**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM THÀNH PHỐ HỒ CHÍ MINH**

**KHOA MÔI TRƯỜNG VÀ TÀI NGUYÊN**

**ĐỀ CƯƠNG Ý TƯỞNG**

CHỦ ĐỀ NGHIÊN CỨU:

**HỆ THỐNG GỢI Ý PHIM HOẠT HÌNH**

**HỌ VÀ TÊN: NGUYỄN THỊ HUỲNH NHƯ**

**MÃ SỐ SINH VIÊN: 20166050**

**CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN**

**NIÊN KHÓA: 2020-2024**

Thành phố Thủ Đức, ngày 01 tháng 06 năm 2023

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC NÔNG LÂM THÀNH PHỐ HỒ CHÍ MINH**

**KHOA MÔI TRƯỜNG VÀ TÀI NGUYÊN**

**ĐỀ CƯƠNG Ý TƯỞNG**

CHỦ ĐỀ NGHIÊN CỨU:

**HỆ THỐNG GỢI Ý PHIM HOẠT HÌNH**

**HỌ VÀ TÊN: NGUYỄN THỊ HUỲNH NHƯ**

**MÃ SỐ SINH VIÊN: 20166050**

**CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN**

**NIÊN KHÓA: 2020-2024**

Thành phố Thủ Đức, ngày 01 tháng 06 năm 2023

****

# **TÓM TẮT**

Đề cương hệ thống gợi ý phim hoạt hình nêu lên một số vấn đề của việc sử dụng thông tin thường gặp trong cuộc sống, một số khái niệm cơ bản, hướng tiếp cận khi xây dựng hệ thống, tìm hiểu một số thuật toán đơn giản của học sâu-thống kê (như: Bayes estimator - Internet Movie Database (IMDB), Sự tương đồng cosin - Measure of Similarity, Thuật toán K-Nearest Neighbors – KNN (thuật toán láng giềng gần/ hàng xóm lân cận)).

Đề cương hệ thống gợi ý phim hoạt hình tập trung vào tìm hiểu ba cách thức để thực hiện quá trình gợi ý trong một hệ thống gợi ý phim hoạt hình là: simple recommenders(đưa ra gợi ý cho người dùng dựa vào mức độ phổ biến), content-based recommenders(đề xuất dựa trên những thể loại tương tự dựa trên một bộ phim hoạt hình cụ thể) và collaborative filtering recommenders(dự đoán “rating” mà người sử dụng sẽ đưa ra một phim hoạt hình dựa trên xếp hạng trước đây và sở thích của những người dùng khác).

# **MỤC LỤC**

[**TÓM TẮT**](#_Toc136690234)

[**MỤC LỤC**](#_Toc136690235)

[**DANH SÁCH HÌNH ẢNH**](#_Toc136690236)

[**DANH SÁCH TỪ TIẾNG ANH**](#_Toc136690237)

[**LỜI CÁM ƠN** 1](#_Toc136690238)

[**ĐẶT VẤN ĐỀ** 2](#_Toc136690239)

[**CHƯƠNG 1: MỞ ĐẦU** 4](#_Toc136690240)

[**1.1.** **Tính cấp thiết của đề tài** 4](#_Toc136690241)

[**1.2.** **Lý do chọn đề tài** 5](#_Toc136690242)

[**1.3.** **Mục tiêu** 5](#_Toc136690243)

[**1.4.** **Đối tượng nghiên cứu** 5](#_Toc136690244)

[**CHƯƠNG 2: TỔNG QUAN** 7](#_Toc136690245)

[**2.1. Cơ sở lý thuyết** 8](#_Toc136690246)

[**2.2. Thông tin về bối cảnh của đề cương ý tưởng:** 9](#_Toc136690247)

[**CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU** 12](#_Toc136690248)

[**3.1. Vật liệu nghiên cứu** 12](#_Toc136690249)

[**3.1.1. Vật liệu nghiên cứu** 12](#_Toc136690250)

[**3.1.2. Các công cụ phần mềm** 12](#_Toc136690251)

[**3.2. Các cách tiếp cận, phương pháp nghiên cứu, phạm vi nghiên cứu** 13](#_Toc136690252)

[**3.*2*.1. Các cách tiếp cận** 14](#_Toc136690253)

[**3.2.2. Phương pháp nghiên cứu** 14](#_Toc136690254)

[**3.2.3*.* Phạm vi nghiên cứu** 15](#_Toc136690255)

[**3.2.4. Thời gian thực hiện** 15](#_Toc136690256)

[**3.3. Phân tích hệ thống gợi ý phim hoạt hình** 15](#_Toc136690257)

[**4.1. Ưu điểm, nhược điểm của các hệ thống gợi ý trên thế giới** 20](#_Toc136690258)

[**4.1.1. Ưu điểm của các hệ thống gợi ý trên thế giới** 20](#_Toc136690259)

[**4.1.2. Nhược điểm của các hệ thống gợi ý trên thế giới** 20](#_Toc136690260)

[**4.2. Ưu điểm, nhược điểm của đề cương ý tưởng** 21](#_Toc136690261)

[**4.2.1. Ưu điểm** 21](#_Toc136690262)

[**4.2.2. Nhược điểm** 21](#_Toc136690263)

[**4.3. Kết quả kỳ vọng** 21](#_Toc136690264)

[**CHƯƠNG 5: KẾT LUẬN** 23](#_Toc136690265)

[**TÀI LIỆU THAM KHẢO** 24](#_Toc136690266)

# **DANH SÁCH HÌNH ẢNH**

**HÌNH 1: Quy trình hoạt động gợi ý đơn giản 13**

**HÌNH 2: Sơ đồ minh họa các hướng tiếp cận hệ thống gợi ý 14**

**HÌNH 3: Quy trình nghiên cứu hệ thống gợi ý 15**

**HÌNH 4: Hệ gợi ý đơn giản (Simple Recommenders) 18**

**HÌNH 5: Hệ thống gợi ý dựa trên nội dung (Content-based Recommenders) 18**

**HÌNH 6: Hệ thống lọc cộng tác (Collaborative filtering Recommenders) 19**

# **DANH SÁCH TỪ TIẾNG ANH**

Bayes estimator - Internet Movie Database (IMDB): Bình quân gia quyền

Measure of Similarity: Sự tương đồng cosin

Thuật toán K-Nearest Neighbors (KNN): Thuật toán láng giềng gần/ hàng xóm lân cận

Simple Recommenders: Hệ thống gợi ý đơn giản

Content-based Recommenders: Hệ thống gợi ý dựa trên nội dung

Collaborative filtering Recommenders(CF): Lọc cộng tác

# **LỜI CÁM ƠN**

Trong quá trình hoàn thành môn học “Phương pháp nghiên cứu khoa học” em đã nhận được sự giúp đỡ và hướng dẫn nhiệt tình của các thầy cô của ngành Hệ thống thông tin Trường đại học Nông Lâm thành phố Hồ Chí Minh để hoàn thành đề cương môn học.

Đặc biệt, em xin bày tỏ sự biết ơn tới thầy Nguyễn Kim Lợi người đã hướng dẫn, giúp đỡ về kiến thức để em có thể hoàn thành đề cương ý tưởng này.

Em xin chân thành cám ơn gia đình, bạn bè đã cổ vũ và giúp đỡ em trong suốt thời gian qua.

Trong suốt quá trình hoàn thiện đề cương vẫn còn những mặt hạn chế và thiếu sót. Em rất mong nhận được ý kiến góp ý và sự chỉ dẫn của các thầy cô giảng viên và các bạn để đề cương ý tưởng hoàn thiện hơn.

Một lần nữa em xin chân thành cám ơn!

