MOVIE RECOMMENDATION SYSTEM IN R

- Hệ thống khuyến nghị phim tự động sử dụng R -

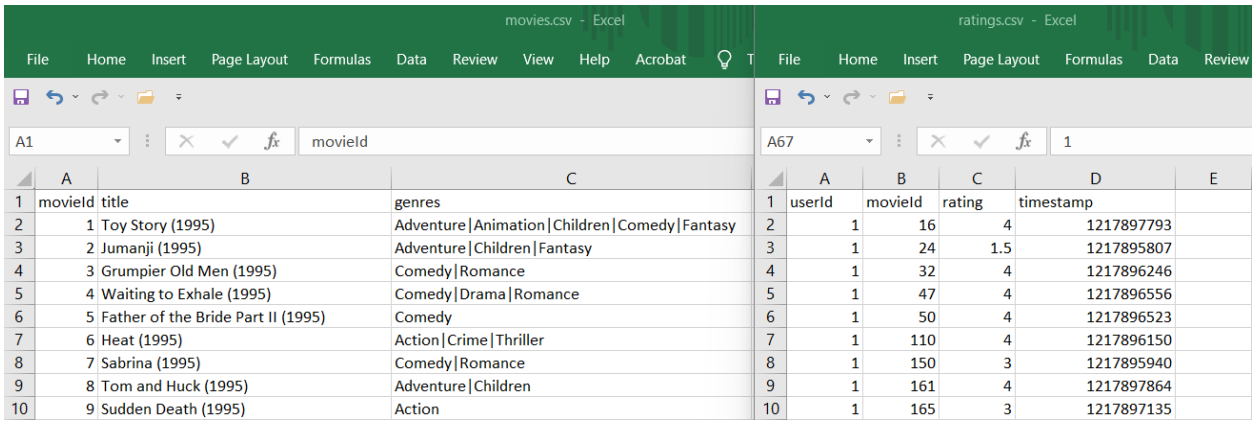
- Môn học: Phân tích dữ liệu với R
- GVHD: Ths.Nguyễn Văn Hồ

STT	Họ và Tên	MSSV
1	Huỳnh Hồng Uyển (Nhóm trưởng)	K204162006
2	Đặng Xuân Mai	K204160667
3	Châu Nguyễn Hương Thảo	K204162000
4	Đào Phương Anh	K204160660
5	Bùi Thu Vân	K204162007

Tóm tắt

Hệ thống khuyến nghị (recommendation systems) đóng một vai trò quan trọng trong thương mại điện tử và các dịch vụ phát trực tuyến. Việc đưa ra đề xuất phù hợp cho sản phẩm, nhạc hoặc phim tiếp theo giúp tăng tỷ lệ giữ chân người dùng và sự hài lòng, dẫn đến tăng trưởng doanh số và lợi nhuận. Mục tiêu chính của dự án máy học này là xây dựng một công cụ đề xuất phim cho người dùng. Dự án R được thiết kế để giúp mọi người hiểu hơn về cách thức hoạt động của hệ thống khuyến nghị. Nhóm nghiên cứu sử dụng phương pháp nghiên cứu tài liệu và phương pháp phân tích tổng hợp. Từ dữ liệu có sẵn, hệ thống đề xuất phân tích các yếu tố khác nhau như mức độ tương tự của người dùng, mức độ tương tự của phim, xếp hạng, v.v. tuy nhiên vẫn còn vài điều hạn chế. Nếu được cải thiện và bổ sung thêm các tệp dữ liệu mới thì hệ thống có thể đề xuất đa dạng phim cho người dùng và nâng cao trải nghiệm khách hàng.

- Dataset: IMDB-Dataset.z
- Minh họa 10 dòng đầu của 2 tệp csv. nhóm sử dụng để phân tích:



movies.csv - Excel

File Home Insert Page Layout Formulas Data Review View Help Acrobat

A1 movieId

	A	B	C
1	movieId	title	genres
2	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
3	2	Jumanji (1995)	Adventure|Children|Fantasy
4	3	Grumpier Old Men (1995)	Comedy|Romance
5	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
6	5	Father of the Bride Part II (1995)	Comedy
7	6	Heat (1995)	Action|Crime|Thriller
8	7	Sabrina (1995)	Comedy|Romance
9	8	Tom and Huck (1995)	Adventure|Children
10	9	Sudden Death (1995)	Action

ratings.csv - Excel

File Home Insert Page Layout Formulas Data Review

A67 1

	A	B	C	D	E
1	userId	movieId	rating	timestamp	
2	1	16	4	1217897793	
3	1	24	1.5	1217895807	
4	1	32	4	1217896246	
5	1	47	4	1217896556	
6	1	50	4	1217896523	
7	1	110	4	1217896150	
8	1	150	3	1217895940	
9	1	161	4	1217897864	
10	1	165	3	1217897135	

TÓM TẮT ĐỀ TÀI
MỤC LỤC
DANH MỤC BẢNG
DANH MỤC HÌNH

MỘT SỐ THUẬT NGỮ

Artificial Intelligence (AI): Trí tuệ nhân tạo

Machine Learning (ML): Học máy

Deep Learning (DL): Học sâu

Supervised Learning: Học có giám sát

Unsupervised Learning: Học không giám sát

Semi-Supervised Learning: Học bán giám sát

Reinforcement Learning: Học củng cố

Recommender System (RSs): Hệ thống khuyến nghị, hệ thống gợi ý

Decision Support System (DSS): Hệ thống hỗ trợ ra quyết định

Collaborative filtering: Lọc cộng tác

Content based filtering: Lọc dựa trên nội dung

CHAPTER 1: OVERVIEW

1. Giới thiệu đề tài

   Hệ thống khuyến nghị (recommendation systems) đóng một vai trò quan trọng trong thương mại điện tử và các dịch vụ phát trực tuyến, chẳng hạn như Netflix, YouTube và Amazon. Việc đưa ra đề xuất phù hợp cho sản phẩm, nhạc hoặc phim tiếp theo giúp tăng tỷ lệ giữ chân người dùng và sự hài lòng, dẫn đến tăng trưởng doanh số và lợi nhuận. Các công ty cạnh tranh vì sự trung thành của khách hàng, qua đó đầu tư vào các hệ thống nhằm nắm bắt và phân tích sở thích của người dùng, đồng thời cung cấp các sản phẩm hoặc dịch vụ có khả năng mua hàng cao hơn.

   Tác động kinh tế giữa mối quan hệ công ty - khách hàng rất rõ ràng: Amazon là công ty bán lẻ trực tuyến lớn nhất tính theo doanh số. Một phần thành công của họ đến từ hệ thống tự đề xuất (recommendation systems) và tiếp thị dựa trên sở thích của người dùng. Năm 2006, Netflix đã đưa ra giải thưởng một triệu đô la cho người hoặc nhóm có thể cải thiện hệ thống đề xuất của họ ít nhất 10%.

   Thông thường, hệ thống đề xuất dựa trên thang điểm đánh giá từ 1 đến 5 hạng hoặc sao, với 1 sao cho biết mức độ hài lòng thấp nhất và 5 sao là mức độ hài lòng cao nhất. Cũng có thể sử dụng các chỉ số khác, chẳng hạn như nhận xét đã đăng trên các mục đã sử dụng trước đó; video, nhạc hoặc liên kết được chia sẻ với bạn bè; tỷ lệ phần trăm phim đã xem hoặc nhạc đã nghe; các trang web đã truy cập và thời gian dành cho mỗi trang; danh mục sản phẩm; và bất kỳ tương tác nào khác với trang web hoặc ứng dụng của công ty đều có thể được sử dụng làm dự đoán.

   Mục tiêu chính của hệ thống đề xuất là giúp người dùng tìm thấy thứ họ muốn dựa trên sở thích và các tương tác trước đó của họ, đồng thời dự đoán xếp hạng cho một bộ phim mới mới. Trong bài nghiên cứu này, nhóm nghiên cứu đã tạo hệ thống đề xuất phim bằng cách sử dụng tập dữ liệu IMDB và áp dụng các bài học trong chương trình HarvardX’s Data Science Professional Certificate 3.

   Tài liệu này được cấu trúc như sau: chương 1 mô tả tập dữ liệu và tóm tắt mục tiêu của dự án và các bước chính đã được thực hiện. Trong chương 2, nhóm nghiên cứu giải thích quy trình và kỹ thuật được sử dụng, chẳng hạn như làm sạch dữ liệu, khám phá và trực quan hóa dữ liệu, bất kỳ thông tin chi tiết nào thu được và cách tiếp cận mô hình hóa. Trong chương 3, nhóm nghiên cứu trình bày các kết quả mô

hình hóa và thảo luận về hiệu suất của mô hình. Kết thúc chương 4 với một bản tóm tắt ngắn gọn về báo cáo, những hạn chế của nó và công việc trong tương lai.

2. Tính cấp thiết

   Xem phim là một hình thức thư giãn, giúp tái tạo năng lượng sau chuỗi ngày áp lực vì công việc hay học tập. Tất cả chúng ta đều cần một thời gian để rút phích cắm ra khỏi cuộc sống và tái tạo năng lượng. Alexis Conason, nhà tâm lý học và nhà nghiên cứu tại Mỹ, cho biết: "Việc xem một bộ phim hoặc chương trình truyền hình hay có thể giúp tâm trí của chúng ta vì nó cho chúng ta cơ hội thoát khỏi những căng thẳng hằng ngày.

   Tuy nhiên, để tìm được một bộ phim ưng ý và hợp "khẩu vị" với mỗi cá nhân khác nhau rất khó, vậy nên, hệ thống khuyến nghị ra đời. Có hai lợi ích của việc sử dụng hệ thống đề xuất: Một mặt, nó có thể giảm lượng lớn công sức tìm kiếm sản phẩm của người dùng và giảm thiểu vấn đề quá tải thông tin. Mặt khác, nó có thể tăng giá trị kinh doanh cho các nhà cung cấp dịch vụ trực tuyến và trở thành nguồn doanh thu quan trọng. Bài nghiên cứu này sẽ giới thiệu những khái niệm cơ bản, các mô hình cổ điển và những bước tiến gần đây của học sâu trong lĩnh vực hệ thống đề xuất, cùng với các ví dụ lập trình, hứa hẹn mang tính thực tiễn, có tính ứng dụng cao và có thể dễ dàng mở rộng sang những lĩnh vực khác.

3. Phương pháp nghiên cứu

   Phương pháp nghiên cứu tài liệu: Nghiên cứu các tài liệu về các đề tài phim, xu hướng đánh giá phim của khán giả và những tài liệu liên quan khác để tiến hành chỉnh sửa, thêm, bớt và chuẩn hóa bộ dữ liệu thuật ngữ.

   Phương pháp phân tích, tổng hợp: Sau khi nghiên cứu các cơ sở lý thuyết thì phân tích các dữ kiện thu thập được, chắt lọc và tổng hợp để định hướng công việc hoàn thành đề tài.

4. Mục tiêu nghiên cứu

   Mục tiêu chính của dự án máy học này là xây dựng một công cụ đề xuất phim cho người dùng. Dự án R được thiết kế để giúp mọi người hiểu hơn về cách thức hoạt động của hệ thống khuyến nghị. Nhóm nghiên cứu sẽ phát triển bộ lọc dựa trên các bộ phim. Ở cuối bài nghiên cứu, nhóm nghiên cứu sẽ đạt được kỹ năng về triển khai R, khoa học dữ liệu và máy học trong một dự án thực tế.

5. Đối tượng và phạm vi nghiên cứu

   5.1 Đối tượng nghiên cứu

Người tham gia đánh giá phim và/hoặc người xem phim, các mô hình nghiên cứu đánh giá và tập dữ liệu các thông tin về phim được đánh giá trên IMDB.

5.2 Phạm vi nghiên cứu
Trang chủ IMDB và các hệ thống liên quan từ năm 1998 đến tháng 7/2019.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

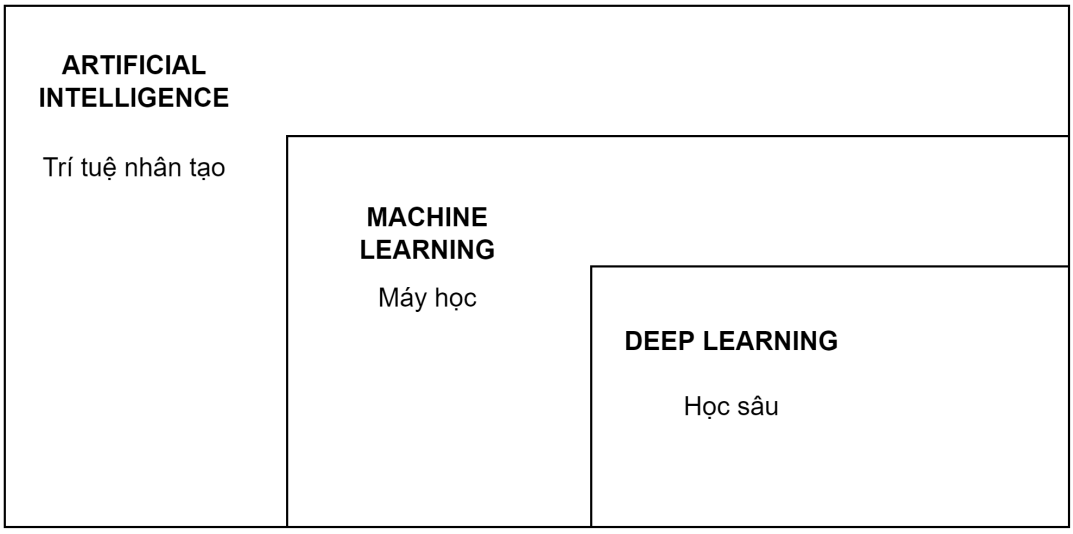
2.1. Một số định nghĩa

2.1.1. Hệ thống khuyến nghị

2.1.1.1. Học máy và một số khái niệm liên quan

Tiếp sau những thành tựu to lớn từ cuộc cách mạng công nghiệp lần thứ ba là công nghệ thông tin, thế giới đang mong ngóng và dốc sức cho cuộc cách mạng công nghiệp lần thứ tư đang diễn ra dựa trên sự kế thừa các nền tảng của cuộc cách mạng số và đem chúng lên một cấp độ vượt trội hơn với sự giúp sức chính từ trí tuệ nhân tạo.

Vì thế trong những năm gần đây, trí tuệ nhân tạo (Artificial intelligence) hay cụ thể hơn là máy học (Machine Learning) rộ lên như sự trỗi dậy thực sự của cuộc cách mạng công nghiệp lần thứ tư. Có thể nói, học máy là một tập con trong trí tuệ nhân tạo. Điều này được hiểu dễ hơn hơn khi chúng ta so sánh chúng như những vòng tròn đồng tâm, với ý tưởng lớn nhất là trí tuệ nhân tạo, sau đó là học máy đang nở rộng và cuối cùng là học sâu (Deep Learning) như sự thúc đẩy nhanh chóng quá trình chuyển đổi số ngày này.



Hình 2.1: Mối quan hệ giữa ba khái niệm

Phải kể đến khi thuật ngữ và lĩnh vực AI được ra đời lần đầu trong hội nghị Dartmouth năm 1956, kể từ đó, thuật ngữ này đã được lặp đi lặp lại như một sự dự báo về chiếc chìa khóa mở ra tương lai xán lạn nhất cho nền văn minh nhân loại, nhưng dường như lúc đó chỉ là những sự dự báo và "ngờ vực" về những gì nó thực sự có thể làm. Cho đến 2015 khi AI bùng nổ vào tạo ra cú hit lớn, nó đem đến vô vàn tiện ích về sự tiện lợi trong kinh tế và phát triển kỹ thuật, và kể từ đó cho đến nay, AI đã được trông thấy với sự phát triển nhanh chưa từng có. Theo đó, sự phát triển của học máy (Machine Learning) cũng ngày càng mạnh mẽ.

