TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ MÔN XỬ LÝ DỮ LIỆU LỚN**

**HỆ THỐNG KHUYẾN NGHỊ**

*Người hướng dẫn*: **Thầy LÊ ANH CƯỜNG**

*Người thực hiện*: **TRƯƠNG NGUYỄN PHI LONG – 51800296**

**NGÔ MINH TOÀN – 51801031**

**DƯƠNG CAO MINH NHẬT – 51704078**

Lớp **: 18050402,** **17050401**

Khoá  **: 22, 21**

**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**



**ĐỒ ÁN CUỐI KỲ MÔN XỬ LÝ DỮ LIỆU LỚN**

**HỆ THỐNG KHUYẾN NGHỊ**

Người hướng dẫn: **Thầy LÊ ANH CƯỜNG**

Người thực hiện: **TRƯƠNG NGUYỄN PHI LONG – 51800296**

**NGÔ MINH TOÀN – 51801031**

**DƯƠNG CAO MINH NHẬT – 51704078**

Lớp **: 18050402, 17050401**

Khoá  **: 22, 21**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2021**

LỜI CẢM ƠN

Xin chân thành cảm ơn thầy Lê Anh Cường phụ trách bộ môn Xử lý dữ liệu lớn đã giảng dạy chúng em nhiều kiến thức, cảm ơn thầy trong quá trình chúng em học tập môn này đã tận tình giải đáp các thắc mắc, hướng dẫn chúng em trong quá trình học tập và thực hiện báo cáo cuối kỳ.

Trong quá trình thực hiện bài báo cáo chúng em vẫn còn nhiều thiếu sót, kiến thức cách hiểu, nhìn nhận đề tài của chúng em còn hạn chế nên rất mong được sự đóng góp ý kiến của thầy để có cơ hội bổ sung thêm kiến thức và bài báo cáo được hoàn thiện hơn.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của thầy Lê Anh Cường. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào chúng tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 21 tháng 4 năm 2021*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Trương Nguyễn Phi Long*

*Ngô Minh Toàn*

*Dương Cao Minh Nhật*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

Dữ liệu lớn (Big data) là việc xử lý một tập hợp dữ liệu rất lớn và phức tạp. Ngày nay tập dữ liệu đang tăng rất nhanh một vì chúng được thu thập bởi số lượng thiết bị internet vạn vật ngày càng rẻ và nhiều, khả năng lưu trữ thông tin của thế giới đã tăng tạo ra các thách thức cho việc khai thác dữ liệu. Trong đó một vấn đề cơ bản về khai thác dữ liệu là hệ thống khuyến nghị.

Hệ thống khuyến nghị là một công cụ mạnh mẽ giúp cải thiện trải nghiệm của khách hàng thông qua các đề xuất được cá nhân hóa. Để đưa ra các đề xuất, nó phân tích hành vi của người dùng, chẳng hạn như xếp hạng có sẵn và lịch sử duyệt web. Một hệ thống giới thiệu được thiết kế tốt có thể làm tăng đáng kể doanh thu của các trang web và ứng dụng thương mại điện tử.

Có nhiều mô hình hệ thống khuyến nghị khác nhau được triển khai. Một hệ thống khuyến nghị thành công phải chính xác trong các dự đoán của nó và đồng thời nhanh chóng. Một khách hàng hài lòng rất có thể sẽ trung thành với trang web hoặc ứng dụng. Vì lý do đó, các mô hình khác nhau phải được đánh giá về độ chính xác và hiệu suất.

Trong bài báo cáo này, chúng ta sẽ nói về hệ thống khuyến nghị mà chủ yếu là mô hình hệ thống lọc dựa trên nội dung (content-based system).

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc70239661)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc70239662)

[TÓM TẮT iv](#_Toc70239663)

[MỤC LỤC 1](#_Toc70239664)

[DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 3](#_Toc70239665)

[CHƯƠNG 1 –GIỚI THIỆU VỀ HỆ THỐNG KHUYẾN NGHỊ 4](#_Toc70239666)

[1.1 Giới thiệu chung về hệ thống khuyến nghị 4](#_Toc70239667)

[1.2 Mục tiêu của recommendation system 5](#_Toc70239668)

[1.3 Phân loại recommendation system 5](#_Toc70239669)

[1.4 Utility Matrix 6](#_Toc70239670)

[1.5 The Long Tail 7](#_Toc70239671)

[1.6 Ứng dụng của hệ thống recommendation systems 8](#_Toc70239672)

[1.7 Xây dựng Utility Matrix 9](#_Toc70239673)

[CHƯƠNG 2 – CONTENT-BASED RECOMMENDATIONS 10](#_Toc70239674)

