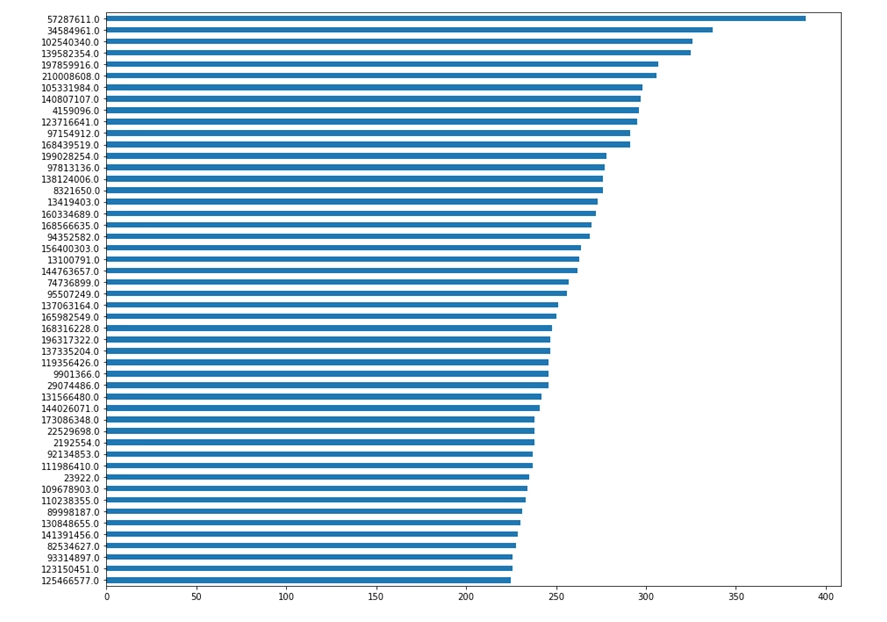


Phân bố top 50 người dùng xem video nhiều nhất trong tháng 11.

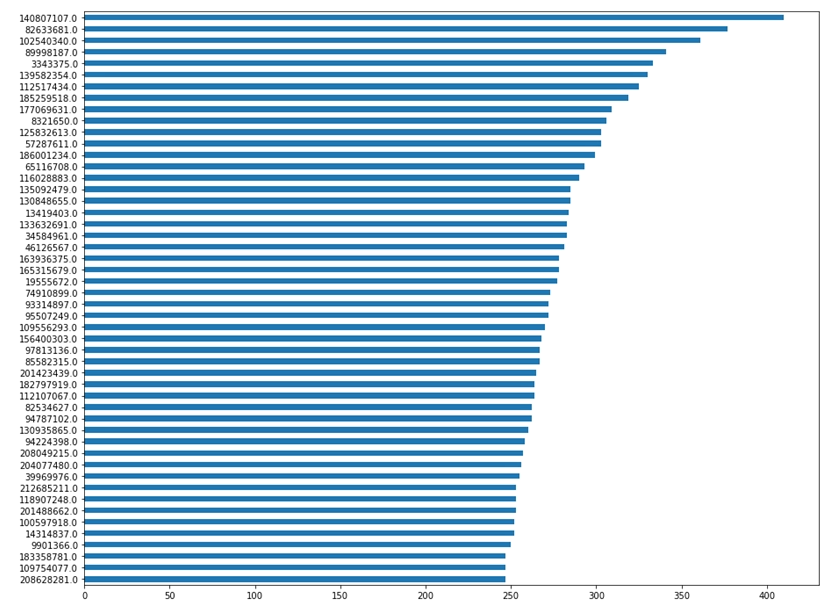
Trong tháng 11, dựa trên thống kê của chúng tôi, hiện có 208.994 người dùng sử dụng video trên TV360. Trong đó, lượng người xem activate là 45.538 và lượng người xem được coi là warm-activate là 163.456. Người dùng được xem là activate nếu xem từ 5 video trở lên trong tháng và người dùng warm-activate là người dùng xem dưới 5 video trong tháng.

Dựa trên biểu đồ 3.10, lượt xem của người dùng xem video trong tháng 11 trong top 50 dao động từ gần 800 – 1450 lượt xem.