Để mô tả học máy, đã có nhiều định nghĩa khác nhau. Theo Arthur Samuel [1] được xem như cha đẻ của thuật ngữ "học máy" lần đầu nói về hai phương pháp như học vẹt hoặc học tổng quát thông qua nghiên cứu xung quanh trò chơi caro. Tiếp đến là một trong những định nghĩa về học máy nổi tiếng nhất theo Tom Mitchell 1997 [2] : " Máy học là một lĩnh vực nghiên cứu những thuật toán áp dụng cho máy tính cho phép các chương trình trên máy tính tự học và cải thiện thông qua kinh nghiệm." Nhưng chung quy lại theo một định nghĩa đơn nhất thì theo Stanford: "Học máy là khoa học để khiến máy tính hành động mà không được lập trình rõ ràng." Từ đó có thể thấy, dù có nhiều định nghĩa khác nhau được đưa ra để mô tả vai trò và định nghĩa của học máy, nhưng về cơ bản thì có chung một điều là học máy đem đến cho máy tính cách thức hoạt động như đúng tên gọi về một "trí tuệ nhân tạo" với phương pháp "học và học" để cải thiện kiến thức, khiến chúng tự trở nên hoàn thiện hơn.

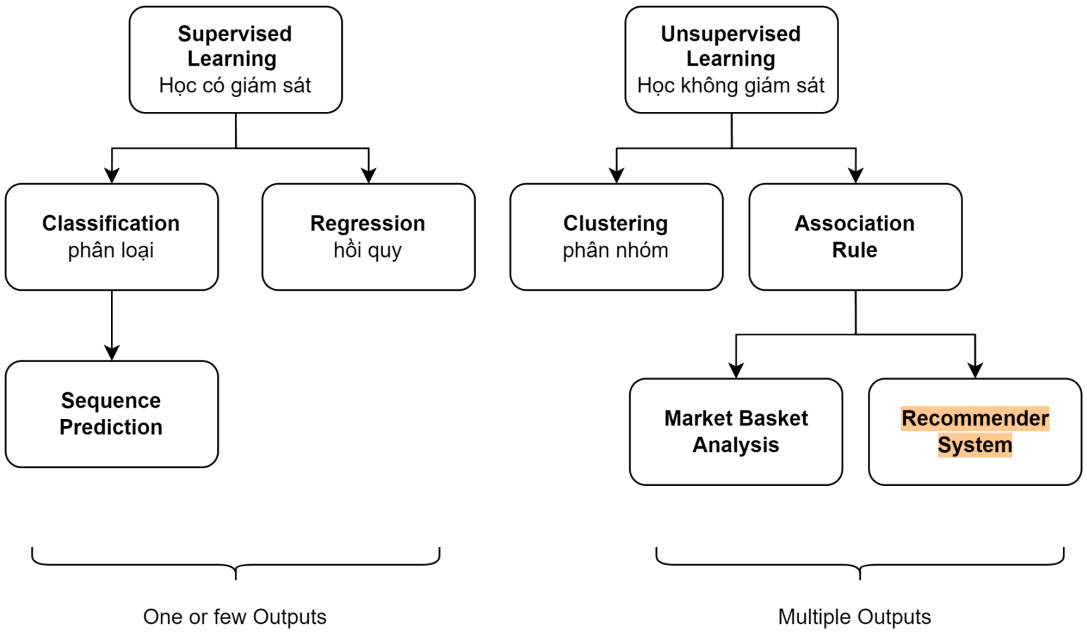
Dựa trên phương thức học, học có giám sát và học không giám sát là hai trong số những phương pháp kỹ thuật tiếp cận cơ bản nhất trong Machine Learning. Và sự khác biệt cơ bản nhất giữa hai kỹ thuật này chính là, một phương pháp thì sử dụng dữ liệu được dán nhãn để giúp dự đoán kết quả (phương pháp học có giám sát) và phương pháp còn lại thì không.

Học có giám sát là thuật toán dự đoán đầu ra của một dữ liệu mới dựa trên cặp dữ liệu đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Học có giám sát dường như là nhóm phổ biến nhất trong các thuật toán Machine Learning. (machinelearningcoban, 2016) Một trong những ví dụ điển hình nhất chính là nhận dạng chữ số được viết bằng tay, như đã biết rằng chữ viết tay của mỗi cá thể sẽ khác nhau, và giả sử có một tập dữ liệu chứa các ảnh về chữ số viết tay của nhiều người khác nhau. Chúng ta đưa các bức ảnh này vào trong một thuật toán và chỉ cho nó biết mỗi bức ảnh tương ứng với chữ số nào. Sau khi thuật toán tạo ra một mô hình, tức một hàm số mà đầu vào là một bức ảnh và đầu ra là một chữ số, khi nhận được một bức ảnh mới mà mô hình chưa nhìn thấy bao giờ, nó sẽ dự đoán bức ảnh đó chứa chữ số nào. Học có giám sát thường áp dụng với các bài toán phân loại và hồi quy.

Học không giám sát thì sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm để thuận tiện trong việc lưu trữ và tính toán. Những thuật toán loại này được gọi là Unsupervised learning vì chúng ta không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào. Ví dụ như khi ta học mà thầy cô chẳng chỉ cho ta biết chữ nào là chữ A chữ nào là chữ B hoặc là không có câu trả lời đúng và không có vị "giáo viên" thật sự nào cả. Phương pháp này được sử dụng cho ba nhiệm vụ chính để phân cụm, kết hợp và giảm kích thước. Đối với nhiệm vụ kết hợp thường được sử dụng cho các công cụ phân tích và đề xuất giỏ thị trường, dọc theo dòng khuyến nghị "Khách hàng đã mua mặt hàng này cũng sẽ có thể mua mặt hàng kia", hoặc chẳng hạn như những người mua món hàng A hay có/ sẽ có/ có thể có khuynh hướng mua món hàng B.

Ngoài ra, còn có các phương pháp khác như học bán giám sát hoặc học củng cố. Học bán giám sát là một phương tiện hữu ích để sử dụng tập dữ liệu đào tạo với cả dữ liệu được gắn nhãn và không được gắn nhãn. Nó đặc biệt hữu ích khi khó trích xuất các tính năng có liên quan từ dữ liệu - và khi có một lưu lượng lớn dữ liệu. [3] "Học bán giám sát có giá trị thực tiễn to lớn. Trong nhiều tác vụ, có rất ít dữ liệu được gắn nhãn. Các nhãn có thể khó lấy vì chúng cần có bộ chú thích từ con người, các thiết bị đặc biệt hoặc các thí nghiệm tốn kém." - Semi-Supervised Learning, 2006.

2.1.1.2. Khái niệm về Hệ thống khuyến nghị



Hình 2.2: Phân nhánh học máy

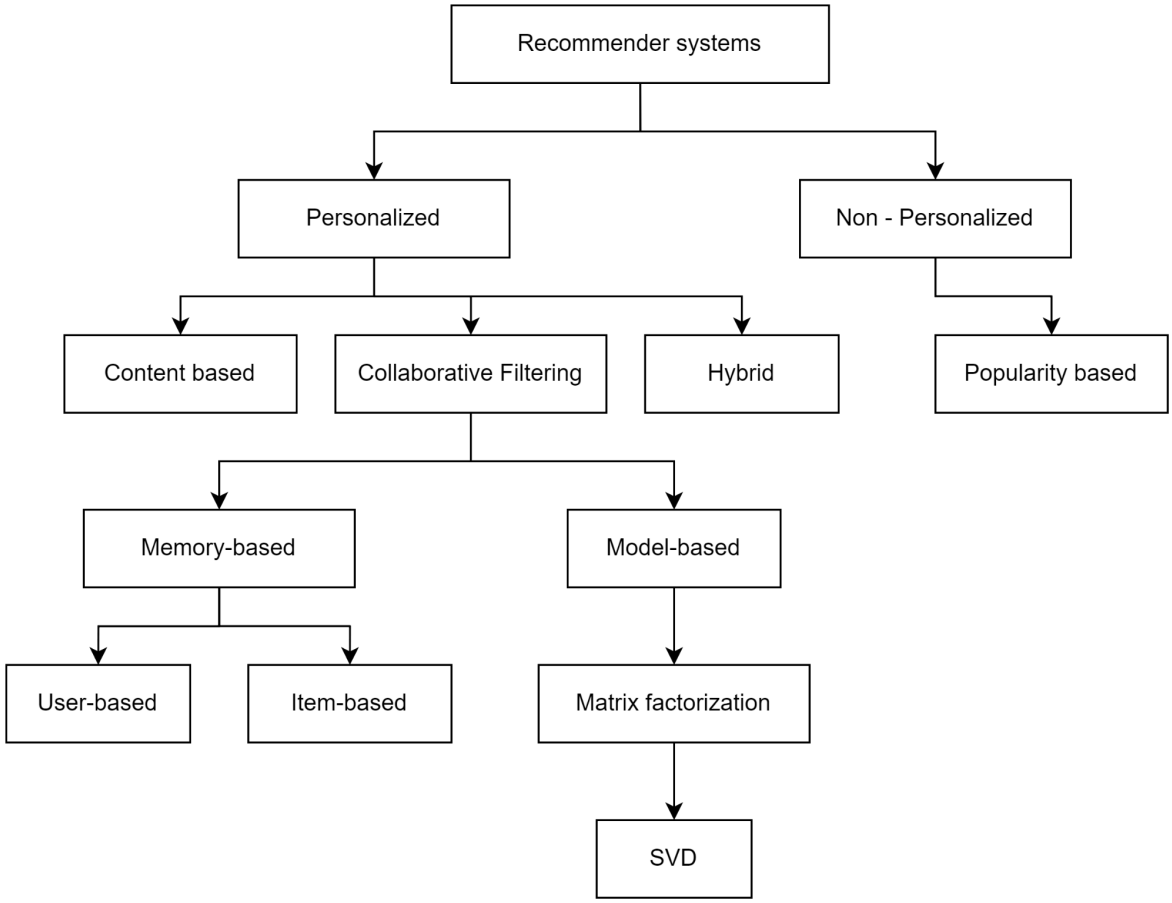
Có thể dễ dàng thấy, hệ khuyến nghị thuộc trong nhánh học máy không giám sát và nằm trong bài toán về Association Rule (quy tắc kết hợp) - đó là bài toán tìm kiếm ra quy luật dựa trên những dữ liệu cho trước.

Về cơ bản, hệ thống khuyến nghị là các công cụ và kỹ thuật phần mềm được sử dụng để truy xuất và lọc thông tin nhằm mục đích cung cấp các đề xuất hữu ích cho người tiêu dùng về các mặt hàng mới dựa trên dữ liệu lịch sử của họ. Hệ thống gợi ý này được triển khai bằng trí tuệ nhân tạo và nó là một danh mục con của Machine Learning và cũng là một danh mục con trong Knowledge-driven DSS (DSS hướng tri thức). (Power, 2007).

Trong một khóa học trực tuyến về Machine Learning, Konstan và Ekstrand (2015) có nói rằng "Hệ thống giới thiệu đã thay đổi cách mọi người tìm kiếm sản phẩm, thông tin và thậm chí cả những người khác." Nó giải thích các thuật toán để lọc dựa trên nội dung, lọc cộng tác user-user, lọc cộng tác item-item và giới thiệu tới người dùng dựa trên các phê bình và tương tác. Hai ứng dụng phổ biến nhất của hệ khuyến nghị là đề xuất các bài báo và cung cấp gợi ý cho khách hàng trực tuyến về những thứ họ cần mua, dựa trên lịch sử mua hàng (Daniel, 2017) Ví dụ: Netflix đề xuất phim cho khách hàng của mình dựa trên lịch sử xem phim của họ hoặc những bộ phim họ chưa xem nhưng những khách

hàng có dữ liệu lịch sử gần giống với họ đã xem. Ngoài ra, công cụ tìm kiếm trên Google cũng sử dụng một dạng hệ thống gợi ý đơn giản.

2.1.1.3. Phân loại trong Hệ thống khuyến nghị

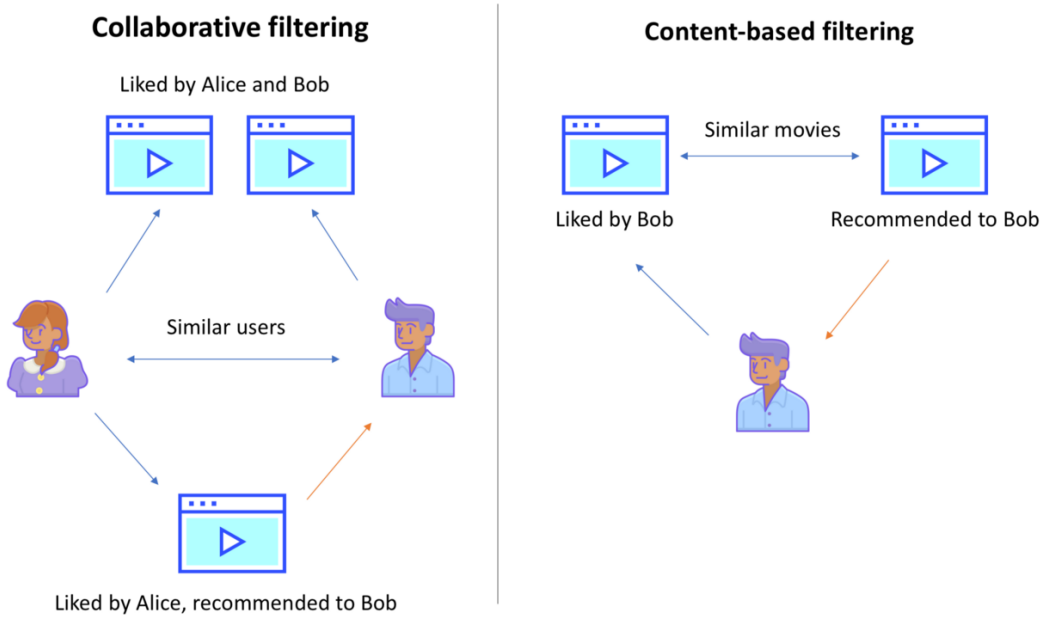


Hình 2.3: Phân nhánh của hệ thống khuyến nghị

Hệ thống khuyến nghị có hai loại, được cá nhân hóa (Personalized) và không được cá nhân hóa (Non- personalized). Trong các đề xuất không được cá nhân hóa, tất cả người dùng đều nhận được các đề xuất giống nhau. Trong các đề xuất được cá nhân hóa, những người dùng hoặc nhóm người dùng khác nhau nhận được các đề xuất khác nhau. Từ nhánh cá nhân hóa trong hệ thống khuyến nghị trên thì có hai loại hệ thống khuyến nghị chính. Leskovec, Rajaraman và Ullman (2014) nói rằng: "Một loại hệ thống đề xuất thì dựa trên nội dung; nó đo lường sự tương đồng bằng cách tìm kiếm các đặc điểm chung của các mặt hàng. Hệ thống đề xuất thứ hai sử dụng tính năng lọc cộng tác; những hệ thống này đo lường mức độ tương tự của người dùng theo mặt hàng họ thích và / hoặc đo lường mức độ tương tự của các mặt hàng bởi những người dùng thích chúng (trang 339).

" Nói theo cách khác, các đề xuất có thể dựa trên các đặc điểm của các mặt hàng đã được khách hàng đó mua, đọc hoặc xem trước đây để xác định độ tương đồng của chúng với các mặt hàng mới. Cách tiếp cận này yêu cầu một số loại dữ liệu liên quan đến hành vi của người dùng. Một cách tiếp cận khác là dựa trên đề xuất từ hành vi của những người dùng khác có nét giống gần nhất với người dùng mục tiêu và sau đó đề xuất các mặt hàng mà những người dùng này đã mua, đọc hoặc xem cho người dùng mục tiêu.