[2.1 Giới thiệu chung về hệ thống khuyến nghị dựa trên nội dung 10](#_Toc70239675)

[2.2 Xây dựng Item Profiles 10](#_Toc70239676)

[2.2.1 TF-IDF 11](#_Toc70239677)

[2.2.2 Sử dụng biểu diễn nhị phân 12](#_Toc70239678)

[2.3 Tìm mô hình cho mỗi user - xây dựng hàm mất mát 12](#_Toc70239679)

[CHƯƠNG 3 – THỰC NGHIỆM 14](#_Toc70239680)

[CHƯƠNG 4 – KẾT LUẬN 16](#_Toc70239681)

[TÀI LIỆU KHAM THẢO 17](#_Toc70239682)

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC CHỮ VIẾT TẮT**

TF-IDF : Term Frequency – Inverse Document Frequency

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

**DANH MỤC HÌNH**

[Hình 1.1 Ma trận tiện ích đại diện cho xếp hạng phim trên thang điểm 1–5 6](#_Toc70236762)

[Hình 1.2 Hiện tượng The Long Tail 8](#_Toc70236763)

[Hình 2.1 Công thức tính TF-IDF 11](#_Toc70236764)

[Hình 2.2 Sử dụng biểu diễn nhị phân 12](#_Toc70236765)

[Hình 3.2 Hệ thống đề xuất các bộ phim liên quan 15](#_Toc70231480)

CHƯƠNG 1 –GIỚI THIỆU VỀ HỆ THỐNG KHUYẾN NGHỊ

1.1 Giới thiệu chung về hệ thống khuyến nghị

Internet đang phát triển nhanh chóng và nhiều nhiệm vụ như giao dịch điện tử và kinh doanh đang được thực hiện trực tuyến. Người dùng có một danh sách khổng lồ các lựa chọn nên việc chọn ra thứ phù hợp nhất theo sở thích của họ có thể là một thách thức. Với sự nổi lên của Youtube, Amazon, Netflix và nhiều dịch vụ web khác, các hệ thống giới thiệu ngày càng có nhiều vị trí trong cuộc sống của chúng ta. Từ thương mại điện tử (gợi ý cho người mua những bài báo mà họ có thể quan tâm dựa trên dự đoán sở thích của họ) đến quảng cáo trực tuyến (gợi ý cho người dùng những nội dung phù hợp dựa trên lịch sử mua hàng, hoặc sở thích của họ), hệ thống giới thiệu ngày nay không thể tránh khỏi trong việc trực tuyến hàng ngày của chúng ta.

Hệ thống khuyến nghị (Recommender systems hoặc Recommendation systems) là các thuật toán có nhiệm vụ chính là cung cấp các khuyến nghị về các mặt hàng cho người dùng. Những đề xuất này đề cập đến các quy trình ra quyết định khác nhau, chẳng hạn như mặt hàng nào để mua, video nào để xem hoặc bài viết nào để đọc. Nói cách khác, hệ thống giới thiệu cung cấp các khuyến nghị, nội dung và dịch vụ được cá nhân hóa cho người dùng.

Hệ thống khuyến nghị đã trở thành một khu vực nghiên cứu độc lập vào giữa những năm 1990, nơi các bài báo đầu tiên về lọc cộng tác (collaborative filtering) bắt đầu xuất hiện. Một yếu tố quan trọng mà các hệ thống khuyến nghị được sử dụng rộng rãi ngày nay là sự dễ dàng khi người dùng có thể cung cấp phản hồi về sở thích của họ thông qua các trang web. Một cách phổ biến mà người dùng có thể cung cấp phản hồi là bằng hệ thống xếp hạng, trong đó họ chỉ định các số hạng(vd: hạng 1,hạng 2,…) để đánh giá một mặt hàng nào đó. Đối tượng mà đề xuất được cung cấp thường được gọi là người dùng và sản phẩm đang được đề xuất cũng được gọi là mục (sách, bài hát, phim, v.v.). Nói cách khác, hệ thống giới thiệu giúp người dùng lọc nội dung của một trang Web bằng cách đưa ra các đề xuất cho họ.

1.2 Mục tiêu của recommendation system

Hệ thống khuyến nghị là các thuật toán nhằm mục đích đề xuất các mục có liên quan cho người dùng, hỗ trợ ra quyết định, cung cấp giải pháp mà không phải khó khăn để tìm kiếm các mục. Các mục tiêu chính của hệ thống khuyến nghị là:

* Giúp mọi người khám phá nội dung mới, xác định những thứ mà mình thích bằng việc cung cấp cho khách hàng những đề xuất về những gì khách hàng có thể muốn mua, dựa trên lịch sử mua hàng, tìm kiếm sản phẩm trong quá khứ của họ.
* Cá nhân hóa trải nghiệm người dùng theo phản hồi của người dùng.
* Mô hình hóa sở thích, ý kiến và hành vi của mọi người.