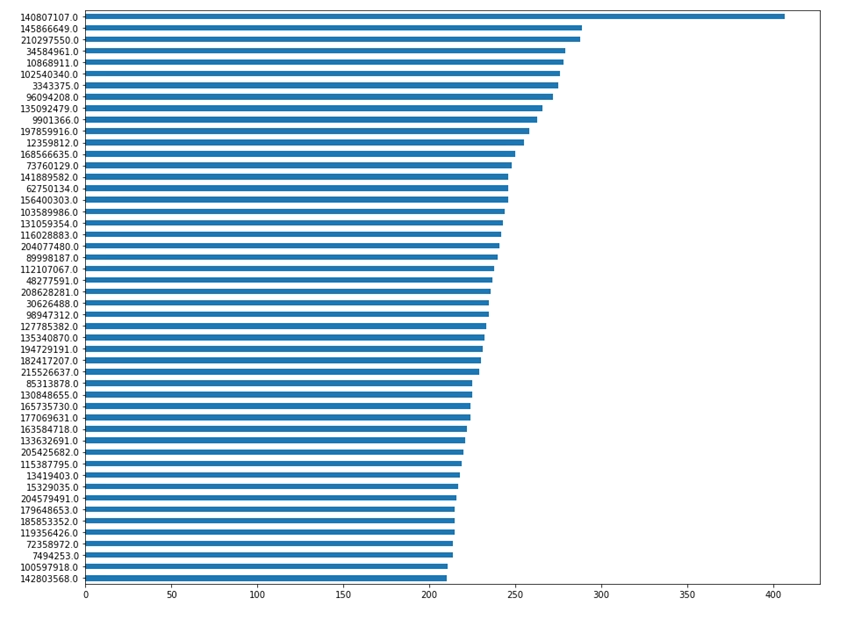
### Phân tích nhóm người dùng nhiều tương tác

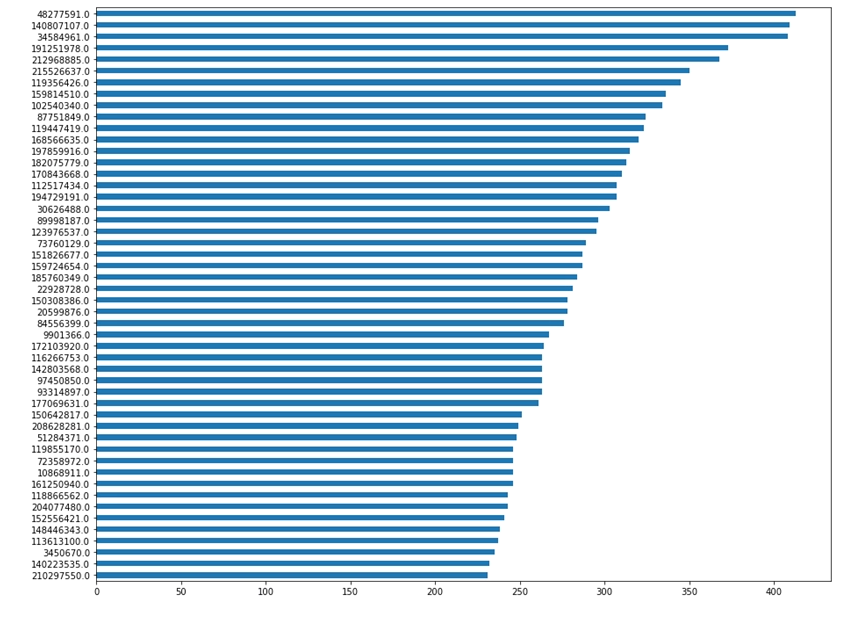


Phân phối người dùng theo tương tác trong tuần 1



Phân phối người dùng theo tương tác trong tuần 2



Phân phối người dùng theo tương tác trong tuần 3  


Phân phối người dùng theo tương tác trong tuần 4

Dựa trên những biểu đồ trên, xếp hạng 50 những người dùng xem video nhiều nhất theo thời gian tính đơn vị tuần dao đông từ 200 đến 400 tương tác. Người có lượt tương tác cao nhất trên các tuần là hơn 400 tơng tác. Lượng tương tác cao và duy trì đều trong nhiều tuần là những người dùng nằm trong top 10. Với lượng tương tác đạt mức cao trên 400 video trên tuần.

