**TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

**🙢🙡🕮🙣🙠**



**BÁO CÁO BÀI TẬP LỚN XỬ LÝ NGÔN NGỮ TỰ NHIÊN**

**ĐỀ TÀI: TÌM HIỂU RECOMMENDER SYSTEM VÀ ÁP DỤNG NLP VÀO RECOMMNER SYSTEM**

**Giảng viên hướng dẫn: PGS.TS. Lê Thanh Hương.**

**Sinh viên thực hiện:**

1. **Nguyễn Thế Thái 20112208**
2. **Trần Hải Nam 20111896**
3. **Nguyễn Trường Thái 20112209**
4. **Nguyễn Thọ Đức 20111422**

**Hà Nội , tháng 11 năm 2015**

MỤC LỤC

[LỜI NÓI ĐẦU 2](#_Toc436229628)

[PHÂN CÔNG CÔNG VIỆC 3](#_Toc436229629)

[DANH SÁCH HÌNH ẢNH 4](#_Toc436229630)

[1. Tổng Quan Về Hệ Thống Gợi Ý (Recommender System). 5](#_Toc436229631)

[1.1. Giới thiệu về hệ thống gợi ý 7](#_Toc436229632)

[1.2. Hướng tiếp cận 8](#_Toc436229633)

[1.2.1. Gợi ý dựa trên nội dung. 8](#_Toc436229634)

[1.2.2. Lọc cộng tác. 8](#_Toc436229635)

[1.2.3. Gợi ý dựa trên thông tin Nhân khẩu học 8](#_Toc436229636)

[1.2.4. Dựa trên tri thức. 9](#_Toc436229637)

[1.2.5. Dựa trên cộng đồng. 9](#_Toc436229638)

[1.2.6. Gợi ý dựa trên hàm tiện ích. 10](#_Toc436229639)

[1.2.7. Hệ thống gợi ý lai. 10](#_Toc436229640)

[1.2.8. So sánh các phương pháp gợi ý. 10](#_Toc436229641)

[II. HỆ THỐNG GỢI Ý PHIM ÁP DỤNG XỬ LÝ NGÔN NGỮ TỰ NHIÊN. 13](#_Toc436229642)

[2.1. Lý do lựa chọn đề tài 13](#_Toc436229643)

[2.2. Phương pháp giải quyết. 13](#_Toc436229644)

[2.2.1. Thuật toán thứ 1: Độ đo tương tự về từ 14](#_Toc436229645)

[2.2.2. Thuật toán thứ 2: đo độ tương tự về thể loại chữ ký chủ đề. 14](#_Toc436229646)

[3. Chương trình DEMO 17](#_Toc436229647)

[TÀI LIỆU THAM KHẢO 18](#_Toc436229648)

# LỜI NÓI ĐẦU

Xử lý ngôn ngữ tự nhiên (natural language processing - NLP) là một nhánh của [trí tuệ nhân tạo](https://vi.wikipedia.org/wiki/Tr%C3%AD_tu%E1%BB%87_nh%C3%A2n_t%E1%BA%A1o) tập trung vào các ứng dụng trên [ngôn ngữ](https://vi.wikipedia.org/wiki/Ng%C3%B4n_ng%E1%BB%AF) của con người. Trong trí tuệ nhân tạo thì xử lý ngôn ngữ tự nhiên là một trong những phần khó nhất vì nó liên quan đến việc phải hiểu ý nghĩa ngôn ngữ-công cụ hoàn hảo nhất của [tư duy](https://vi.wikipedia.org/wiki/T%C6%B0_duy) và [giao tiếp](https://vi.wikipedia.org/wiki/Giao_ti%E1%BA%BFp).

**Hệ thống gợi ý (Recommendation System)** là ứng dụng rõ ràng nhất và hữu dụng nhất của Khoa học dữ liệu (Data Science). Bạn có thể đã gặp ứng dụng của hệ thống này khi qua các trang Amazon, Netflix, Vinabook … Ngoài nội dung mà bạn đang xem, các trang nói trên còn hiển thị các nội dung liên quan, giữ người dùng ở lại trên trang web lâu hơn, tăng được số lượng khách hàng tiềm năng.

Xử lý ngôn ngữ tự nhiên và hệ thống gợi ý là hai lĩnh vực có liên hệ rất mật thiết với nhau. Hai lĩnh vực này cùng bổ trợ cho nhau để phát triển, giúp những nhà phát triển tạo ra những hệ thống hoạt động thông minh và hiệu quả. Trong bài tập lớn này chúng em sẽ trình bày những tìm hiểu của mình về Hệ thống gợi ý và trình mày một ứng dụng về việc sử dụng xử lý ngôn ngữ tự nhiên để xây dựng một hệ thống gợi ý phim.

Báo cáo của chúng em gồm có 3 phần chính:

* Tổng Quan Về Hệ Thống Gợi Ý (Recommender System).
* Hệ thống gợi ý phim áp dụng xử lý ngôn ngữ tự nhiên
* Chương trình DEMO

PHÂN CÔNG CÔNG VIỆC

Nguyễn Trường Thái: Tìm hiểu và xây dựng chương trình DEMO

Nguyễn Thế Thái: Tìm hiểu về Hệ thống gợi ý.