1.3 Phân loại recommendation system

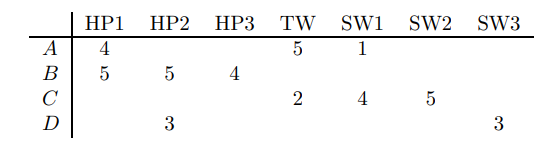
Mục đích của hệ thống giới thiệu là đề xuất các mặt hàng có liên quan cho người dùng. Để đạt được nhiệm vụ này, tồn tại hai loại phương pháp chính: phương pháp lọc cộng tác (collaborative filtering) và phương pháp dựa trên nội dung (content based). Hãy thảo luận ngắn gọn về hai mô hình chính này.

* Hệ thống gợi ý dựa theo nội dung (Content-based systems) kiểm tra thuộc tính của các mặt hàng được đề nghị. Ý tưởng phương pháp này là cố gắng xây dựng một mô hình giải thích các tương tác giữa người dùng và mục. Ví dụ: nếu người dùng đã xem nhiều phim gia đình, thì đề xuất một bộ phim được trong cơ sở dữ liệu là có thể loại “gia đình”.
* Hệ thống gợi ý dựa theo lọc cộng tác (collaboractive filtering) là phương pháp chỉ dựa trên các tương tác trong quá khứ được ghi lại giữa người dùng và các mặt hàng để đưa ra các đề xuất mới. Hệ thống đề xuất các mặt hàng dựa trên các biện pháp tương đồng giữa người dùng hoặc các mặt hàng. Các mục được đề xuất cho người dùng là những mục được những người dùng tương tự ưa thích.

Trong bài báo cáo này, ta sẽ tập trung vào hệ thống content-based systems.

1.4 Utility Matrix

Trong một hệ thống khuyến nghị, có hai lớp thực thể là người dùng (user) và mục (item). Người dùng có các tùy chọn cho các mục nhất định và các tùy chọn này sẽ được đưa ra khỏi dữ liệu. Bản thân dữ liệu này sẽ được biểu diễn dưới dạng ma trận Utility Matrix, cung cấp cho mỗi người dùng một cặp mục, một giá trị đại diện cho những gì đã biết về mức độ ưa thích của người dùng đối với mặt hàng đó. Giá trị đến từ một tập hợp có thứ tự, ví dụ: số 1–5 đại diện cho số sao người dùng đưa ra làm xếp hạng cho mặt hàng đó.



Hình 1. Ma trận tiện ích đại diện cho xếp hạng phim trên thang điểm 1–5

Trong hình trên, chúng ta thấy một Utility Matrix biểu thị xếp hạng phim của người dùng trên thang điểm 1–5. Khoảng trống thể hiện tình huống người dùng không đánh giá phim. Tên các phim lần lượt là HP1, HP2, HP3, TW, SW1, SW2 và SW3. Người dùng được thể hiện bằng các chữ A, B, C và D.

Mục tiêu của hệ thống khuyến nghị là dự đoán các khoảng trống trong Utility Matrix. Như để dự đoán người dùng A có thích SW2 ta xem xét sự giống nhau giữa SW1 và SW2, ta kết luận rằng vì A không thích SW1 nên họ cũng không thích SW2. Với nhiều dữ liệu hơn, chúng ta có thể nhận thấy những người đã xếp hạng cả SW1 và SW2 có xu hướng xếp hạng chúng tương tự nhau. Từ đó có thể kết luận rằng A cũng sẽ cho SW2 một xếp hạng thấp như xếp hạng của A cho SW1 .

Không nhất thiết phải dự đoán mọi mục trống trong ma trận mà thay vào đó chỉ cần phát hiện ra một số mục trong mỗi hàng có khả năng được xếp hạng cao.