Trong nghiên cứu này sẽ ứng dụng phương pháp collaborative filtering (lọc cộng tác) nên phương pháp này sẽ được chú trọng giải thích hơn. Thuật ngữ "lọc cộng tác" được đặt ra bởi Goldberg và cộng sự (1992) người đã đề xuất rằng quá trình lọc thông tin sẽ trở nên hiệu quả hơn khi có sự tham gia của con người. Trong khi lọc theo nội dung là đề xuất các mục dựa trên sự so sánh giữa nội dung của các mục và dữ liệu hồ sơ người dùng thì đối với phương pháp lọc cộng tác, các đề xuất mới cho người dùng sẽ dựa trên việc lọc các mục mà người dùng có thể thích dựa trên cơ sở của những người dùng có hành vi gần tương tự [4]. Nó hoạt động bằng cách tìm kiếm một nhóm lớn các người dùng và tìm một nhóm nhỏ các người dùng có thị hiếu tương tự như một người dùng cụ thể. Nó xem xét các mục họ thích và kết hợp chúng để tạo ra một danh sách đề xuất được xếp hạng.

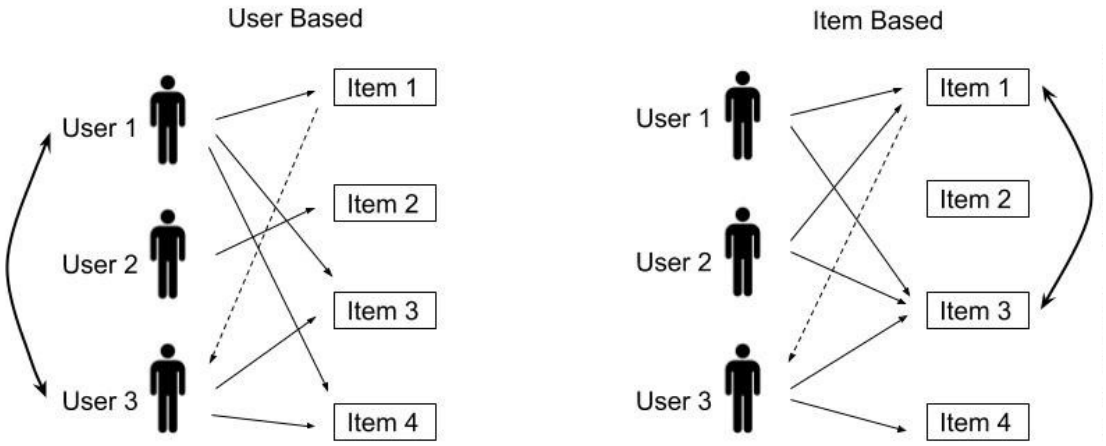


Hình 2.4: Phân loại Collaborative filtering và Content-based filtering (Ubajaka CJ, 2020)

Phương pháp	Dữ liệu đầu vào	Tiến trình xử lý
Lọc cộng tác (Collaborative filtering)	Các đánh giá của người dùng (User) đối với đối tượng (Item)	Tìm kiếm, lọc và nhận ra người dùng A có hành vi (sở thích) gần tương tự nhất với một số người dùng B → gợi ý cho người dùng A.
Lọc theo nội dung (Content-based filtering)	Các đặc điểm của đối tượng (Item)	Tạo ra mô hình mô tả sở thích người dùng, sau đó đánh giá mức độ ưa thích đối với đối tượng → gợi ý những đối tượng gần, tương đồng, dự đoán mà người dùng có thể thích.

Bảng 2.1: So sánh Lọc cộng tác và Lọc theo nội dung

Quy trình lọc cộng tác chia thành hai nhánh chính: lọc mục người dùng (user) và lọc mục- mục (item-item). Lọc mục người dùng đưa một người dùng cụ thể, tìm những người dùng tương tự với người dùng đó dựa trên mức độ tương tự về hành vi và đề xuất giống với người dùng tương tự đó. Ngược lại, tính năng lọc item-item sẽ lấy một mục, tìm những người dùng đã thích mục đó và tìm các mục khác mà những người dùng đó hoặc những người dùng tương tự cũng thích sau đó đề xuất.



Hình 2.5: Phân loại UBCF và IBCF (Ankita Prasad, 2020)

2.1.1.4. Vai trò của Hệ thống khuyến nghị

- **Đối với khách hàng**

Vai trò của hệ thống khuyến nghị phải nói là rất lớn bởi số lượng các sản phẩm, dịch vụ hoặc nội dung được cung cấp dường như quá nhiều và người dùng cũng khó có thể tìm được thứ mà bản thân cần trong mênh mông sản phẩm trên thị trường. Hệ thống khuyến nghị ra đời như một phương thuốc giúp khắc phục những thách thức liên quan đến quá tải thông tin và dường như là công cụ đặc biệt có giá trị cho những người dùng "thiếu kinh nghiệm" khi tham gia vào các quy trình cần đưa ra quyết định một cách chuyên sâu hơn.

Như vậy, hệ thống khuyến nghị đã giúp người dùng tiết kiệm thời gian, tăng tốc độ tìm kiếm và giúp người dùng truy cập tới nội dung họ quan tâm một cách dễ dàng hơn, đồng thời, gợi ý tới người dùng những đề xuất mới mà trước đây họ chưa từng biết đến.

- **Đối với doanh nghiệp**

Với khả năng của hệ thống khuyến nghị, các doanh nghiệp sử dụng chúng để giới thiệu sản phẩm mới hoặc phù hợp nhất tới người tiêu dùng, từ đó có thể giúp doanh nghiệp gia tăng doanh số nhờ các ưu đãi, sản phẩm, dịch vụ được khuyến nghị một cách cá nhân hóa, nâng cao trải nghiệm khách hàng. Điều này cải thiện lợi thế cạnh tranh của doanh nghiệp và giảm thiểu tỷ lệ khách hàng rời bỏ (churn rate) mà đến với đối thủ cạnh

tranh khi họ nhận thấy doanh nghiệp mới hiểu nhu cầu của họ và cung cấp cho họ những thứ họ muốn hơn là doanh nghiệp hiện tại.

Hệ thống khuyến nghị đã trở thành một thành phần không thể thiếu của các nền tảng trực tuyến cung cấp đa dạng các loại hình dịch vụ, từ các website thương mại điện tử tới nền tảng đào tạo trực tuyến. Năm 2013, báo cáo của McKinsey & Company cho thấy rằng 35% số lần mua hàng trên Amazon trực tiếp đến từ khả năng độc đáo của trang web là cung cấp các bài đánh giá sản phẩm tương tự được hỗ trợ bởi các thuật toán và mô hình dự đoán (hệ thống khuyến nghị). [5]

2.1.1.5. Ứng dụng của hệ thống khuyến nghị

Các ứng dụng của hệ thống giới thiệu bao gồm giới thiệu phim, nhạc, chương trình truyền hình, sách, tài liệu, trang web, hội nghị, danh lam thắng cảnh du lịch, tài liệu học tập, và liên quan đến các lĩnh vực thương mại điện tử, học tập điện tử, thư viện điện tử và dịch vụ kinh doanh điện tử.

Và những hệ thống khuyến nghị lớn như là Netflix, YouTube, Tinder và Amazon … Hệ thống thu hút người dùng bằng các đề xuất có liên quan dựa trên các lựa chọn mà họ đưa ra. Các hệ thống khuyến nghị có thể coi là một trong những ứng dụng thành công của trí tuệ nhân tạo và với dịch vụ trực tuyến ngày càng phát triển mạnh mẽ và dữ liệu ngày càng lớn như hiện nay thì tầm quan trọng của hệ khuyến nghị càng được nâng cao.

2.1.2. Các phần mềm và công cụ liên quan

2.1.2.1. Ngôn ngữ lập trình và thống kê

Hệ thống khuyến nghị thường được xây dựng bởi rất nhiều ngôn ngữ lập trình khác nhau, nhưng hiện nay chủ yếu có thể kể đến như là ngôn ngữ lập trình Python và ngôn ngữ thống kê R. Cụ thể trong nghiên cứu này, nhóm tác giả sẽ sử dụng R làm ngôn ngữ chính xuyên suốt.

2.1.2.2. Thư viện được sử dụng chính

- **Thư viện recommenderlab**

Thư viện recommenderlab là một gói khá phổ trong R hỗ trợ trong các nghiên cứu, phân tích và xây dựng hệ khuyến nghị dựa trên phương pháp chủ yếu là lọc cộng tác. Ý tưởng là dựa trên dữ liệu xếp hạng của nhiều người dùng cho nhiều mục (ví dụ: 1 đến 5 sao cho các bộ phim được lấy trực tiếp từ người dùng), người ta có thể dự đoán xếp hạng

của người dùng cho một mặt hàng mà cô ấy hoặc anh ta không biết hoặc tạo cho người dùng cái gọi là danh sách top-N gồm các mặt hàng được đề xuất. [6] Điều này hoàn toàn tương đồng với bản chất của lọc cộng tác như đã được định nghĩa ở trên.

- **Thư viện reshape2**

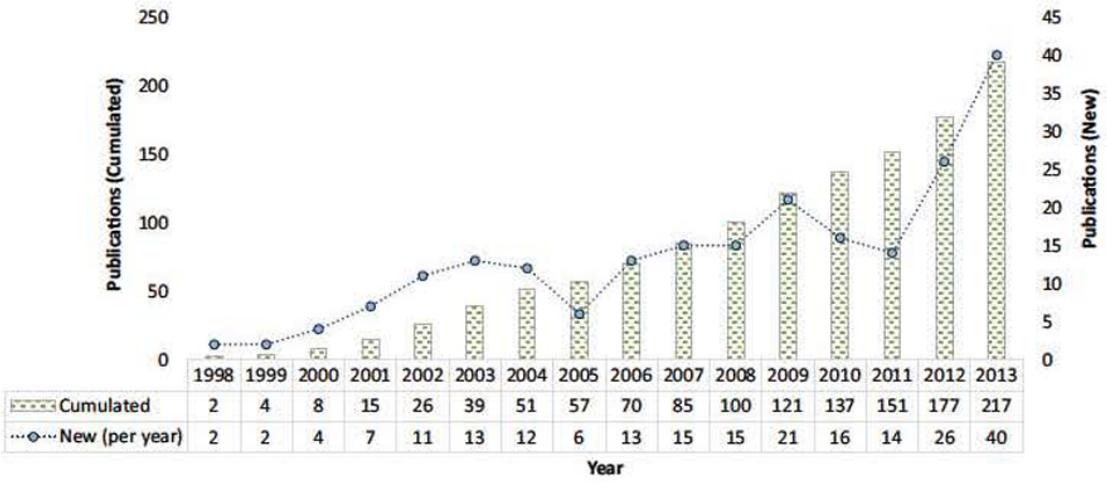
Reshape2 là một thư viện trong ngôn ngữ thống kê R được viết bởi Hadley Wickham hỗ trợ người dùng có thể dễ dàng chuyển đổi dữ liệu giữa các định dạng rộng và dài. Đây là một gói hỗ trợ cho việc xử lý dữ liệu.

- **Thư viện ggplot2**

Thư viện ggplot2 là một thư viện hỗ trợ trực quan hóa rất tốt trong R, đây có thể coi là một trong những thư viện quan trọng nhất trong R. Dựa trên thư viện này có thể vẽ rất nhiều biểu đồ dạng bar chart, line, plot, density, candle chart,pie,… và rất nhiều các đồ thị khác.Ngoài ra ggplot2 còn cho phép người dùng tùy chỉnh màu sắc, kích cỡ, theme, title, … để đồ thị được đẹp hơn. ggplot2 hỗ trợ rất nhiều trong việc thể hiện kết quả giúp những nhà nghiên cứu có cái nhìn tổng quan hơn.

2.2. Các nghiên cứu liên quan

Trong các công trình nghiên cứu khoa học trên thế giới và các tạp chí nổi tiếng về lĩnh vực số nói chung và trí tuệ nhân tạo nói riêng, đề tài về các hệ thống khuyến nghị, cách xây dựng, những lợi ích chúng đem lại đã là một đề tài rất phổ biến trong việc nghiên cứu.



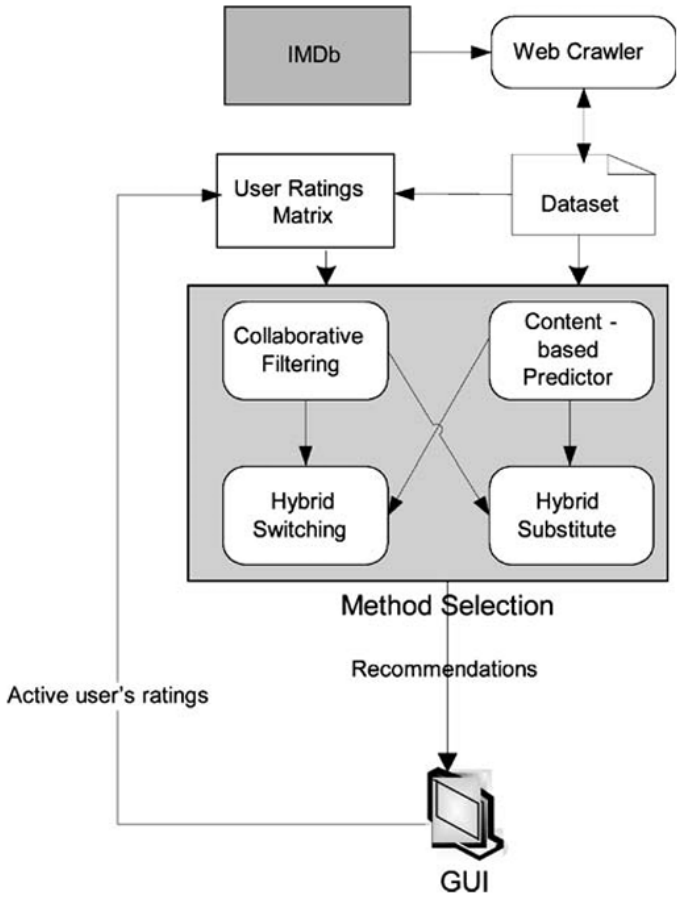
	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Cumulated	2	4	8	15	26	39	51	57	70	85	100	121	137	151	177	217
New (per year)	2	2	4	7	11	13	12	6	13	15	15	21	16	14	26	40

Hình 2.6: Ấn phẩm hàng năm tới 2013 trong lĩnh vực nghiên cứu hệ thống khuyến nghị. [7]

Năm 1998, Giles và cộng sự. đã giới thiệu nghiên cứu đầu tiên về hệ thống khuyến nghị như một phần của dự án CiteSeer [8]. Kể từ đó, ít nhất 216 bài báo liên quan đến 120 cách tiếp cận hệ thống khuyến nghị trên giấy nghiên cứu đã được xuất bản. Điều đó cho thấy sự phát triển mạnh và độ quan tâm đối với hệ thống khuyến nghị đã tăng cao. Điều này dễ hiểu vì nhu cầu thương mại, giao dịch trực tuyến đã nở rộng. Và các bài báo liên quan đến phương pháp lọc cộng tác cũng đã tăng dần. Điều đó được chứng minh qua các nghiên cứu sau:

Đầu tiên là nghiên cứu của Sharma và cộng sự (2013) rằng bài báo này đã đưa ra một số đánh giá chung cho hệ thống khuyến nghị. Các phương pháp tiếp cận có thể được phân loại thành ba phần là lọc cộng tác, khuyến nghị kết hợp và khuyến nghị dựa trên nội dung. Hơn nữa, các vấn đề khác về hệ khuyến nghị cũng được thảo luận trong bài báo này. Bài báo này cho một cái nhìn chung về hệ thống khuyến nghị.

George và cộng sự (2008) đã trình bày phương pháp tiếp cận hybrid (kết hợp) dựa trên nội dung (content based) và lọc cộng tác (collaborative). Kỹ thuật này được sử dụng trong hệ khuyến nghị đề xuất. Cách tiếp cận được trình bày cung cấp đánh giá thực nghiệm của kỹ thuật kết hợp đối với các kỹ thuật hiện có của lọc cộng tác và dựa trên nội dung và cung cấp các kết luận hữu ích về kỹ thuật này.