Trần Hải Nam: Tìm hiểu về cách áp dụng NLP vào RS

Nguyễn Thọ Đức: Tìm hiểu thuật toán và xây dựng chương trình demo.

# DANH SÁCH HÌNH ẢNH

Hình 1: Phương thức hoạt động của một hệ thống gợi ý cơ bản

Hình 2: Hệ gợi ý của AMAZON

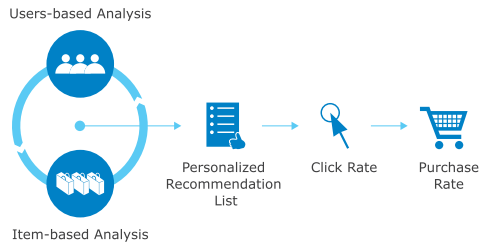
Hình 3: Hệ gợi ý của NETFLIX

Hình 4: Hệ gợi ý của YARPP

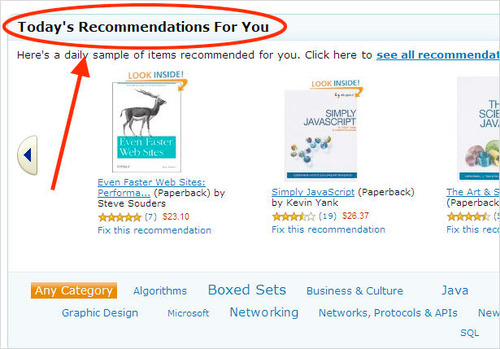
Hình 5: Điểm số về độ tương tự trung bình đối với các thuật toán

# Tổng Quan Về Hệ Thống Gợi Ý (Recommender System).

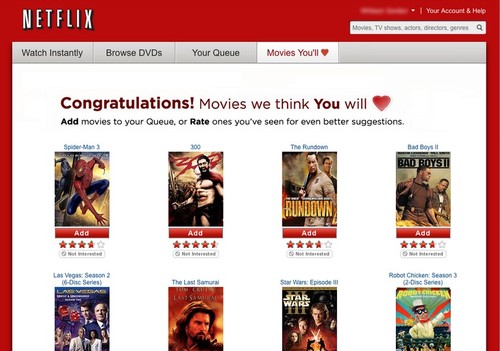
Hệ thống gợi ý (recommder system – RS) là một hệ thống tích hợp mở rộng và không thể thiếu đối với bất cứ sản phẩm internet nào hiện nay. Dù xuất hiện với quy mô lớn hay nhỏ thì RS đóng vai trò quan trọng giúp sản phẩm hoàn thiện hơn.



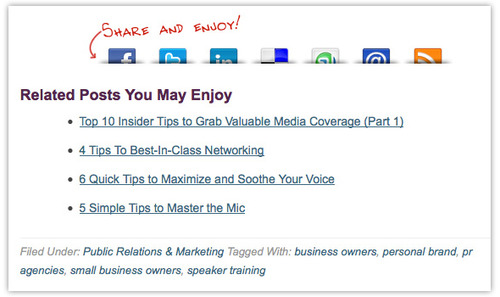
Hình 1: Phương thức hoạt động của một hệ thống gợi ý cơ bản



Hình 2: Hệ gợi ý của AMAZON



Hình 3: Hệ gợi ý của NETFLIX



Hình 4: Hệ gợi ý của YARPP

Ba ví dụ trên lần lượt là hệ thống gợi ý của Amazon, Netflix và YARPP. Với mức độ quy mô khác nhau nhưng đều tập trung vào mục tiêu hiểu và dự đoán được thị hiếu người dùng chính xác nhất.

## Giới thiệu về hệ thống gợi ý

Hệ thống gợi ý (Recommender systems hoặc Recommendation systems) là một phân lớp của hệ thống lọc thông tin, tìm cách dự đoán trước mức độ ưu tiên của người dùng với đối tượng sản phẩm. Hệ thống gợi ý đã trở nên rất phổ biến trong những năm gần đây và được áp dụng trong rất nhiều lĩnh vực khác nhau như âm nhạc, phim ảnh, thương mại, tin tức, sách báo, truy vấn tìm kiếm, …

