TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**MÔN:**

**XỬ LÝ DỮ LIỆU LỚN**

**TÌM HIỂU VỀ BÀI TOÁN RECOMMENDATION SYSTEM**

*Người hướng dẫn*: **GV.Lê Anh Cường**

*Người thực hiện*: **PHAN LÊ HOÀI NAM - 51800905**

**MẠCH TRUNG TÍN –**

**LƯU HUY THÔNG – 51800631**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**MÔN:**

**XỬ LÝ DỮ LIỆU LỚN**

**TÌM HIỂU VỀ BÀI TOÁN RECOMMENDATION SYSTEM**

Người hướng dẫn: **GV. LÊ ANH CƯỜNG**

Người thực hiện: **PHAN LÊ HOÀI NAM**

**MẠCH TRUNG TÍN**

**LƯU HUY THÔNG**

Khoá  **: 22**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Lời đầu tiên, em xin trân trọng cảm ơn thầy Lê Anh cường, thầy là người đã trao cho bọn em những kiến thức nền tảng để giúp chúng em vững vàng, là bước đệm để chúng em có thể hoàn thành được đồ án này. Dù không được học trực tiếp nhưng với sự chỉ bảo cũng như sự cố gắng của các bạn trong nhóm, chúng em xin được trình bày báo cáo về Recommendation System. Và trong một bài báo cáo luôn luôn sẽ có những sai sót không đáng có, mong thầy chỉ bảo. Cuối cùng, em xin chúc thầy sức khỏe, thành công trong công việc để còn có thể đưa những bài giảng hay, những kiến thức thú vị tới toàn thể những sinh viên theo học thầy.

TÓM TẮT

Ở trong cuộc sống này, chúng ta sẽ có những lúc phân vân với những sự lựa chọn, sự đắn đo. Vậy, đưa ra sự quyết định sẽ rất khó khăn nếu như chúng ta không có những sự gợi ý hợp lý. Các gợi ý luôn được dựa trên các thông tin được thu thập, chúng ta sử dụng Recommendation System như là một cách để có thể giải quyết vấn đề này.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc87549646)

[TÓM TẮT ii](#_Toc87549647)

[MỤC LỤC 1](#_Toc87549648)

[CHƯƠNG 1 – TÌM HIỂU VỀ MỤC TIÊU BÀI TOÁN VÀ CÁC PHƯƠNG PHÁP GIẢI QUYẾT BÀI TOÁN 2](#_Toc87549649)

[1.1 Khái niệm 2](#_Toc87549650)

[1.2 Phân loại 2](#_Toc87549651)

[1.2.1 Content-Based recommendation system: 2](#_Toc87549652)

[1.2.2 Collaborative filtering: 2](#_Toc87549653)

[1.3 Tìm hiểu về Utility matrix 2](#_Toc87549654)

[1.4 Hướng tiếp cận bài toán 4](#_Toc87549655)

[1.5 Những bước cần làm để xây dựng một hệ thống gợi ý 4](#_Toc87549656)

[1.5.1 Thu thập nguồn dữ liệu: 4](#_Toc87549657)

[1.5.2 Chuẩn hóa dữ liệu: 5](#_Toc87549658)

[1.5.3 Chạy mô hình dữ liệu: 5](#_Toc87549659)

[1.5.4 Đánh giá mô hình dữ liệu: 5](#_Toc87549660)

[CHƯƠNG 2 – CƠ SỞ LÝ THUYẾT 6](#_Toc87549661)

[2.1 Công cụ 6](#_Toc87549662)

[2.2 Lý thuyết 6](#_Toc87549663)

[CHƯƠNG 3 – THỰC NGHIỆM 7](#_Toc87549664)

[4.1 Ưu điểm 7](#_Toc87549665)

[4.2 Nhược điểm 7](#_Toc87549666)

CHƯƠNG 1 – TÌM HIỂU VỀ MỤC TIÊU BÀI TOÁN VÀ CÁC PHƯƠNG PHÁP GIẢI QUYẾT BÀI TOÁN

1.1 Khái niệm

Recommender system là hệ thống có khả năng đưa ra những gợi ý cho người dùng những sản phẩm tương tự cho một sản phẩm mà họ đang xem.

Ví dụ: người mua đang xem các mặt hàng dầu gội thì hệ thống sẽ gợi ý dầu xã,....