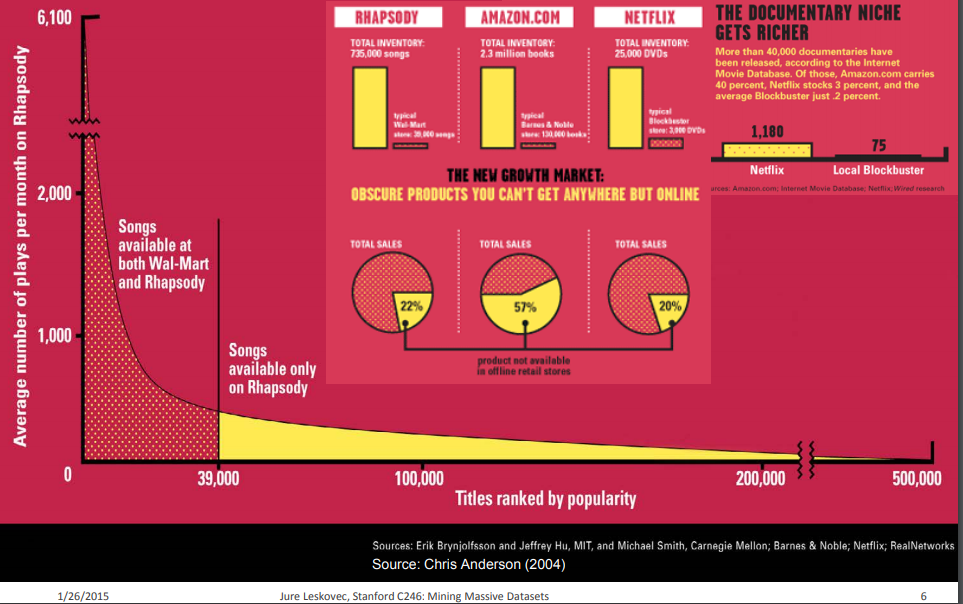
Trong hầu hết các ứng dụng, hệ thống đề xuất không cung cấp cho người dùng bảng xếp hạng của tất cả các mục, mà đề xuất một số mục mà người dùng có thể đánh giá cao hoặc chỉ tìm một tập hợp con có các mục xếp hạng cao nhất.

1.5 The Long Tail

Trước khi thảo luận về các ứng dụng chính của hệ thống khuyến nghị, chúng ta hãy suy ngẫm về The Long Tail khiến hệ thống khuyến nghị trở nên cần thiết. Trong thương mại những sản phẩm bán chạy nhất chỉ chiếm phần nhỏ của tổng số sản phẩm.

Các cửa hàng thực có không gian hạn chế thường là khu trưng bày và kho nên chỉ có thể hiển thị cho khách hàng một phần nhỏ trong số tất cả các sản phẩm, những sản phẩm được trưng ra mang tính phổ biến chưa chắc đã phù hợp với một khách hàng cụ thể. Các cửa hàng online trực tuyến có thể trưng ra mọi sản phẩm cho khách hàng.

Phần lớn doanh thu của các cửa hàng thực đến từ phần nhỏ số sản phẩm phổ biến nhất. Do phần nhỏ các sản phẩm tạo ra phần lớn doanh thu và các sản phẩm khác chỉ tạo ít doanh thu nên ta có một hiện tượng đó là long tail phenomenon nghĩa là phần đuôi dài của những sản phẩm ít phổ biến.



Hình 1. Hiện tượng The Long Tail

Các mặt hàng được xếp thứ tự trên trục hoành tùy theo mức độ phổ biến của chúng. Các tổ chức vật lý chỉ cung cấp các mặt hàng phổ biến nhất, trong khi các tổ chức trực tuyến tương ứng cung cấp toàn bộ các mặt hàng.

Hiện tượng The Long Tail buộc các tổ chức trực tuyến phải giới thiệu các mặt hàng cho người dùng cá nhân. Không thể trình bày tất cả các mặt hàng có sẵn cho người dùng như các tổ chức vật lý.

1.6 Ứng dụng của hệ thống recommendation systems

Chúng ta nói đến một số ứng dụng quan trọng của hệ thống khuyến nghị đó là:

* Đề xuất Sản phẩm: Những đề xuất này không phải ngẫu nhiên mà dựa trên quyết định mua hàng của những khách hàng tương tự hoặc dựa trên các kỹ thuật khác.
* Đề xuất Phim: Netflix cung cấp cho khách hàng các đề xuất về phim mà họ có thể thích. Các khuyến nghị này dựa trên xếp hạng do người dùng cung cấp.
* Đề xuất Bài báo: Các dịch vụ tin tức đã cố gắng xác định các bài báo mà độc giả quan tâm, dựa trên các bài báo mà họ đã đọc trong quá khứ. Sự giống nhau có thể dựa trên sự giống nhau của các từ quan trọng trong tài liệu hoặc trên các bài báo được đọc bởi những người có sở thích đọc tương tự. Các nguyên tắc tương tự cũng áp dụng cho việc đề xuất các blog trong số hàng triệu blog có sẵn, video trên YouTube,…

1.7 Xây dựng Utility Matrix

Nếu không có Utility Matrix, hầu như không thể giới thiệu các mặt hàng. Nhưng việc thu thập dữ liệu để từ đó xây dựng ma trận thường gặp nhiều khó khăn. Có hai cách tiếp cận chung để khám phá giá trị mà người dùng đặt trên các mặt hàng:

* Chúng ta có thể yêu cầu người dùng xếp hạng các mặt hàng. Xếp hạng phim thường được lấy theo cách này và một số cửa hàng trực tuyến cố gắng lấy xếp hạng từ người mua. Các trang web cung cấp nội dung, như một số trang tin tức hoặc YouTube cũng yêu cầu người dùng xếp hạng các mục. Cách tiếp cận này bị hạn chế vì người dùng thường không sẵn sàng cung cấp phản hồi và thông tin có thể bị sai lệch bởi thực tế là nó đến từ những người sẵn sàng cung cấp xếp hạng.
* Với cách thứ hai chúng ta có thể đưa ra suy luận từ hành vi của người dùng. Ví dụ nếu người dùng mua một sản phẩm tại Amazon, xem phim trên YouTube, Netflix hoặc đọc một bài báo từ trang web nào đó, thì người dùng có thể “thích” mặt hàng này. Lưu ý rằng hệ thống loại này chỉ có một giá trị là 1 có nghĩa là người dùng thích mặt hàng đó 0 nghĩa là chưa có thông tin.

CHƯƠNG 2 – CONTENT-BASED RECOMMENDATIONS

Hệ thống khuyến nghị mà bài báo cáo này nói đến sẽ chủ yếu dựa trên phương pháp content-based. Chương 2 của bài báo cáo sẽ tập trung chủ yếu vào phương pháp này, giới thiệu phương pháp cũng như cách để tạo một Content-based systems.

2.1 Giới thiệu chung về hệ thống khuyến nghị dựa trên nội dung

Content-based systems còn được gọi là hệ thống lọc dựa trên nội dung đề xuất các mục cho người dùng dựa trên sự so sánh giữa nội dung của các mục và hồ sơ người dùng . Ý tưởng của các phương pháp dựa trên nội dung là cố gắng xây dựng một mô hình, dựa trên các “tính năng” có sẵn, giải thích các tương tác giữa người dùng và mục được quan sát. Thuật toán content-based gồm 2 bước: Biểu diễn các item dưới dạng vec-tơ thuộc tính và từ đó tìm mô hình cho mỗi user.

Hệ thống content-based có các ưu điểm:

* Độc lập: Hệ thống có thể cung cấp xếp hạng độc quyền được người dùng để xây dựng hồ sơ.
* Tính minh bạch: Hệ thống có thể xác minh các đề xuất cho người dùng.
* Dự đoán hiệu quả cho các mặt hàng mới: Hệ thống có thể đề xuất các mặt hàng mà chúng chưa được xếp hạng.

Nhược điểm của hệ thống content-based là:

* Đề xuất rõ ràng: Nhiều trường hợp cần đề xuất rõ ràng do sử dụng từ khóa hoặc nội dung.
* Dự đoán không hiệu quả cho người dùng mới: Chất lượng phụ thuộc vào độ lớn dữ liệu lịch sử thao tác của người sử dụng.

2.2 Xây dựng Item Profiles

Hệ thống content-based dựa trên nội dung của mỗi item nên chúng ta cần xây dựng một bộ hộ sơ cho mỗi item tức feature vector cho mỗi item. Trong một số trường hợp feature vector được trực tiếp trích xuất từ item sau đó tính độ tương đồng giữa các feature vector của chúng để xác định các mục có nội dung tương tự. Ví dụ, xem xét các features của một bộ phim có thể được sử dụng trong các Recommendation Systems:

* Diễn viên của phim: Một số người thích phim có diễn viên họ thích.
* Đạo diễn: Một số người thích phim của một đạo diễn.
* Năm sản xuất bộ phim: Một số người thích phim cũ hoặc phim mới.
* Thể loại hoặc loại phim chung: Một số người chỉ thích phim hài,…

Xây dựng feature vector có 2 phương pháp chính là TF-IDF và biểu diễn nhị phân. Ta sẽ chỉ giới thiệu sơ về biểu diễn nhị phân và tập trung nói về TF-IDF do bài thực nghiệm trong bài báo cáo này sẽ sử dụng phương pháp đó là TF-IDF.

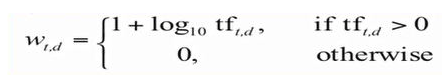
2.2.1 TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency) là 1 kĩ thuật khai phá dữ liệu để phân loại văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong văn bản nó tăng phụ thuộc vào số lần từ xuất hiện trong văn bản bù lại bởi tần suất của từ đó trong dữ liệu.

