**Mục lục**

[**1. Mục tiêu của bài toán** 2](#_Toc179963405)

[**1.1. Đầu vào của giải pháp** 2](#_Toc179963406)

[**1.2. Đầu ra của giải pháp** 2](#_Toc179963407)

[**2. Đề xuất kiến trúc giải pháp** 2](#_Toc179963408)

[**2.1. Kiến trúc chung** 2](#_Toc179963409)

[**2.2. Mô tả các nhóm giải pháp** 2](#_Toc179963410)

[**3. Chỉ số đánh giá giải pháp được đề xuất** 3](#_Toc179963411)

[**4. Nhóm giải pháp: Thống kê độ phổ biến** 4](#_Toc179963412)

[**4.1. Mô tả chi tiết dữ liệu đầu vào** 4](#_Toc179963413)

[**4.2. Mô tả chi tiết dữ liệu đầu ra** 4](#_Toc179963414)

[**4.3. Huấn luyện mô hình** 4](#_Toc179963415)

[**4.4. Kết quả mô hình cuối** 4](#_Toc179963416)

[**4.5. Các hạn chế và đề xuất cải tiến** 4](#_Toc179963417)

[**5. Nhóm giải pháp: Sử dụng lọc cộng tác** 5](#_Toc179963418)

[**5.1. Mô tả chi tiết dữ liệu đầu vào** 5](#_Toc179963419)

[**5.2. Mô tả chi tiết dữ liệu đầu ra** 5](#_Toc179963420)

[**5.3. Khám phá dữ liệu** 5](#_Toc179963421)

[**5.4. Chia tập** 5](#_Toc179963422)

[**5.5. Chuẩn bị dữ liệu** 5](#_Toc179963423)

[**5.6. Huấn luyện mô hình** 5](#_Toc179963424)

[**5.7. Chi tiết tracking model** 6](#_Toc179963425)

[**5.8. Kết quả mô hình cuối** 6](#_Toc179963426)

[**5.9. Các hạn chế và đề xuất cải tiến** 6](#_Toc179963427)

[**6. Nhóm giải pháp: Sử dụng ma trận nhân tử** 6](#_Toc179963428)

[**6.1. Mô tả chi tiết dữ liệu đầu vào** 6](#_Toc179963429)

[**6.2. Mô tả chi tiết dữ liệu đầu ra** 6](#_Toc179963430)

[**6.3. Khám phá dữ liệu** 7](#_Toc179963431)

[**6.4. Chia tập** 7](#_Toc179963432)

[**6.5. Chuẩn bị dữ liệu** 7](#_Toc179963433)

[**6.6. Huấn luyện mô hình** 7](#_Toc179963434)

[**6.7. Chi tiết tracking model** 7](#_Toc179963435)

[**6.8. Kết quả mô hình cuối** 7](#_Toc179963436)

[**6.9. Các hạn chế và đề xuất cải tiến** 7](#_Toc179963437)

**1. Mục tiêu của bài toán**

Bài toán xây dựng hệ tư vấn dựa trên thông tin và hành vi sử dụng của người dùng nhằm dự đoán và đưa ra danh sách các gợi ý tiếp theo phù hợp. Những giải pháp được đề xuất trong tài liệu này sẽ được sử dụng làm mô hình baseline cho quá trình phát triển trong tương lai.

## **1.1. Đầu vào của giải pháp**

Những dữ liệu được đưa vào đầu vào của mô hình bao gồm:

* Dữ liệu thông tin phim, video, chương trình
* Dữ liệu thông tin người dùng
* Dữ liệu thông tin tương tác giữa người dùng với các mục

## **1.2. Đầu ra của giải pháp**

Danh sách những phim, video, chương trình phù hợp nhất cho người dùng.

# **2. Đề xuất kiến trúc giải pháp**

## **2.1. Kiến trúc chung**

Những mô hình được lựa chọn làm baseline cho quy trình phát triển hệ tư vấn bao gồm:

* Mô hình dựa trên thống kê độ phổ biến: Đây là một mô hình dễ triển khai và đặc biệt hữu ích cho những người dùng mới nhờ khả năng đề xuất các mục đồng thời ngay lập tức. Đồng thời, khả năng tính toán nhanh chóng của mô hình phù hợp với điều kiện đề xuất thời gian thực.
* Mô hình lọc cộng tác: Mô hình có thể tận dụng thông tin hành vi người dùng để tìm những mục hoặc người dùng tương tự để cá nhân hóa các đề xuất. Đây là mô hình dễ dàng mở rộng khi dữ liệu lớn, phù hợp với các hệ tư vấn thời gian thực có lượng người dùng và sản phẩm đa dạng.
* Mô hình ma trận nhân tử: Đây là mô hình đem lại hiệu quả với dữ liệu thưa, khai thác thông tin từ những tương tác có sẵn, từ đó đưa ra các đề xuất mang tính cá nhân hóa cao hơn. Dù yêu cầu tính toán cao hơn, mô hình đem lại hiệu năng tốt với quy mô lớn và có thể được tối ưu hóa để sử dụng trong thời gian thực.