Thủ Đức,ngày 01 tháng 06 năm 2023

Nguyễn Thị Huỳnh Như

# **ĐẶT VẤN ĐỀ**

Nhìn vào bức tranh lớn của thế giới hiện nay, trong cuộc sống hàng ngày, con người phụ thuộc vào lời giới thiệu từ người khác, bằng lời nói, đánh giá phim và sách, thư giới thiệu, hướng dẫn nhà hàng và khách sạn hoặc bằng một số phương tiện khác (Resnick, 1997). Sự phát triển bùng nổ và sự đa dạng của thông tin có sẵn trên Web và sự ra đời nhanh chóng của các dịch vụ kinh doanh điện tử mới ( mua sản phẩm, giải trí, giáo dục, v.v.) khiến người dùng dễ choáng ngợp trước lượng thông tin lớn như vậy, khiến họ đưa ra những quyết định sai lầm. Dữ liệu năm 2022 mới nhất cho thấy Google đã xử lý hơn 99.000 lượt tìm kiếm mỗi giây. Tương đương việc tạo ra hơn 8,5 tỷ lượt tìm kiếm mỗi ngày (Internet Live Stats, 2022). Đối với vấn đề gây quá tải thông tin cho người dùng, hệ thống gợi ý trong những năm trở lại đây là phương tiện có thể giúp người dùng giải quyết được vấn đề này.

Hệ thống gợi ý là công cụ được sử dụng để tương tác với các không gian thông tin phức tạp. Hệ thống gợi ý là một loại ứng dụng dựa trên Website cung cấp một chế độ xem được cá nhân hóa cụ thể cung cấp cho người dùng các đề xuất được cá nhân hóa về nội dung mà họ có thể quan tâm. Các đề xuất được cá nhân hóa được cung cấp dưới dạng danh sách các mục được xếp hạng. Khi thực hiện xếp hạng, hệ thống gợi ý cố gắng dự đoán đâu là sản phẩm hoặc dịch vụ phù hợp nhất, dựa trên sở thích và ràng buộc của người dùng trong hệ cơ sở quản trị dữ liệu. Hệ thống gợi ý thu thập hoặc có thể cập nhật nhanh chóng theo thời gian thực các thông tin từ người dùng thông qua các tùy chọn của họ,hay các cuộc khảo sát, phỏng vấn, được thể hiện rõ ràng,… Từ các dữ liệu người dùng đó các nhà cung cấp dịch vụ chuyển đổi dữ liệu theo nhiều mục tiêu đánh giá đề có thể gợi ý cho người dùng một cách nhanh chóng. Và ở đây các thuật toán được kết hợp linh hoạt với nhau để hoạt động giúp hỗ trợ người dùng đưa ra quyết định một cách tự động.

Theo một nghiên cứu của McKinsey, hệ thống gợi ý giúp bạn hiểu thị những thứ mà người dùng của bạn yêu thích, những thống kê phản ánh những con số thực tế từ việc mua sắm trực tuyến thông qua nhà hàng đến ứng dụng video, người giới thiệu chiếm: 40% lượt cài ứng dụng trên Google Play, 60% thời gian xem trên YouTube, 35% mua hàng trên Amazon, 75% lượt xem phim trên Netflix (Ian MacKenzie, et al., 2013).

Năm 2009, Netflix đã treo thưởng 1.000.000 USD cho 10% cải tiến thuật toán gợi ý. Đây là một bước nhảy vọt cho sự phát triển của lĩnh vực này. Điều này khá thú vị để làm tiền đề để bắt đầu viết đề cương tìm hiểu về lĩnh vực.

Phim hoạt hình gắn liền với tuổi thơ của rất nhiều người, lĩnh vực này phát triển khá nhanh nên số lượng phim hoạt hình ngày càng nhiều, đa dạng thể loại, nội dung, hình thức,... với khối lượng dữ liệu lớn thông tin làm người dùng khá bối rối khi đưa ra lựa chọn. Đề cương hy vọng sẽ đưa ra một vài gợi ý đơn giản dựa trên các chỉ số đánh giá, sở thích, sự tương quan giữa các thư mục dữ liệu- người dùng, người dùng- thư mục dữ liệu,... giúp hỗ trợ việc đưa ra quyết định của người sử dụng.

Tóm lại, đề cương hệ thống gợi ý phim hoạt hình áp dụng các thuật toán cơ bản của học sâu-thống kê (như: Bayes estimator - Internet Movie Database (IMDB), Sự tương đồng cosin - Measure of Similarity, Thuật toán K-Nearest Neighbors – KNN (thuật toán láng giềng gần/ hàng xóm lân cận)) để gợi ý tên phim hoạt hình cho người dùng bằng một số cách tiếp cận đơn giản, hỗ trợ việc ra quyết định trở nên dễ dàng hơn.

# **CHƯƠNG 1: MỞ ĐẦU**

Cuộc sống hàng ngày luôn có các tình huống cần phải đưa ra quyết định. Hôm nay bạn muốn xem gì? Chẳng hạn như bạn muốn tìm một bộ phim hoạt hình gì vào ngày chủ nhật? Khi công việc học tập mệt mỏi bạn muốn xem một bộ phim hoạt hình vui nhộn để thư giãn? Hôm nay bạn muốn coi phim hoạt hình với thông điệp cuộc sống ý nghĩa? Bạn muốn xem các bộ phim hoạt hình tình cảm? Bạn cần những thông tin gì để lựa chọn được một bộ phim hoạt hình đáp ứng các yêu cầu của mình?

Những lời gợi ý cho bạn thường đến từ bạn bè, các lượt đánh giá, lượt xem và bình luận của những người khác đã xem trước đây, … giúp bạn một phần trong việc đưa ra quyết định. Nhưng khi hệ thống gợi ý xuất hiện thì mọi thứ trở nên đơn giản hơn. Ví dụ như YouTube gợi ý dựa trên trình duyệt trước đây đã cung cấp các bộ phim hoạt hình với nội dung nghe nhìn làm người dùng hài lòng với tỷ lệ tương đối cao hay dựa vào sự ưu tiên để đưa ra gợi ý.

So với giải pháp thông thường dựa trên đề xuất của bạn bè, trang web được đề cập ở trên đưa ra đề xuất dựa trên thư viện video lớn nhất thế giới. Vì cái nhìn sâu sắc tập thể của bạn bè chúng ta chỉ là một phần nhỏ trong số đó, nên nó có một lợi thế không thể đánh bại. Bằng cách này, chúng tôi có thể khám phá các bài hát của các ban nhạc mà chúng tôi rất có thể chưa từng gặp theo bất kỳ cách nào khác. Vì sử dụng phương pháp này, không chỉ những người quen của chúng tôi có thể cung cấp nội dung cho chúng tôi mà tất cả mọi người trên thế giới đều làm như vậy – không chủ ý – thông qua hệ thống giới thiệu (Apáthy, 2019).

Hệ thống gợi ý có nhiều mặt hạn chế riêng nên: Chúng ta nên coi những điều này nhiều hơn như một cơ hội, có thể giúp chúng ta quyết định và tiết kiệm thời gian trong một số câu hỏi nhất định – ít mang tính cá nhân hơn – hoặc tìm cách giải trí mà chúng ta có thể chưa bao giờ khám phá nếu không có những hệ thống này (Apáthy, 2019).

* 1. **Tính cấp thiết của đề tài**

Thế giới công nghệ ngày càng phát triển nhanh chóng. Phim hoạt hình đóng góp một phần quan trọng trong việc giải trí của người dùng. Đáp ứng nhu cầu tìm kiếm ngày càng tăng của người dùng, việc gợi ý một tên phim hoạt hình nào đó đáp ứng đủ nhu cầu tìm kiếm của người dùng một cách nhanh chóng, tức thời. Là động lực để thực hiện đề cương hệ thống gợi ý cung cấp các thông tin người dùng cần.

* 1. **Lý do chọn đề tài**

Đáp ứng yêu cầu của môn học: phương pháp nghiên cứu khoa học.

Đưa ra các gợi ý ở quy mô cá nhân.

Tìm hiểu sự cá nhân hóa trong các hệ thống gợi ý.

Tìm hiểu thêm ứng dụng của các thuật toán và bài toán thực tế.

* 1. **Mục tiêu**

Mục tiêu: đưa ra gợi ý các bộ phim hoạt hình nhanh chóng cho người dùng, cung cấp theo sự thiên vị của người dùng về các bộ phim hoạt hình. Có thể giới thiệu cho người dùng với thị hiếu độc đáo, có thể giới thiệu các bộ phim hoạt hình dựa trên sự ưu ái.

Mục tiêu chính: tiết kiệm thời gian trong việc tìm kiếm các bộ phim hoạt hình, mang tính cá nhân theo nội dung mà người dùng có thể quan tâm.