Hình 2.7: Mô hình khuyến nghị dựa trên phương pháp hybrid của George và cộng sự

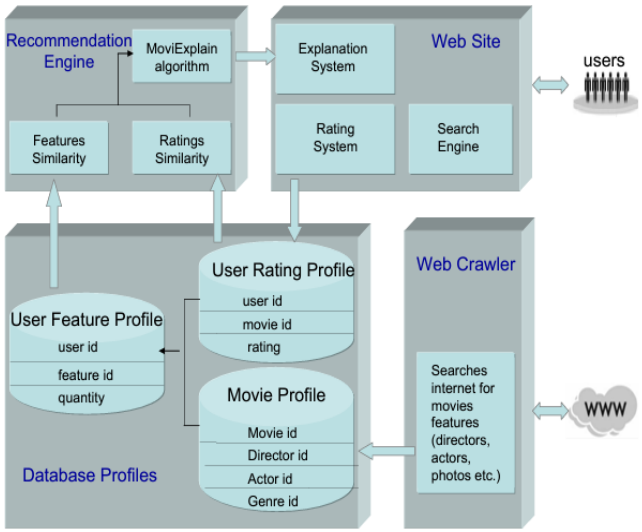
Trong nghiên cứu của Fernandez và cộng sự (2014) đã trình bày một hệ khuyến nghị cho những người dùng đến rạp chiếu phim hoặc rạp hát. Phương pháp được đề xuất sử dụng thuật toán Slope One. Tác giả đi sâu vào nghiên cứu áp dụng các kỹ thuật này vào các vấn đề cụ thể, bao gồm giải pháp cho các vấn đề thực tế và một số cải tiến áp dụng cho các trường hợp. Nghiên cứu này có đặc điểm chung khi sử dụng dữ liệu là điểm Imdb.

Trong nghiên cứu khác của Li and Yamada (2004), tác giả đã sử dụng thuật toán học quy nạp đã được đề xuất và thuật toán này được áp dụng cho quy trình khuyến nghị. Trong công việc này, cây quyết định đã được xây dựng thay vì sử dụng sự tương đồng giữa người dùng và người dùng. Cây quyết định này hiển thị tham chiếu người dùng. Có thể nói rằng các đề xuất được thực hiện bằng cách phân loại cây quyết định. Các kết quả cho thấy rằng kỹ thuật được trình bày phù hợp với việc khảo sát các vấn đề quy mô rất lớn và có thể ước tính được các câu hỏi chất lượng cao. Ở đây, tác giả sử dụng phương pháp content-based, khác với phương pháp collaborative mà trong nghiên cứu này sẽ

hướng đến, nhưng bài của tác giả có nhiều điều đáng tham khảo và việc khi một tập dữ liệu ở quy mô lớn hơn.

Theo đó, còn có bài báo của Christakou và cộng sự (2005). Bài báo đề cập và sử dụng một kỹ thuật phân cụm đã được đề xuất mà phụ thuộc vào phân tích bán giám sát (semi - supervised). Trong bài báo này, cách tiếp cận đã trình bày được sử dụng để tạo ra một hệ thống đề xuất các bộ phim có sự kết hợp giữa thông tin kinh tế và nội dung. Hệ thống đại diện được kiểm tra trên Movie Lens DS, cung cấp các khuyến nghị về độ chính xác cao. Bài báo này cho một cái nhìn khác và thêm nhiều kiến thức liên quan đến phân tích bán giám sát, một kỹ thuật khá hay và dễ áp dụng khi ta không thể gắn nhãn cho tất cả các dữ liệu hoặc dữ liệu quá lớn.

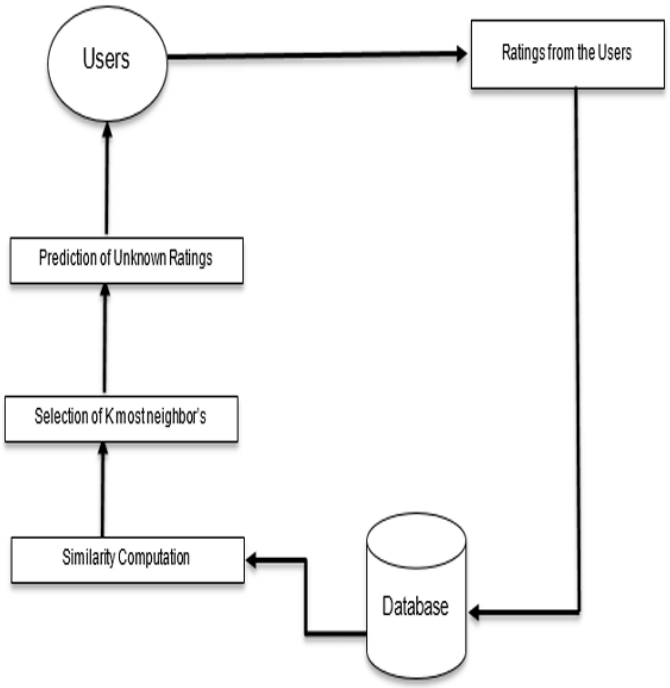
Trong bài nghiên cứu của Symeonidis và cộng sự (2009) cho thấy những kết quả khả quan vượt mong đợi với đề xuất cho người dùng rất tốt. Bài báo sử dụng phương pháp phân tích kết hợp (hybrid) và sử dụng một cách tiếp cận được gọi là Lọc cộng tác tăng cường nội dung (Content-Boosted Collaborative Filtering). Ý tưởng cơ bản của CBCF là sử dụng các dự đoán dựa trên nội dung để lấp đầy ma trận xếp hạng mục của người dùng.



Hình 2.8: Hệ thống khuyến nghị MoviExplain của Symeonidis và cộng sự

P.Abhilash (2018) đã triển khai phương pháp của mình, mô tả phương pháp lọc cộng tác dựa trên mục. Tác giả đã xây dựng các đề xuất dựa trên dữ liệu lịch sử. Đầu tiên ông đã xác định ma trận xếp hạng mặt hàng của người dùng và xem xét các mối quan hệ cho nhiều mặt hàng, sau đó sử dụng các mối quan hệ này để tính toán các đề xuất phù

hợp cho từng người dùng. Tác giả đã sử dụng tập dữ liệu user-item database của Netflix. Để đánh giá mô hình, tác giả đã sử dụng Mean Absolute Method (MAE). Trong bài báo này, tác giả đã tiếp cận bằng phương pháp lọc cộng tác dựa trên mục (item), có nhiều điều đáng tham khảo.



Hình 2.9: Mô hình hệ thống khuyến nghị dựa trên lọc cộng tác của Abhilash

CHƯƠNG 3: ĐỀ XUẤT MÔ HÌNH NGHIÊN CỨU

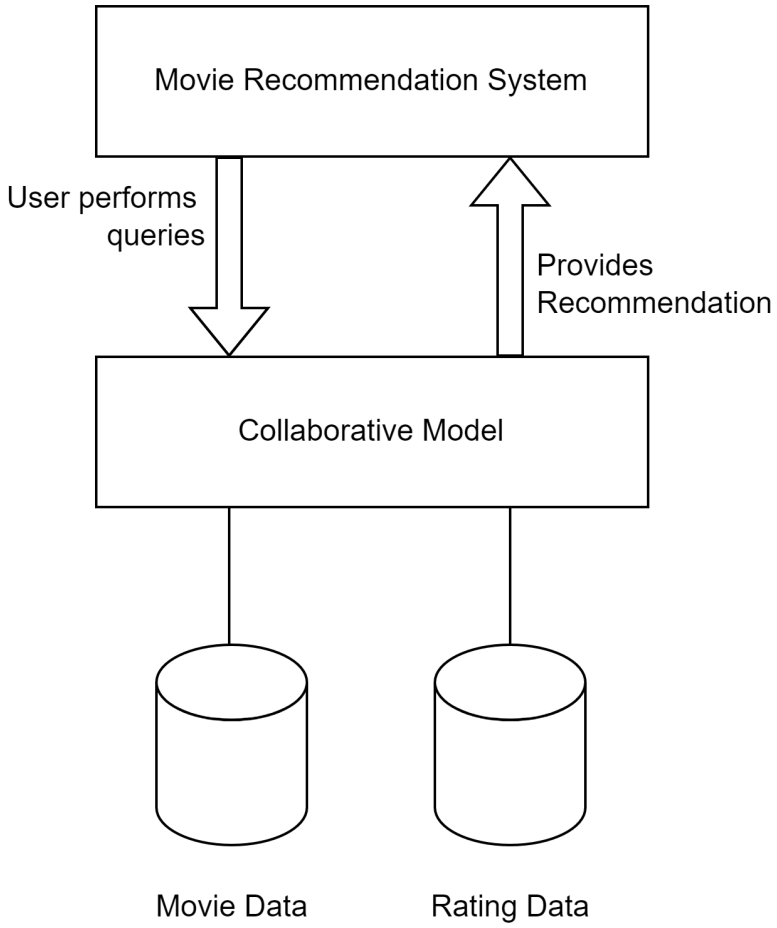
3.1. Giả thuyết nghiên cứu.

Với mục tiêu xây dựng thành công mô hình hệ khuyến nghị phim nhằm nâng cao trải nghiệm của người dùng, hệ thống được đề xuất cần áp dụng các kỹ thuật chọn lọc chuyên nghiệp, ngoài cách lọc ngẫu nhiên truyền thống. Như đã đề cập tại khung lý thuyết liên quan, hệ khuyến nghị (Recommender system) thường sử dụng một hoặc kết hợp cả 2 mô hình Chọn lọc theo Cộng tác (Collaborative Filtering) và Chọn lọc theo nội dung (Content Filtering), tuy nhiên tại đây, chúng tôi chỉ tập trung triển khai mô hình Collaborative Filtering dựa trên 2 mô hình con: Chọn lọc theo yếu tố sản phẩm (Item-based collaborative filtering - IBCF) và Chọn lọc theo người dùng (Users-based collaborative filtering - UBCF). (P.Abhilash 2018)

3.2. Mô hình và phương pháp thực hiện nghiên cứu.

3.2.1. Mô hình

Kế thừa từ kỹ thuật lọc cộng tác theo yếu tố sản phẩm (Item-based Collaborative Filtering Technique) của Abhilash năm 2018. Nghiên cứu của chúng tôi sử dụng tập dữ liệu của IMDB (Internet Movie Database là một trang cơ sở dữ liệu trực tuyến về điện ảnh thế giới _ Wikipedia) đủ lớn với hơn 100 nghìn dòng dữ liệu, trong đó lưu trữ thông tin cơ bản về bộ phim (movie data) và thông tin đánh giá của người dùng về một hoặc nhiều phim (rating data). Đồng thời, một số thuật toán khác để đánh giá điểm tương đồng (điều kiện quan trọng để đưa ra đề xuất trong mô hình khuyến nghị) như Cosine similarity và K-Nearest-Neighbor (KNN).



Hình 3.1: Cấu trúc sơ bộ của Mô hình khuyến nghị phim (Movie Recommendation system) _ *Nguồn*

Vòng đời hệ thống khuyến nghị khởi chạy dựa trên điểm đánh giá phim từ những người dùng (tập dữ liệu rating), từ đó thông qua các thuật toán cho ra kết quả tương thích với những phim (IBCF) mà người dùng nhận định tốt (có điểm đánh giá - rating cao) dựa trên yếu tố thể loại phim (tập dữ liệu movie) và hơn cả đó là đưa ra danh sách khuyến nghị dựa trên tính tương đồng giữa những người dùng (UBCF).

3.2.2. Phương pháp nghiên cứu.

Để hoàn tất việc gợi ý mục tin, đặc biệt là đối với ngành công nghiệp phim hiện đại, nhìn chung tập dữ liệu cần đảm bảo 2 yếu tố là người dùng (user) và sản phẩm (movie). Đề cập đến tập dữ liệu IMDB mà nhóm sử dụng đã đáp ứng được yêu cầu trên, đồng thời tập dữ liệu được sử dụng còn mang ý nghĩa phân loại: nhóm dữ liệu rõ ràng (có thể sử dụng ngay trong lúc triển khai mô hình) và dữ liệu ẩn (dựa trên dữ liệu rõ ràng để

đưa ra các yếu tố như xu hướng người dùng hay movie biases), tất cả đều mang ý nghĩa trong công tác đề ra kết quả dự đoán. (Bùi Văn Minh)

Nghiên cứu của nhóm ứng dụng lọc cộng tác theo yếu tố sản phẩm (item based collaborative filtering) để đưa ra khuyến nghị, vì lý do "khẩu vị" thường thức của người dùng luôn biến đổi, nhưng nhìn chung yếu tố then chốt của sản phẩm (movie) không quá khác biệt theo thời gian. Dưới đây là quy trình bắt buộc để dự án diễn ra một cách hiệu quả.

- Tải dữ liệu:
    - Tiến hành thu thập và tải dữ liệu, đồng thời thực hiện một vài thao tác để có thể quan sát tổng quát về các biến cũng như ý nghĩa của tập dữ liệu nghiên cứu (IMDB dataset). Tiến hành sáp nhập 2 tập dữ liệu (movie và rating)
- Phân tách dữ liệu:
    - Loại bỏ các yếu tố dư thừa trong tập dữ liệu.
- Tiền xử lý dữ liệu:
    - Tiến hành chuyển đổi định dạng biến và biến thể loại (genres) được tách riêng có định dạng ma trận nhị phân 0-1 (vì một phim có thể có nhiều thể loại).

Việc làm tiếp theo đó chính là xác định độ tương thích giữa các đối tượng người dùng (users) và phim (movies) bằng Cosine similarity.

3.3. Tổng quan về quy trình thực hiện.

To provide an overview about the entire project, this chart below would descript and explain about our working process to build Movie Recommendation system based on IBCF technique.

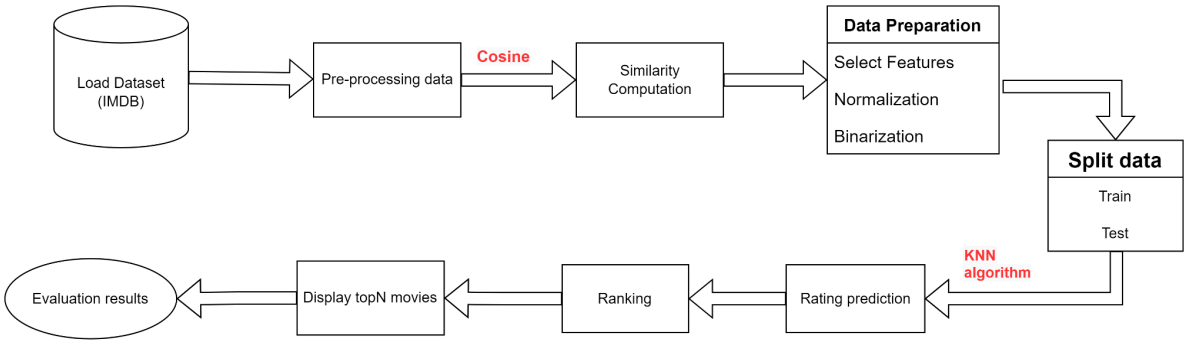


Fig. Workflow of Movie Recommender System based on Item-based Collaborative Filtering (IBCF).

The Movie Recommender System model based on the Item-based Collaborative Filtering (IBCF) technique is built in the following step-by-step sequence:

- Step 1: Download and import IMDB data into RStudio platform, including 2 movie and rating files.
- Step 2: Perform Pre-processing data by selecting useful variables and converting the data type of each variable to meet the model's requirements.
- Step 3: Use Cosine similarity technique to calculate the similarity between users, repeat with movie object (movie).
- Step 4: Data Preparation serves the model through 3 basic steps:
    - Selecting the appropriate variables (Select Features) to run the models.
    - Normalization to limit data bias in the analysis implementation.
    - Binarization by building a matrix containing only 2 values 0 and 1 for the rating element in the set. (0 if rating > 3 and vice versa)
- Step 5: Split data on a ratio of 80-20, to see how effective the model is (80% of observational data is provided for training data) and 20% leaf for testing (testing data).
- Step 6: Apply K-Nearest-Neighbour (KNN) algorithm to determine the utility clustering matrix and similar items.
- Step 7: Arrange the results obtained in step 6 in ascending order by similarity score.
- Step 8: Suggestions to users. (In this case, we would choose top 10 the most compatible movies from the total recommended list)
- Step 9: Evaluate proposed results by using RMSE, MSE and MAE method.