## **2.2. Mô tả các nhóm giải pháp**

|  |  |  |  |
| --- | --- | --- | --- |
| Nhóm giải pháp | Đặc điểm dữ liệu đầu vào | Đặc điểm dữ liệu đầu ra | Mô tả giải pháp |
| Thống kê độ phổ biến | - Thông tin tương tác giữa người dùng và mục  - Thông tin phim, video, chương trình | Danh sách phim, video, chương trình phù hợp cho người dùng | Mô hình thu thập thông tin số lượng cũng như điểm đánh giá của từng phim để sắp xếp độ phổ biến và gợi ý cho người dùng. |
| Lọc cộng tác | Mô hình tính toán độ tương đồng giữa các người dùng và các mục để đưa ra gợi ý có độ tương đồng cao. |
| Ma trận nhân tử | Mô hình điền các giá trị còn thiếu trong dữ liệu ma trận tương tác đầu vào giữa người dùng và mục, từ đó đưa ra gợi ý mới có điểm số cao. |

# **3. Chỉ số đánh giá giải pháp được đề xuất**

|  |  |  |
| --- | --- | --- |
| Tên chỉ số đánh giá | Định nghĩa chỉ số đánh giá | Nhóm giải pháp sử dụng chỉ số đánh giá |
| RMSE | Đo lường độ chính xác của mô hình dự đoán bằng cách đo độ lệch chuẩn của sai số. |  |
| Precision@K | Đo lường độ chính xác của đề xuất dựa trên số lượng mục phù hợp với người dùng |  |
| Recall@K | Đo lường khả năng tìm ra các mẫu tích cực, cho biết tỉ lệ mô hình có thể tìm được các kết quả thực sự mong muốn. |  |
| NDCG@K | Đo lường độ chính xác cũng như thứ tự sắp xếp của các mục đề xuất. |  |
| Thời gian thực thi | Đo lường thời gian huấn luyện và xử lý dữ liệu |  |
| Đánh giá người dùng | Người dùng thử nghiệm và đánh giá chất lượng của gợi ý |  |

# **4. Nhóm giải pháp: Thống kê độ phổ biến**

## **4.1. Mô tả chi tiết dữ liệu đầu vào**

## **4.2. Mô tả chi tiết dữ liệu đầu ra**

Đầu ra của mô hình là danh sách 5 mục mà mô hình đánh giá là phù hợp nhất để đề xuất cho người dùng.

## **4.3. Huấn luyện mô hình**

Vì một số phim phổ biến hơn nhiều so với những phim còn lại, nên theo trực giác, nhiều người dùng có thể thích những phim phổ biến này hơn. Từ quan sát đó, nó truyền cảm hứng cho một cách tiếp cận đơn giản để cung cấp các đề xuất dựa trên mức độ phổ biến (tức là số lượng đánh giá) của các phim.

Để xây dựng hệ gợi ý dựa trên thống kê độ phổ biến dựa trên số lượng cũng như giá trị trung bình của đánh giá người dùng, hệ thống sẽ sử dụng công thức tính đánh giá có trọng số (WR) của IMDB để xếp hạng những mục phổ biến nhất trong tập dữ liệu:

Trong đó:

* là số lượng đánh giá của mục
* là số lượng đánh giá tối thiểu để được coi là một mục phổ biến. Trong thiết lập hệ thống, những mục thuộc top 10% nhiều đánh giá nhất trên toàn bộ tập dữ liệu sẽ được coi là mục phổ biến.
* là điểm đánh giá trung bình của mục
* là điểm đánh giá trung bình của toàn bộ tập dữ liệu

## **4.4. Kết quả mô hình cuối**

Mô hình được huấn luyện trên tập train và đánh giá trên tập test. Kết quả huấn luyện của mô hình được trình bày dưới bảng sau:

## **4.5. Các hạn chế và đề xuất cải tiến**

Những hạn chế của mô hình gợi ý dựa trên thống kê độ phổ biến bao gồm:

* Thiếu tính cá nhân hóa: Một danh sách những mục phổ biến nhất, loại trừ những mục đã được theo dõi, sẽ được đề xuất cho người dùng. Điều này dẫn đến việc danh sách gợi ý của toàn bộ người dùng gần như sẽ tương đồng với nhau.
* Bỏ qua thông tin thời gian: Những nội dung xu hướng phổ biến trong thời gian gần sẽ không được đề xuất do không đạt điều kiện cần.