Mục tiêu phụ: tìm hiểu các thuật toán liên quan, cách thức đưa ra các gợi ý.

Mục tiêu ứng dụng: tìm hiểu ứng dụng của các thuật toán học sâu- thống kê trong thực tế.

* 1. **Đối tượng nghiên cứu**

Sử dụng nguồn dữ liệu có sẵn nên chỉ có thể đáp ứng một số mục tiêu, nguồn dữ liệu là phim hoạt hình Nhật Bản(anime).

Nguồn dữ liệu thu thập: Anime Recommendations Database, (Kaggle, 2017).

Bối cảnh: Bộ dữ liệu này chứa thông tin về dữ liệu sở thích của người dùng từ 73.516 người dùng trên 12.294 anime. Mỗi người dùng có thể thêm phim hoạt hình vào danh sách đã hoàn thành của họ và xếp hạng cho nó và bộ dữ liệu này là tập hợp của những xếp hạng đó (Kaggle, 2017)

Mô tả dữ liệu:

* anime.csv: có 12294 hàng, 7 cột( anime\_id, name, genre, type, episodes, rating, members).
* anime\_id: id duy nhất của myanimelist.net xác định một anime.
* Name: tên đầy đủ của anime.
* Genre: danh sách thể loại được phân tách bằng dấu phẩy cho anime này.
* Type: phim, TV, OVA, v.v.
* Episodes: có bao nhiêu tập trong chương trình này. (1 nếu là phim).
* Rating: xếp hạng trung bình trên 10 cho anime này.
* Members: số thành viên cộng đồng nằm trong

"nhóm" của anime này.

* rating.csv: có 7813737 hàng, 3 cột( user\_id, anime\_id, rating).
* user\_id: id người dùng được tạo ngẫu nhiên không thể nhận dạng.
* anime\_id: anime mà người dùng này đã xếp hạng.
* rating: xếp hạng trong số 10 người dùng này đã chỉ định (-1 nếu người dùng đã xem nhưng không chỉ định xếp hạng).

# **CHƯƠNG 2: TỔNG QUAN**

Trên thế giới, khái niệm hệ thống gợi ý bắt đầu xuất hiện từ năm 1955. Trong những năm đầu hệ thống gợi ý phát triển chưa có điểm nổi bật. Đến năm 2006, Netflix tạo ra bước đột phá trong lĩnh vực này. Tính cá nhân hóa được xem là tâm trung của sự phát triển.Các nhà phát triển có sự vượt bậc nhờ hệ thống gợi ý như: Amazon.com, play.google.com, netflix.com, Alibaba.com, youtube.com, facebook.com, ebay.com, MovieFinder.com…Mang lại doanh thu tăng nhanh đáng kể.

Tại Việt Nam: ở giai đoạn đầu phần lớn các hệ thống gợi ý chỉ dừng ở mức gợi ý chung chung, chưa có tính cá nhân hóa cao. Trong một thập niên gần đây các hệ thống gợi ý của các trang thương mại điện tử phát triển ở Việt Nam ngày càng hoàn thiện hơn có thể kể đến: lazada.vn, Vatgia.com, Shopee.com, www.fptshop.com.vn, Thegioididong.com, Tiki.vn, …

Đáp ứng yêu cầu của môn học: phương pháp nghiên cứu khoa học là tiền đề đề thực hiện đề cương ý tưởng Hệ thống gợi ý phim hoạt hình đặt ra mục tiêu là đưa ra gợi ý tên phim hoạt hình theo một số thuật toán học sâu-thống kê cơ bản trên nền tảng Anaconda bằng ngôn ngữ Python. Kết quả kỳ vọng sẽ đưa ra các gợi ý tên phim hoạt hình đơn giản giúp hỗ trợ đưa quyết định, tìm hiểu thêm một số ứng dụng của thuật toán học sâu-thống kê trong cuộc sống. Về mặt ý nghĩa của đề cương có thể đáp ứng yêu câu môn học đề ra, hiểu thêm về sự góp mặt của thuật toán trong cuộc sống hằng ngày, tiếp cận thực tế từ khoa học, áp dụng khoa học vào thực tế như thế nào.

## **2.1. Cơ sở lý thuyết**

Hệ thống đề xuất đang được sử dụng bởi số lượng ngày càng tăng của các trang web thương mại điện tử để giúp người tiêu dùng tìm sản phẩm để mua. Những gì bắt đầu như một sự mới lạ đã trở thành một công cụ kinh doanh nghiêm túc. Các hệ thống gợi ý sử dụng kiến thức về sản phẩm - hoặc kiến thức được mã hóa thủ công do các chuyên gia cung cấp hoặc kiến thức “khai thác” học được từ hành vi của người tiêu dung - để hướng dẫn người tiêu dùng vượt qua nhiệm vụ khó khăn thường là tìm kiếm sản phẩm mà họ sẽ thích (Schafer, B., et al., 2000).

Lĩnh vực này, được đặt tên vào năm 1995, đã phát triển vượt bậc trong nhiều vấn đề được giải quyết và kỹ thuật được sử dụng, cũng như trong các ứng dụng thực tế của nó. Nghiên cứu hệ thống gợi ý đã kết hợp một loạt nhiều kỹ thuật trí tuệ nhân tạo bao gồm học máy, khai thác dữ liệu, mô hình hóa người dùng, dựa trên trường hợp lý luận, và hạn chế sự hài lòng, trong số những người khác. Các đề xuất được cá nhân hóa là một phần quan trọng của nhiều ứng dụng thương mại điện tử trực tuyến như Amazon.com, Netflix và Pandora (Burke, 2011).

Một hệ thống gợi ý ( hay còn gọi là hệ thống khuyên dùng, tiếng Anh: recommendation system) là một lớp con của hệ thống lọc thông tin, tìm kiếm dự đoán "đánh giá" hoặc "ưa thích" của người dùng với một sản phẩm hoặc đối tượng nào đó (Ricci, 2011).

Người giới thiệu các hệ thống dựa trên lọc cộng tác dự đoán sở thích của người dùng đối với các sản phẩm hoặc dịch vụ bằng cách tìm hiểu quá khứ mối quan hệ người dùng-mục từ một nhóm người dùng có cùng sở thích và sở thích (Mohd Abdul Hameed et al., 2012).

Các hệ thống gợi ý có nguồn gốc từ "Usenet", một hệ thống thảo luận phân tán trên toàn thế giới bắt nguồn từ Đại học Duke vào cuối những năm 1970. Usenet hoạt động ở định dạng máy khách/máy chủ, cho phép đầu vào của người dùng được phân loại thành các "nhóm tin" cụ thể. Trong Usenet, các bài đăng của người dùng được phân loại thành các nhóm tin này, sau đó được chia thành các danh mục con nếu cần (Smith, 2012).

Trên Internet, nơi có quá nhiều lựa chọn, cần phải lọc, ưu tiên và cung cấp thông tin liên quan một cách hiệu quả để giảm bớt vấn đề quá tải thông tin, vốn đã tạo ra một vấn đề tiềm ẩn đối với nhiều người dùng Internet. Các hệ thống đề xuất giải quyết vấn đề này bằng cách tìm kiếm thông qua khối lượng lớn thông tin được tạo động để cung cấp cho người dùng nội dung và dịch vụ được cá nhân hóa (Isinkaye, 2015).

Hệ thống gợi ý: một hệ thống lọc thông tin hỗ trợ người dùng trong một tình huống ra quyết định nhất định bằng cách thu hẹp tập hợp các tùy chọn có thể và ưu tiên các yếu tố của nó trong một ngữ cảnh cụ thể. Việc sắp xếp thứ tự ưu tiên có thể dựa trên sở thích được thể hiện rõ ràng hoặc ẩn ý của người dùng và cũng dựa trên hành vi trước đó của người dùng có sở thích tương tự (Apáthy, 2019).

Hệ thống gợi ý là một tập hợp con của máy học sử dụng dữ liệu để giúp người dùng tìm thấy sản phẩm và nội dung. Các trang web và dịch vụ phát trực tuyến sử dụng hệ thống gợi ý để tạo các trang và nội dung “dành cho bạn” hoặc “bạn cũng có thể thích” (Shetty, 2023).