3.4. Sơ lược về tập dữ liệu (IMDB dataset).

The datasets used in this project are from an open database source. The IMDB dataset includes 2 files in Excel format, movie and rating datasets.

Movie dataset stores 3 characteristics that are used to describe and identify a movie :

- movieId: The identifier of a movie
- title: Movie's name
- genres: Genre of movies

Note: 1 movie can belong to 1 or more genres at the same time and 1 genre also has many movies.

The rating file summarizes each user's rating by their rating and has 4 variables:

- userId: the identifier of a user
- movieId: the identifier of a movie
- Rating: User rating on a scale of (0.5 - 5), the valid interval is 0.5 for the measurement range.
- timestamp: timestamp the user published the review.

In addition, the data set is also recorded the association between the two data sets by movieId key.

CHƯƠNG 4. PHÂN TÍCH DỮ LIỆU VÀ KẾT QUẢ NGHIÊN CỨU

1. Bộ dữ liệu

Để xây dựng hệ thống đề xuất của mình, nhóm nghiên cứu đã sử dụng bộ dữ liệu IMDB từ MovieLens bao gồm movies.csv và ratings.csv. Áp dụng mô hình đã được sử dụng trong Dự án Hệ thống Đề xuất của mình tại đây. Dữ liệu này bao gồm 105339 ratings (xếp hạng) được áp dụng trên 10329 movies (phim).

MoviesRecommendationSystem.R | movies | ratings

Filter

	userId	movieId	rating	timestamp
1	1	16	4.0	1217897793
2	1	24	1.5	1217895807
3	1	32	4.0	1217896246
4	1	47	4.0	1217896556
5	1	50	4.0	1217896523
6	1	110	4.0	1217896150
7	1	150	3.0	1217895940
8	1	161	4.0	1217897864
9	1	165	3.0	1217897135
10	1	204	0.5	1217895786
11	1	223	4.0	1217897795
12	1	256	0.5	1217895764
13	1	260	4.5	1217895864
14	1	261	1.5	1217895750
15	1	277	0.5	1217895772
16	1	296	4.0	1217896125
17	1	318	4.0	1217895860
18	1	349	4.5	1217897058
19	1	356	3.0	1217896231
20	1	377	2.5	1217896373

Showing 1 to 20 of 105,339 entries, 4 total columns

20 dòng đầu của bộ dữ liệu movies.

MoviesRecommendationSystem.R | movies | ratings

Filter

	movieId	title	genres
1	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	2	Jumanji (1995)	Adventure|Children|Fantasy
3	3	Grumpier Old Men (1995)	Comedy|Romance
4	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	5	Father of the Bride Part II (1995)	Comedy
6	6	Heat (1995)	Action|Crime|Thriller
7	7	Sabrina (1995)	Comedy|Romance
8	8	Tom and Huck (1995)	Adventure|Children
9	9	Sudden Death (1995)	Action
10	10	GoldenEye (1995)	Action|Adventure|Thriller
11	11	American President, The (1995)	Comedy|Drama|Romance
12	12	Dracula: Dead and Loving It (1995)	Comedy|Horror
13	13	Balto (1995)	Adventure|Animation|Children
14	14	Nixon (1995)	Drama
15	15	Cutthroat Island (1995)	Action|Adventure|Romance
16	16	Casino (1995)	Crime|Drama
17	17	Sense and Sensibility (1995)	Drama|Romance
18	18	Four Rooms (1995)	Comedy
19	19	Ace Ventura: When Nature Calls (1995)	Comedy
20	20	Money Train (1995)	Action|Comedy|Crime|Drama|Thriller

Showing 1 to 20 of 10,329 entries, 3 total columns

20 dòng đầu của bộ dữ liệu ratings.

2. Mô tả thống kê

Sau khi truy xuất dữ liệu từ 2 tệp csv thành 2 dataframe: movies.csv thành movies và ratings.csv thành ratings, chúng ta có thể mô tả thống kê bằng hàm summary() và hàm head() để biểu diễn 6 dòng đầu tiên của từng dataframe.

```
> summary(movies)
    movieId          title              genres
 Min.   :     1   Length:10329       Length:10329
 1st Qu.:  3240   Class :character   Class :character
 Median :  7088   Mode  :character   Mode  :character
 Mean   : 31924
 3rd Qu.: 59900
 Max.   :149532
> head(movies)
  movieId                               title                                        genres
1       1                    Toy Story (1995) Adventure|Animation|Children|Comedy|Fantasy
2       2                      Jumanji (1995)                  Adventure|Children|Fantasy
3       3             Grumpier Old Men (1995)                              Comedy|Romance
4       4            Waiting to Exhale (1995)                        Comedy|Drama|Romance
5       5 Father of the Bride Part II (1995)                                      Comedy
6       6                            Heat (1995)                       Action|Crime|Thriller
```

```
> summary(ratings)
     userId          movieId           rating        timestamp
 Min.   :  1.0   Min.   :     1   Min.   :0.500   Min.   :8.286e+08
 1st Qu.:192.0   1st Qu.:  1073   1st Qu.:3.000   1st Qu.:9.711e+08
 Median :383.0   Median :  2497   Median :3.500   Median :1.115e+09
 Mean   :364.9   Mean   : 13381   Mean   :3.517   Mean   :1.130e+09
 3rd Qu.:557.0   3rd Qu.:  5991   3rd Qu.:4.000   3rd Qu.:1.275e+09
 Max.   :668.0   Max.   :149532   Max.   :5.000   Max.   :1.452e+09
> head(ratings)
  userId movieId rating  timestamp
1      1      16    4.0 1217897793
2      1      24    1.5 1217895807
3      1      32    4.0 1217896246
4      1      47    4.0 1217896556
5      1      50    4.0 1217896523
6      1     110    4.0 1217896150
>
```

Bảng trên cho thấy cột userId và movieId tồn tại các giá trị là các số nguyên. Hơn nữa, chúng ta cần chuyển đổi các thể loại tồn tại trong các dataframe thành định dạng dễ sử dụng hơn. Để làm như vậy, trước tiên nhóm nghiên cứu sẽ tạo một ma trận bao gồm các thể loại tương ứng cho mỗi bộ phim ở phần tiếp theo.

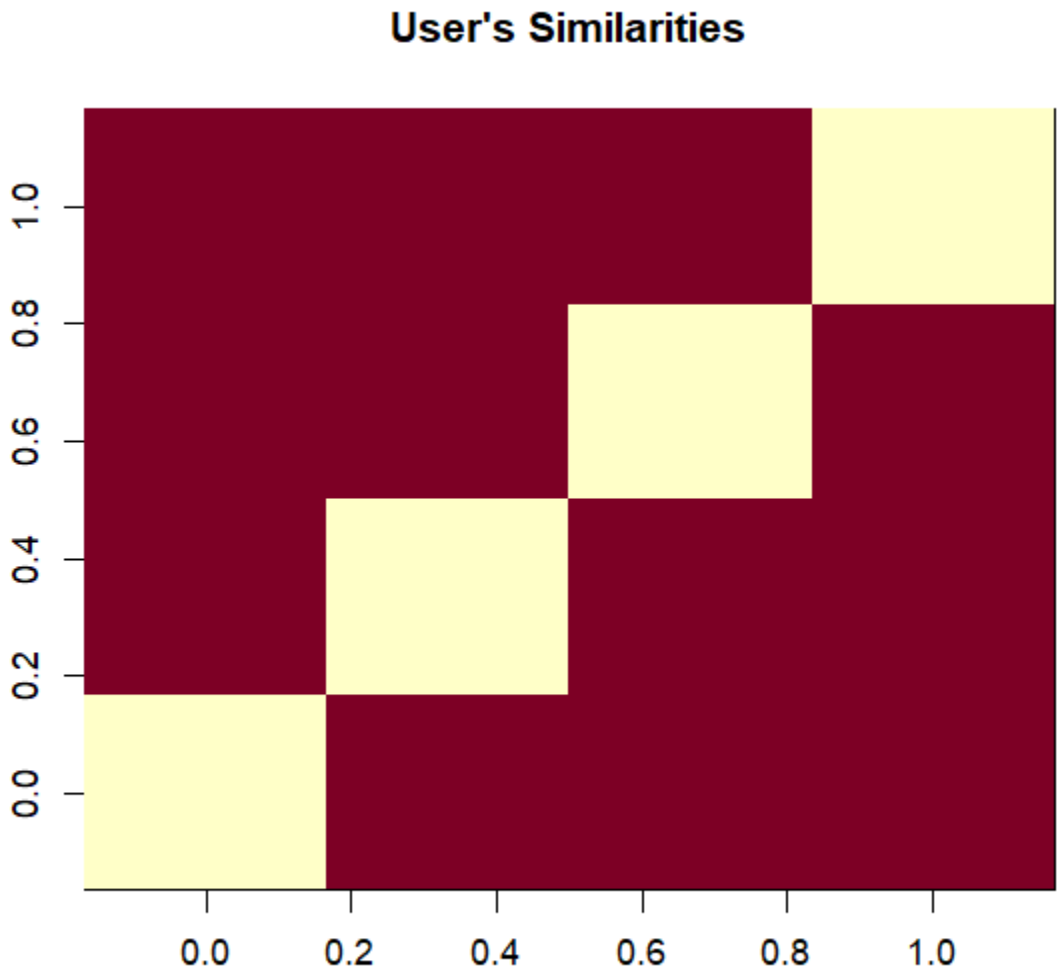
3. Kết quả mô hình - Phương pháp - Thuật toán

3.1. Collaborative Filtering

Có thể được hiểu là khám phá những dữ liệu tương đồng, nó liên quan đến việc đề xuất phim cho người dùng dựa trên việc thu thập sở thích từ nhiều người dùng khác nhau. Ví dụ: nếu người dùng A thích xem phim hành động và người dùng B cũng vậy, thì những bộ phim mà người dùng B sẽ xem trong tương lai sẽ được giới thiệu cho người A và ngược lại. Do đó, việc giới thiệu phim phụ thuộc vào việc tạo ra mối quan hệ tương đồng giữa hai người dùng.

Với sự trợ giúp của Recommenderlab, ta có thể tính toán được các điểm tương đồng bằng cách sử dụng các toán tử khác nhau như cosine, pearson cũng như jaccard. Ở nghiên cứu này, nhóm nghiên cứu chọn phương pháp cosine.

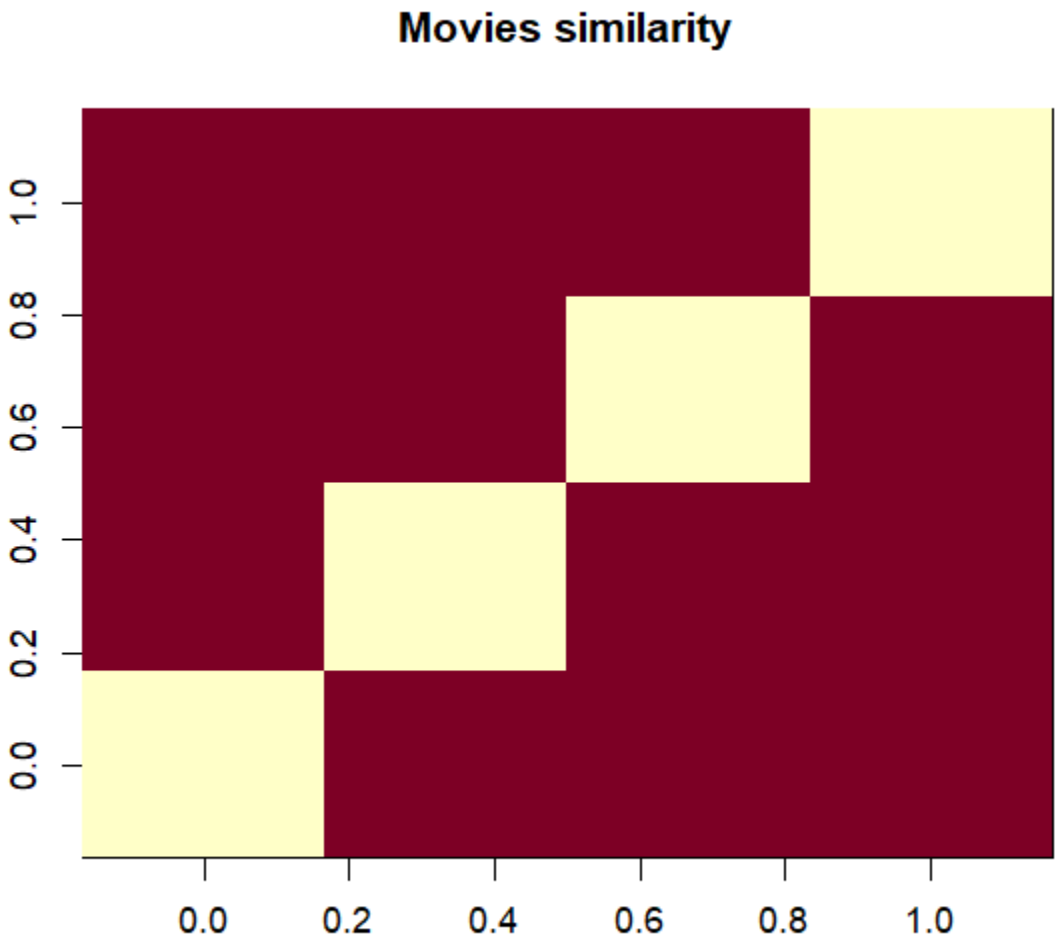
```
> # Exploring Similar Data
> # Suggesting movies to the users that are based on collecting preferences from many other users.
> similarity_mat <- similarity(ratingMatrix[1:4, ], method = "cosine", which = "users")
> as.matrix(similarity_mat)
          1         2         3         4
1 0.0000000 0.9760860 0.9641723 0.9914398
2 0.9760860 0.0000000 0.9925732 0.9374253
3 0.9641723 0.9925732 0.0000000 0.9888968
4 0.9914398 0.9374253 0.9888968 0.0000000
> image(as.matrix(similarity_mat), main = "User's Similarities")
>
```



Trong ma trận trên, mỗi hàng và cột đại diện cho một người dùng. Nhóm đã lấy bốn người dùng và mỗi ô trong ma trận này đại diện cho sự giống nhau được chia sẻ giữa hai người dùng.

Sau đó chúng ta đi xác định sự giống nhau trong các lựa chọn giữa các bộ phim.