### Phân tích nhóm người dùng ít tương tác

Tương tự như hiện trạng của phim, nhưng người dùng sử dụng video ít tương tác chiếm phần lớn. Đa phần chỉ hoạt động với 1 tương tác với video. Qua đó, việc gợi ý để hấp dẫn người dùng trở nên khó khăn hơn. Việc gợi ý này có thể thông qua gợi ý chéo thông qua tương tác người dùng trên các tab khác như phim hoặc kênh truyền hình.

## Đánh giá hiện trạng dữ liệu

Đối với hiện trạng dữ liệu, vấn đề lớn nhất hiện nay của video chính là phân loại thể loại video đang còn hỗn loạn. Video có những trường mang tính thể loại nhưng bị sai mục đích hoặc chưa đầy đủ. Một số hiện trạng như sau:

* Category\_name: mang 1 thể loại cho 1 video.
* Sub\_category\_name: mang nhiều thể loại con nhưng đan lẫn với các dữ liệu mang tính trending hoặc tagging

Lượng người dùng tương tác với video đạt mức tốt lên đến 400 lượt tương tác trên tuần đối với nhóm đầu.

# KHẢO SÁT MÔ HÌNH

## Mô hình RecVAE

### Cơ sở lý thuyết

Phương pháp được đề xuất, RecVAE, là một mô hình bộ mã hóa tự động biến thể mới cho các đề xuất top-N có phản hồi ngầm. Nó giới thiệu một số ý tưởng mới để cải thiện các mô hình hiện có, bao gồm:

* Một phân phối trước tổng hợp mới cho các mã tiềm ẩn
* Một cách tiếp cận mới để thiết lập siêu tham số β cho khung β-VAE
* Một cách tiếp cận mới để đào tạo dựa trên các cập nhật xen kẽ

RecVAE đã được chứng minh là vượt trội đáng kể so với các mô hình dựa trên bộ mã hóa tự động được đề xuất trước đây trong đánh giá thử nghiệm.

Mô hình Mult-VAE là một biến thể của Bộ mã hóa tự động biến đổi (VAE) được sử dụng để lọc cộng tác, một kỹ thuật được sử dụng trong các hệ thống đề xuất. Ý tưởng cơ bản của mô hình Mult-VAE tương tự như VAE, nhưng nó sử dụng phân phối đa thức làm hàm khả năng thay vì phân phối Gaussian và Bernoulli thường được sử dụng trong VAE. Trong mô hình Mult-VAE, biểu diễn tiềm ẩn k chiều, ký hiệu là zu, được lấy mẫu cho mỗi người dùng u. Biểu hiện tiềm ẩn này nắm bắt các đặc điểm hoặc sở thích cơ bản của người dùng. Biểu diễn tiềm ẩn zu sau đó được biến đổi bằng cách sử dụng hàm fθ : Rk → R|I|, trong đó θ biểu thị các tham số của hàm biến đổi. Biểu diễn được chuyển đổi được sử dụng để tạo ra các đề xuất cho người dùng. Lịch sử phản hồi xu của người dùng u, bao gồm các tương tác nu như nhấp chuột hoặc mua hàng, được giả sử là được rút ra từ phân phối đa thức. Phân phối này thể hiện xác suất mỗi mục được người dùng tương tác. Mô hình Mult-VAE nhằm mục đích tìm hiểu các tham số θ giúp tối đa hóa khả năng xảy ra lịch sử phản hồi được quan sát xu với biểu diễn tiềm ẩn zu và hàm biến đổi fθ.

So sánh với VAE: Sự khác biệt chính giữa mô hình Mult-VAE và VAE nằm ở việc lựa chọn hàm khả năng. Trong khi VAE thường sử dụng phân phối Gaussian hoặc Bernoulli thì Mult-VAE sử dụng phân phối đa thức. Phân phối đa thức phù hợp để mô hình hóa dữ liệu rời rạc, chẳng hạn như sự tương tác giữa người dùng và các mục trong hệ thống đề xuất. Bằng cách sử dụng phân phối đa thức, mô hình Mult-VAE có thể nắm bắt tốt hơn các mẫu và tùy chọn trong tương tác giữa người dùng và mục, dẫn đến các đề xuất được cải thiện.