Tóm lại, hệ thống gợi ý là một kỹ thuật của trí tuệ nhân tạo, hỗ trợ con người trong việc ra quyết định, sử dụng các phương pháp lọc theo mục tiêu, mục đích của người sử dụng. Cá nhân hóa thông qua việc các nhà phát triển thu thập thông tin và phân tích về cá nhân để tạo ra các gợi ý.

## **2.2. Thông tin về bối cảnh của đề cương ý tưởng:**

Đáp ứng yêu cầu đề ra của môn học. Đề cương ý tưởng hệ thống gợi ý phim hoạt hình dựa vào bộ dữ liệu có sẵn để thực hành các thuật toán của học sâu\_thống kê để đưa ra các gợi ý. Với sự đa dạng của các thông tin phim hoạt hình làm người dùng bối rối việc tìm hiểu về một hệ thống gợi ý đáp ứng nhu cầu tìm kiếm của người dùng là đều tất yếu.

Thuật toán của Netflix, CineMatch là hệ thống đề xuất thành công nhất cho việc bán phim trực tuyến vào đầu những năm 2000. Nó là một chất xúc tác quan trọng cho những nghiên cứu như vậy và lĩnh vực khoa học – vốn chỉ bắt đầu tồn tại độc lập vào những năm 90 – bắt đầu phát triển nhanh chóng. Thách thức của Giải thưởng Netflix năm 2006 là tạo ra một thuật toán đề xuất dựa trên 100 triệu bài đánh giá phim do họ cung cấp, đưa ra các đề xuất tốt hơn ít nhất 10% so với kết quả của CineMatch. Giải thưởng 1 triệu đô la năm 2009 đã được trao cho một giải pháp bao gồm 107 thuật toán khác nhau và kết hợp các đề xuất của chúng tùy thuộc vào hoàn cảnh (Ian MacKenzie, et al., 2013).

Thuật toán CineMatch hướng dẫn các máy chủ của Netflix xử lý thông tin từ cơ sở dữ liệu của nó để xác định những bộ phim mà khách hàng có thể sẽ thích (Algorithm Hall of Fame, 2023).

Thuật toán tính đến:

* Bản thân các bộ phim, được sắp xếp thành các nhóm phim thông thường.
* Xếp hạng của khách hàng, phim đã thuê và hàng đợi hiện tại.
* Xếp hạng tổng hợp của tất cả người dùng Netflix.

Tại sao nghiên cứu này cần được thực hiện? Nghiên cứu này được thực hiện vì đáp ứng yêu cầu đề ra của môn học và tìm hiểu ứng dụng của thuật toán học sâu\_thống kê.

Điều gì mà nghiên cứu này sẽ mang lại? Xây dựng một hệ thống gợi ý phim hoạt hình đơn giản ở quy mô cá nhân.

Với bộ dữ liệu có sẵn đề cương tập trung vào một hệ thống gợi ý phim hoạt hình dựa vào lượt đánh giá, các chỉ số đo sự quan tâm, yêu thích của khán giả đối với một bộ phim, hoặc theo thể loại phim tương tự,…Thông qua các thuật toán của học sâu\_thông kê(như: Bayes estimator - Internet Movie Database (IMDB), Sự tương đồng cosin - Measure of Similarity, Thuật toán K-Nearest Neighbors – KNN (thuật toán láng giềng gần/ hàng xóm lân cận)), đưa ra các gợi ý từ ngẫu nhiên đến có chủ đích theo mục tiêu của người dùng, hay theo xu hướng phổ biến của bộ phim hoạt hình, hoặc các tương tự liên quan. Đề cương tìm hiểu khả năng ứng dụng của học sâu\_thông kê ở quy mô nhỏ.

# **CHƯƠNG 3: PHƯƠNG PHÁP NGHIÊN CỨU**

## **3.1. Vật liệu nghiên cứu**

### **3.1.1. Vật liệu nghiên cứu**

Bộ dữ liệu là bộ dữ liệu thứ cấp: Anime Recommendations Database, (Kaggle, 2017).

Mô tả dữ liệu:

* + anime.csv: có 12294 hàng, 7 cột( anime\_id, name, genre, type, episodes, rating, members).
* anime\_id: id duy nhất của myanimelist.net xác định một anime.
* Name: tên đầy đủ của anime.
* Genre: danh sách thể loại được phân tách bằng dấu phẩy cho anime này.
* Type: phim, TV, OVA, v.v.
* Episodes: có bao nhiêu tập trong chương trình này. (1 nếu là phim).
* Rating: xếp hạng trung bình trên 10 cho anime này.
* Members: số thành viên cộng đồng nằm trong

"nhóm" của anime này.

* + rating.csv: có 7813737 hàng, 3 cột( user\_id, anime\_id, rating).
* user\_id: id người dùng được tạo ngẫu nhiên không thể nhận dạng.
* anime\_id: anime mà người dùng này đã xếp hạng.
* rating: xếp hạng trong số 10 người dùng này đã chỉ định (-1 nếu người dùng đã xem nhưng không chỉ định xếp hạng).

### **3.1.2. Các công cụ phần mềm**

Anaconda: phát hành 17 tháng 7 năm 2012 vào được cập nhật thêm qua từng năm. Anaconda là bản phân phối ngôn ngữ lập trình Python và R cho tính toán khoa học, với mục đích đơn giản hóa việc quản lý và triển khai gói. Bản phân phối bao gồm các gói khoa học dữ liệu phù hợp với Windows, Linux và macOS. Anaconda Distribution hay Anaconda Individual Edition là những sản phẩm miễn phí thuộc Anaconda, Inc., trong khi các sản phẩm khác của công ty là Anaconda Team Edition và Anaconda Enterprise Edition đều miễn phí. (Got It Vietnam, 2021)

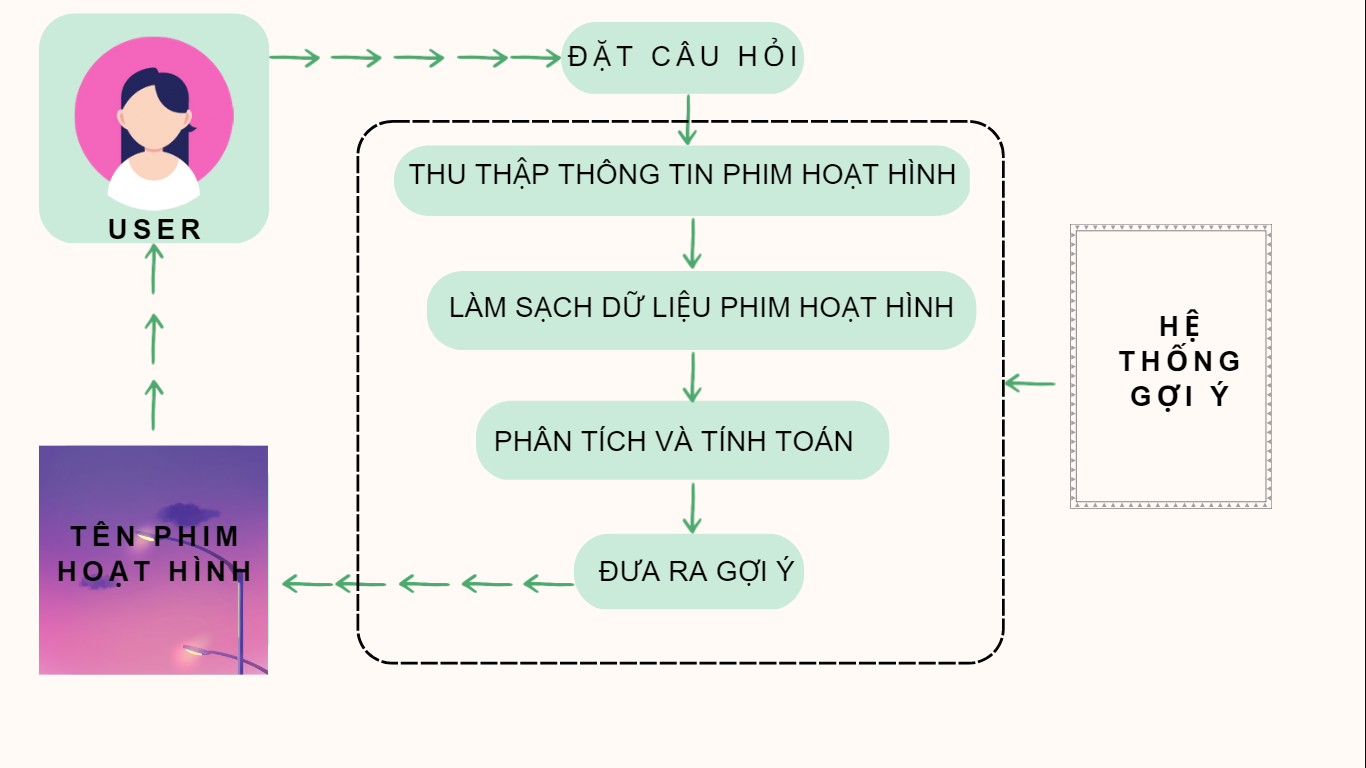
Python 3.11: được phát hành tháng 10 năm 2022. Python 3.11.1 là bản phát hành chính mới nhất của ngôn ngữ lập trình Python và nó chứa nhiều tính năng và tối ưu hóa mới. (PSF, 2022).