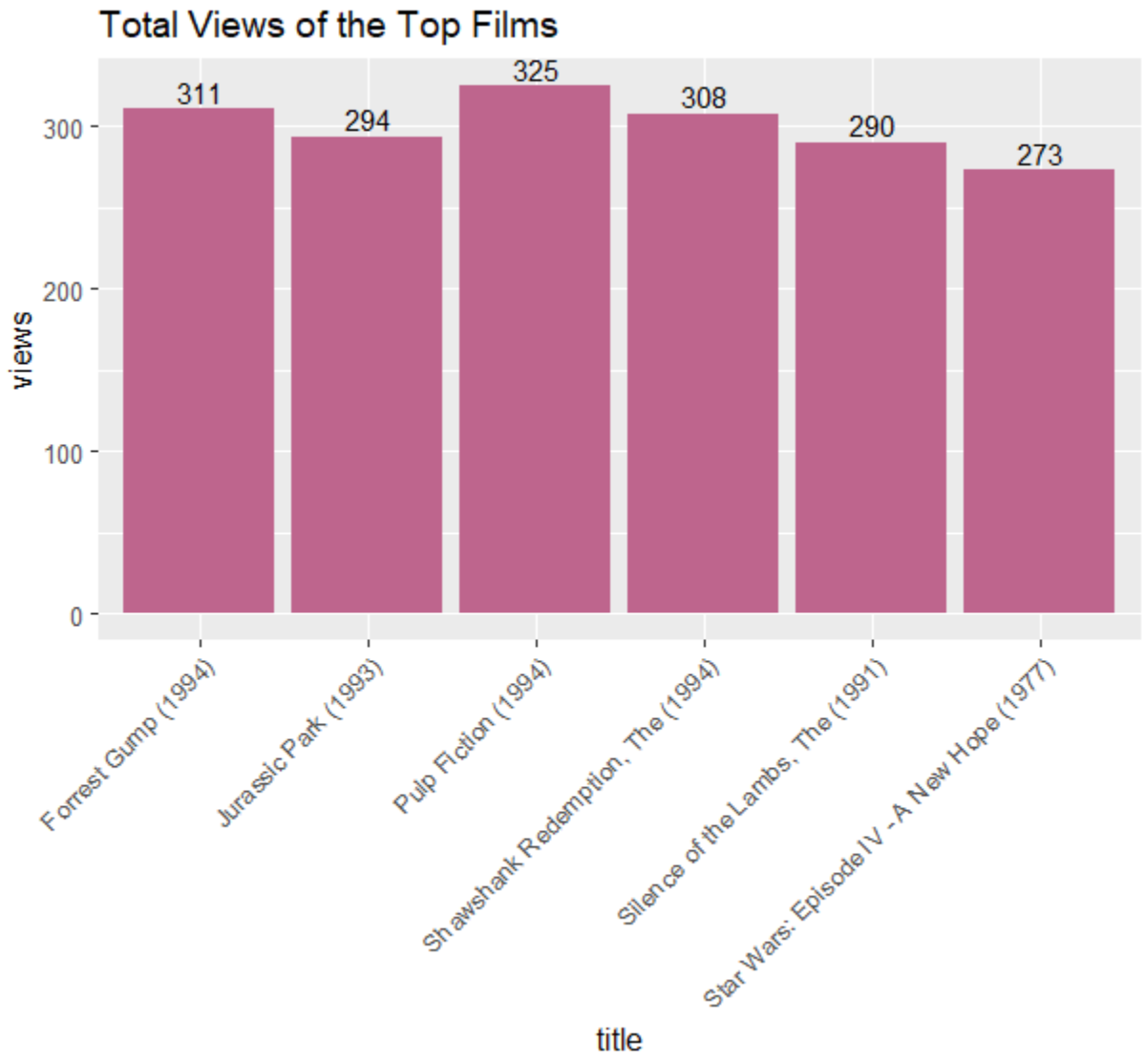
```
> movie_similarity <- similarity(ratingMatrix[, 1:4], method = "cosine", which = "items")
> as.matrix(movie_similarity)
          1         2         3         4
1 0.0000000 0.9669732 0.9559341 0.9101276
2 0.9669732 0.0000000 0.9658757 0.9412416
3 0.9559341 0.9658757 0.0000000 0.9864877
4 0.9101276 0.9412416 0.9864877 0.0000000
> image(as.matrix(movie_similarity), main = "Movies similarity")
>
```



Trước tiên, ta đếm số lượt xem trong một bộ phim và sau đó sắp xếp chúng trong một bảng để nhóm chúng theo thứ tự giảm dần.

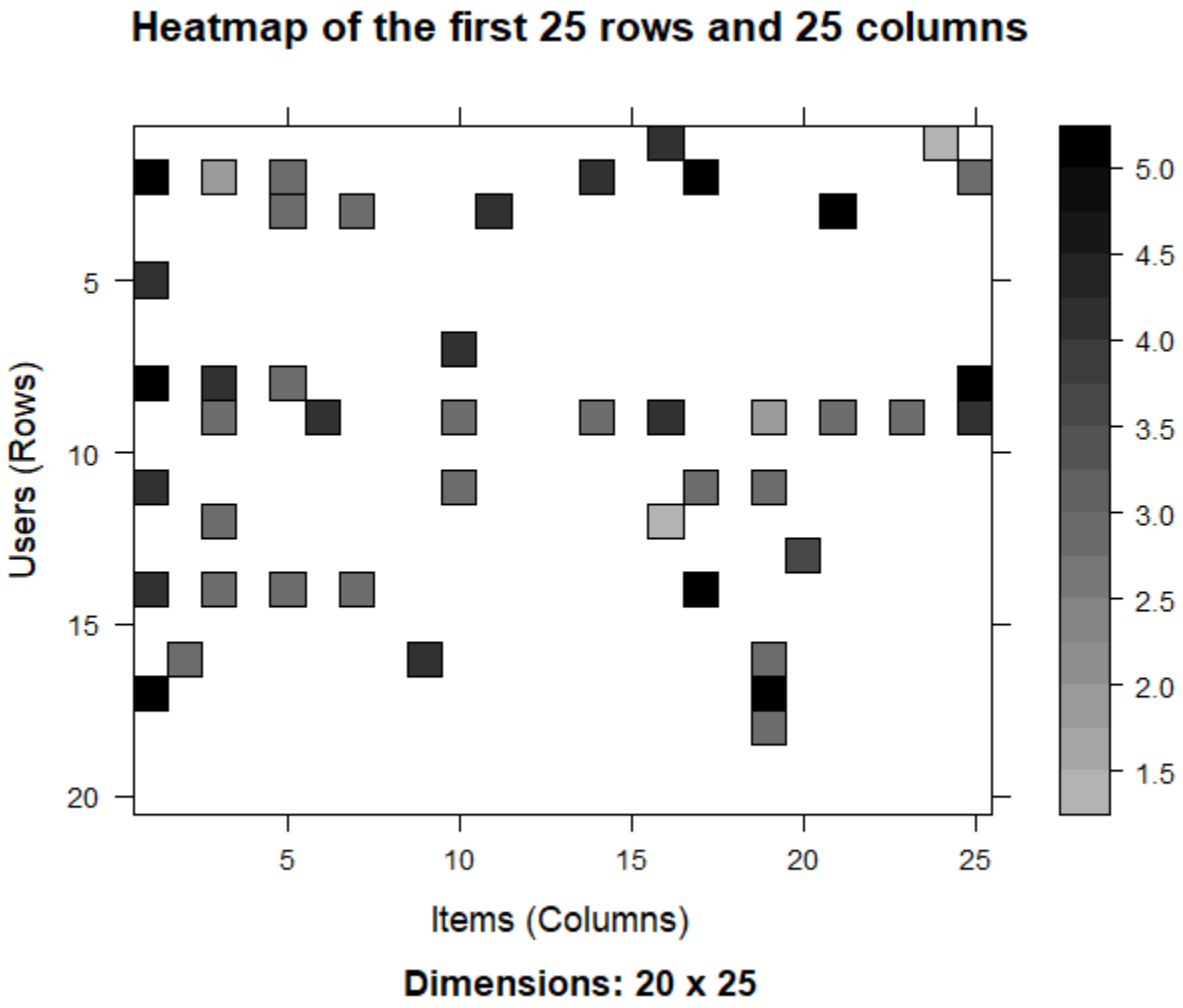
```
> # Most Viewed Movies Visualization
> movie_views <- colCounts(ratingMatrix) # Count views for each movie
> table_views <- data.frame(movie = names(movie_views),
+                           views = movie_views) # Create dataframe of views
> table_views <- table_views[order(table_views$views,
+                                  decreasing = TRUE), ] # Sort by number of views
> table_views$title <- NA
> for (index in 1:10325){
+   table_views[index,3] <- as.character(subset(movies,
+                                               movies$movieId == table_views[index,1])$title)
+ }
> table_views[1:6,]
    movie views                                    title
296   296   325                      Pulp Fiction (1994)
356   356   311                      Forrest Gump (1994)
318   318   308          Shawshank Redemption, The (1994)
480   480   294                     Jurassic Park (1993)
593   593   290          Silence of the Lambs, The (1991)
260   260   273 Star Wars: Episode IV - A New Hope (1977)
>
```

Đây là biểu đồ cột thể hiện những bộ phim được xem nhiều nhất trong tập dữ liệu, được thực hiện bằng ggplot2.



Từ biểu đồ trên ta thấy được bộ phim được xem nhiều nhất là Pulp Fiction, ở vị trí thứ 2 là Forrest Gump.

Bây giờ, chúng ta sẽ sử dụng bản đồ nhiệt cho việc xếp hạng phim, chứa 25 hàng và 25 cột đầu tiên.



3.2. Chuẩn bị dữ liệu

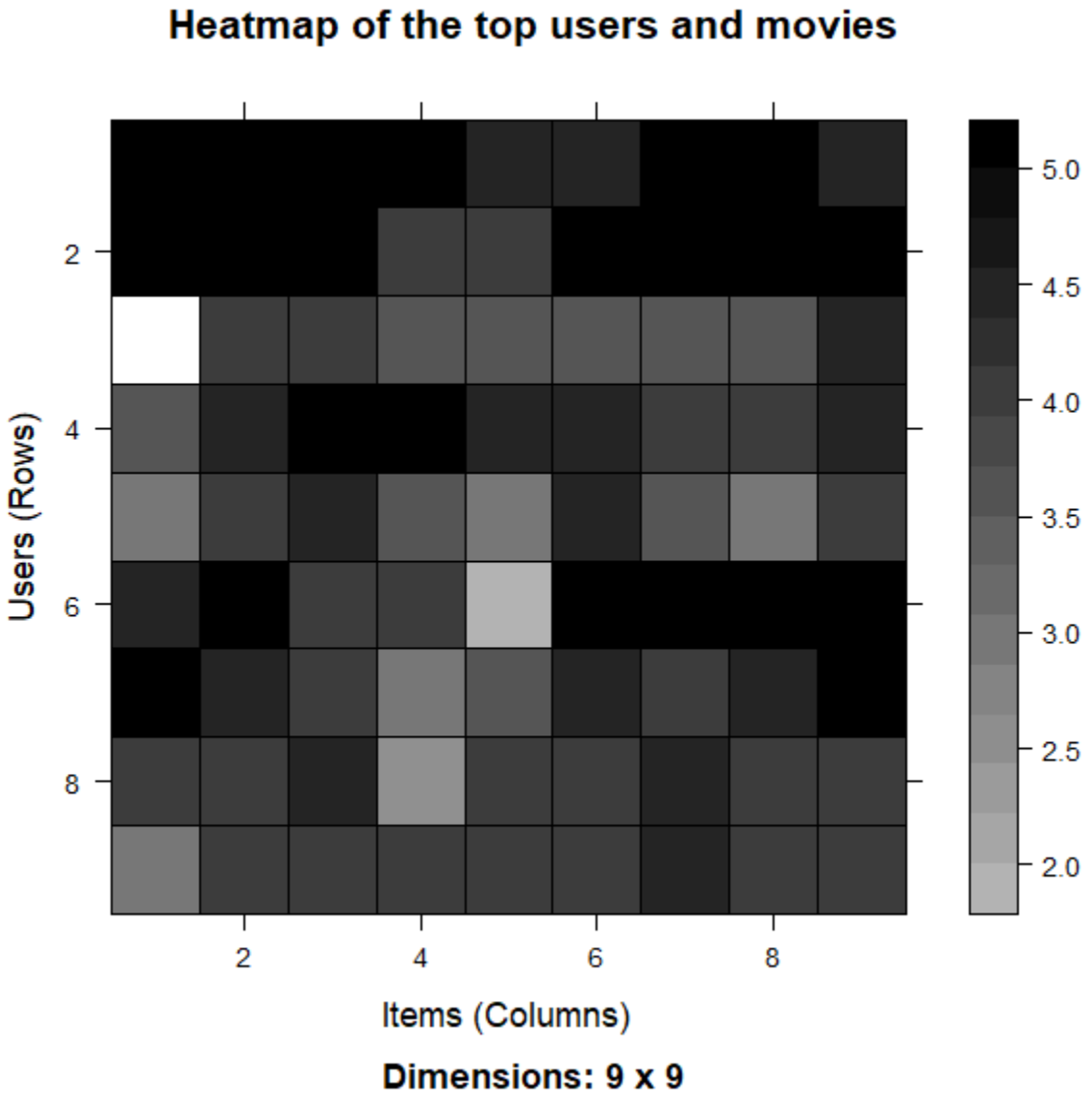
Để tìm kiếm dữ liệu hữu ích trong tập dữ liệu, nhóm đã đặt ngưỡng cho số lượng người dùng tối thiểu đã xếp hạng phim là 50 và tương tự đối với số lượt xem tối thiểu trên mỗi phim. Bằng cách này, nhóm đã lọc được danh sách các bộ phim đã được xem từ những bộ phim ít được xem nhất.

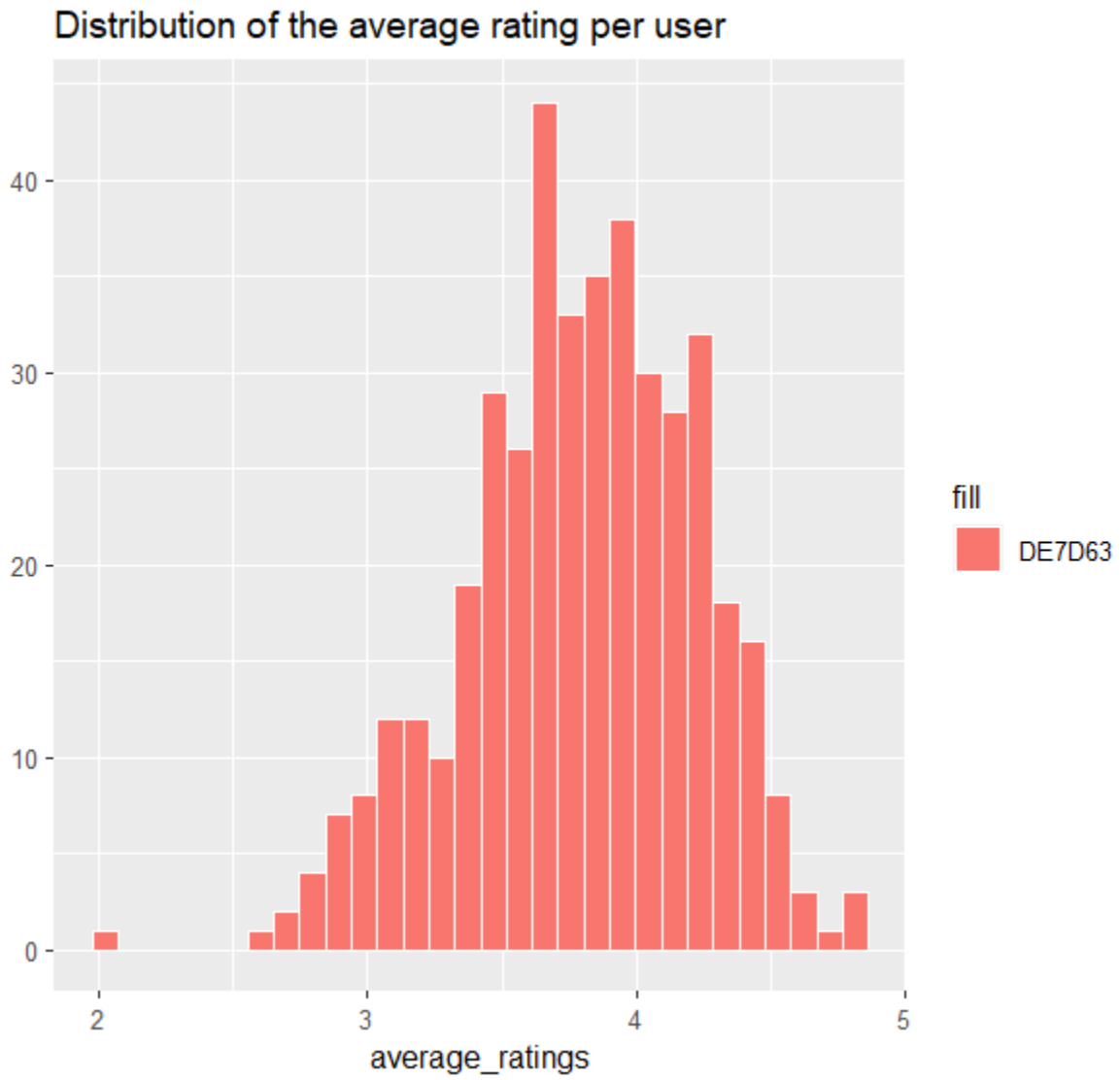
```
> # Performing Data Preparation
> movie_ratings <- ratingMatrix[rowCounts(ratingMatrix) > 50,
+                               colCounts(ratingMatrix) > 50]
> movie_ratings
420 x 447 rating matrix of class 'realRatingMatrix' with 38341 ratings.
>
```

Kết quả đầu ra của movies_ratings cho thấy 420 người dùng và 447 phim trái ngược với 668 người dùng và 10325 phim trước đó.