Thuật toán recommendation được phát triển và ứng dụng rộng rãi trong nhiều doanh nghiệp thuộc đa dạng các lĩnh vực khác nhau như: thương mại điện tử, tài chính, ngân hàng, kinh doanh, phim ảnh,....

1.2 Phân loại

**Trong thuật toán Recommendation system thường được chia thành hai nhóm lớn:**

1.2.1 Content-Based recommendation system:

Cách này dựa vào thuộc tính của sản phẩm như tên, nhà sản xuất, giá cả, mô tả,.... Để đưa ra các sản phẩm tương tự nhau.

Cách tiếp cận này yêu cầu việc sắp xếp các items vào từng nhóm hoặc đi tìm các đặc trưng của từng item. Tuy nhiên, có những item không có nhóm cụ thể và việc xác định nhóm hoặc đặc trưng của từng item đôi khi là bất khả thi.

1.2.2 Collaborative filtering:

Cách này dựa vào thói quen tiêu dùng của những users có xu hướng tương tự trong lịch sử để gợi ý ra các sản phẩm cho người dùng.

Cách tiếp cận này đòi hỏi có một tập dữ liệu lịch sử đủ lớn để có thể phán đoán, đưa ra gợi ý cho người tiêu dùng.

1.3 Tìm hiểu về Utility matrix

Có 2 thực thể chính trong bài toán Recommendation Systems là users và items. Mỗi user sẽ có một *mức độ quan tâm* tới từng item khác nhau. Mức độ quan tâm này. Nếu đã biết trước, sẽ được gán cho một giá trị ứng với mỗi cặp user-item. Giả sử rằng mức độ quan tâm được đo bằng giá trị user rate cho item (rating). Tập hợp tất cả các ratings bao gồm cả những giá trị chưa biết cần được dự đoán, ta sẽ có một ma trận gọi là Utility matrix.

|  |  |  |  |
| --- | --- | --- | --- |
|  | User1 | User2 | User3 |
| Con bướm xuân | 4 | 5 | 1 |
| Hạ trắng | 3 | ? | ? |
| Tuổi hồng | 1 | 2 | ? |

Bảng 1.1 Utility matrix

Trong hình trên có 3 user User1, User2, User3 và 3 bài hát. Các ô màu xanh thể hiện việc một user đã ratings từ 0-5. Các ô ? màu xám tương ứng với các ô chưa có dữ liệu. Công việc của Recommendation systems là dự đoán giá trị tại các ô màu xám này, từ đó đưa ra gợi ý cho người dùng. Vì vậy, đôi khi Recommendation systems đôi khi cũng được coi là bài toán matrix completion (hoàn thiện ma trận).

Trong ví dụ trên, dễ thấy có 2 thể loại nhạc khác nhau: bài đầu là nhạc trẻ và 2 bài sau là bolero. Từ dữ liệu trên, ta có thể thấy rằng User1 và User2 có gu âm nhạc gần giống nhau, nên hệ thống sẽ gợi ý bài hạ trắng cho User2.

Thông thường, có rất nhiều users và items trong hệ thống, mỗi user sẽ chỉ rate một số lượng rất nhỏ các item, thậm chí có những user không rate item nào (với những users này thì cách tốt nhất là gợi ý các items phổ biến, thịnh hành). Vì vậy, lượng ô màu xám của utility matrix trong các bài toán đó thường rất lớn, còn các ô đã được điền là một số rất nhỏ.

Ta có thể thấy, càng nhiều ô được điền thì độ chính xác của hệ thống sẽ càng được cải thiện. Vì vậy, các hệ thống luôn hỏi người dùng về sự quan tâm của họ tới sản phẩm và muốn người dùng đánh giá càng nhiều sản phẩm càng tốt. Việc đánh giá các sản phẩm không những giúp dùng khác biết được chất lượng của sản phẩm mà còn giúp hệ thống biết được sở thích của người dùng, qua đó có chính sách quảng cáo hợp lý.