Excel 2019: giúp ghi lại, trình bày các thông tin với dạng bảng và nhiều tiện ích khác. Excel 2019 cho Windows nâng cấp từ một phiên bản Excel trước đó, bạn vẫn sẽ sở hữu mọi tính năng quen thuộc—cũng như sẽ thấy một số tính năng mới! (Microsoft, 2018).

.csv : Tệp CSV (các giá trị được phân tách bằng dấu phẩy) là tệp văn bản có định dạng cụ thể cho phép lưu dữ liệu ở định dạng có cấu trúc bảng (Google, 2021).

## **3.2. Các cách tiếp cận, phương pháp nghiên cứu, phạm vi nghiên cứu**

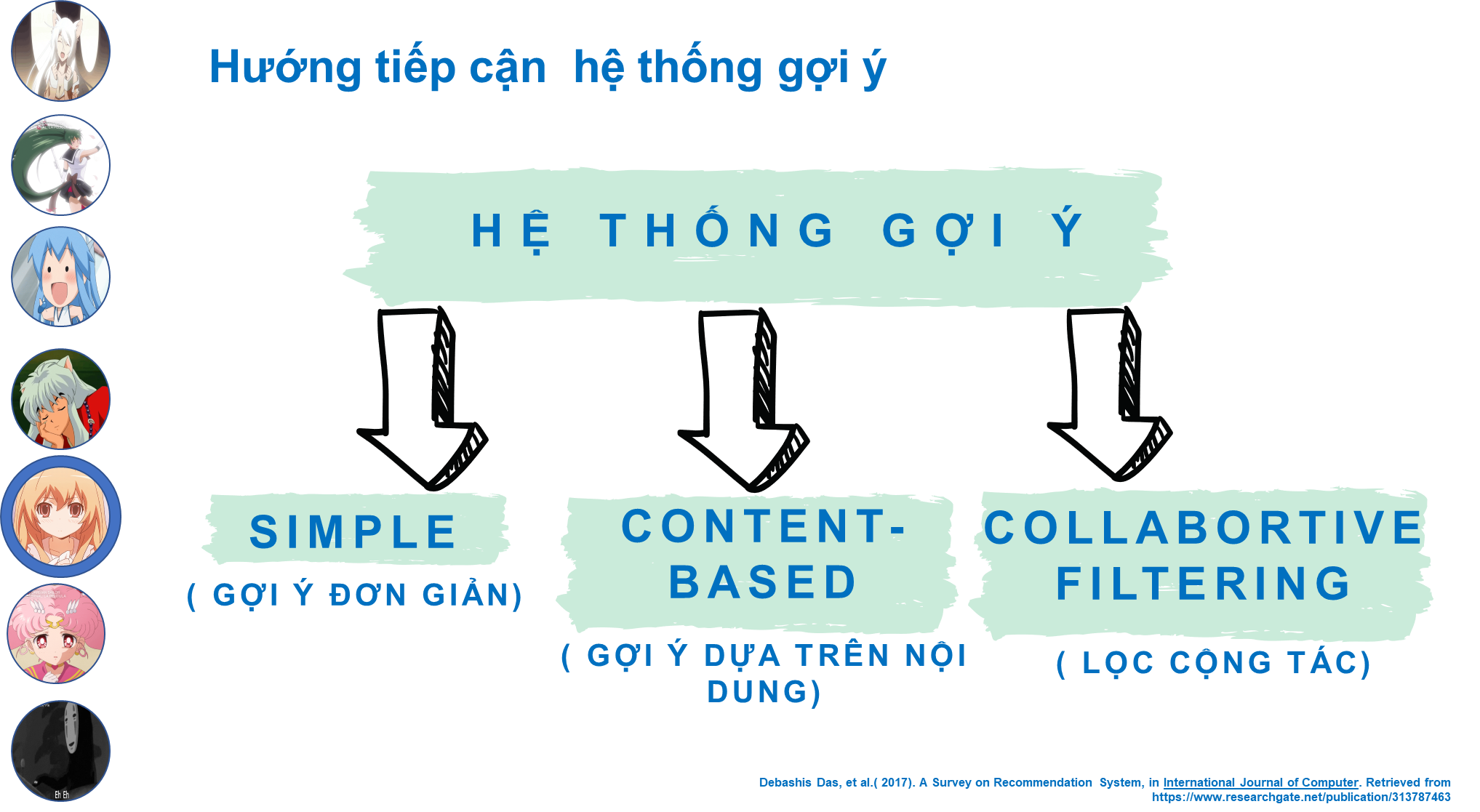
Thiết kế tổng quát dữ liệu đầu vào và đầu ra của đề cương.



**HÌNH 1: Quy trình hoạt động gợi ý đơn giản**

### **3.*2*.1. Các cách tiếp cận**

* + Tiếp cận dựa trên mức độ phổ biến (sample): trả lời câu hỏi gợi ý cho ai đó để làm gì?
  + Tiếp cận dựa trên nội dung (content-based): trả lời câu hỏi thư mục dữ liệu gợi ý gì cho người sử dụng?
  + Tiếp cận dựa trên lọc cộng tác (collabrative filterling- CF): trả lời câu hỏi có ai có sở thích tương tự mình không?



**HÌNH 2: Sơ đồ minh họa các hướng tiếp cận hệ thống gợi ý**

### **3.2.2. Phương pháp nghiên cứu**

Bước 1: Đặt câu hỏi để đưa ra quyết định trên dữ liệu.

Bước 2: Chuẩn bị dữ liệu để khám phá.

Bước 3: Xử lý dữ liệu từ bẩn sang sạch.

Bước 4: Phân tích dữ liệu đề trả lời câu hỏi.

Bước 5: Chia sẻ và trực quan hóa dữ liệu.

Bước 6: Trình bày kết quả.

A picture containing text, screenshot, font, design

Description automatically generated

**HÌNH 3: Quy trình nghiên cứu hệ thống gợi ý**

### **3.2.3*.* Phạm vi nghiên cứu**

Phạm vi nghiên cứu: ở mức độ nhỏ, quy mô cá nhân, đáp ứng nhu cầu môn học đề ra.

Sử dụng bộ dữ liệu có sẵn: Anime Recommendations Database (Union, 2017).

### **3.2.4. Thời gian thực hiện**

Từ ngày 02 tháng 02 năm 2023 đến ngày 09 tháng 06 năm 2023.

## **3.3. Phân tích hệ thống gợi ý phim hoạt hình**

Bước 1: Đặt câu hỏi để đưa ra quyết định trên dữ liệu.

- Bạn muốn xem gì?

- Từ dữ liệu có thể đưa ra các gợi ý theo xu hướng phim hoạt hình gì?

- Các thông tin từ bộ dữ liệu sử dụng các chỉ số đánh giá liên quan có thể đưa ra các gợi ý gì?

- Sử dụng dữ liệu có thể đáp ứng các yêu cầu nào?

- Các đặc trưng thống kê của dữ liệu cho biết những gì?