Để quan sát được rõ những người dùng có liên quan, ta cần phác thảo ma trận như dưới đây.

```
> # Delineate matrix of relevant users
> minimum_movies<- quantile(rowCounts(movie_ratings), 0.98)
> minimum_users <- quantile(colCounts(movie_ratings), 0.98)
> image(movie_ratings[rowCounts(movie_ratings) > minimum_movies,
+                     colCounts(movie_ratings) > minimum_users],
+       main = "Heatmap of the top users and movies")
>
```



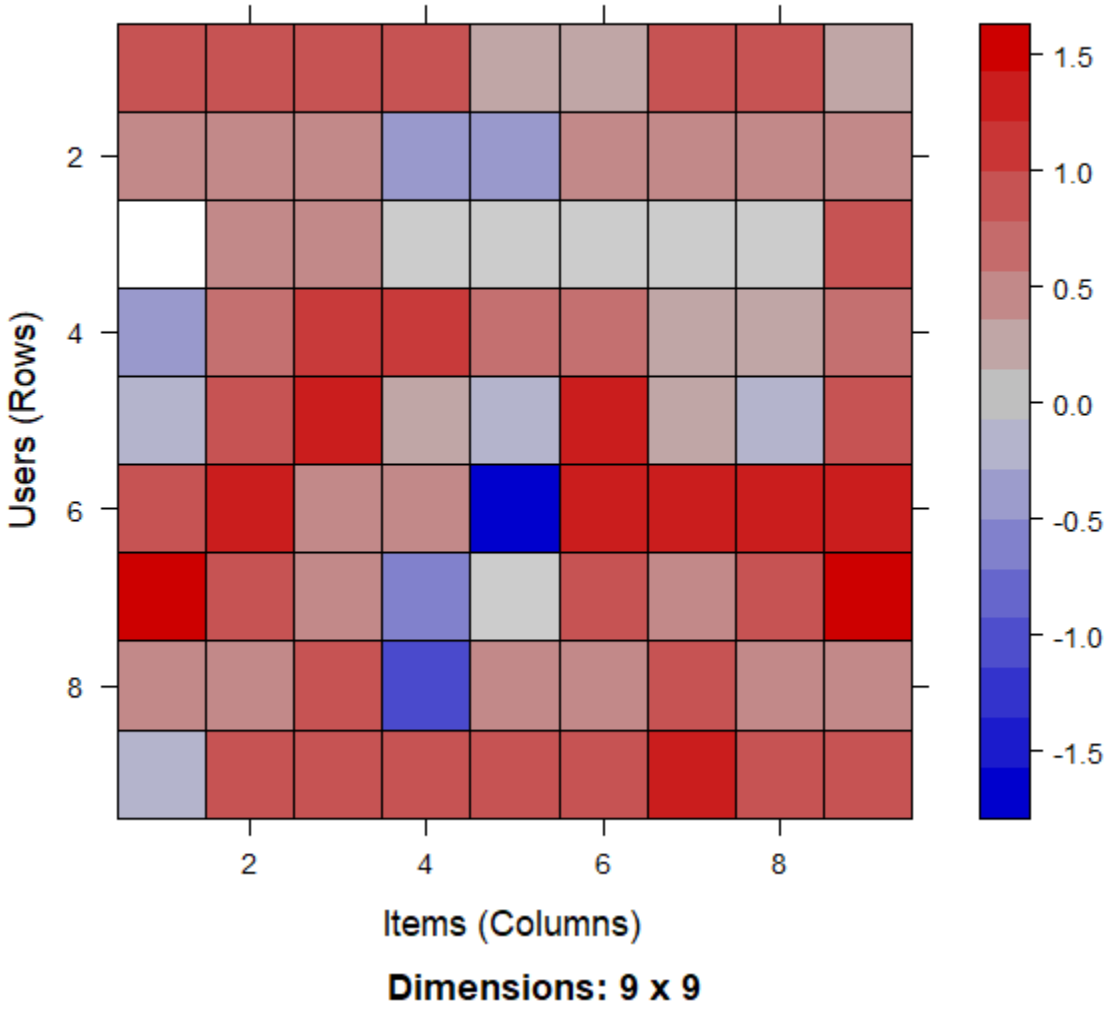


Biểu đồ về phân phối về đánh giá trung bình trên mỗi người dùng.

3.2.1. Chuẩn hóa dữ liệu

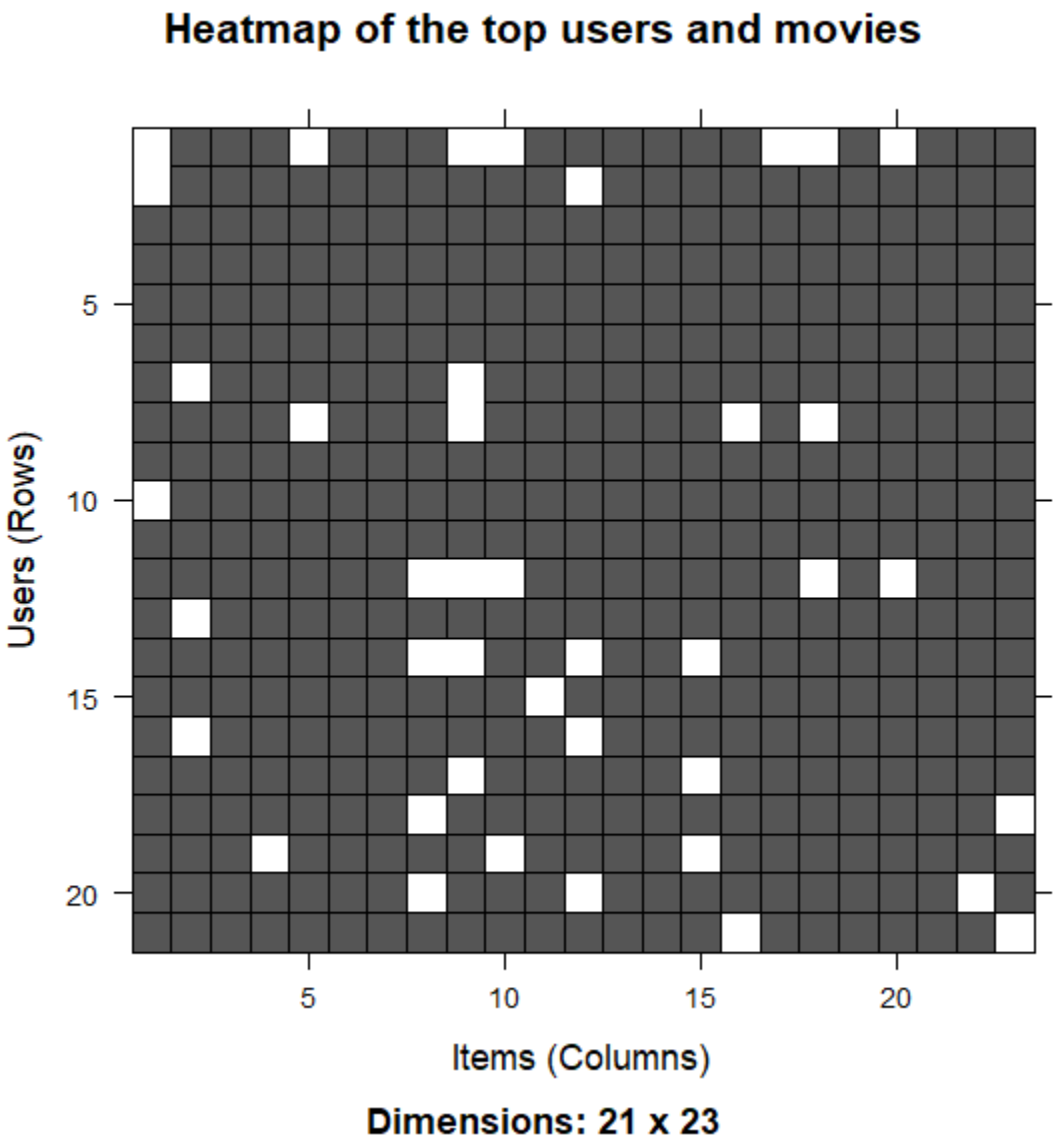
Trong trường hợp của một số người dùng, họ xếp hạng cao hoặc thấp cho tất cả các bộ phim đã xem sẽ gây nên sai lệch trong khi triển khai mô hình. Để loại bỏ điều này, ta cần tiến hành chuẩn hóa dữ liệu.

Chuẩn hóa là một thủ tục chuẩn bị dữ liệu để chuẩn hóa các giá trị số trong một cột thành một giá trị tỷ lệ chung. Điều này được thực hiện theo cách mà không có sự biến dạng trong phạm vi giá trị. Chuẩn hóa chuyển đổi giá trị trung bình của cột xếp hạng thành 0. Sau đó, nhóm sẽ vẽ một bản đồ nhiệt mô tả các xếp hạng đã được chuẩn hóa.



3.2.2. Nhị phân dữ liệu

Trong bước cuối cùng của quá trình chuẩn bị dữ liệu trong nghiên cứu này, nhóm nghiên cứu sẽ mã hóa dữ liệu. Phân chia dữ liệu có nghĩa là chúng ta có hai giá trị riêng biệt 1 và 0, điều này sẽ cho phép hệ thống đề xuất hoạt động hiệu quả hơn. Một ma trận được xác định theo quy tắc bao gồm 1 nếu xếp hạng trên 3 và nếu không, nó sẽ là 0.



3.3. Collaborative Filtering System

Hệ thống này sẽ tìm thấy điểm giống nhau trong các mục dựa trên xếp hạng của mọi người về chúng. Đầu tiên, thuật toán xây dựng một bảng các mặt hàng tương tự của những khách hàng đã mua chúng thành một tổ hợp các mặt hàng tương tự. Điều này sau đó được đưa vào hệ thống khuyến nghị.

Sự giống nhau giữa các sản phẩm đơn lẻ và các sản phẩm liên quan có thể được xác định bằng thuật toán sau:

- Đối với mỗi Mặt hàng i1 có trong danh mục sản phẩm, do khách hàng C mua.
- Với mỗi mặt hàng i2 cũng được mua bởi khách hàng C.
- Tạo bản ghi rằng khách hàng đã mua mặt hàng i1 và i2.
- Tính độ giống nhau giữa i1 và i2.

Nhóm nghiên cứu sẽ xây dựng hệ thống lọc này bằng cách chia nhỏ tập dữ liệu thành 80% tập huấn luyện và 20% tập kiểm tra.

```
> # Collaborative Filtering System
> sampled_data<- sample(x = c(TRUE, FALSE),
+                       size = nrow(movie_ratings),
+                       replace = TRUE,
+                       prob = c(0.8, 0.2))
> training_data <- movie_ratings[sampled_data, ]
> testing_data <- movie_ratings[!sampled_data, ]
>
```

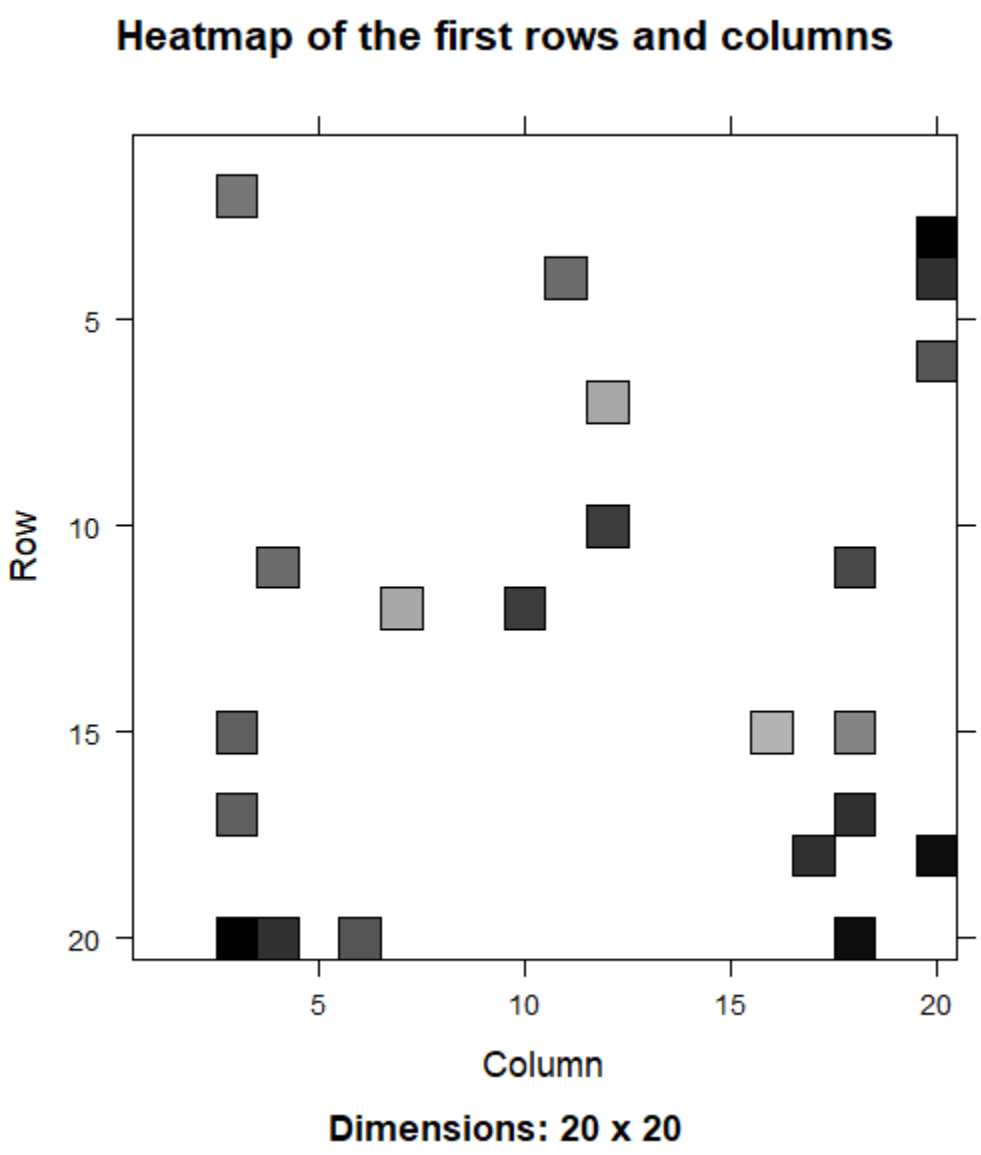
3.4. Xây dựng Hệ thống khuyến nghị phim

Sử dụng phương pháp cosine là phương pháp mặc định để khám phá các tham số khác nhau dựa trên các mặt hàng của Bộ lọc cộng tác. Trong bước đầu tiên, k biểu thị số lượng các mặt hàng để tính toán các điểm tương đồng của chúng. Ở đây, k bằng 30. Do đó, thuật toán bây giờ sẽ xác định k mặt hàng giống nhau nhất và lưu trữ số của chúng.

```
> # Building the Recommendation System using R
> recommendation_system <- recommenderRegistry$get_entries(dataType ="realRatingMatrix")
> recommendation_system$IBCF_realRatingMatrix$parameters
$k
[1] 30

$method
[1] "Cosine"

$normalize
[1] "center"

$normalize_sim_matrix
[1] FALSE

$alpha
[1] 0.5

$na_as_zero
[1] FALSE

>
> recommend_model <- Recommender(data = training_data,
+                                method = "IBCF",
+                                parameter = list(k = 30))
> recommend_model
Recommender of type 'IBCF' for 'realRatingMatrix'
learned using 338 users.
> class(recommend_model)
[1] "Recommender"
attr(,"package")
[1] "recommenderlab"
>
```