Ưu điểm của Multi-VAE: Mô hình Mult-VAE đã cho thấy kết quả xuất sắc đối với các đề xuất top-N, vượt trội so với các mô hình dựa trên bộ mã hóa tự động khác trong các tác vụ lọc cộng tác. Việc sử dụng phân phối đa thức làm hàm khả năng cho phép Mult-VAE mô hình hóa tốt hơn bản chất rời rạc của các tương tác giữa người dùng và vật phẩm, dẫn đến các đề xuất chính xác hơn. Mô hình Mult-VAE rất linh hoạt và có thể được áp dụng cho nhiều bộ dữ liệu lọc cộng tác khác nhau, khiến nó trở thành lựa chọn linh hoạt cho các hệ thống đề xuất.

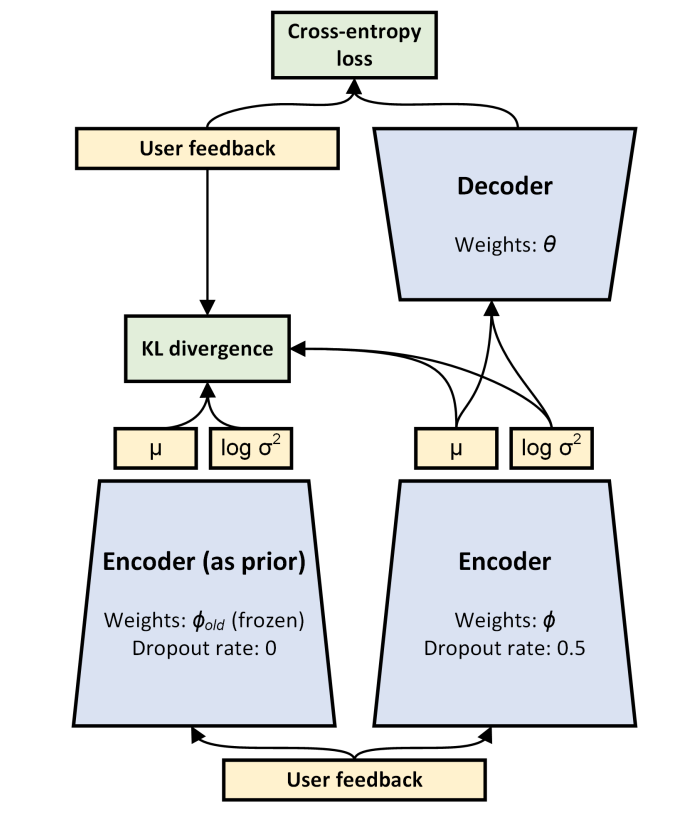
Kiến trúc mô hình của RecVAE: RecVAE là mô hình bộ mã hóa tự động biến thiên (VAE) dành cho các đề xuất top-N có phản hồi ngầm. Nó dựa trên mô hình Mult-VAE nhưng giới thiệu một số ý tưởng mới để cải thiện nó. Kiến trúc chung của RecVAE bao gồm một bộ mã hóa tự động biến thiên khử nhiễu, với sự phân phối trước tổng hợp mới cho các mã tiềm ẩn. Mạng suy luận của RecVAE sử dụng các lớp được kết nối dày đặc, các hàm kích hoạt chuyển động và chuẩn hóa lớp. Mạng giải mã là một lớp tuyến tính đơn giản với kích hoạt softmax. Nhiễu đầu vào cho bộ mã hóa tự động khử nhiễu và nhiễu biến đổi tiềm ẩn của tích hợp Monte Carlo đóng các vai trò khác nhau trong mô hình. Mô hình này vượt trội hơn các mô hình dựa trên bộ mã hóa tự động trước đây, bao gồm Mult-VAE và RaCT, về chất lượng đề xuất trên các bộ dữ liệu lọc cộng tác cổ điển.