- Với bộ dữ liệu Anime Recommendations Database có sẵn có thể thực hiện được một số kỳ vọng như việc thiết kế thuật toán theo mức độ phổ biến( anime ( rating), anime (members), rating ( rating) ), hay theo mức độ liên quan trong một nhóm người,…

Bước 2: Chuẩn bị dữ liệu để bắt đầu phân tích.

- Cần thu thập để phân tích là việc hết sức quan trọng, để có một bộ dữ liệu đáp ứng nhu cầu đề ra cần rất nhiều thời gian. Có một số cách để thu thập dữ liệu như: phỏng vấn, khảo sát nhóm đại diện trong dân số, nguồn dữ liệu có sẵn, … đề cương sử dụng bộ dữ liệu có sẵn và miễn phí. Có thể xem lại mục 3.1.1. để hiểu rõ bộ dữ liệu.

- Tìm hiểu các kiểu dữ liệu và định dạng dữ liệu.

- Hiểu về cách truy cập cơ sở dữ liệu và trích xuất, lọc và sắp xếp dữ liệu chứa trong đó.

- Xác định các loại sai lệch khác nhau trong dữ liệu có thể ảnh hưởng đến đảm bảo độ tin cậy của dữ liệu như: tính toàn vẹn, kiểu dữ liệu trong cột không đồng bộ, dữ liệu khiếm khuyết, mất dữ liệu, trùng lặp dữ liệu,… cần được làm sạch trước khi sử dụng.

- Bộ dữ liệu có thể trả lời câu hỏi gì?

- Bất kỳ vấn đề gì có thể xảy ra với dữ liệu?

Bước 3: Xử lý dữ liệu từ bẩn sang sạch

- Việc sử lý dữ liệu bẩn rất quan trọng trước khi tiến hành phân tích: việc còn tồn tại những dữ liệu bẩn có thể tạo ra những ảnh hưởng đến mức độ chính xác và độ tin cậy của bộ dữ liệu. Làm sạch dữ liệu là cơ bản để đảm bảo khi dữ liệu được sử dụng, các kết luận đúng được đưa ra đáp ứng yêu cầu của người dùng.

- Kiểm tra tính toàn vẹn của dữ liệu.

- Kiểm tra các sai lệch trong dữ liệu và giải quyết nó.

- Kiểm tra và làm sạch dữ liệu bằng Anaconda.

- Các yếu tố và tầm quan trọng của việc làm sạch dữ liệu quyết định được các gợi ý tên phim hoạt hình từ bộ dữ liệu có thể đưa ra có thể tin tưởng được, độ chính xác càng gia tăng độ tin cậy của người dùng đối với hệ thống.

Bước 4: Phân tích dữ liệu đề trả lời câu hỏi.

- Cách sắp xếp dữ liệu theo mục tiêu để phân tích.

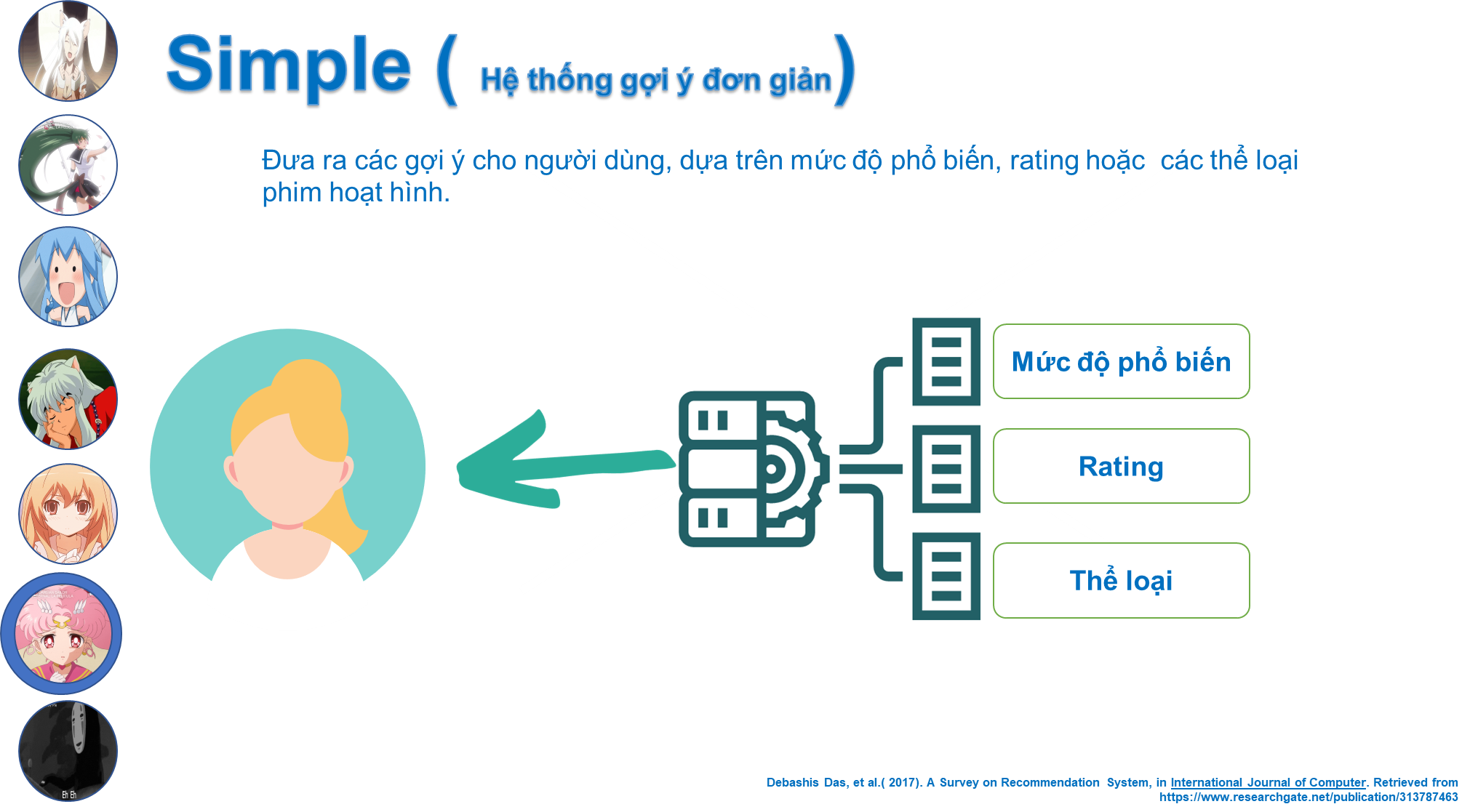
- Các quy trình định dạng và điều chỉnh dữ liệu để phù hợp với mục đích hướng tới cần giải quyết thông qua excel hoặc csv .

- Sử dụng các công thức và thuật toán để tính toán dữ liệu để sàn lọc các vấn đề quan tâm.

- Trực quan hóa sinh động bằng thư viện vẽ biểu đồ của ngôn ngữ lập trình python để dữ liệu thêm sinh động dễ hình dung hơn.