Một số lý do cần hệ thống gợi ý

* ***Tăng doanh số bán hàng.*** Đây có lẽ là chức năng quan trọng nhất của hệ thống gợi ý thương mại, ví dụ có thể bán thêm nhiều sản phẩm hơn trong cùng một bộ sản phẩm so với bất kì hình thức gợi ý nào khác. Mục tiêu này hoàn toàn có thể đạt được vì các sản phẩm gợi ýhoàn toàn phù hợp với nhu cầu và mong muốn của khách hàng. Ứng dụng phi thương mại cũng có chung mục tiêu này. Ví dụ một hệ thống báo điện tử cần tăng số lượng tin tức độc giả đọc từ trang của mình. Nhìn chung, đứng ở quan điểm của nhà cung cấp dịch vụ, hệ thống gợi ý giúp tối đa hoá tỉ lệ chuyển đổi, tức là, số lượng người dùng chấp nhận các gợi ý và tiêu thụ sản phẩm so với số lượng khách tiêu thụ thông tin bình thường khác.
* ***Bán nhiều sản phẩm đa dạng hơn***. Một chức năng nổi bật của hệ thống gợi ý là giúp người dùng có thể chọn lựa sản phẩm mà có thể rất khó tìm thấy nếu không thông qua việc gợi ý chính xác. Ví dụ, trong một hệ thống gợi ý phim như của Netflix, người cung cấp dịch vụ muốn cho thuê tất cả số đĩa phim họ có thay vì chỉ tập trung vào các phim được nhiều người quan tâm. Nếu không có hệ thống gợi ý thì việc tập trung quảng cáo vào một số tựa phim ít người quan tâm sẽ mang lại rủi ro cho nhà cung cấp dịch vụ vì chúng có thể không phù hợp với số đông người dùng. Ngược lại, hệ thống gợi ý hoàn toàn có thể tư vấn bộ phim cho đúng người dùng quan tâm.
* ***Gia tăng sự thoả mãn của khách hàng***. Một hệ thống gợi ý được thiết kế tốt cũng làm gia tăng trải nghiệm khách hàng trên một website hoặc một ứng dụng. Người dùng cảm thấy việc gợi ý là có ích và có liên quan, cùng với thiết kế tương tác phù hợp, người dùng sẽ thích sử dụng hệ thống. Điều này làm gia tăng đánh giá tích cực từ chủ quan người dùng. Từ đó gia tăng việc sử dụng hệ thống và tỉ lệ các gợi ý được chấp nhận.
* ***Tăng người sử dụng trung thành***. Hệ thống phải phân biệt được ai là khách hàng cũ của mình và đối xử với họ như một người dùng đem lại giá trị. Đây là một chức năng phổ biến của hệ thống gợi ý bởi vì hệ thống gợi ý tận dụng được các thông tin có được từ những tương tác trước đây của người sử dụng, ví dụ đánh giá của người dùng về các sản phẩm. Do đó, người dùng tương tác với trang web càng nhiều, mô hình người dùng được hệ thống phác thảo lại càng chính xác và từ đó hệ thống có thể nắm được sở thích người dùng và kết quả của hệ thống gợi ý càng được tối ưu hơn.
* ***Hiểu rõ hơn khách hàng muốn gì***. Một chức năng quan trọng khác của hệ thống gợi ý là mô tả được sở thích người dùng, hoặc từ thu thập rõ ràng từ hành vi người dùng hoặc dựa trên dự đoán. Nhà cung cấp dịch vụ có thể quyết định tái sử dụng những thông tin này cho một số mục tiêu khác nhưcải thiện số tồn sản phẩm tại cửa hàng hoặc năng lực sản xuất. Ví dụ, một tổ chức quản lý tour du lịch có thể thiết kế tour phù hợp với một nhóm khách hàng mới dựa trên dữ liệu từ nhóm khách hàng cũ.

## Hướng tiếp cận

Có rất nhiều hướng tiếp cận khác nhau để giải quyết bài toán gợi ý.

### Gợi ý dựa trên nội dung.

Là phương pháp sử dụng mô tả của sản phẩm để suy luận độ phù hợp của chúng tới người sử dụng. Mỗi người sử dụng được mô hình thông qua một hồ sơ. Hệ thống sẽ sử dụng hồ sơ của người sử dụng để đoán trước sản phẩm đó có phù hợp với họ hay không.

Gợi ý dựa trên nội dung có thể đưa ra những sản phẩm mới nhất tới người sử dụng thông qua việc mô tả nội dung của sản phẩm với các yêu cầu người sử dụng.

Tuy nhiên, hạn chế của phương pháp này là việc quản lý những người sử dụng mới. Đối với những người sử dụng này, hệ thống đòi hỏi phải có đầy đủ thông tin về các lĩnh vực họ quan tâm và ưa thích để xây dựng các tập phân loại có chất lượng tốt.

### Lọc cộng tác.

Quá trình lọc cộng tác là hướng tiếp cận thông dụng nhất hiện nay. Hướng tiếp cận lọc cộng tác tập hợp các đánh giá của người sử dụng đối với một sản phẩm và sử dụng chúng để phân nhóm người dùng thành các nhóm có cùng chung sự đánh giá. Hệ thống sử dụng lọc cộng tác dự đoán mỗi quan tâm của người dùng đối với một sản phẩm dựa trên đánh giá của người sử dụng khác trong cùng nhóm phân loại với họ. Khác với kỹ thuật lọc thông tin khác, kỹ thuật lọc cộng tác không khai thác dữ liệu mô tả sản phẩm (hoặc là ăn bản, hoặc là đặc điểm). mục tiêu chính của lọc cộng tác là tìm ra những người sử dụng có cùng chung sở thích.

Tuy nhiên, lọc cộng tác cũng gặp vấn đề như đánh giá sớm, đánh giá không chính xác và rải rác của người dùng. Sản phẩm mới không thể được gợi ý tới người sử dụng cho đến khi có một số lượng các đánh giá phù hợp với các sản phẩm này. Vấn đề đánh giá rải rác này xảy ra khi một số ít người cùng đánh giá một số sản phẩm.

