TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**DỰ ÁN CUỐI KỲ**

**XỬ LÝ DỮ LIỆU LỚN**

**RECOMMENDER SYSTEM**

*Người hướng dẫn*: **GV. LÊ ANH CƯỜNG**

*Người thực hiện*: **NGUYỄN TUẤN ANH – 51800840**

**LÊ QUANG PHỤC – 51703160**

**LƯU NGUYÊN CHƯƠNG - 51603034**

*Khoá* **:22,21,20**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**DỰ ÁN CUỐI KỲ**

**XỬ LÝ DỮ LIỆU LỚN**

**RECOMMENDER SYSTEM**

*Người hướng dẫn*: **GV. LÊ ANH CƯỜNG**

*Người thực hiện*: **NGUYỄN TUẤN ANH – 51800840**

**LÊ QUANG PHỤC – 51703160**

**LƯU NGUYÊN CHƯƠNG - 51603034**

*Khoá* **:22,21,20**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2020**

# LỜI CẢM ƠN

Nhóm em xin gửi lời cảm ơn chân thành đến các thầy cô Khoa Công nghệ thông tin trường Đại học Tôn Đức Thắng.

Trong thời gian học tập và rèn luyện tại trường Đại học Tôn Đức Thắng, nhóm em đã nhận được rất nhiều sự giúp đỡ tận tình của quý thầy cô. Thầy cô là người đã hướng dẫn và truyền cảm hứng cho nhóm em trong học tập từ những ngày đầu tiên bước chân vào môi trường đại học. Bên cạnh đó, quý thầy cô còn truyền đạt cho nhóm em rất nhiều kiến thức hay và bổ ích, rèn luyện những kĩ năng hữu ích cho công việc trong tương lai. Với lòng biết ơn sâu sắc từ tận đáy lòng, em xin chân thành gửi lời cảm ơn đến quý thầy cô, đặc biệt là Thầy Lê Anh Cường – người đã trực tiếp dẫn dắt nhóm em trong môn Công nghệ thông tin trong Quản lý quan hệ khách hàng.

Trong quá trình làm báo cáo, nhóm em khó có thể tránh khỏi những sai sót, rất mong quý thầy cô có thể đóng góp ý kiến để nhóm em có được thêm nhiều bài học và rút kinh nghiệm cho lần sau hoàn thành tốt hơn.

Lời cuối cùng em xin chân thành cảm ơn và gửi lời chúc tốt đẹp nhất đến quý thầy cô đã tạo cơ hội cho em có thể trau dồi bản thân trong môn học này.

# ĐỒ ÁN ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

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

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

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

Tp. Hồ Chí Minh, ngày tháng năm

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Lê Quang Phục*

*Nguyễn Tuấn Anh*

*Lưu Nguyên Chương*

# PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

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

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

# TÓM TẮT

# MỤC LỤC

Contents

[LỜI CẢM ƠN i](#_Toc71725426)

[ĐỒ ÁN ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG ii](#_Toc71725427)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc71725428)

[TÓM TẮT iv](#_Toc71725429)

[MỤC LỤC 1](#_Toc71725430)

[I. Phân tích hệ thống khuyến nghị (recommender systems) 2](#_Toc71725431)

[1.1. Trình bày về bài toán 2](#_Toc71725432)

[1.2. Dữ bài về dữ liệu 3](#_Toc71725433)

[1.2.1. Về book.csv 4](#_Toc71725434)

[1.2.2. Về ratings.csv 4](#_Toc71725435)

[1.3. Mô hình bài toán 5](#_Toc71725436)

[1.3.1. Mô hình phổ biến 5](#_Toc71725437)

[1.3.2. Mô hình lọc dữ liệu nội dung 5](#_Toc71725438)

[1.3.3. Mô hình lọc công tác 6](#_Toc71725439)

[1.3.4. Dữ liệu hóa ma trận 6](#_Toc71725440)

[1.3.5. Mô hình kết hợp 7](#_Toc71725441)

[1.4. Trình bài các subtask 7](#_Toc71725442)

[1.5. Matrix Fractorization 7](#_Toc71725443)

# Phân tích hệ thống khuyến nghị (recommender systems)

## Trình bày về bài toán

Mục tiêu hệ thống khuyến nghị là giới thiệu các mặt hàng có liên quan cho người dùng, dựa trên sở thích của họ. Mức độ ưa thích và mức độ liên quan là chủ quan và chúng thường được suy ra bởi các mặt hàng mà người dùng đã sử dụng trước đó.