TF (Term Frequency) là số lần term, nội dung đó xuất hiện trong văn bản, một số term có thể xuất hiện nhiều lần trong một văn bản. Ta nói ngắn gọn như sau: TF(t, d) = (Số lần từ t xuất hiện trong văn bản d) / (tổng số từ trong d)

IDF (Inverse Document Frequency) đánh giá độ quan trọng của một term. Ta có thể nói gọn như sau: IDF(t, D) = log\_e (Tổng văn bản trong tập D/ Số văn bản có từ t )

Nhưng trong khi tính toán TF-IDF, log được sử dụng để làm giảm ảnh hưởng của các từ tần suất cao. Ví dụ: TF = 3 so với TF = 4 khác biệt rất lớn so với TF = 10 so với TF = 1000. Nói cách khác, mức độ liên quan của một từ trong tài liệu không thể được đo lường như một số liệu đếm đơn giản và do đó có phương trình dưới đây:



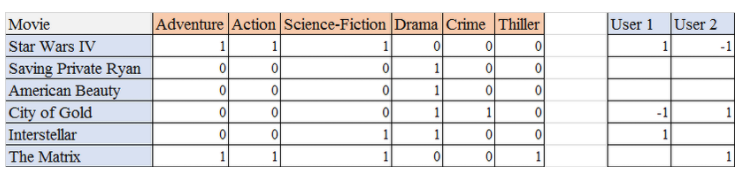
Hình 2. Công thức tính TF-IDF

Sau khi tính toán TF-IDF để xác định mục nào gần nhau hơn, gần với hồ sơ người dùng hơn ta sử dụng độ tương đồng cosin, mô hình không gian vectơ tính toán độ gần nhau dựa trên góc giữa các vectơ.

2.2.2 Sử dụng biểu diễn nhị phân

Một cách khác để xây dựng feature vector chính là sử dụng biểu diễn nhị phân. Để dễ hình hung hơn ta có ví dụ sau:

Ví dụ có ta có danh sách 6 bộ phim và một hồ sơ người dùng. Giá trị 0 thể hiện bộ phim đó không thuộc thể loại ở cột tương ứng, giá trị 1 thì thể hiện thuộc thể loại ở cột tương ứng. Giá trị 1 ở cột user thể hiện sự quan tâm ở cột tương ứng, giá trị -1 thể hiện sự không quan tâm, và trống thể hiện user chưa đánh giá bộ phim đó. Như hình dưới đây User 1 không quan tâm bộ phim City of Gold, còn User 2 thì có, tương tự như vậy ở những dòng khác.



Hình 2. Sử dụng biểu diễn nhị phân

2.3 Tìm mô hình cho mỗi user - xây dựng hàm mất mát

Bài toán đi tìm mô hình cho mỗi user có thể được coi là một bài toán Regression hoặc Classification. Dữ liệu training để xây dựng mô hình là các cặp (item profile, ratings). Mục tiêu của hệ thống khuyến nghị là dự đoán các khoảng trống trong ma trận tiện ích bằng việc điền vào các giá trị thiếu khi áp dụng mô hình nào đó lên chúng.

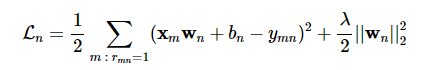
Việc lựa chọn mô hình nào tuỳ thuộc vào hệ thống chúng ta hướng tới. Trong báo cáo này ta sẽ lấy ví dụ về mô hình tuyến tính.

Mô hình tuyến tính:

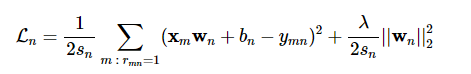
Giả sử rằng users là N, items là M, Y ma trận user-item. Trong đó thành phần ở y(n,m) là mức độ quan tâm của user thứ n lên sản phẩm thứ m mà hệ thống đã thu thập được. Ma trận Y bị thiếu các giá trị và có nhiều ô trống. R là ma trận rated hoặc not rated thể hiện một user đã đánh giá item đó chưa.

Giả sử rằng ta có thể tìm được một mô hình cho mỗi user có thể tính được bằng một hàm tuyến tính: ymn = xmwn +bn (Xm là một vector hàng đặc trưng của item m)

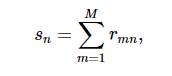
Xét một user thứ n, nếu training set là tập hợp các thành phần đã được điền của yn, ta có thể xây dựng hàm mất mát tương tự:



Trung bình cộng của lỗi được dùng, và mất mát Ln được viết lại thành:



Trong đó Sn là số lượng các items mà user thứ n đã đánh giá:



Vì hàm mục tiêu chỉ phụ thuộc vào các items đã đánh giá, ta có thể rút gọn nó như sau:



Đây chính là hàm mất mát của Ridge Regression.