Những đề xuất cải tiến cho mô hình:

* Kết hợp với các mô hình khác: Mô hình hỗn hợp có thể được xây dựng để học sở thích người dùng theo các thuộc tính mục, và đưa ra đề xuất phổ biến dựa trên những thuộc tính đó.
* Chia khoảng thời gian: Hệ thống đề xuất những mục phổ biến theo tháng/theo ngày thay vì chỉ sử dụng toàn lịch sử thời gian.

# **5. Nhóm giải pháp: Sử dụng lọc cộng tác**

## **5.1. Mô tả chi tiết dữ liệu đầu vào**

## **5.2. Mô tả chi tiết dữ liệu đầu ra**

Đầu ra của mô hình là danh sách 5 mục mà mô hình đánh giá là phù hợp nhất để đề xuất cho người dùng.

## **5.3. Khám phá dữ liệu**

## **5.4. Chia tập**

## **5.5. Chuẩn bị dữ liệu**

## **5.6. Huấn luyện mô hình**

Mặc dù đều mang đặc điểm chung là sử dụng thông tin từ người dùng hoặc mục để dự đoán những sản phẩm hoặc nội dung mà người dùng có thể quan tâm, nhóm giải pháp này được chia ra làm hai hướng xây dựng chính:

* Lọc cộng tác hướng người dùng: Lọc cộng tác hướng người dùng dựa trên việc tìm kiếm những người dùng có sở thích tương tự để gợi ý các mục mà một người dùng cụ thể có thể thích. Hệ thống sẽ phân tích hành vi hoặc đánh giá của một người dùng và tìm kiếm những người dùng có lịch sử tương tự (dựa trên các mục mà họ đã đánh giá hoặc tương tác). Sau đó, hệ thống sẽ đề xuất các mục mà những người dùng tương tự đã đánh giá cao nhưng người dùng hiện tại chưa tương tác.
* Lọc cộng tác hướng mục (phim): Lọc cộng tác hướng mục tập trung vào sự tương đồng giữa các mục, hay trong trường hợp này là phim, tức là gợi ý những mục tương tự mà người dùng có thể thích dựa trên các mục mà họ đã đánh giá cao trước đó. Hệ thống sẽ phân tích các mục mà người dùng đã tương tác và tìm những mục tương tự về nội dung hoặc hành vi của người dùng. Sau đó, nó sẽ gợi ý các mục tương tự cho người dùng dựa trên các mục mà họ đã đánh giá cao hoặc đã tương tác.

Hai cách tiếp cận phổ biến để đo lường sự tương đồng giữa hai người dùng/hai mục là độ tương đồng Cosine và độ tương quan Pearson. Ngoài ra, nếu độ tương đồng Cosine dựa trên ma trận xếp hạng lấy trung bình làm trọng tâm, một biến thể mới tên AdjustedCosine xuất hiện và sẽ được sử dụng cho phương pháp lọc cộng tác hướng mục.

Một số mục rất phổ biến và có xu hướng ít phân biệt với các mục khác, vậy nên tần suất nghịch đảo (IDF) sẽ được sử dụng để tính trọng số cho mỗi mục, điểm. BM25, một phiên bản nâng cấp của TF-IDF cũng sẽ được sử dụng. Ngoài ra, một thực tế phổ biến để tăng tầm quan trọng tương đối của điểm số tương đồng là khuếch đại hàm mũ.

## **5.7. Chi tiết tracking model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phương pháp** | **Độ đo tương đồng** | **Trọng số** | **Độ khuếch đại** | **mean\_center** |
| Lọc cộng tác hướng người dùng | Cosine | - | - | - |
| Cosine | IDF | - | - |
| Cosine | BM25 | - | - |
| Pearson | IDF | - | - |
| Pearson | BM25 | - | - |
| Pearson | - | 5.0 | - |
| Pearson | - | 0.5 | - |
| Lọc cộng tác hướng mục | Cosine | - | - | - |
| AdjustedCosine | - | - | True |
| Pearson | - | - | - |
| Pearson | - | - | True |

## **5.8. Kết quả mô hình cuối**

Mô hình được huấn luyện trên tập train và đánh giá trên tập test. Kết quả huấn luyện của mô hình được trình bày dưới bảng sau:

## **5.9. Các hạn chế và đề xuất cải tiến**

Những hạn chế của mô hình gợi ý sử dụng lọc cộng tác bao gồm:

* Cold-start: Lọc cộng tác yêu cầu dữ liệu người dùng và sản phẩm để hoạt động. Tuy nhiên, khi có người dùng hoặc sản phẩm mới, hệ thống sẽ không có đủ thông tin để gợi ý chính xác.
* Khả năng mở rộng: Khi lượng dữ liệu và số lượng người dùng tăng, việc tính toán ma trận tương tác giữa người dùng và sản phẩm sẽ tốn rất nhiều tài nguyên, gây khó khăn trong việc mở rộng hệ thống.