Mô hình Mult-VAE là một biến thể của Bộ mã hóa tự động biến đổi (VAE) được sử dụng để lọc cộng tác, đây là một kỹ thuật được sử dụng trong các hệ thống đề xuất. Trong mô hình Mult-VAE, hàm khả năng khác với các phân bố Gaussian và Bernoulli thường được sử dụng trong VAE. Thay vào đó, nó sử dụng phân phối đa thức làm hàm khả năng. Mô hình tổng quát của Mult-VAE bắt đầu bằng cách lấy mẫu biểu diễn tiềm ẩn k chiều (zu) cho người dùng (u). Biểu diễn tiềm ẩn này sau đó được chuyển đổi bằng cách sử dụng hàm (fθ) được tham số hóa bởi θ. Lịch sử phản hồi (xu) của người dùng, bao gồm các tương tác như nhấp chuột hoặc mua hàng, được giả định là được rút ra từ phân phối đa thức.

(1)

(2)

Các phương trình (1) và (2) biểu thị quá trình tổng hợp của Mult-VAE. Phương trình (1) cho thấy biểu diễn tiềm ẩn (zu) được lấy mẫu từ phân phối Gaussian với giá trị trung bình là 0 và ma trận hiệp phương sai nhận dạng (). Phương trình (2) cho thấy lịch sử phản hồi (xu) được rút ra từ phân phối đa thức với các tham số (nu) và (π(zu)), trong đó () thu được bằng cách áp dụng hàm softmax cho biểu diễn tiềm ẩn được chuyển đổi ().



Kiến trúc của RecVAE

Điều quan trọng cần lưu ý là các mô hình lọc cộng tác cổ điển cũng tuân theo sơ đồ tương tự, nhưng có hàm tuyến tính (fθ). Tính linh hoạt bổ sung của Mult-VAE đến từ việc tham số hóa hàm (fθ) với mạng thần kinh có tham số θ. Điều này cho phép thực hiện các phép biến đổi phức tạp hơn và phi tuyến tính hơn của biểu diễn tiềm ẩn.

Nhìn chung, mô hình Mult-VAE là một biến thể của VAE sử dụng phân phối đa thức làm hàm khả năng và tham số hóa hàm biến đổi bằng mạng nơ-ron. Điều này mang lại sự linh hoạt hơn và cho phép lập mô hình phản hồi của người dùng tốt hơn trong các hệ thống đề xuất.

**Đầu vào và đầu ra của Multi-VAE**

Mô hình Mult-VAE lấy các vectơ thưa chiều cao làm cả đầu vào và đầu ra. Các vectơ thưa chiều cao đề cập đến các vectơ có số lượng kích thước (đặc điểm) lớn nhưng hầu hết các giá trị trong vectơ đều bằng 0 hoặc bị thiếu. Trong ngữ cảnh lọc cộng tác, các vectơ này thể hiện sự tương tác giữa người dùng và mục, trong đó mỗi thứ nguyên tương ứng với một mục và giá trị thể hiện sở thích hoặc xếp hạng của người dùng đối với mục đó. Do các bộ dữ liệu lọc cộng tác thường thưa thớt, với hầu hết người dùng chỉ tương tác với một tập hợp con nhỏ các mục, nên các vectơ đầu vào và đầu ra của Mult-VAE cũng thưa thớt.

Mult-VAE sử dụng phần sau gần đúng được khấu hao được chia sẻ, giúp điều chỉnh quá trình học tập. Chính quy hóa là một kỹ thuật được sử dụng để ngăn chặn việc trang bị quá mức và cải thiện tính tổng quát hóa của mô hình. Tuy nhiên, các cập nhật sau cho một số phần của dữ liệu được quan sát có thể ảnh hưởng tiêu cực đến các tham số biến thiên tương ứng với các phần khác của dữ liệu. Điều này có nghĩa là khi cập nhật mô hình dựa trên dữ liệu mới quan sát được, các tham số đã học trước đó đối với một số phần của dữ liệu có thể bị “quên” hoặc ghi đè, dẫn đến mất ổn định trong quá trình huấn luyện. Hiệu ứng này tương tự như hiệu ứng "quên" nổi tiếng trong học tập tăng cường, trong đó mô hình quên những hành vi tốt đã học trước đó khi học những hành vi mới. Chính quy hóa các tham số mô hình Công việc trước đây về học tăng cường dựa trên chính sách đã chỉ ra rằng việc chuẩn hóa các tham số mô hình bằng cách đưa chúng đến gần hơn với các tham số từ kỷ nguyên trước có thể giúp ngăn chặn hiệu ứng quên. Trong học tăng cường, điểm cuối cùng của mô hình dự kiến sẽ tăng dần theo thời gian, cho thấy mô hình đang học tập và cải thiện. Bằng cách thường xuyên hóa các tham số của mô hình, điểm cuối cùng có thể tăng trơn tru hơn, giúp mô hình không quên những hành vi tốt đã học được ở các kỷ nguyên trước. Kỹ thuật chính quy hóa này giúp ổn định quá trình đào tạo và cải thiện hiệu suất tổng thể của mô hình.