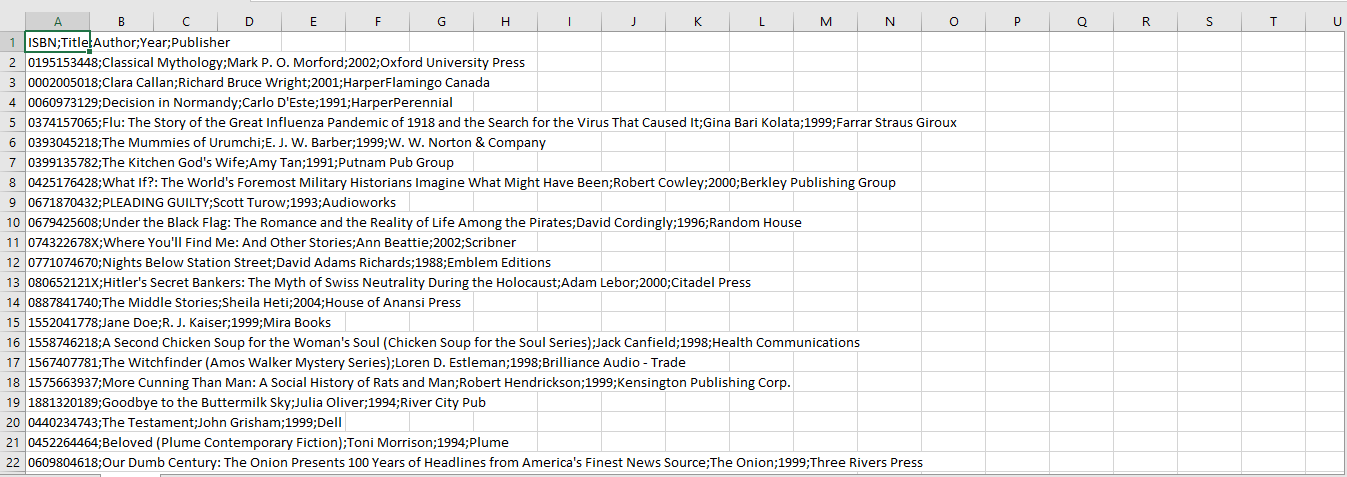
Các phương thức cho recommender systems:

* Collaborative filtering (lọc cộng tác): phương pháp này đưa ra dự đoán tự động (lọc) về sở thích của người dùng bằng cách thu thập thông tin sở thích hoặc sở thích từ nhiều người dùng (cộng tác). Giả thiết cơ bản của phương pháp lọc cộng tác là nếu một người A có cùng quan điểm với người B về một tập hợp các mục, thì A có nhiều khả năng có ý kiến của B đối với một mục nhất định hơn là ý kiến của một người được chọn ngẫu nhiên.
* Content-Based Filtering (lọc dựa trên nội dung): Phương pháp này chỉ sử dụng thông tin về mô tả và thuộc tính của các mặt hàng mà người dùng đã sử dụng trước đó để lập mô hình tùy chọn của người dùng. Nói cách khác, các thuật toán này cố gắng đề xuất các mục tương tự như những mục mà người dùng đã thích trong quá khứ (hoặc đang kiểm tra ở hiện tại). Đặc biệt, các items khác nhau được so sánh với các mục được người dùng đánh giá trước đó và các items phù hợp nhất được đề xuất.
* Hybrid method (phương pháp kết hợp): Nghiên cứu gần đây đã chứng minh rằng phương pháp kết hợp, kết hợp lọc cộng tác và lọc dựa trên nội dung có thể hiệu quả hơn phương pháp thuần túy trong một số trường hợp. Các phương pháp này cũng có thể được sử dụng để khắc phục một số vấn đề thường gặp trong hệ thống khuyến cáo như khởi động nguội và vấn đề thưa thớt.

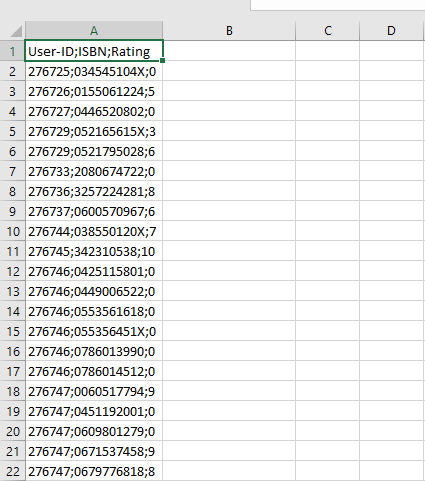
## Dữ bài về dữ liệu

Dùng 2 tệp dữ liệu book.csv và ratings.csv

**Book.csv:**



**ratings.csv:**



### Về book.csv

Chứa thông tin về các dữ liệu sách được chia sẻ trong nền tảng. Mỗi loại sách có ISBN (Số sách tiêu chuẩn quốc tế), Title (Tên Sách), Author (Tác Giả), Year (Năm xuất bản), Publisher (Nhà Xuất Bản).

### Về ratings.csv

Chứa nhật ký các tương tác của người dùng được chia sẻ.