### Gợi ý dựa trên thông tin Nhân khẩu học

Là phương pháp sử dụng dữ liệu nhân khẩu học của người sử dụng để xác định mối quan hệ giữa người sử dụng và sản phẩm, phân thành nhiều nhóm người sử dụng khác nhau, các thành viên trong mỗi nhóm có cùng đặc điểm về nhân khẩu học. Dữ liệu về nhân khẩu học có thể là dữ liệu hồ sơ (như tên, số điện thoại…), dữ liệu địa lý (như zip, code, nước,…), dữ liệu về tình trạng cá nhân (như giới tính, tuổi), các dữ liệu về thể hiện sự ưa thích cá nhân (như phong cách sống…) và các dữ liệu đăng ký (đăng ký chào và bán hàng,…)… Một nhóm người có cùng đặc điểm về nhân khẩu học sẽ được hệ thống gợi ý cùng một tập sản phẩm.

Gợi ý dựa trên thông tin nhân khẩu học có tác dụng tốt với các sản phẩm mới. Tuy nhiên phương pháp này có hạn chế đối với những người sử dụng mới, khi thông tin nhân khẩu học của họ chưa có được đầy đủ.

Tuy nhiên phương pháp này đòi hỏi yêu cầu về dữ liệu nhân khẩu học của người dùng. Trong khi các thông tin này lại tương đối nhạy cảm với người sử dụng, dữ liệu loại này thường không chính xác và đầy đủ, do đó không có nhiều hệ thống gợi ý sử dụng hướng tiếp cận này.

### Dựa trên tri thức.

Hệ thống gợi ý dựa trên tri thức dựa trên miền kiến thức chuyên biệt để xác định chính xác xem các tính năng của sản phẩm có phù hợp đúng với nhu cầu và sở thích của người dùng hay không hay nói ngắn gọn là sản phẩm gợi ý đưa ra hữu ích cho người dùng tới mức nào. Hệ thống gợi ý dựa trên tri thức nổi bật là dựa trên trường hợp. Trong các hệ thống này một hàm số tương đồng sẽ ước tính sự tương đồng giữa nhu cầu người dùng (mô tả bài toán) và sự gợi ý. Ở đây hệ số tương đồng có thể được hiểu trực tiếp là công dụng của sự gợi ý cho người dùng. Hệ thống dựa trên ràng buộc là một loại khác của hệ thống gợi ý dựa trên tri thức. Về mặt kiến thức sử dụng, hai hệ thống tương đương nhau: yêu cầu người dùng được thu thập; việc sửa chữa cho các yêu cầu không phù hợp được tự động đề xuất trong trường hợp không có giải pháp nào khả thi; và kết quả gợi ý được diễn giải ra. Điểm khác biệt cơ bản nằm ở cách giải pháp được tính toán ra. Hệ thống gợi ý dựa trên trường hợp xác định gợi ý dựa trên thông số tương đồng cơ bản trong khi hệ thống gợi ý dựa trên ràng buộc chủ yếu khai thác kiến thức đã được định sẵn mà chứa các luật rõ ràng về liên hệ giữa nhu cầu người dùng và tính năng của sản phẩm. Hệ thống dựa trên tri thức có xu hướng hoạt động tốt hơn vào giai đoạn đầu triển khai nhưng nếu không được trang bị thành phần học kiến thức thì hệ thống này sẽ bị vượt qua nhanh chóng bỏ các phương pháp trực diện như phân tích tương tác người người máy (ví dụ trong lọc cộng tác).

### Dựa trên cộng đồng.

Loại hệ thống gợi ý này dựa trên sở thích của bạn bè của người dùng. Kĩ thuật này dựa trên châm ngôn “Hãy nói cho tôi bạn của anh, tôi sẽ cho anh biết anh là người như thế nào”. Các nghiên cứu cho thấy, mọi người thường có xu hướng tin tưởng gợi ý từ bạn bè mình hơn gợi ý từ những người họ không biết chính xác. Quan sát này cộng với việc phát triển nhanh chóng của các mạng xã hội đang tạo ra nhu cầu lớn trong hệ thống dựa trên cộng đồng hay còn gọi là hệ thống gợi ý xã hội. Loại hệ thống gợi ý này yêu cầu thông tin về mối liên hệ giữa người dùng vô vàn sở thích của bạn bè họ. Việc gợi ý dựa trên đánh giá của bạn bè người dùng. Thật sự, hệ thống gợi ý này đi theo việc phát triển của mạng xã hội và cho phép lấy dữ liệu một cách đơn giản và toàn diện dựa trên các mối quan hệ xã hội của người dùng.

### Gợi ý dựa trên hàm tiện ích.

Phương pháp này dựa trên lý thuyết ra quyết định sử dụng hàm tiện ích. Giả sử hàm tiện ích của người sử dụng được cho trước, khi lựa chọn một trong hai quyết định nào đó thì người sử dụng sẽ lựa chọn quyết định nào có giá trị tiện ích cao hơn. Hệ thống gợi ý dựa trên sự tiện ích xây dựng các hàm tiện ích tương ứng với sự quan trọng của các thuộ tính có ích, phục vụ cho việc đưa ra các gợi ý. Gợi ý dựa trên hàm tiện ích không phải xây dựng tiểu sử của người sử dụng nên không gặp vấn đề về sản phẩm mới và người dùng.

Hạn chế lớn nhất trong trường hợp số lượng các thuộc tính được bao hàm trong hàm tiện ích là quá lớn thì cách tiếp cận này là không thực tế, hạn chế lớn nhất của cách tiếp cận gợi ý dựa trên hàm tiện ích là phải xây dựng hàm tiện ích cho mỗi cá thể một cách chính xác. Việc xây dựng hàm chức năng thường hoàn thành trong một hoàn cảnh rõ ràng.

### Hệ thống gợi ý lai.

Là phương pháp sử dụng cách kế hợp hai hay nhiều phương pháp gợi ý trên cùng một hệ thống. Mỗi phương pháp gợi ý có những thế mạnh và hạn chế riêng. Vì vậy để hệ thống đạt hiệu quả cao trong quá trình sử dụng, việc lai các phương pháp gợi ý là giải pháp tối ưu nhất để giảm hạn chế tới mức thấp nhất. Phương pháp lọc cộng tác thường được sử dụng để lai với một số phương pháp khác, ví dụ lọc cộng tác thường gặp vấn đề với sản phẩm mới vì chúng chưa có dữ liệu về đánh giá nhưng các hệ thống dựa trên nội dung thì lại không gặp phải vấn đề này. Tuy nhiên, việc kết hợp giữa các phương pháp gợi ý gặp nhiều khó khăn khi thực hiện.

### So sánh các phương pháp gợi ý.

|  |  |  |
| --- | --- | --- |
| Phương pháp gợi ý | Ưu điểm | Nhược điểm |
| Lọc cộng tác | -chất lượng tăng theo thời gian  -không cần tri thức miền | -vấn đề người sử dụng mới  -vấn đề sản phẩm mới  -phụ thuộc vào dữ liệu lịch sử  -tính ổn định và mềm dẻo |
| Lọc dựa trên nội dung | -chất lượng cải thiện theo thời gian  -không cần miền tri thức | -vấn đề người sử dụng mới  -phụ thuộc vào lịch sử dữ liệu  -tính ổn định và tính mềm dẻo |
| Lọc theo nhân khẩu học | -không cần tập hợp danh sách đánh giá từ người dùng  -chất lượng cải thiện theo thời gian | -Phải tập hợp thông tin nhân khẩu học  -phụ thuộc vào dữ liệu lịch sử  -tính ổn định và tính mềm dẻo. |
| Lọc dựa trên hàm tiện ích | -có thể sử dụng các đặc điểm không thuộc sản phẩm | -người dùng cần nhập hàm tiện ích  -khó thực hiện |
| Lọc dựa trên cơ sở tri thức | -yêu cầu người sử dụng ở mức thấp  -không bao hàm vấn đề khởi đầu | -cần phải thu nhận tri thức  -chất lượng gợi ý phụ thuộc vào tri thức |

Hệ thống gợi ý sử dụng lọc cộng tác phụ thuộc vào các đánh giá chồng nhau của người sử dụng. Khi không gian đánh giá thưa thớt (một sản phẩm có rất ít người đánh giá) thì phương pháp này gặp nhiều khó khăn. Phương pháp gợi ý dựa trên lọc cộng tác không thể đưa ra gợi ý cho những người dùng chưa từng đánh giá một sản phẩm nào đó. Một vấn đề khác trong hệ thống gợi ý sử dụng lọc cộng tác là vấn đề người sử dụng mới bởi vì các gợi ý chỉ được đưa ra khi mà hệ thống so sánh các đánh giá của người dùng này với các đánh giá của người dùng khác. Do đó người sử dụng mới, với một số ít các đánh giá sẽ không được phân nhóm một cách chính xác. Vấn đề sản phẩm mới cũng là một hạn chế của phương pháp lọc cộng tác. Phương pháp lọc cộng tác vẫn được sử dụng rộng rãi hiện nay và rất thành công.MỘt lợi thế của phương pháp này chính là hệ thống không cần phải dựa trên các đặc điểm, đặc tính của sản phẩm để đưa ra gợi ý. Thêm nữa, phương pháp này có thể đưa ra các gợi ý tình cờ vì các gợi ý được đưa ra không cần phải dựa vào nội dung của sản phẩm.

Phương pháp gợi ý dựa trên thông tin nhân khẩu học không gặp phải vấn đề người sử dụng mới do không đòi hỏi một danh sách các đánh giá từ người sử dụng. Nhưng thay vào đó là vấn đề thu thập thông tin nhân khẩu học của người sử dụng. Với tính chất nhạy cảm về tính riêng tư, các thông tin này càng khó để thu thập được chính xác.

Phương pháp gợi ý dựa trên nội dung gặp vấn đề “khởi động” vì hệ thống cần phải tập hợp đủ các đánh giá để có thể phân loại một cách chính xác. Cũng giống như phương pháp lọc cộng tác, phương pháp gợi ý dựa trên nội dung cũng gặp phải vấn đề giới hạn các đặc điểm của sản phẩm có liên quan tới mục tiêu gợi ý.

Phương pháp gợi ý dựa trên hàm tiện ích và cơ sở tri thức không gặp phải vấn đề thư athowts dữ liệu và “khởi động”. Kỹ thuật lọc dựa trên hàm tiện ích đòi hỏi hệ thống cần phải có các hàm tiện ích hoàn chỉnh đối với tất cả các đặc điểm của đối tượng đang xem xét. Một lợi ích của hướng tiếp cận này có thể liên kết nhiều nhân tố ngoài những đặc điểm chính của sản phẩm, ví dụ như thời gian vận chuyển, điều kiện bảo hành. Tuy nhiên, sự linh hoạt này cũng có một số nhược điểm. Người dùng cần phải xây dựng hàm tiện ích hoàn chỉnh và cần phải đánh giá sự quan trọng của các đặc tính của sản phẩm. Thông thường, công việc này là khó khăn và mất nhiều công sức.

Phương pháp lọc dựa trên cơ sở tri thức có hạn chế là cần phải có “thu nhận tri thức”. Ngoài hạn chế đó, hệ gợi ý dựa trên cơ sở tri thức có một số lợi thế lớn. Phương pháp này có thể đưa ra các gợi ý “sâu”, ít phụ thuộc vào người sử dụng hơn phương pháp dựa trên hàm tiện ích, không gặp phải vấn đề “khởi động”.

Cho tới nay, có rất nhiều hướng tiếp cận khác nhau để xây dựng hệ gợi ý. Trong các phương pháp đó thì hai phương phap gợi ý dựa trên nội dung và lọc cộng tác được sử dụng rộng rãi nhất. Các ứng dụng của hướng tiếp cận dựa trên nội dung được giới hạn trong một số ít lĩnh vực như gợi ý các bài viết mới trên web, phương pháp gợi ý dựa trên lọc cộng tác được sử dụng rộng rãi hơn, trong hầu hết các ứng dụng gợi ý.

# II. HỆ THỐNG GỢI Ý PHIM ÁP DỤNG XỬ LÝ NGÔN NGỮ TỰ NHIÊN.

## 2.1. Lý do lựa chọn đề tài

Trong khi hệ thống gợi ý đang làm khá hiệu quả trong việc gợi ý dựa trên nội dung, nhưng với những người dùng mới (có ít thông tin và sở thích) thì hiệu suất gợi ý đúng giảm đáng kể. Những tình huống như vậy ko phải là hiếm.

Việc này dẫn tới việc cần phải có một ý tưởng mới để phát triển hệ thống với những trường hợp như vậy trở nên hiệu quả hơn. Nghiên cứu đầu tiên là về sử dụng Xử lý ngôn ngữ tự nhiên đã giúp hệ thống gợi ý tốt hơnmà không quá phụ thuộc vào thông tin của người sử dụng.

Hệ thống IMDb (Internet Movie Database) đã sử dụng một thuật toán “ Sử dụng đặc trưng người dùng như điểm vote, thể loại phim, title, keywords, và quan trọng nhất là đề nghị người dùng sử dụng chức năng gợi ý tự động của hệ thống. ”. Do vậy, Hệ thống yêu cầu rất nhiều sự hỗ trợ từ người dùng trong việc nhập các từ khóa…, việc làm này tốn rất nhiều thời gian. Bằng việc sử dụng các kỹ thuật thống kê trong NLP(Natural Language Processing) hệ thống hy vọng tìm ra những cách gợi ý hoàn toàn tự động để sinh ra những gợi ý chính xác cho người sử dụng.

Hệ thống sẽ sử dụng một phép đo lường gọi là độ tương tự. Ví dụ: Với một bộ phim, hệ thống sẽ tìm kiếm trong cơ sở dữ liệu và sinh ra “độ tương tự” cho mỗi bộ phim với bộ phim đó. Gợi ý của hệ thống sẽ tập trung vào những bộ phim mà có “độ tương tự” gần nhất với bộ phim mà người dùng đánh giá cao. Trong việc tạo ra 1 phương pháp đo “độ tương tự”, hệ thống kết hợp đặc trưng của một bộ phim “đạo diễn, tác giả, diễn viên…” với việc phân tích tóm tắt cốt truyện dựa trên NLP.

## 2.2. Phương pháp giải quyết.

Hệ thống gợi ý coi việc giới hạn về thông tin người dùng là một trong những nội dung đo độ tương tự. Đối với một bộ phim, Chúng ta quan tâm tới cấu trúc của một bộ phim (đạo diễn, diễn viên, tác giả…), cũng như đoạn văn bản mô tả tóm tắt bộ phim. Vì các thông tin là tuyến tính, hệ thống sẽ đo độ tương tự cho từng loại thông tin.Như vậy hệ thống sẽ có các điểm tương tự về đạo diễn, diễn viên, tác giả…., và độ tương tự về nội dung tóm tắt của bộ phim.

Đối với các thông tin có cấu trúc, hệ thống sử dụng phương pháp tính điểm đơn giản là tính độ trùng nhau giữa 2 thông tin. Ví dụ: Sự giống nhau về dàn cast giữa 2 bộ phim là bằng (***tổng số cast chung/tổng số cast của cả 2 bộ phim***).

Tiếp theo, đối với thông tin không có cấu trúc, chúng ta có 2 thuật toán áp dụng NLP để tính độ tương tự giữa 2 thông tin đó. Phương pháp đầu tiên, chúng ta sử dụng tiêu chuẩn tính độ tương tự trong không gian vector từ(word-space Vector) thường được sử dụng trong việc khôi phục thông tin. Phương pháp thứ 2 là việc sử dụng chữ ký chủ đề để tạo bộ vector từ dựa trên chủ đề của phim, nó sẽ được sử dụng như một điểm để so sánh giữa 2 phim.