Nhìn chung, ề những thách thức và sự mất ổn định tiềm ẩn trong quá trình đào tạo mô hình Mult-VAE do các vectơ đầu vào và đầu ra thưa thớt có chiều cao, cũng như hiệu ứng quên tiềm ẩn trong quá trình cập nhật sau. Nó gợi ý rằng việc thường xuyên hóa các tham số mô hình có thể giúp giảm thiểu những vấn đề này và cải thiện hiệu suất của mô hình.

### Quy mô dữ liệu

Dữ liệu đầu vào:

|  |  |  |
| --- | --- | --- |
| Tập huấn luyện | 121.150 người dùng | |
| Tập item | 69.737 item | |
| Tập người dùng kiểm thử | Masked | 16.163 người dùng |
| No masked | 18.243 người dùng |
| Tập người dùng warm | Masked | 3.022 người dùng |
| No masked | 3.574 người dùng |

### Kết quả mô hình

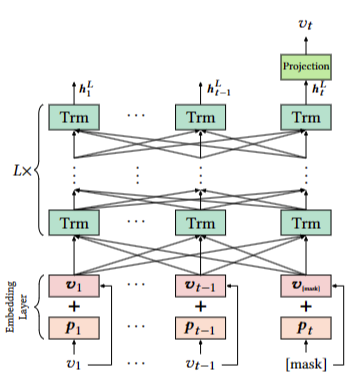
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Tập kiểm thử** | **Tập warm** | **Tập cold** |
| **Baseline** | 0.36 | | |
| **Masked** | 0.1841 | 0.2394 | 0.1055 |
| **No Masked** | **0.2521** | **0.3733** | **0.1004** |

## Mô hình Bert4Rec

### Cơ sở lý thuyết

Việc lập mô hình sở thích động của người dùng từ các hành vi lịch sử của họ là một thách thức và quan trọng đối với các hệ thống đề xuất. Các phương pháp trước đây sử dụng các mô hình một chiều từ trái sang phải để mã hóa các tương tác lịch sử của người dùng, nhưng các mô hình này có những hạn chế về sức mạnh của các biểu diễn ẩn và giả định về các chuỗi có thứ tự cứng nhắc. Để giải quyết những hạn chế này, bài viết đề xuất BERT4Rec, một mô hình đề xuất tuần tự sử dụng khả năng tự chú ý hai chiều sâu sắc để mô hình hóa chuỗi hành vi của người dùng. BERT4Rec sử dụng mục tiêu Cloze để huấn luyện mô hình hai chiều, dự đoán các mục bị che trong chuỗi bằng cách điều chỉnh ngữ cảnh bên trái và bên phải của chúng. Mô hình học cách trình bày hai chiều để đưa ra đề xuất bằng cách kết hợp thông tin từ cả bên trái và bên phải của từng mục trong hành vi lịch sử của người dùng. Kết quả thử nghiệm trên bốn bộ dữ liệu điểm chuẩn cho thấy BERT4Rec vượt trội hơn nhiều mô hình tuần tự tiên tiến khác nhau một cách nhất quán.

BERT4Rec là mô hình khuyến nghị tuần tự. Nó sử dụng khả năng tự chú ý hai chiều sâu sắc để mô hình hóa chuỗi hành vi của người dùng. Nó giải quyết các hạn chế của mô hình một chiều bằng cách cho phép từng mục trong hành vi lịch sử của người dùng kết hợp thông tin từ cả bên trái và bên phải. Mô hình sử dụng mục tiêu Cloze để dự đoán các mục được che dấu ngẫu nhiên trong chuỗi, dựa trên bối cảnh bên trái và bên phải của chúng. BERT4Rec vượt trội hơn nhiều mô hình tuần tự tiên tiến khác nhau về hiệu suất đề xuất.



Kiến trúc của mô hình BERT4REC

Quá trình tính toán các biểu diễn ẩn cho từng vị trí trong chuỗi đầu vào bằng cách sử dụng lớp Transformer.

Chuỗi đầu vào có độ dài 't' và mục tiêu là tính toán các biểu diễn ẩn 'hl' cho từng vị trí 'i' cùng một lúc. Các biểu diễn ẩn 'hl' được tính toán lặp lại ở mỗi lớp 'l' của mô hình Transformer. Các biểu diễn ẩn 'hl' được xếp chồng lên nhau thành một ma trận 'Hl' có kích thước 't × d', trong đó 'd' biểu thị số chiều của các biểu diễn ẩn. Chức năng chú ý được áp dụng đồng thời cho tất cả các vị trí, cho phép tính toán hiệu quả. Lớp Transformer bao gồm hai lớp phụ: Mạng tự chú ý nhiều đầu và Mạng chuyển tiếp nguồn cấp dữ liệu theo vị trí. Lớp con Tự chú ý nhiều đầu chịu trách nhiệm nắm bắt sự phụ thuộc giữa các vị trí khác nhau trong chuỗi đầu vào. Nó tính toán trọng số chú ý cho từng vị trí bằng cách xem xét mối quan hệ giữa tất cả các vị trí trong chuỗi. Lớp con Mạng chuyển tiếp nguồn cấp dữ liệu theo vị trí áp dụng một phép biến đổi phi tuyến tính cho biểu diễn ẩn của từng vị trí một cách độc lập. Lớp con này giúp nắm bắt các mẫu và tương tác phức tạp trong biểu diễn ẩn của từng vị trí.

Lớp Transformer trong mô hình BERT4Rec không có bất kỳ mô-đun lặp lại hoặc tích chập nào, có nghĩa là nó không có cơ chế tích hợp để xem xét thứ tự của chuỗi đầu vào. Đây là một vấn đề vì trong đề xuất tuần tự, thứ tự tương tác lịch sử của người dùng rất quan trọng trong việc hiểu sở thích của họ và đưa ra đề xuất chính xác.

Để giải quyết vấn đề này, mô hình BERT4Rec đưa các phần nhúng vị trí vào các phần nhúng mục đầu vào ở cuối ngăn xếp lớp Transformer. Các phần nhúng vị trí này giúp mô hình mã hóa thông tin tuần tự của đầu vào bằng cách chỉ ra vị trí của từng mục trong chuỗi. Biểu diễn đầu vào với các phần nhúng mục và vị trí. Đối với mỗi mục vi trong chuỗi đầu vào, biểu diễn đầu vào h0i của nó được xây dựng bằng cách tính tổng mục vi nhúng tương ứng và pi nhúng vị trí. Mục nhúng vi là một vectơ d chiều đại diện cho mục vi, trong khi pi nhúng vị trí cũng là một vectơ d chiều đại diện cho chỉ số vị trí i của mục trong chuỗi. Tổng của hai phần nhúng này tạo ra biểu diễn ban đầu h0i cho mục vi, mục này sẽ được xử lý thêm bởi lớp Transformer. Trong mô hình BERT4Rec, các phần nhúng vị trí có thể học được được sử dụng thay vì các phần nhúng hình sin cố định để có hiệu suất tốt hơn. Ma trận nhúng vị trí P là ma trận có kích thước N x d, trong đó N là độ dài câu tối đa mà mô hình có thể xử lý và d là số chiều của các phần nhúng. Các phần nhúng vị trí có thể học được cho phép mô hình xác định phần nào của chuỗi đầu vào mà nó đang xử lý, giúp mô hình nắm bắt thông tin tuần tự một cách hiệu quả. Tuy nhiên, việc sử dụng các phần nhúng vị trí cũng đặt ra một hạn chế về độ dài câu tối đa N mà mô hình có thể xử lý. Nếu chuỗi đầu vào [v1, ..., vt] dài hơn N thì cần cắt bớt N mục cuối cùng [vu t −N +1, ..., vt]. Việc cắt bớt này đảm bảo rằng mô hình chỉ xem xét N mục gần đây nhất trong các tương tác lịch sử của người dùng, đây là một cách tiếp cận thực tế khi xem xét các hạn chế của việc nhúng vị trí.