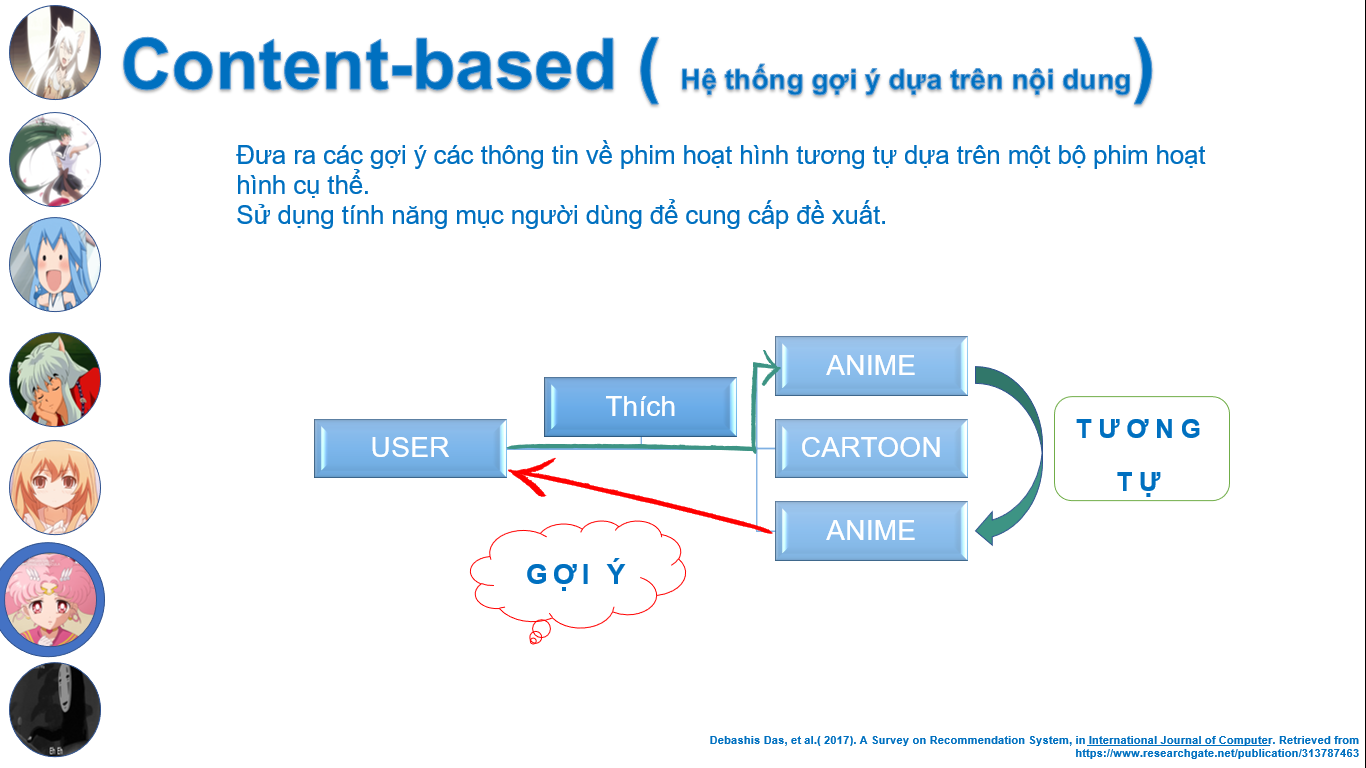
Tiến hành một số gợi ý cơ bản dựa trên các cách tiếp cận:

* Simple Recommenders: đưa ra các gợi ý cho người dùng, dựa trên mức độ phổ biến hoặc các thể loại phim hoạt hình (Debashis Das, et al., 2017).
* Mô tả:



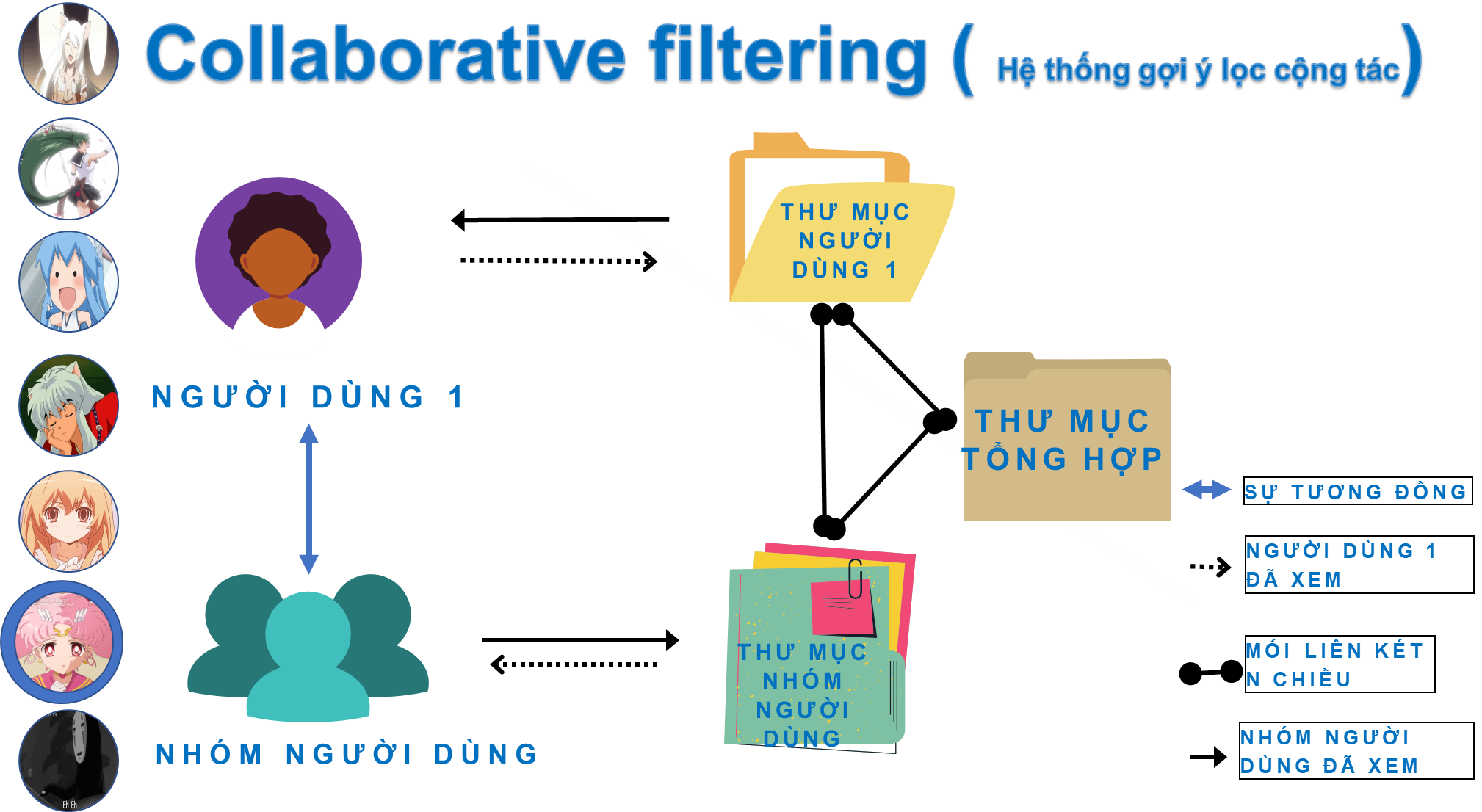
**HÌNH 4: Hệ gợi ý đơn giản (Simple Recommenders)**

* + Content-based Recommenders: gợi ý các thông tin về phim hoạt hình tương tự dựa trên một bộ phim hoạt hình cụ thể (Debashis Das, et al., 2017).
* Mô tả:



**HÌNH 5: Hệ thống gợi ý dựa trên nội dung (Content-based Recommenders)**

* + Collaborative filtering Recommenders: những hệ thống này cố gắng dự đoán “ratings” mà người dùng sẽ đưa ra cho một bộ phim hoạt hình dựa trên các chỉ số đánh giá, xếp hạng trước đây và sở thích của những người dùng khác (Debashis Das, et al., 2017).
  + Mô tả:



**HÌNH 6: Hệ thống lọc cộng tác (Collaborative filtering Recommenders)**

Bước 5: Chia sẻ và trực quan hóa dữ liệu.

- Phát hiện vấn vấn đề liên quan.

- Định hướng câu chuyện dữ liệu của hệ thống gợi ý phim hoạt hình.

- Trực quan hóa dữ liệu và chia sẻ.

Bước 6: Trình bày kết quả.

- Kết luận cuối cùng cho các phân tích là có đưa ra được kết quả gợi ý tên phim hoạt hình không?

- Có thể mở rộng và bổ sung gì trong việc gợi ý tên phim hoạt hình?

**CHƯƠNG 4: KẾT QUẢ KỲ VỌNG**

## **4.1. Ưu điểm, nhược điểm của các hệ thống gợi ý trên thế giới**

### **4.1.1. Ưu điểm của các hệ thống gợi ý trên thế giới**

* + - Tăng doanh số bán hàng và lợi nhuận cho các doanh nghiệp bằng cách giới thiệu sản phẩm hoặc dịch vụ tốt nhất cho khách hàng và có thể giúp các doanh nghiệp tăng doanh số bán hàng và cải thiện trải nghiệm khách hàng.
    - Giảm thời gian và công sức giúp người dùng tìm kiếm thông tin sản phẩm phù hợp một cách dễ dàng, nhanh chóng và giảm thiểu sự bối rối khi phải chọn từ hàng ngàn sản phẩm.
    - Hệ thống gợi ý có khả năng tùy chỉnh và cung cấp các gợi ý sản phẩm phù hợp với sở thích và nhu cầu của từng người dùng.
    - Hệ thống gợi ý có khả năng học hỏi từ hành vi của người dùng, giúp cải thiện chất lượng gợi ý theo thời gian.
    - Hệ thống gợi ý có khả năng phân tích và xử lý các dữ liệu lớn và phức tạp để tạo ra các gợi ý chính xác.