1.4 Hướng tiếp cận bài toán

Có một điều dễ nhận thấy thì *phương pháp gợi ý dựa trên nội dung* đòi hỏi chúng ta phải thu thập rất nhiều thông tin về các *mục tin tương tự* . Chính việc xác định xem một mục tin nào là tương tự với mục tin hiện tại đòi hỏi chúng ta phải thu thập và phần tích, xử lý toàn bộ các mục tin trong cơ sở dữ liệu. Tuy nhiên với *phương pháp lọc công tác* chúng ta không cần quá nhiều thông tin. Đơn giản chỉ là item\_id của item hiện tại, các user\_id và các feedback trên item đó mà thôi nên thực tế thì *phương pháp lọc cộng tác* được sử dụng phổ biến hơn để xây dựng các hệ thống gợi ý

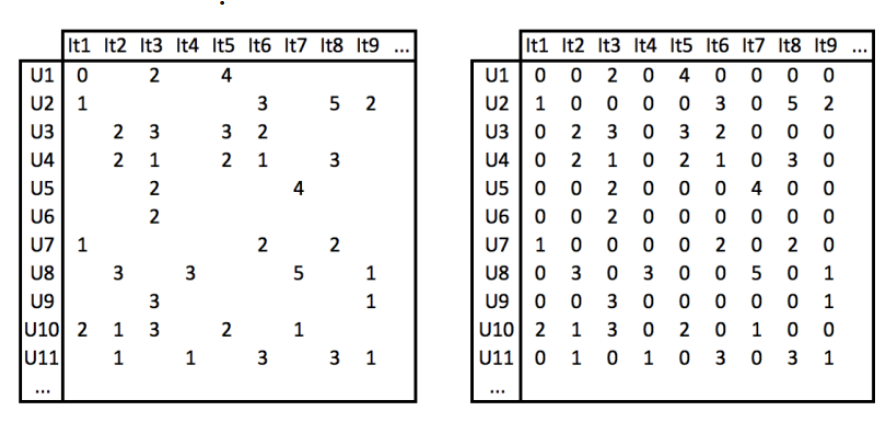
1.5 Những bước cần làm để xây dựng một hệ thống gợi ý

1.5.1 Thu thập nguồn dữ liệu:

Nếu chúng ta đơn giản chỉ quan tâm đến việc *rating* của user với item thì vấn đề trở nên khá đơn giản, dữ liệu của chúng ta đã có sẵn trong DB. Tuy nhiên tùy vào bài toán cụ thể mà không phải lúc nào những chỉ số của chúng ta là tường mình và có sẵn và chính vì thế chúng ta cần phải có một kế hoạch để thu thập các chỉ số thể hiện mối tương quan này trước khi chúng ta định xây dựng một hệ thống gợi ý.

Sau quá trình thu thập chúng ta có rất nhiều dữ liệu ở các phiên làm việc khác nhau tương ứng với các thao tác khác nhau của một user đối với item. Sau khi xử lý bằng các xử lý toàn học không đi sâu ở đây, chúng ta sẽ thu được một chỉ số duy nhất giữa một cặp user-item. Việc cần làm tiếp theo đó là chuẩn hóa dữ liệu.

1.5.2 Chuẩn hóa dữ liệu:



Hình 1.1 Chuẩn hóa dữ liệu

Ma trận dữ liệu của chúng ta chủ yếu là ma trận thưa tức là số lượng dữ liệu còn trống là rất nhiều, chính vì thế nên chúng ta cần phải chuẩn hóa đống dữ liệu này mới có thể áp dụng các *thuật toán học máy* trên đó được. Sau khi chuẩn hóa dữ liệu, chúng ta sẽ xây dựng mô hình học máy để tính toán độ tương tự giữa các user. Mình sẽ nói chi tiết phần này trong các bài viết tiếp theo

1.5.3 Chạy mô hình dữ liệu:

Sau khi lựa chọn được mô hình phù hợp chúng ta sẽ tiến hành chạy mô hình đó và lựa chọn ra top N item sử dụng để gợi ý cho người dùng

1.5.4 Đánh giá mô hình dữ liệu:

Cũng như các bài toán học máy khác, chúng ta cũng cần đánh giá mô hình dựa trên tập dữ liệu kiểm tra. Tuy nhiên do đặc thù của bài toán sẽ có những phương pháp đánh giá khác nhau. Do phạm vi bài viết quá dài nên hiện tại mình chưa trình bày kĩ ở đây. Hẹn gặp các bạn trong các bài viết tiếp theo

CHƯƠNG 2 – CƠ SỞ LÝ THUYẾT

2.1 Công cụ

Phần mềm sử dụng:

Ngôn ngữ:

2.2 Lý thuyết

Để viết được ứng dụng này cần có những kiến thức cơ bản sau đây:

* Kiến thức về Recommendation System.

CHƯƠNG 3 – THỰC NGHIỆM

4.1 Ưu điểm

4.2 Nhược điểm

**TÀI LIỆU THAM KHẢO**