Những đề xuất cải tiến cho mô hình:

* Cải tiến bằng các mô hình dựa trên học sâu: Các mô hình học sâu (Autoencoder, NCF…) có thể giúp mô hình nắm bắt được các mối quan hệ phức tạp giữa người dùng và mục, tăng cường hiệu quả của hệ thống.
* Tích hợp thông tin ngữ cảnh: Bổ sung các yếu tố ngữ cảnh về thời gian vào hệ thống gợi ý có thể giúp gợi ý phù hợp hơn với nhu cầu thực tế của người dùng tại thời điểm cụ thể.

# **6. Nhóm giải pháp: Sử dụng ma trận nhân tử**

## **6.1. Mô tả chi tiết dữ liệu đầu vào**

## **6.2. Mô tả chi tiết dữ liệu đầu ra**

Đầu ra của mô hình là danh sách 5 mục mà mô hình đánh giá là phù hợp nhất để đề xuất cho người dùng.

## **6.3. Khám phá dữ liệu**

## **6.4. Chia tập**

## **6.5. Chuẩn bị dữ liệu**

## **6.6. Huấn luyện mô hình**

Nhóm giải pháp này được chia ra làm hai hướng phân tích.

* Phân tích ma trận nhân tử: Ma trận ban đầu thường là ma trận người dùng-mục, các giá trị trong ma trận thể hiện đánh giá của người dùng đối với các mục. Ma trận gốc R (*m x n*) sẽ được chia làm 2 ma trận nhỏ hơn P (*m x k*) và Q (*n x k*) lần lượt đại diện cho người dùng và mục. Mỗi người dùng và mục sẽ được biểu diễn bằng vector đặc trưng trong không gian chiều k. Từ đó, mức độ đánh giá của người dùng đối với mục có thể được tính bằng tích vô hướng giữa hai vector này.
* Phân tích ma trận nhân tử không âm: Mô hình và quá trình phân tích ma trận nhân tử của phương này tương tự với phương pháp phân tích ma trận nhân tử thông thường. Điểm khác biệt duy nhất ở đây nằm ở giá trị trong ma trận, trong đó toàn bộ giá trị xuất hiện trong ma trận đều sẽ không âm.

Những chỉ số ảnh hưởng đến kết quả mô hình phân tích ma trận nhân tử được đánh giá bao gồm:

* k: số đặc trưng ẩn của mô hình, kiểm soát số lượng chiều của ma trận phân tích
* max\_iter: số vòng lặp tối đa, giới hạn số lần thuật toán tối ưu được thực hiện
* lambda\_reg: tham số chuẩn hóa, ngăn sự quá khớp dữ liệu
* bias: ghi lại hành vi cơ sở, cải thiện độ chính xác gợi ý của mô hình

## **6.7. Chi tiết tracking model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phương pháp** | **k** | **max\_iter** | **lambda\_reg** | **bias** |
| Phân tích ma trận nhân tử | 10 | 100 | - | - |
| 10 | 20 | - | - |
| 10 | 20 | 0.01 | - |
| 10 | 20 | 0.01 | True |
| Phân tích ma trận nhân tử không âm | 10 | 100 | - | - |
| 10 | 20 | - | - |
| 10 | 20 | 0.01 | - |
| 10 | 20 | 0.01 | True |

## **6.8. Kết quả mô hình cuối**

Mô hình được huấn luyện trên tập train và đánh giá trên tập test. Kết quả huấn luyện của mô hình được trình bày dưới bảng sau:

## **6.9. Các hạn chế và đề xuất cải tiến**

Những hạn chế của mô hình gợi ý sử dụng ma trận nhân tử bao gồm:

* Cold-start: Tương tự với lọc cộng tác, ma trận nhân tử yêu cầu dữ liệu người dùng và sản phẩm để hoạt động. Tuy nhiên, khi có người dùng hoặc sản phẩm mới, hệ thống sẽ không có đủ thông tin để gợi ý chính xác.
* Thiếu khai thác thông tin nội dung: Ma trận nhân tử truyền thống chỉ dựa trên ma trận tương tác người dùng-mục mà không tận dụng thêm các thông tin bổ sung từ người dùng hoặc mục.

Những đề xuất cải tiến cho mô hình:

* Sử dụng thông tin bổ sung: Factorization Machines (FM) là một phương pháp mở rộng của ma trận nhân tử giúp khai thác cả dữ liệu tương tác lẫn thông tin bổ sung.
* Sử dụng biến thể Bayesian: Các kỹ thuật dựa trên mô hình Bayesian như Bayesian Personalized Ranking (BPR) có thể giúp hệ thống học cách xếp hạng các mục một cách hiệu quả hơn, đặc biệt trong trường hợp thiếu dữ liệu đánh giá rõ ràng.