### Quy mô dữ liệu

|  |  |
| --- | --- |
| Tập người dùng | 580.859 người dùng |
| Tập item | 40.390 item |
| Tập huấn luyện | 9.096.287 tương tác |
| Tập kiểm thử | 321.905 tương tác |

### Kết quả mô hình

|  |  |  |
| --- | --- | --- |
|  | **Tập kiểm thử** | |
|  | **Recall@50** | **NDCG@50** |
| **Bert4REC** | 0.3206 | 0.1215 |

# Đề xuất cải tiến

Đối với những hạn chế và nhược điểm của hệ thống hiện tại. Chúng tôi tổng hợp lại những vấn đề và các giải pháp tương ứng cho từng vấn đề như sau:

* Vấn đề thiên kiến phổ biến: Mô hình ưu tiên hiển thị hoặc đề xuất những mục được xem nhiều, mua nhiều, hoặc được đánh giá cao bởi đa số người dùng. Điều này có thể dẫn đến việc tạo ra một vòng tuần hoàn, trong đó các mục phổ biến trở nên ngày càng phổ biến hơn, trong khi các mục ít phổ biến ít có cơ hội được thấy hoặc được khám phá bởi người dùng.
* Vấn đề về hỗn loạn dữ liệu thể loại: Hệ thống đang có dữ liệu hỗn loạn về dữ liệu thể loại video với các thể loại chưa được gán chính xác.
* Các chiến lược ensemble
* Các vấn đề cold-start: Các vấn đề về cold-start item và cold-start user
* Giải pháp cá nhân hoá cho từng người dùng

**Giải pháp cho các vấn đề:**

Thống nhất **cách đánh giá mới để phù hợp** với các kịch bản và cơ chế hệ thống đang triển khai. Hệ thống đang triển khai việc sử dụng dữ liệu quá khứ để dự đoán cho những sản phẩm ở tương lai gần. Vậy nên chiến lược kiểm thử cần cho mô hình học hết các hành vi của người dùng có hoạt động trong hệ thống (không tính người dùng ở tập warm và tập cold) sau đó, gợi ý cho các dự đoán tương lai.

**Vấn đề về phân loại thể loại video:**

* Đề xuất phân loại thể loại video theo bộ thể loại video chung. Việc định nghĩa bộ thể loại video chung cũng sẽ được thực hiện và thống nhất. Mỗi video sẽ được gán đa nhãn.
* Đề xuất xây dựng bài toán gán nhãn tagging cho video. Việc phân loại tagging cho video sẽ giúp cải thiện hiệu suất tìm kiếm, tăng khả năng gợi ý video cho người dùng.
* Đề xuất xây dựng bài toán gán nhãn trending cho video. Việc gán nhãn trending cho video có thể ứng dụng cho nhiều bài toán như gợi ý video dựa trên trending hay gợi ý đề xuất tìm kiếm dựa trên các nhãn trending.

**Vấn đề về sản phẩm mới cho người dùng:**

* Chúng tôi đề xuất cách giải quyết vấn đề này bằng việc sử dụng kết quả từ việc embedding cho các item sau đó tìm ra những sản phẩm có độ tương đồng cao. Từ đó, sử dụng danh sách sản phẩm đưa qua mô hình học sâu để dự đoán khả năng yêu thích của mô hình.
* Việc đánh giá embedding có thể mở rộng cải thiện trong tương lai bằng việc biểu diễn thêm các nội dung của video hoặc các đặc trưng video như: thời lượng, thumball.

**Vấn đề về người dùng mới cho dịch vụ**:

* Đề xuất giải pháp khuyến nghị chéo dựa trên các hành vi người dùng dựa trên lịch sử tương tác với các dịch vụ khác
* Đề xuất gợi ý dựa trên hồ sơ embedding của người dùng để phân loại gợi ý phù hợp với người dùng mới

**Đề xuất giải pháp ensemble,** hiện tại chúng tôi đã thử nghiệm các kịch bản và phương pháp kết hợp mô hình. Chúng tôi đề xuất kết quả A/B testing áp dụng cho kết quả trong tháng tới để tiến hành triển khai và tối ưu mô hình ensemble.