### 2.2.1. Thuật toán thứ 1: Độ đo tương tự về từ

Thuật toán thứ nhất là sử dụng độ tương tự về từ. Mỗi tóm tắt cốt truyện được chuyển thành một vector có các thành phần là số nhị phân, mỗi thành phần vector đại diện cho một từ cụ thể có trong văn bản tóm tắt bộ phim. Sự giống nhau giữa 2 bộ phim là việc tính khoảng cách cosin giữa 2 vector đại diện.

### 2.2.2. Thuật toán thứ 2: đo độ tương tự về thể loại chữ ký chủ đề.

Thuật toán này bỏ qua các từ đơn lẻ, riêng biệt, không có ý nghĩa trong tóm tắt cốt truyện ra và thay vào đó nó sẽ tập trung vào những từ, cụm từ có ý nghĩa trong việc xác định phim đó thuộc về một thể loại phim cụ thể nào.

Như vậy chúng ta có thể nói bộ phim X là tương tự bộ phim thứ Y không phải vì 2 bộ phim đó có những từ tương tự, nhưng ta vẫn nói 2 bộ phim đó tương tự nhau vì nó cùng là phim kinh dị, và càng không phải tương tự với một bộ phim khoa học viễn tưởng. Ở đây chúng ta sử dụng một phép so sánh đơn giản hơn thay vì sử dụng việc phân tích các từ. Các thuật toán cho việc này như sau:

Chúng ta tạo ra chữ ký chủ đề cho mỗi thể loại theo quy định IMDB. Chữ ký viết là danh sách thuật ngữ mà trọng bằng cách chỉ định quy định thời hạn là các thể loại. Bảng 1 cho thấy ví dụ về các chủ đề chữ ký cho 2 thể loại.

|  |  |  |  |
| --- | --- | --- | --- |
| Sci-Fi | | Horror | |
| 1689.4 | Alien | 1142.5 | Vampire |
| 1155.9 | Planet | 519.2 | Blood |
| 1094.1 | Space | 489.9 | Horror |
| 886.3 | Scientist | 387.1 | Monster |

Chúng ta mở rộng danh sách các thể loại bằng cách sử dụng phân cấp chủ đề tập hợp để tạo ra các thể loại mới mà không phải là mô tả trong IMDB. Những hạng mục mới được tạo ra các thể loại phát hiện đầu tiên mà là tương tự nhau. Sát nhập những thể loại với nhau, và sau đó tạo ra chữ ký đề tài cho các thể loại mới được sát nhập. Bảng 2 cho thấy ví dụ về chữ ký đề tài cho hai của mở rộng danh mục thể loại.

|  |  |  |  |
| --- | --- | --- | --- |
| Comedy/Romance | | Action/ Adventure | |
| 747.3 | Love | 375.0 | Jungle |
| 452.9 | Marry | 207.5 | Pirate |
| 293.7 | Comedy | 179.8 | Sword |
| 70.9 | Sex | 66.5 | Smuggler |

Chúng ta sử dụng chữ ký chủ đề trong những thể loại mở rộng như mô tả vector của các thể loại phim mô hình. Đối với Ví dụ, chữ ký đề tài cho các thể loại kinh dị là đại diện vector các mô hình kinh dị trong phim, trong đó mỗi tính năng là một lời từ chủ đề danh sách chữ ký và mỗi giá trị là trọng lượng của từ.

Sau đó chúng ta tạo ra một đại diện vector cho mỗi bộ phim. Như vậy mỗi tính năng trong các vector là cosin tương đồng của bộ phim với một trong những bộ phim thể loại mô hình. Như vậy đại diện cho bộ phim “Alien” cho ta một ví dụ: H = 0,63 orror, Action = 0,73, Action / Adventure = 40, Hài / Lãng mạng = 0,12.

Sự giống nhau giữa bất kỳ hai bộ phim được tính bằng sự tương đồng cosin giữa 2 vector.

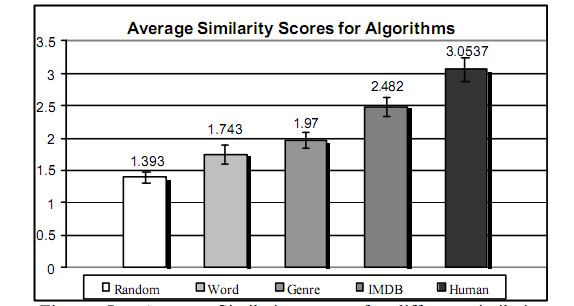
Trong khi chữ ký đề tài và các phân cấp nhóm chủ đề không được thực hiện trong thời gian thực, Chúng ta có thể tính ẩn một lần và sau đó các vector cơ quan đại diện của bộ phim có thể được lưu trữ. Bởi vì có là thể loại ít hơn từ được sử dụng. Tóm lại, tính thực tế của những tương đồng giữa phim sẽ làm cho thuật toán 2 nhanh hơn thuậ toán 1.