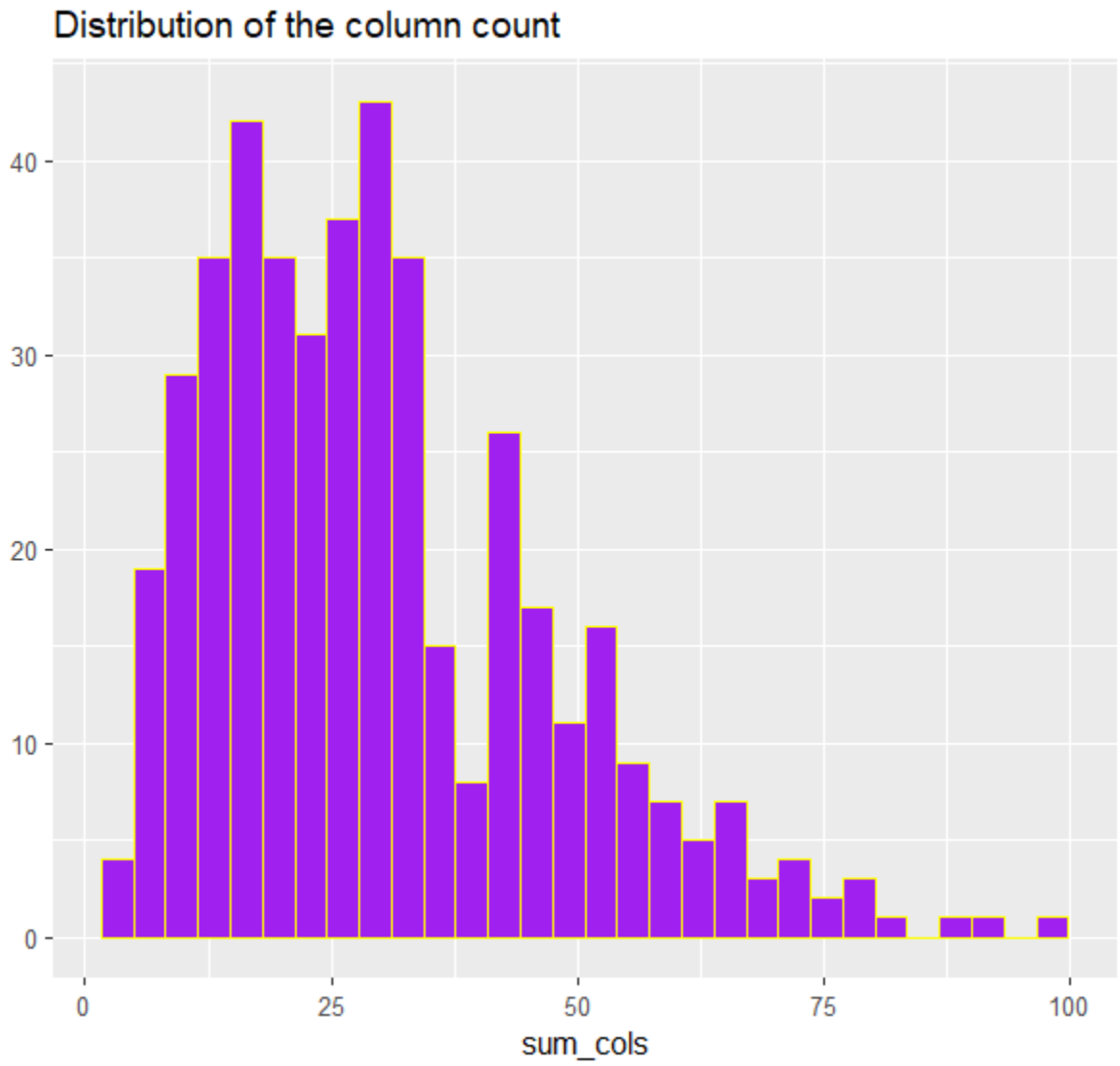
Sử dụng hàm getModel () sẽ truy xuất ra được recommen_model. Sau đó, nhóm sẽ tìm lớp và kích thước của ma trận tương tự được chứa trong model_info. Kế đến, chúng tôi sẽ tạo một bản đồ nhiệt chứa 20 mục hàng đầu và quan sát sự tương đồng được chia sẻ giữa chúng.

```
> # 20 Items
> model_info <- getModel(recommend_model)
> class(model_info$sim)
[1] "dgCMatrix"
attr(,"package")
[1] "Matrix"
> dim(model_info$sim)
[1] 447 447
> top_items <- 20
> image(model_info$sim[1:top_items, 1:top_items],
+       main = "Heatmap of the first rows and columns")
>
```



Dimensions: 20 x 20

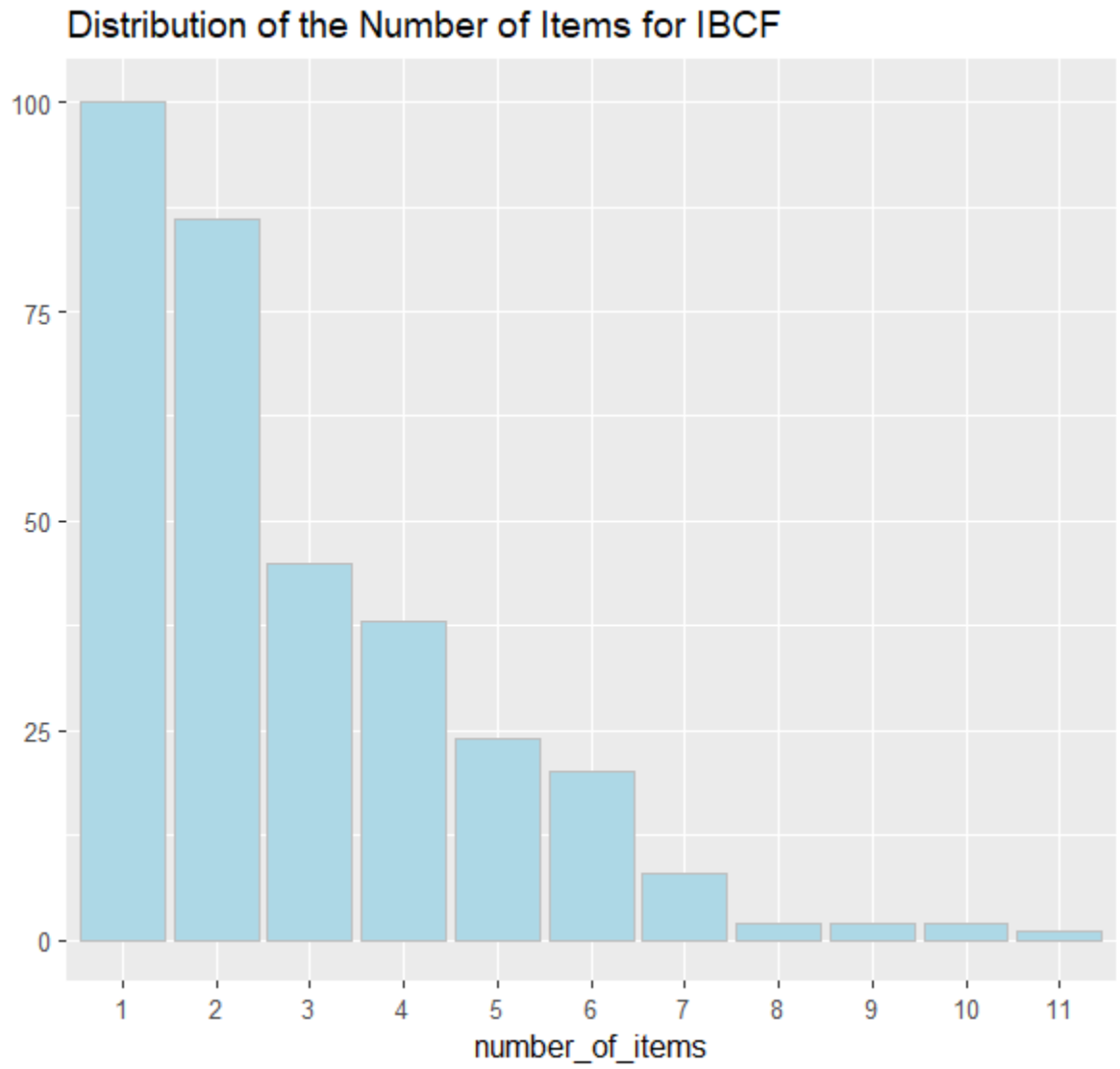
Trong bước tiếp theo, chúng ta sẽ thực hiện tính tổng các hàng và cột có độ giống nhau của các đối tượng trên 0. Chúng được thể hiện thông qua phân phối như sau:



Biến top_recommendations được khởi tạo thành 10, chỉ định số lượng phim cho mỗi người dùng. Sau đó, hàm dự đoán () được dùng để xác định các bộ phim tương tự nhau và xếp hạng chúng một cách thích hợp.

```
> # Create a top_recommendations variable which will be initialized to 10, specifying the number of films to each user.
> top_recommendations <- 10 # The number of items to recommend to each user
> predicted_recommendations <- predict(object = recommend_model,
+                                      newdata = testing_data,
+                                      n = top_recommendations)
> predicted_recommendations
Recommendations as 'topNList' with n = 10 for 82 users.
>
```

```
> user1 <- predicted_recommendations@items[[1]] # Recommendation for the first user
> movies_user1 <- predicted_recommendations@itemLabels[user1]
> movies_user2 <- movies_user1
> for (index in 1:10){
+   movies_user2[index] <- as.character(subset(movies,
+                                              movies$movieId == movies_user1[index])$title)
+ }
> movies_user2
 [1] "WALL·E (2008)"                          "English Patient, The (1996)"
 [3] "Airplane! (1980)"                       "One Flew Over the Cuckoo's Nest (1975)"
 [5] "M*A*S*H (a.k.a. MASH) (1970)"           "Hangover, The (2009)"
 [7] "Great Escape, The (1963)"               "Wizard of Oz, The (1939)"
 [9] "Citizen Kane (1941)"                    "Slumdog Millionaire (2008)"
>
```



```
> # Perform Number of items
> number_of_items_sorted <- sort(number_of_items, decreasing = TRUE)
> number_of_items_top <- head(number_of_items_sorted, n = 4)
> table_top <- data.frame(as.integer(names(number_of_items_top)),
+                         number_of_items_top)
> for(i in 1:4) {
+   table_top[i,1] <- as.character(subset(movies,
+                                         movies$movieId == table_top[i,1])$title)
+ }
>
> colnames(table_top) <- c("Movie Title", "No. of Items")
> head(table_top)
                      Movie Title No. of Items
161           Crimson Tide (1995)           11
497 Much Ado About Nothing (1993)           10
919      Wizard of Oz, The (1939)           10
16                  Casino (1995)            9
>
```

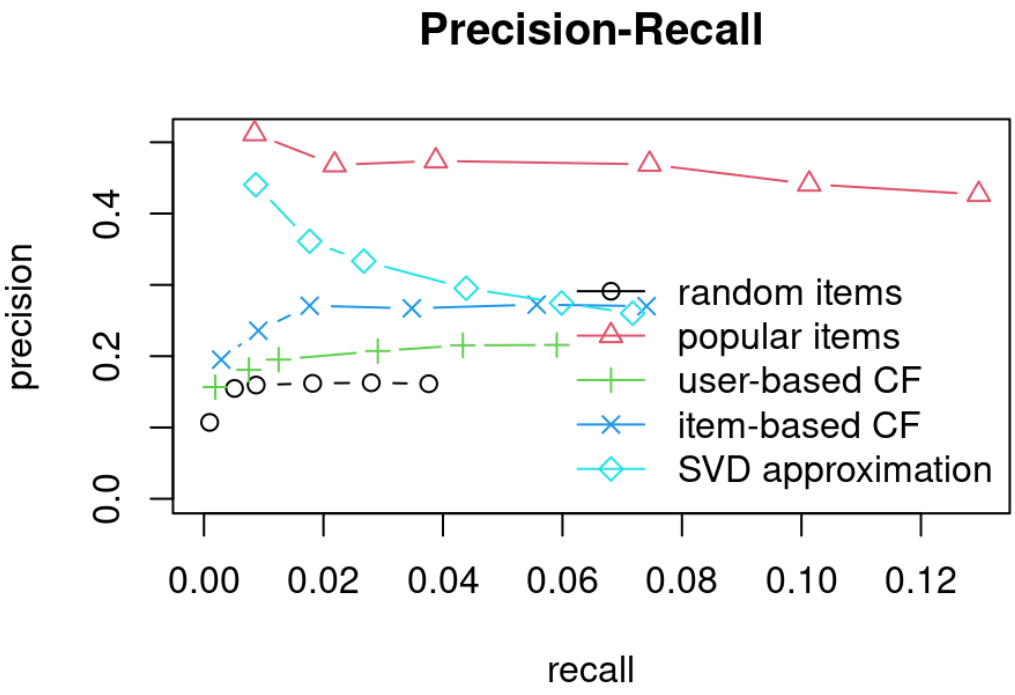
CHƯƠNG 5: TỔNG KẾT

1. Kết quả

Đây là kết quả sau khi phân tích dữ liệu:

```
          RMSE       MSE       MAE
[1,] 0.9398581 0.8833333 0.8333333
[2,] 1.5000000 2.2500000 1.2500000
[3,] 0.5773503 0.3333333 0.3333333
[4,] 1.0606602 1.1250000 0.7500000
[5,] 1.4604262 2.1328448 1.1156078
[6,] 0.8970970 0.8047830 0.6058958
```

Vì cả 3 giá trị này đều có miền giá trị từ [0,+∞], ta thấy các giá trị được dự đoán khá nhỏ, từ đó cho thấy sự chính xác của mô hình đã được triển khai. Có thể thấy, với top 6 khách hàng được áp dụng mô hình thì kết quả đánh giá RMSE, MSE và MAE khá cao. Từ đó kết luận độ chính xác của mô hình không cao.



Chúng ta có thể thấy rằng độ chính xác của mô hình Item-based CF cao hơn độ chính các của mô hình user-based CF. Việc gợi ý cho khách hàng được đánh giá cao nhất khi sử dụng mô hình phân tích dựa trên các "popular items". Cũng cần lưu ý rằng cả UBCF và IBCF đều có những hạn chế - ví dụ: khi xử lý những người dùng không chọn phim nào hoặc các bộ phim mới chưa được chọn lần nào.

2. Kết luận

Hệ thống đề xuất phim là hệ thống đề xuất bộ phim tiếp theo cho người dùng. Collaborative filtering được sử dụng cho việc này. Từ dữ liệu có sẵn, hệ thống này phân tích các yếu tố khác nhau như mức độ tương tự của người dùng, mức độ tương tự của phim, xếp hạng, v.v.

3. Hạn chế của nghiên cứu

Độ chính xác của hệ thống đề xuất chưa cao; chưa đa dạng tệp dữ liệu.

4. Hướng phát triển trong tương lai

Nếu được cải thiện và bổ sung thêm các tệp dữ liệu mới thì hệ thống có thể đề xuất đa dạng phim cho người dùng và nâng cao trải nghiệm khách hàng.

LỜI KẾT

DANH MỤC TÀI LIỆU THAM KHẢO