### **4.1.2. Nhược điểm của các hệ thống gợi ý trên thế giới**

* + - Một số hệ thống gợi ý có khả năng tạo ra các gợi ý giới hạn, không khuyến khích người dùng khám phá những sản phẩm mới, đưa ra gợi ý về các sản phẩm tương tự hoặc không quen thuộc, bị ảnh hưởng bởi các yếu tố ngoài tầm kiểm soát.
    - Các vấn đề liên quan đến bảo mật thông tin cá nhân của người dùng có thể xảy ra nếu hệ thống gợi ý thu thập dữ liệu và sử dụng thông tin cá nhân một cách không đúng đắn.
    - Việc thu phí hoặc quảng cáo trong hệ thống này có thể gây ra ảnh hưởng tiêu cực đến tính khách quan của các gợi ý sản phẩm.
    - Một số người dùng có thể không muốn chia sẻ thông tin cá nhân của mình để nhận được các gợi ý sản phẩm. Từ đó dẫn đến vấn đề phát sinh là hệ thống gợi ý không có đủ dữ liệu để xây dựng mô hình hoặc khi dữ liệu không chính xác gây sự thiếu hiểu biết về người dùng.
    - Hệ thống gợi ý có khả năng gây ra hiện tượng "hộp kín thông tin" (information bubble), khi người dùng chỉ nhận được các gợi ý giống nhau và không được tiếp cận với các sản phẩm hoặc ý kiến khác.
    - Hệ thống gợi ý có thể gây ra sự phiền toái cho người dùng nếu không được cài đặt hoặc điều chỉnh đúng cách. Hệ thống gợi ý có thể gặp vấn đề khi không hiển thị được các sản phẩm phù hợp với người dùng, hoặc khi hiển thị quá nhiều sản phẩm.
    - Một số hệ thống gợi ý có thể gây ra sự phiền toái cho người dùng khi chúng cố gắng quá mức để cung cấp các gợi ý. Hệ thống gợi ý có thể bị ảnh hưởng bởi các yếu tố không mong muốn, chẳng hạn như sự thiên vị trong dữ liệu hoặc việc lạm dụng thông tin cá nhân của người dùng.

## **4.2. Ưu điểm, nhược điểm của đề cương ý tưởng**

### **4.2.1. Ưu điểm**

* + - Sử dụng bộ dữ liệu có sẵn và miễn phí nên tiết kiệm được khá nhiều thời gian và công sức.
    - Bộ dữ liệu có tính toàn vẹn có thể sử dụng.
    - Có tính minh bạch.
    - Tránh được một số lỗi sai lệch khi thu thập bộ dữ liệu.
    - Bộ dữ liệu đã qua xử lý nên có tính chính xác có thể tin cậy khi sử dụng.

### **4.2.2. Nhược điểm**

* + - Hạn chế về mặt thu thập dữ liệu: vì sử dụng bộ dữ liệu có sẵn chỉ đáp ứng một số mục tiêu cơ bản đề ra ở mức độ đơn giản nhất định,..
    - Bộ dữ liệu có tính lỗi thời vì thu thập từ năm 2017.

## **4.3. Kết quả kỳ vọng**

* + Kết quả kỳ vọng từ các phương hướng tiếp cận có thể đưa ra các gợi ý tên phim hoạt hình.
  + Áp dụng được một số thuật toán(như: Bayes estimator - Internet Movie Database (IMDB), Sự tương đồng cosin - Measure of Similarity, Thuật toán K-Nearest Neighbors – KNN (thuật toán láng giềng gần/ hàng xóm lân cận)) để đưa ra gợi ý đơn giản.
  + Biểu diễn biểu đồ tạo sự trực quan sinh động, giúp dễ tiếp cận với đề cương.

# **CHƯƠNG 5: KẾT LUẬN**

Đề cương giải quyết được yêu cầu của môn học phương pháp nghiên cứu khoa học đề ra. Đề cương mô tả một số hướng tiếp cận phổ biến của hệ thống gợi ý. Áp dụng một số thuật toán cơ bản của học sâu- thống kê(như: Bayes estimator - Internet Movie Database (IMDB), Sự tương đồng cosin - Measure of Similarity, Thuật toán K-Nearest Neighbors – KNN (thuật toán láng giềng gần/ hàng xóm lân cận)) để đưa ra gợi ý tên phim hoạt hình.

Tìm hiểu thêm ứng dụng thực tế của khoa học vào cuộc sống. Đề cương nghiên cứu theo các phương hướng tiếp cận đơn giản để gợi ý cho người dùng tên của một số phim hoạt hình bằng cách áp dụng một số thuật toán cơ bản thông dụng bằng ngôn ngữ lập trình Python. Tóm lại, đề cương hướng tới một hệ thống gợi ý đơn giản ở mức độ quy mô cá nhân.

# **TÀI LIỆU THAM KHẢO**

Algorithm Hall of Fame. (2023). *Cinematch*. Được truy lục từ Algorithm Hall of Fame: https://www.algorithmhalloffame.org/algorithms/cinematch/

Apáthy, S. (2019). *History of Recommender Systems.* Được truy lục từ ONESPIRE: https://onespire.net/news/history-of-recommender-systems/

Burke, R. F. (2011). Recommender systems: An overview. *AI Magazine, 31(4), 17-38.*

Debashis Das, et al. (2017). A Survey on Recommendation System. *International Journal of Computer*. Được truy lục từ https://www.researchgate.net/publication/313787463

Google. (2021, 8 10). *About Google Ads optimization score*. Được truy lục từ Google: https://support.google.com/google-ads/answer/9004364?hl=en

Got It AI. (2021, 10 20). *Anaconda là gì? Tìm hiểu nền tảng khoa học dữ liệu phổ biến nhất*. Được truy lục từ Got It AI: https://vn.got-it.ai/blog/anaconda-la-gi-tim-hieu-nen-tang-khoa-hoc-du-lieu-pho-bien-nhat

Ian MacKenzie, et al. (2013, 10 19). How retailers can keep up with consumers. *McKinsey & Company*.

Internet Live Stats. (2022, 6 30). *Internet Live Stats*. Được truy lục từ Internet Live Stats: http://www.internetlivestats.com/

Isinkaye, F. O. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal, 16(3), 261-273*.

Microsoft. (2018). *Có gì mới trong Excel 2019 cho Windows*. Được truy lục từ Microsoft: https://support.microsoft.com/vi-vn/office/c%C3%B3-g%C3%AC-m%E1%BB%9Bi-trong-excel-2019-for-windows-5a201203-1155-4055-82a5-82bf0994631f

Python Software Foundation . (2021). *Python 3.11.1*. Được truy lục từ Python Software Foundation : https://www.python.org/downloads/release/python-3111/

Resnick, P. &. (1997). Recommender systems. *Communications of the ACM, 40(3), 56.*

Ricci, F. R. (2011). Introduction to Recommender Systems. Trong L. R. In F. Ricci, *Recommender Systems Handbook (pp. 1-35)* (trang Springer US).

Schafer, B., et al. (2000). E-Commerce Recommendation Applications. *Data Mining and Knowledge Discovery, 4(1), 115-153.* Được truy lục từ https://doi.org/10.1023/A:1009804230409

Shetty, B. (2023, 3 2). *Recommender Systems: What Are They And How Do They Work?* Được truy lục từ Built In: https://builtin.com/data-science/recommender-systems

Smith, J. (2012). *Recommendation system*. Được truy lục từ Citizendium.

Union, C. (2017). *Anime Recommendations Database*. Được truy lục từ Kaggle: https://www.kaggle.com/datasets/CooperUnion/anime-recommendations-database?resource=download&select=rating.csv