Kết Quả

Để kiểm tra chất lượng của các thuật toán. Chúng ta chọn ngẫu nhiên 10 phim từ một tập hợp con của các IMDB2. Sau đó chúng ta tính toán sự tương tự giữa 10 bộ phim sử dụng cả 2 thuật toán ở trên. Chúng ta tải về 10 bộ phim đã được gợi ý cho những bộ phim trên được IMDB post trên website. Chúng ta tiếp tục biên soạn mười ngẫu nhiên bộ phim được lựa chọn và 5 bộ phim được đánh giá bởi một con người là tương tự nhau cho mỗi mục tiêu. Để tăng khả năng rằng các bộ phim được lựa chọn sẽ nhìn thấy bởi các thẩm phán, chúng ta chỉ phim gồm các xếp hạng IMDB được dựa trên nhiều hơn 1000 phiếu bầu. Bảng sau cho thấy ví dụ Ofthe phim được lựa chọn bởi mỗi phương pháp cho một bộ phim mục tiêu: “French Kiss”.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ag.1 : Words | Ag.2: genres | Human | IMDB | Random |
| Man in the Moon | When Harry Met Sally | Sleepless in Seattle | Forget Paris | 10 Commandments |
| Wyatt Earp | Chasing Amy | When Harry Met Sally | Amer. Werewolf in Paris | Rambo |
| Story of Us | Shakespeare in Love | To Catch a Thief | Prelude to a Kiss | Naked Gun 2 ½ |
| Silverado | Anniversary Party | Forget Paris | Moulin Rouge! | All President’s Men |
| Always | Shop Around the Corner | Pretty Woman | Charade | Last of the Mohicans |

Chúng ta sau đó trình bày các danh mục để bốn người trong ngẫu nhiên trật và yêu cầu họ đánh giá như thế nào tương tự như mỗi bộ phim là để các mục tiêu trên thang điểm 1-4. Lưu ý rằng trả lời cho những cuộc khảo sát đã không hoàn thành; những người được hỏi làm bản án trên trung bình 51,2 % so sánh đó của họ không có luôn luôn được nhìn thấy nhau trong những bộ phim. Tỷ lệ này là rất cao xem xét các nhiệm vụ và chỉ ra các giám khảo trình độ chuyên môn. Kết quả khảo sát được thể hiện trong hình sau:



Hình 5: Điểm số về độ tương tự trung bình đối với các thuật toán

Kiểm tra t sinh viên giữa mỗi thuật toán chương trình sự khác biệt đáng kể tại p < 0,01 mức trừ giữa các thuật toán ngẫu nhiên và từ không gian, đó là ý nghĩa ở mức p < 0,05 độ. Theo dự kiến, ngẫu nhiên và phim tạo ra con người nhận được tồi tệ nhất và tốt nhất bản án tương tự, tương ứng. Chữ ký chủ đề thể loại dựa trên số liệu tương tự là tốt hơn đáng kể so với phương pháp tiếp cận từ ngẫu nhiên. Tuy nhiên, cả hai đều kém hiệu quả hơn so với các thuật toán được sử dụng bởi IMDB, mà lợi dụng các khuyến nghị người dùng và người dùng tạo khóa. Thật thú vị, con người tạo ra phim trung bình ít hơn so với số điểm tương đồng tối đa. Chỉ rõ tính chất cực kỳ chủ quan của nhiệm vụ.

# 3. Chương trình DEMO

* Thông tin về dữ liệu: Dữ liệu được lấy từ trang web IMDB.COM về 1682 bộ phim. Mỗi bộ phim gồm các trường dữ liệu về:
  + - Tên phim
    - Năm ra mắt
    - Mô tả về thể loại phim
    - URL về mô tả chi tiết của bộ phim
    - Điểm đánh giá về bộ phim đó
  + Chức năng chính:
    - Input: Danh sách một bộ phim yêu thích của USER
    - Ouput: Danh sách gợi ý các bộ phim khác dành cho USER đó
  + Hệ thống đưa ra gợi ý dựa trên:
    - Độ tương tự về thể loại với những bộ phim thuộc sở thích của USER.
    - Kế đến là điểm đánh giá về bộ phim đó.

🡪 Hai yếu tố này được liên hệ khá mật thiết với nhau trong quá trình xử lý của chương trình.

# TÀI LIỆU THAM KHẢO

1. Balabanovic, M., and Shoham, Y. Fab: Content based, collaborative recommendation. *Communications of the ACM,* 40(3), (1997), 66-72.

2. ***Fleischman***, M. *Text Classification in Reduced Dimensionsusing Topic Signatures and Hierarchical Topic Clustering*, Masters Project, University of Southern California, 2002.

3. ***Jurafsky, D. and Martin, J.****Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall, 2000.

4. ***Lin, C.-Y. and E.H. Hovy***. The Automated Acquisition of Topic Signatures for Text Summarization. Proc. of the *COLING Conference*. Strasbourg, France, 2000.

5. ***Melville, P., Mooney, R. J., and Nagarajan***, R. Contentboosted collaborative filtering for improved recommendations. Proc. of the *18th National Conference onArtificial Intelligence*, 2002.

6. ***Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S., Konstan, J.A., & Riedl, J***. Getting to Know You: Learning New User Preferences in Recommender Systems. Proc. Of the *2002 International Conference on Intelligent UserInterfaces*, San Francisco, CA, (2002), pp. 127-134.