Các giá trị eventType:

* User-ID: ID người dùng
* ISBN: Số sách tiêu chuẩn quốc tế
* Ratings: Đánh giá sách

## Mô hình bài toán

### Mô hình phổ biến

Một cách tiếp cận cơ sở phổ biến (và thường khó đánh bại) là mô hình Mức độ phổ biến. Mô hình này không thực sự được cá nhân hóa - nó chỉ đơn giản là đề xuất cho người dùng những mặt hàng phổ biến nhất mà người dùng chưa tiêu thụ trước đó.

Vì sự phổ biến chiếm "sự khôn ngoan của đám đông", nó thường cung cấp các đề xuất tốt, thường là thú vị đối với hầu hết mọi người.

Mục tiêu chính của hệ thống giới thiệu là tận dụng các mặt hàng lâu dài cho những người dùng có sở thích rất cụ thể, điều này vượt xa kỹ thuật đơn giản này.

### Mô hình lọc dữ liệu nội dung

Phương pháp lọc dựa trên nội dung tận dụng mô tả hoặc thuộc tính từ các mặt hàng mà người dùng đã tương tác để đề xuất các mặt hàng tương tự. Nó chỉ phụ thuộc vào những lựa chọn trước đó của người dùng, làm cho phương pháp này trở nên mạnh mẽ để tránh sự cố khởi động nguội.

Đối với các mục văn bản, như bài báo, tin tức và sách, thật đơn giản khi sử dụng văn bản thô để tạo hồ sơ mục và hồ sơ người dùng.

Ở đây sử dụng một kỹ thuật rất phổ biến trong truy xuất thông tin (công cụ tìm kiếm) tên là TF-IDF. Kỹ thuật này chuyển đổi văn bản không có cấu trúc thành cấu trúc vectơ, trong đó mỗi từ được thể hiện bằng một vị trí trong vectơ và giá trị đo mức độ liên quan của một từ nhất định đối với một bài báo. Vì tất cả các mục sẽ được thể hiện trong cùng một Mô hình không gian vectơ, nên nó phải tính toán sự tương đồng giữa các bài báo.

Để lập mô hình hồ sơ người dùng, lấy tất cả các hồ sơ mặt hàng mà người dùng đã tương tác và tính trung bình chúng. Mức trung bình được tính theo độ mạnh tương tác, nói cách khác, các bài viết mà người dùng đã tương tác nhiều nhất (ví dụ: đã thích hoặc đã nhận xét) sẽ có độ mạnh cao hơn trong hồ sơ người dùng cuối cùng.

### Mô hình lọc công tác

Có hai chiến lược chiến khải:

* Dựa trên bộ nhớ: Cách tiếp cận này sử dụng bộ nhớ của các tương tác của người dùng trước đó để tính toán các điểm tương đồng của người dùng dựa trên các mục mà họ đã tương tác (cách tiếp cận dựa trên người dùng) hoặc tính các điểm tương đồng dựa trên những người dùng đã tương tác với họ (cách tiếp cận dựa trên mục).
* Dựa trên mô hình: Cách tiếp cận này, các mô hình được phát triển bằng cách sử dụng các thuật toán học máy khác nhau để giới thiệu các mục cho người dùng. Có nhiều thuật toán CF dựa trên mô hình, như mạng nơ-ron, mạng bayesian, mô hình phân cụm và mô hình yếu tố tiềm ẩn như Phân tích giá trị số ít (SVD) và phân tích ngữ nghĩa tiềm ẩn theo xác suất.

### Dữ liệu hóa ma trận

Mô hình yếu tố tiềm ẩn nén ma trận mục người dùng thành một biểu diễn chiều thấp về yếu tố tiềm ẩn. Một lợi thế của việc sử dụng phương pháp này là thay vì có một ma trận chiều cao chứa nhiều giá trị bị thiếu, chúng ta sẽ xử lý một ma trận nhỏ hơn nhiều trong không gian chiều thấp hơn.

Có một số lợi thế với mô hình này. Nó xử lý độ thưa thớt của ma trận gốc tốt hơn so với ma trận dựa trên bộ nhớ. Ngoài ra, so sánh độ giống nhau trên ma trận kết quả có khả năng mở rộng hơn nhiều, đặc biệt là trong việc xử lý các bộ dữ liệu thưa thớt lớn.

Một quyết định quan trọng là số lượng yếu tố ảnh hưởng đến ma trận người dùng-mục. Số lượng yếu tố càng cao thì việc phân tích nhân tử trong tái tạo ma trận ban đầu càng chính xác. Do đó, nếu mô hình được phép ghi nhớ quá nhiều chi tiết của ma trận ban đầu, nó có thể không tổng quát hóa tốt cho dữ liệu mà nó không được đào tạo. Giảm số lượng yếu tố làm tăng khả năng tổng quát hóa của mô hình.