CHƯƠNG 3 – THỰC NGHIỆM

Hệ thống khuyến nghị (Recommender systems) với mục đích là đề xuất các mặt hàng có liên quan cho người dùng. Từ bài báo cáo ta có hai mô hình chính để giải quyết bài toán của hệ thống khuyến nghị đó là Collaborative và Content-based.

Hệ thống khuyến nghị trong phần thực nghiệm của bài báo cáo sẽ chủ yếu dựa trên phương pháp Content-base. Apache Spark là một framework mã nguồn mở tính toán cụm, phát triển vào năm 2009 bởi AMPLab. Spark cung cấp một giao diện để lập trình toàn bộ các cụm với tính song song dữ liệu ngầm, xử lý dữ liệu theo thời gian thực, vừa nhận dữ liệu từ các nguồn vừa thực hiện việc xử lý dữ liệu vừa nhận và khả năng chịu lỗi. Bằng việc áp dụng PySpark để giải quyết bài toán ta có thể xử lý song song dữ liệu và cải thiện hiệu suất thực thi của chương trình.

Dưới đây là đường link source code của hệ thống khuyến nghị sử dụng PySpark và theo cách thông thường:

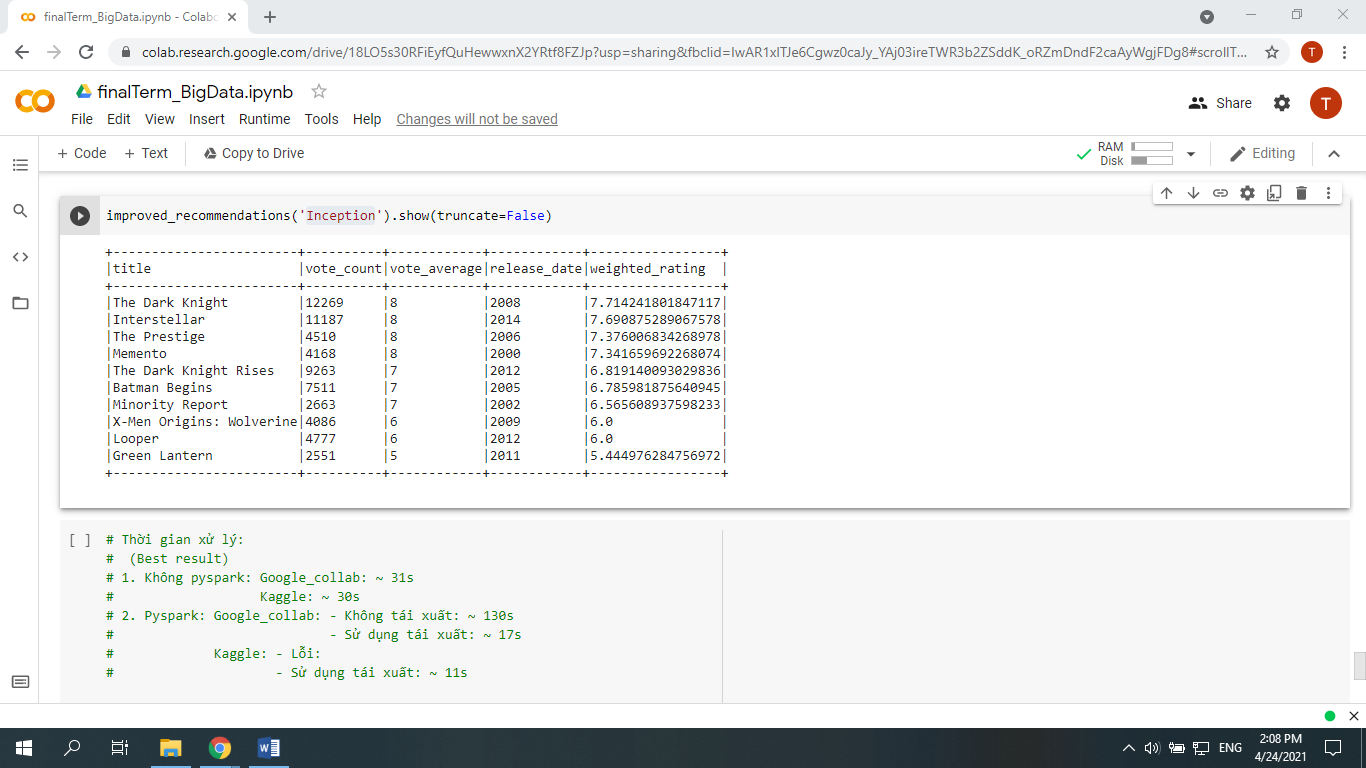
<https://colab.research.google.com/drive/18LO5s30RFiEyfQuHewwxnX2YRtf8FZJp?usp=sharing&fbclid=IwAR1xlTJe6Cgwz0caJy_YAj03ireTWR3b2ZSddK_oRZmDndF2caAyWgjFDg8>

Dữ liệu của bài thực nghiệm này được lấy từ kaggle bao gồm những dữ liệu về các bộ phim như tên bộ phim, id , thể loại, ngôn ngữ, độ phổ biến (credits.csv, ratings.csv, movies\_metadata.csv…). Truy cập đường link sau đây để tải về dataset :

<https://www.kaggle.com/rounakbanik/the-movies-dataset>

Input của bài toán: Đầu tiên chúng ta cần có dữ liệu về các users, items, feedback.

Output của bài toán: Các đề xuất hoặc dự đoán bộ phim liên quan mà người dùng có thể thích hơn.



Hình 3. Hệ thống đề xuất các bộ phim liên quan

Hệ thống khuyến nghị trong bài báo cáo sẽ sử dụng phương pháp gợi ý dựa theo nội dung (Content-based): Thuật toán này sử dụng thông tin mô tả của item, biểu diễn các item dưới dạng vec-tơ. Sau đó dùng các vec-tơ này để tìm mô hình của mỗi user hay ma trận trọng số của user với mỗi item và từ đó đề xuất hoặc dự đoán loại mặt hàng mà người dùng có thể thích hơn.

CHƯƠNG 4 – KẾT LUẬN

Dữ liệu lớn (Big data) là việc xử lý một tập hợp dữ liệu rất lớn và phức tạp. Ngày nay tập dữ liệu đang tăng rất nhanh vì chúng được thu thập bởi số lượng thiết bị internet vạn vật ngày càng rẻ và nhiều, khả năng lưu trữ thông tin của thế giới đã tăng tạo ra các thách thức như phân tích, thu thập, giám sát dữ liệu, tìm kiếm, chia sẻ, lưu trữ,…

Hệ thống đề xuất là một cơ chế lọc thông tin cố gắng dự đoán xếp hạng mà người dùng sẽ cho một sản phẩm cụ thể. Hệ thống đề xuất mở ra cơ hội mới để truy xuất thông tin được cá nhân hóa trên Internet. Nó cũng giúp giảm vấn đề quá tải thông tin, cho phép người dùng có quyền truy cập vào các sản phẩm, dịch vụ mà không có sẵn trên hệ thống. Từ bài báo cáo ta biết được hai kỹ thuật truyền thống cho các hệ thống khuyến nghị và nêu bật những điểm mạnh và thách thức của chúng từ đó cải thiện kỹ thuật đề xuất.

Apache Spark là một framework mã nguồn mở tính toán cụm. Spark cung cấp một giao diện để lập trình toàn bộ các cụm với tính song song dữ liệu ngầm, xử lý dữ liệu theo thời gian thực, vừa nhận dữ liệu từ các nguồn vừa thực hiện việc xử lý dữ liệu vừa nhận và khả năng chịu lỗi. Áp dụng spark vào bài toán hệ thống khuyến nghị ta thấy được hiệu suất thực thi đã tăng lên. Qua bài báo cáo ta thấy thuật toán song song khó viết hơn so với những thuật toán tuần tự, vì sự tương tranh tạo ra nhiều lớp mới tiềm tàng các lỗi phần mềm, trong đó lỗi điều kiện ganh đua là phổ biến nhất.

TÀI LIỆU KHAM THẢO

1. <http://snap.stanford.edu/class/cs246-2015/slides/07-recsys1.pdf>
2. <http://infolab.stanford.edu/~ullman/mmds/ch9.pdf>
3. <https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada>
4. <https://viblo.asia/p/gioi-thieu-ve-he-thong-goi-y-recommender-systems-hoac-recommendation-systems-maGK78yOZj2>
5. <https://viblo.asia/p/tim-hieu-ve-content-based-filtering-phuong-phap-goi-y-dua-theo-noi-dung-phan-1-V3m5WGBg5O7>
6. <https://machinelearningcoban.com/2017/05/17/contentbasedrecommendersys/>
7. <https://nguyenvanhieu.vn/tf-idf-la-gi/>
8. <https://quyv.wordpress.com/2016/07/30/bag-of-words-tf-idf/>
9. Pasquale Lops, Marco de Gemmis, Giovanni Semeraro, “Content-based Recommender Systems: State of the Art and Trends”, 2011
10. Content Based Recommendations - Stanford University

<https://www.youtube.com/watch?v=2uxXPzm-7FY>

1. Theodosios Siomos, “Parallel Implementation of Basic Recommendation Algorithms”, 2018