### Mô hình kết hợp

Điều gì sẽ xảy ra nếu chúng ta kết hợp các phương pháp Lọc cộng tác và Lọc dựa trên nội dung?

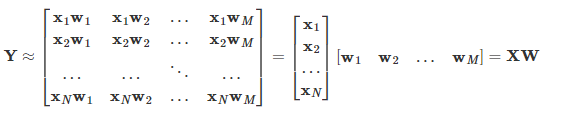
Điều đó có cung cấp cho chúng tôi các đề xuất chính xác hơn không?

Trên thực tế, các phương pháp kết hợp đã hoạt động tốt hơn các phương pháp tiếp cận riêng lẻ trong nhiều nghiên cứu và được các nhà nghiên cứu và thực hành sử dụng rộng rãi.

Hãy xây dựng một phương pháp kết hợp đơn giản, dưới dạng một tập hợp lấy trung bình có trọng số của điểm CF chuẩn hóa với điểm Dựa trên nội dung và xếp hạng theo điểm kết quả.

## Trình bài các subtask

## Matrix Fractorization



Khi đó mỗi dòng của ma trận X đại diện cho một véc tơ nhân tố ẩn của một item, đó là những nhân tố bất kì, rất trừu tượng mà chúng ta không nên đặt tên cho chúng. Mỗi cột của ma trận W đại diện cho một véc tơ các hệ số thể hiện mức độ yêu thích của user đối với các nhân tố ẩn. Số lượng nhân tố ẩn thông thường là một số có giá trị rất nhỏ so với số lượng user và item nên dung lượng cần lưu trữ đối với 2 ma trận X và W là rất nhỏ so với lưu trữ toàn bộ ma trận.

Sau khi tìm được các ma trận items X và ma trận users W, giá trị ước lượng rating của một user j lên một item i sẽ chính bằng tích:

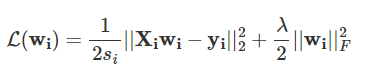


Như vậy giá trị dự báo được tính toán đơn giản hơn so với Neighborhood-based collaborative Filtering vì chỉ cần thực hiện phép nhân véc tơ mà không cần phải cộng với trung bình cột để chuyển về giá trị gốc.

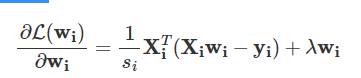
Qúa trình dự báo hệ số cho mô hình hồi qui của mỗi user tương tự như phương pháp content-based filtering. Nhưng có sự kết hợp giữa tìm nghiệm tối ưu của ma trận items và ma trận users. Qúa trình này được thực hiện xen kẽ nhau nên không chỉ tận dụng được các thông tin là đầu vào của users mà còn tận dụng được sự giống nhau trong sở thích của các users. Chính vì thế phương pháp mới được xếp vào nhóm collaborative filtering.

**Thuật Toán:**

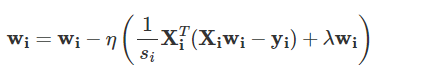
Về bản chất bài toán của chúng ta có thể chia thành N bài toán nhỏ và mỗi bài toán tương ứng với đi tìm nghiệm tối ưu cho một user. Khi đó phương trình mất mát trên một user sẽ là:



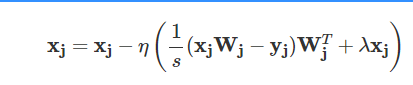
Đạo hàm đối với một user sẽ là:



Một quá trình cập nhật nghiệm theo gradient descent sẽ được thực hiện như sau:

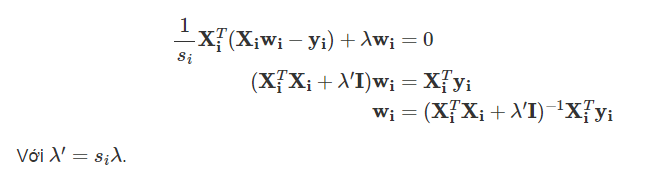


Hoàn toàn tương tự ta cũng suy ra phương trình cập nhật nghiệm theo gradient descent trên xj:



Qúa trình huấn luyện theo gradient descent sẽ khá lâu vì tốc độ hội tụ phụ thuộc vào learning rate.

Trong khi ta nhận thấy hàm loss function trên mỗi user hoặc item có thể tính toán được nghiệm tối ưu thông qua giải phương trình đạo hàm. Qúa trình hội tụ xen kẽ X và W khi tính theo phương trình nghiệm sẽ nhanh hơn rất nhiều so với gradient descent. Phương pháp này có tên là ALS (Alternating Least Square). Giải phương trình đạo hàm đối với user thứ i:



Tương tự, nghiệm tối ưu đối với item thứ j:



Chúng ta sẽ thực hiện vòng lặp xen kẽ như sau:

