BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC HUẾ
TRƯỜNG ĐẠI HỌC KHOA HỌC

NGUYỄN THỊ PHƯỢNG

NGHIÊN CỨU KỸ THUẬT LỌC CỘNG TÁC VÀ ỨNG DỤNG XÂY DỰNG HỆ THỐNG GỢI Ý BÁN SÁCH TRỰC TUYẾN

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH
MÃ SỐ: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC
ĐỊNH HƯỚNG NGHIÊN CỨU

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGS.TS. LÊ MẠNH THẠNH

Thừa Thiên Huế, 2016

LỜI CAM ĐOAN

Tôi xin cam đoan đây là kết quả nghiên cứu của riêng cá nhân tôi. Các số liệu, kết quả trình bày trong luận văn là trung thực. Tất cả các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin hoàn toàn chịu trách nhiệm theo qui định cho lời cam đoan của mình.

Huế, ngày tháng năm 2016

Người cam đoan

Nguyễn Thị Phượng

LỜI CẢM ƠN

Trong quá trình học tập chương trình cao học Khoa Học Máy Tính tại trường Đại học Khoa Học – Đại học Huế và đặc biệt là trong quá trong làm luận văn tốt nghiệp của mình, tôi đã nhận được sự quan tâm, giúp đỡ rất nhiều của các thầy cô, gia đình, bạn bè và các cơ quan chuyên môn.

Trước hết, tôi xin bày tỏ lòng biết ơn sâu sắc đến thầy giáo PGS.TS. Lê Mạnh Thạnh – người đã trực tiếp hướng dẫn tôi trong quá trình thực hiện luận văn. Bằng sự tận tâm và nhiệt huyết của mình, thầy đã giúp tôi rất nhiều để bản thân vượt qua những khó khăn nhất định và hoàn thành luận văn đúng thời hạn.

Xin cùng được bày tỏ lòng biết ơn chân thành tới quý thầy giáo, cô giáo – những người đã mang lại cho tôi nhiều kiến thức hay về Khoa học Máy Tính cũng như các lĩnh vực khác và các kiến thức bổ trợ khác cho luận văn tốt nghiệp của mình trong suốt hai năm qua.

Nhân đây, tôi cũng muốn gửi lời cảm ơn chân thành đến Ban Giám Hiệu nhà trường, phòng Đào Tạo Sau Đại học, Ban Chủ Nhiệm Khoa khoa Công Nghệ Thông Tin đã tạo nhiều điều kiện cho chúng tôi trong quá trình học tập.

Cuối cùng, tôi xin gửi lời tri ân sâu sắc đến gia đình, bạn bè và người thân đã luôn luôn động viên và khuyến khích tôi trong quá trình học tập và thực hiện luận văn của mình.

Huế, ngày tháng năm 2016

Nguyễn Thị Phượng

MỤC LỤC

DANH MỤC CÁC BẢNG

Trang

DANH MỤC CÁC HÌNH

Trang

DANH MỤC CÁC CHỮ VIẾT TẮT

BFD	Sơ đồ chức năng kinh doanh (Business Function Diagram)
CF	Lọc cộng tác (Collaborative Filtering)
DFD	Mô hình luồng dữ liệu (Data flow Diagram)
IPTV	Truyền hình giao thức Internet (Internet Protocol Television)
IR	Lọc thông tin (Information Filtering)
NN	Láng giềng gần nhất (Nearest neighbors)
TT	Thông tin

MỞ ĐẦU

1. LÝ DO CHỌN ĐỀ TÀI

Trong thời đại công nghệ thông tin hiện nay, internet với các tiện ích của nó đang có ảnh hưởng lớn đối với đại bộ phận người sử dụng mạng. Với hàng triệu thông tin được đưa lên internet mỗi ngày, trong nhiều trường hợp người dùng cần đưa ra các lựa chọn dựa trên những ý kiến hay lời khuyên của mọi người xung quanh, có thể qua lời nói, các bản đánh giá sản phẩm, khảo sát thị trường, thư giới thiệu… điều này dẫn tới yêu cầu phải có các phương pháp tự động thu thập thông tin và đưa ra lời khuyên để hỗ trợ cho các phương pháp truyền thống trên, người dùng cần có sự gợi ý kịp thời để có thể tìm kiếm thông tin một cách chính xác và tiết kiệm tối đa thời gian, một khi dữ liệu càng lớn thì sự gợi ý càng có vai trò quan trọng. Hệ thống gới ý (Recommender System) là một giải pháp như vậy. Hệ thống này đưa ra gợi ý, đưa ra những mục thông tin phù hợp cho người dùng bằng cách dựa vào dữ liệu về hành vi đã thực hiện trong quá khứ của họ để dự đoán những mục thông tin mới trong tương lai mà người dùng có thể thích, hoặc dựa trên tổng hợp ý kiến của những người dùng khác. Hệ thống gợi ý đã trở thành một ứng dụng quan trọng và thu hút được sự quan tâm lớn của các nhà nghiên cứu cũng như các doanh nghiệp kinh doanh lớn qua mạng.

Trong hệ thống gợi ý, lọc cộng tác là một kỹ thuật được dùng để đánh giá độ quan tâm của người dùng trên sản phẩm mới. Kỹ thuật này được áp dụng thành công trong nhiều ứng dụng. Trong các hệ thống lọc cộng tác, sở thích của người dùng trên các sản phẩm mới được dự đoán dựa trên dữ liệu về sở thích của người dùng – sản phẩm (hoặc đánh giá của người dùng trên sản phẩm) trong quá khứ. Nó có thể xem như là một hệ gợi ý tự động bằng cách dựa trên sự tương tự giữa những người dùng hoặc giữa những sản phẩm trong hệ thống và đưa ra dự đoán sự quan tâm của người dùng tới một sản phẩm, hoặc đưa ra gợi ý một sản phẩm mới cho người dùng nào đó.

Hệ thống gợi ý thực sự cần thiết cho một website mua bán hàng hóa với số lượng hàng hóa khổng lồ, số lượng chủng loại mặt hàng lớn cùng với vô số thông tin về mặt hàng để giúp khách hàng nắm bắt thông tin mà họ tìm kiếm. Hệ thống có thể đưa ra những mục thông tin phù hợp cho người dùng, giúp người dùng dễ dàng lựa chọn những sản phẩm phù hợp với họ nhất.

Do đó, tôi thực hiện đề tài "Nghiên cứu kỹ thuật lọc cộng tác và ứng dụng xây dựng hệ thống gợi ý bán sách trực tuyến" với mục tiêu nghiên cứu lý thuyết về hệ gợi ý, các kỹ thuật của hệ gợi ý, đặc biệt là phương pháp lọc cộng tác và kỹ thuật láng giềng thuộc phương pháp lọc cộng tác. Tiếp đến, đề tài tập trung xây dựng website gợi ý sách sử dụng kỹ thuật láng giềng của phương pháp lọc cộng tác, phân tích, đánh giá hiệu quả của việc ứng dụng hệ gợi ý trong việc triển khai xây dựng website.

2. TỔNG QUAN TÀI LIỆU

Hiện nay, đã có khá nhiều bài viết nghiên cứu về hệ thống gợi ý cũng như việc sử dụng hệ gợi ý cho lọc cộng tác, chẳng hạn như trong nước có bài báo: "Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc cộng tác" (Nguyễn Hùng Dũng, Nguyễn Thái Nghe) [1] nói về giải thuật lọc cộng tác và việc tích hợp giải thuật lọc cộng tác vào hệ thống bán hàng trực tuyến.

Trên thế giới cũng có nhiều nghiên cứu nói về vấn đề này, như bài báo "Recommender Systems" (Prem Melville and Vikas Sindhwani) [5] nói về định nghĩa, cấu trúc cũng như các phương pháp của hệ thống gợi ý; Các bài báo "Recommendation System Based on Collaborative Filtering" (Zheng Wen) [2]; "Item-based Collaborative Filtering Recommendation Algorithms" (Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl) [3], "Collaborative Filtering Recommender Systems" (J. Ben Schafer, Dan Freankowski, Jon Herlocker, and Shilad Sen) [4]; "Collaborative Filtering Recommender Systems" (Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan) [6] tập trung nói về các khái niệm cơ bản, chức năng của các phương pháp của hệ thống gợi ý, đặc biệt là phương

pháp lọc cộng tác và kỹ thuật láng giềng, đánh giá hệ thống, những thách thức cũng như hướng phát triển trong tương lai... "Learning Collaborative Filtering and Its Application to People to People Recommendation in Social Networks" (Xiongcai Cai, Michael Bain, Alfred Krzywicki, Wayne Wobcke, Yang Sok Kim, Paul Compton and Ashesh Mahidadia) [7] nói về mô hình đóng góp người dùng tương tự và ứng dụng trong mạng xã hội…

Mặc dù vậy, việc xây dựng một hệ thống gợi ý hoàn chỉnh, có tích hợp giải thuật lọc cộng tác để gợi ý cho người dùng vẫn chưa được quan tâm.

3. MỤC TIÊU NGHIÊN CỨU

- Nghiên cứu kỹ thuật lọc cộng tác và kỹ thuật láng giềng thuộc phương pháp lọc cộng tác.

- Xây dựng website gợi ý sách sử dụng kỹ thuật lọc cộng tác.

4. ĐỐI TƯỢNG NGHIÊN CỨU

- Kỹ thật lọc cộng tác và kỹ thuật láng giềng thuộc phương pháp lọc cộng tác.

- Ứng dụng kỹ thuật lọc cộng tác trong xây dựng website.

5. PHƯƠNG PHÁP NGHIÊN CỨU

5.1. Nghiên cứu lý thuyết

- Nghiên cứu cơ sở lý thuyết về hệ gợi ý, kỹ thuật lọc cộng tác.

- Nghiên cứu, phân tích các tài liệu tiếng Việt và quốc tế liên quan đến ứng dụng lọc cộng tác trong xây dựng hệ thống bán hàng trực tuyến.

5.2. Nghiên cứu thực nghiệm

- Thu thập dữ liệu thực tế.

- Xây dựng hệ thống gợi ý bán hàng trực tuyến.

6. PHẠM VI NGHIÊN CỨU

Nghiên cứu tổng quan về lý thuyết gợi ý, cơ sở lý thuyết kỹ thuật lọc cộng tác và cơ sở dữ liệu thực nghiệm để xây dựng hệ thống gợi ý bán sách trực tuyến.

7. CẤU TRÚC LUẬN VĂN

Chương 1 trình bày tổng quan về hệ gợi ý – Recommender System. Chương này sẽ giới thiệu tổng quan về hệ gợi ý, các chức năng, dữ liệu và các nguồn kiến thức, các phương pháp và các ứng dụng, đánh giá của hệ gợi ý.

Chương 2 trình bày phương pháp lọc cộng tác và kỹ thuật láng giềng (Neighborhood-based). Chương này đi vào tìm hiểu sâu về phương pháp lọc cộng tác và kỹ thuật láng giềng (Neighborhood-based) thuộc phương pháp lọc cộng tác.

Chương 3 trình bày về xây dựng hệ thống bán sách trực tuyến dựa trên kỹ thuật láng giềng của phương pháp lọc cộng tác. Nội dung chương này đi vào phân tích hệ gợi ý được sử dụng trong luận văn, phân tích và thiết kế hệ thống, các kỹ thuật ứng dụng trong luận văn, giới thiệu demo chương trình, rút ra các kết luận.

Chương 1. TỔNG QUAN VỀ HỆ THỐNG GỢI Ý

1.1. GIỚI THIỆU VỀ HỆ THỐNG GỢI Ý

Hệ thống gợi ý là hệ thống bao gồm các kỹ thuật và công cụ phần mềm nhằm đưa ra những gợi ý cho người sử dụng, đáp ứng nhu cầu của họ về một sản phẩm, dịch vụ nào đó trên Internet. Những gợi ý được cung cấp nhằm hỗ trợ người sử dụng đưa ra quyết định lựa chọn những sản phẩm, dịch vụ phù hợp với nhu cầu và thị hiếu của mình, chẳng hạn như: Mua sản phẩm nào, nghe thể loại nhạc gì hay tin tức trực tuyến nào nên đọc,...

Trong những năm gần đây, hệ thống gợi ý là một phương tiện có giá trị để giải quyết với vấn đề quá tải thông tin. Đích cuối cùng mà hệ thống gợi ý muốn hướng tới là hướng dẫn cho một người dùng mới về các sản phẩm chưa hoặc không được xem trước đó nhưng lại có liên quan đến tác vụ hiện hành của người dùng. Theo yêu cầu của người dùng, nó có thể được khớp nối hay không phụ thuộc vào phương pháp tiếp cận gợi ý theo bối cảnh và nhu cầu người dùng. Hệ thống gợi ý đưa ra các gợi ý sử dụng các biến thể khác nhau của kiến thức và dữ liệu người dùng, các sản phẩm có sẵn và các giao dịch trước đó được lưu trong cơ sở dữ liệu tùy biến. Sau đó người dùng có thể duyệt các gợi ý: Chấp nhận hay không và ngay lập tức đưa ra thông tin phản hồi ngầm hay rõ ràng ở giai đoạn tiếp theo. Tất cả những hành động và phản hồi của người dùng được lưu trữ trong cơ sở dữ liệu và có thể được sử dụng để đưa ra các gợi ý mới trong sự tương tác với người sử dụng hệ thống tiếp theo.

Một vài ứng dụng nổi tiếng về hệ thống gợi ý như: Gợi ý sản phẩm Amazon, hệ gợi ý phim của NetFlix. Hệ thống gợi ý đã chứng minh được ý nghĩa to lớn trong việc giúp người sử dụng trực tuyến giải quyết với tình trạng quá tải thông tin. Chính vì vậy, hệ thống gợi ý trở thành một trong những công cụ mạnh mẽ và phổ biến trong thương mại điện tử và trên nhiều lĩnh vực khác.

Hình 1.1. Hệ gợi ý của trang web Amazon.com.

Trong hầu hết các trường hợp, bài toán gợi ý được coi là bài toán ước lượng xếp hạng (Rating) của các Sản phẩm (Phim, cd, nhà hàng . . .) chưa được người dùng xem xét. Việc ước lượng này thường dựa trên những đánh giá đã có của chính người dùng đó hoặc từ những người dùng khác. Những Sản phẩm có xếp hạng cao nhất sẽ được dùng để gợi ý. Từ đó người dùng có những lựa chọn thích hợp với nhu cầu và thị hiếu của mình.

Một cách hình thức, bài toán gợi ý được mô tả như sau:

Gọi U là tập các người dùng (Users) của hệ thống.

Gọi I là toàn bộ không gian đối tượng sản phẩm (Items).

Hàm r(u,i) là đánh giá (độ phù hợp) của người dùng u với sản phẩm i.

Vậy bài toán là sự ánh xạ r: UxI → R. Trong đó R chính là tập hợp các đối tượng được đưa ra giới thiệu.

Tập R sẽ được sắp xếp theo thứ tự giảm dần của r. Công việc chính của giải thuật là đi tìm giá trị hàm r=f(u, i), với r lớn nhất là sản phẩm i được người dùng u ưa thích nhất.

1.2. CHỨC NĂNG CỦA HỆ THỐNG GỢI Ý

Trước hết, chúng ta phải phân biệt giữa vai trò hệ gợi ý của nhà cung cấp so với vai trò hệ gợi ý của người sử dụng. Ví dụ, một hệ thống gợi ý du lịch thường được giới thiệu bởi một trung gian du lịch hoặc một tổ chức quản lý để tăng doanh thu của nó qua việc cho thuê phòng khách sạn nhiều hơn hoặc để tăng số lượng khách du lịch. Trong khi đó, động cơ của người sử dụng khi truy cập vào hai hệ thống là tìm một khách sạn phù hợp với nhu cầu, túi tiền cùng các sự kiện thú vị/các điểm hấp dẫn khi đến thăm một điểm đến.

Dưới đây là một số chức năng của hệ thống.

- **Đối với nhà cung cấp:**

- *Tăng số lượng các mặt hàng bán ra cho các hệ thống thương mại điện tử:* Đây có lẽ là chức năng quan trọng nhất của hệ thống gợi ý. Thay vì người dùng chỉ mua một sản phẩm mà họ cần, họ được gợi ý mua những sản phẩm 'có thể họ cũng quan tâm' mà bản thân họ không nhận ra. Hệ thống gợi ý tìm ra những 'mối quan tâm ẩn'. Bằng cách đó, hệ thống gợi ý làm gia tăng nhu cầu của người dùng và gia tăng số lượng mặt hàng bán ra.Tương tự đối với các hệ thống phi thương mại (Như các trang báo), hệ thống gợi ý sẽ giúp người dùng tiếp cận với nhiều đối tượng thông tin mang tính đa chiều và được nhiều người quan tâm hơn.

- *Bán các mặt hàng đa dạng hơn trên các hệ thống thương mại điện tử:* Đây là chức năng quan trọng thứ hai của hệ thống gợi ý. Hầu hết các hệ thống thương mại đều có các mặt hàng hết sức là đa dạng và phong phú. Khi nắm bắt được nhu cầu của người dùng, hệ thống gợi ý dễ dàng mang đến sự đa dạng trong sự lựa chọn hàng hóa. Từ đó đòi hỏi các hệ thống thương mại điện tử cung cấp nhiều mặt hàng đa dạng và phù hợp với người sử dụng hơn. Ví dụ, trong một hệ gợi ý phim như Netflix, các nhà cung cấp dịch vụ quan tâm đến việc cho thuê tất cả các đĩa DVD trong danh mục, không chỉ các phim phổ biến nhất. Điều này có thể là khó khăn nếu như không có một hệ gợi ý gợi ý, các nhà cung cấp dịch vụ có thể gặp rủi ro nếu như quảng cáo mà

không để ý đến việc phim có phù hợp với sở thích của một người dùng cụ thể nào đó không. Dó đó, hệ gợi ý sẽ là một gợi ý hay để quảng cáo cho loại phim không phổ biến cho người sử dụng.

- *Tăng sự hài lòng người dùng:* Vai trò chủ đạo của hệ thống gợi ý là hiểu nhu cầu của người dùng, gợi ý cho họ những thứ họ cần... Người dùng sẽ tìm thấy các gợi ý thú vị, có hiệu quả, chính xác, gợi ý kịp thời và một giao diện đẹp có thể tối ưu việc sử dụng và làm tăng sự hài lòng của người dùng trong hệ thống. Chính vì vậy hệ thống gợi ý tăng sự hài lòng của người dùng trên hệ thống và lựa chọn ưu tiên khi họ có những băn khoăn hoặc khi chưa có kiến thức về sản phẩm.

- *Tăng độ tin cậy, độ trung thực của người dùng:* Một khi hệ thống gợi ý cho người dùng những lựa chọn và họ hài lòng về những gợi ý đó thì lòng tin của họ đối với hệ thống (Nơi mà giúp họ tìm ra những thứ họ thực sự quan tâm) được nâng lên một cách đáng kể. Đây thật sự là một điều thích thú và thu hút người dùng. Có một điểm quan trọng là hệ thống gợi ý hoạt động dựa trên những xếp hạng thật từ chính bản thân người dùng trong quá khứ. Do đó, khi người dùng càng tin cậy vào hệ thống, đưa ra những đánh giá trung thực cho các sản phẩm, hệ thống sẽ mang lại cho người dùng nhiều gợi ý chính xác hơn, phù hợp với nhu cầu, sở thích của họ.

- *Hiểu rõ hơn về những gì người dùng muốn:* Đây là một chức năng quan trọng khác của hệ thống gợi ý được thừa kế từ nhiều ứng dụng khác nhau là thu thập hoặc dự đoán sở thích người dùng thông qua hệ thống. Điều này giúp cho các nhà phát triển dịch vụ có thể quyết định tái sử dụng các sản phẩm theo mục tiêu cải thiện quản lý cửa hàng hoặc tiến hành sản xuất.

- **Đối với người sử dụng:**

- *Tìm ra một số sản phẩm tốt nhất*: Hệ thống gợi ý tới người dùng một số sản phẩm được xếp hạng và dự đoán số người dùng khác thích chúng. Đây là chức năng chính mà nhiều hệ thống thương mại điện tử sử dụng.

- *Tìm ra tất cả sản phẩm tốt*: Gợi ý tất cả sản phẩm mà có thể làm hài lòng nhu cầu của khách hàng. Trong nhiều trường hợp không đủ cơ sở để đưa ra các sản phẩm tốt nhất. Điều này chỉ đúng khi số lượng sản phẩm liên quan tương đối nhỏ hoặc khi hệ gợi ý là chức năng quan trọng trong ứng dụng tài chính và y tế.

- *Gợi ý liên tục*: Thay vì tập trung vào tạo gợi ý đơn, các hệ thống gợi ý tạo các gợi ý liên tục tới người dùng cho tới khi họ tìm được sản phẩm mong muốn.

- *Gợi ý một nhóm sản phẩm*: Đề xuất một nhóm các sản phẩm mà tương đương nhau. Ví dụ như kế hoạch du lịch có thể là gồm nhiều điểm đến, các dịch vụ nơi ở, các sự kiện hấp dẫn. Từ quan điểm của người dùng những lựa chọn khác nhau có thể được xem xét và được lựa chọn một điểm đến du lịch hợp lý.

- *Chỉ duyệt tìm*: Trong tác vụ này, người dùng duyệt các danh mục mà không có ý định mua sản phẩm nào, tác vụ này đưa ra gợi ý giúp người dùng duyệt tìm các sản phẩm có nhiều khả năng thuộc vào phạm vi sở thích của người dùng với phiêm truy cập xác định. Đây là tác vụ được hỗ trợ bởi các kỹ thuật đa phương tiện.

- *Tìm kiếm các gợi ý tin tưởng*: Một số người dùng không tin tưởng vào các hệ thống gợi ý, họ tham gia vào hệ thống để thấy được các hệ thống này đưa ra gợi ý tốt tới mức nào. Do đó, một số hệ thống có thể đưa ra các chức năng chính xác để cho phép họ thử nghiệm hành vi của họ, ngoài các yêu cầu gợi ý.

- *Cải thiện hồ sơ cá nhân người dùng*: Người dùng có khả năng cung cấp thông tin, những gì họ thích, không thích với hệ thống gợi ý. Điều này là hết sức cần thiết để đưa ra các gợi ý mang tính chất cá nhân hóa. Nếu như hệ thống không xác định tri thức về người dùng đang hoạt động thì nó chỉ có thể đưa ra các gợi ý giống nhau.

- *Bày tỏ ý kiến của mình*: Một số người dùng có thể không quan tâm tới các gợi ý, đúng hơn, những gì quan trọng với họ là được góp ý kiến, đánh giá về sản phẩm, giúp ích người khác khi lựa chọn sản phẩm này.

- *Tác động tới những người dùng khác*: Trong hệ gợi ý trên web, có nhiều người tham gia với mục tiêu của họ là tác động tới hệ gợi ý, dẫn tới ảnh hưởng tới người dùng khác khi mua một sản phẩm cụ thể (Thông qua đánh giá sản phẩm,…). Tác động của họ có thể thúc đẩy hoặc gây bất lợi cho sản phẩm.

1.3. DỮ LIỆU VÀ CÁC NGUỒN TRI THỨC

Hệ gợi ý là hệ thống xử lý thông tin thu thập từ các loại dữ liệu khác nhau để xây dựng các gợi ý. Dữ liệu chủ yếu là về các mặt hàng cần gợi ý và người dùng sẽ nhận được các gợi ý này. Tuy nhiên, dữ liệu và các nguồn tri thức sẵn có cho các hệ thống gợi ý có thể rất đa dạng. Trong bất kỳ trường hợp nào, dữ liệu được sử dụng bởi hệ gợi ý thuộc ba loại: sản phẩm (Item), người sử dụng (User), và các giao dịch (Transactions), đó chính là quan hệ giữa người sử dụng và các mặt hàng.

Sản phẩm (Item): Sản phẩm là các đối tượng được gợi ý. Các sản phẩm này đặc trưng bởi tiện ích và giá trị của nó. Giá trị của một sản phẩm có thể là tích cực nếu hữu ích cho người sử dụng, hoặc tiêu cực nếu sản phẩm không phù hợp với người sử dụng.

Sản phẩm có giá trị thấp là: tin tức, các trang web, sách, đĩa CD, phim. Sản phẩm có giá trị lớn hơn là: máy ảnh kỹ thuật số, điện thoại di động, máy tính cá nhân,… . Các sản phẩm phức tạp nhất như là những chính sách bảo hiểm, tài chính đầu tư, gợi ý du lịch, công việc… Hệ gợi ý có thể sử dụng một loạt các thuộc tính và các tính năng của các sản phẩm.

Ví dụ trong một hệ thống gợi ý phim, thể loại (Hài, kinh dị,...) cũng như tên tuổi các đạo diễn và diễn viên có thể được sử dụng để mô tả một bộ phim và là đặc điểm nổi bật của nó.

Người sử dụng (User): Người sử dụng của một hệ gợi ý có thể có các đặc điểm và mục tiêu rất đa dạng. Để cá nhân hóa các gợi ý và hỗ trợ tương tác giữa máy tính và con người, hệ gợi ý khai thác một loạt các thông tin về người sử dụng. Thông tin này có thể được cấu trúc theo nhiều cách khác nhau và hệ thống sẽ lựa chọn những thông tin nào phụ thuộc vào kỹ thuật gợi ý.

Ví dụ: Trong lọc cộng tác, người sử dụng được mô hình hóa bởi một danh sách đơn giản có chứa các đánh giá được cung cấp bởi người sử dụng đối với một số mặt hàng. Các dữ liệu người dùng này sẽ được sử dụng để tạo thành mô hình mã hóa sở thích và nhu cầu người sử dụng.

Giao dịch (Transaction): Giao dịch là sự tương tác giữa một người dùng và hệ gợi ý. Nó lưu trữ dữ liệu, thông tin đăng nhập quan trọng được tạo ra trong quá trình tương tác giữa con người - máy tính và có ích cho thuật toán gợi ý mà hệ thống đang sử dụng.

Ví dụ: một bản ghi giao dịch có thể chứa một tham chiếu đến mặt hàng được lựa chọn bởi người sử dụng và một mô tả về bối cảnh (mục tiêu người sử dụng/truy vấn) cho gợi ý cụ thể. Nếu có sẵn, giao dịch cũng có thể bao gồm một thông tin phản hồi rõ ràng của người sử dụng đã cung cấp, chẳng hạn như đánh giá cho các sản phẩm được chọn.

Trong thực tế, xếp hạng là hình thức phổ biến nhất của các dữ liệu giao dịch trong hệ gợi ý. Những đánh giá này có thể được thu thập một cách rõ ràng hoặc ngầm định. Người dùng sẽ được yêu cầu cung cấp ý kiến của mình về một sản phẩm theo một thang đánh giá. Thang đánh giá có thể là một trong các dạng sau:

- Xếp hạng số từ 1-5 sao (được dùng trong trang web Amazon.com).
- Xếp hạng theo thứ tự, chẳng hạn: "hoàn toàn đồng ý",...
- Xếp hạng đơn giản. Ví dụ: tốt hay xấu, thích hay không thích,…
- Hoặc có đánh giá hoặc không đánh giá.

1.4. CÁC PHƯƠNG PHÁP XÂY DỰNG HỆ THỐNG GỢI Ý

1.4.1. Phương pháp gợi ý dựa trên nội dung

Các phương pháp tiếp cận dựa trên nội dung gợi ý dựa trên việc tính năng của các mặt hàng có thể có ích trong việc giới thiệu chúng. Với cách tiếp cận này, các tính năng của các mặt hàng và sở thích riêng của người sử dụng là những yếu tố duy nhất ảnh hưởng đến việc gợi ý cho người sử dụng.

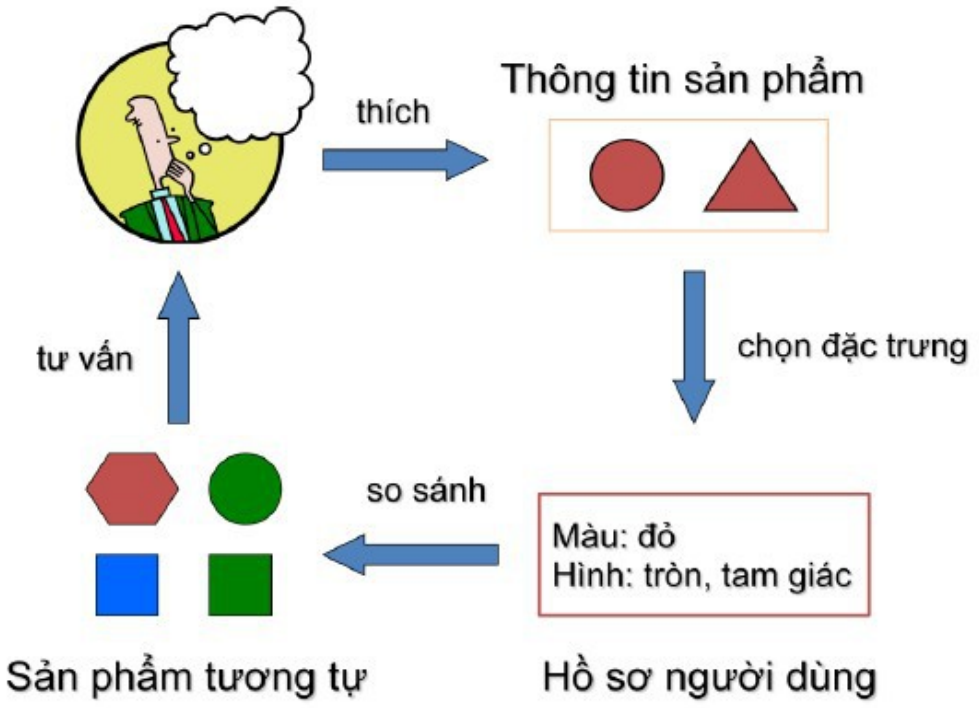
Trong phương pháp lọc dựa trên nội dung, hệ thống sẽ phân tích và so sánh nội dung của các thông tin, các mặt hàng từ đó đánh giá khả năng người dùng sẽ thích mặt hàng đó. Phương pháp lọc dựa trên nội dung dựa trên nguyên lý người dùng thích mặt hàng hay thông tin A sẽ thích mặt hàng hay thông tin B tương tự với mặt hàng A.

Phương pháp lọc dựa trên nội dung còn so sánh nội dung của mặt hàng hay thông tin với sở thích mà người dùng cung cấp. Chẳng hạn người dùng quan tâm tới những thông tin về chứng khoán thì sẽ được gợi ý những bản tin nói về chứng khoán…

Nói cách khác, trong những phương pháp gợi ý dựa trên nội dung, chỉ số đánh giá dự đoán của người dùng u đối với đối tượng i thường được ước lượng dựa vào những chỉ số dự đoán của người dùng u đó đối với những đối tượng tương tự với đối tượng i. Sự tương tự giữa hai đối tượng i và i’ được tính toán tùy theo nội dung của chúng.

Ví dụ trong hệ thống gợi ý phim dựa trên nội dung, để gợi ý những bộ phim cho người dùng u, hệ thống cố gắng tìm hiểu những sở thích của người dùng bằng cách phân tích những điểm tương đồng về mặt nội dung của những bộ phim mà người dùng u đã từng đánh giá trong quá khứ. Khi đó, chỉ những bộ phim nào có độ tương tự cao, phù hợp với sở thích của người dùng mới được hệ thống gợi ý.

Hướng tiếp cận dựa trên nội dung bắt nguồn từ những nghiên cứu về thu thập thông tin (Information Retrieval) và lọc thông tin (Information Filtering). Do đó, rất nhiều hệ thống dựa trên nội dung hiện nay tập trung vào tư vấn các đối tượng chứa dữ liệu văn bản như tin tức, website. Những tiến bộ so với hướng tiếp cận cũ của IR là do việc sử dụng hồ sơ về người dùng (chứa thông tin về sở thích, nhu cầu,..). Hồ sơ này được xây dựng dựa trên những thông tin được người dùng cung cấp trực tiếp (khi trả lời khảo sát) hoặc gián tiếp (do khai phá thông tin từ các giao dịch của người dùng).



Hình 1.2. Phương pháp tiếp cận dựa trên nội dung.

Gợi ý dựa trên nội dung có những ưu điểm:

- Đầu tiên, nó không yêu cầu số lượng người sử dụng lớn để đạt được độ chính xác đề nghị hợp lý.
- Ngoài ra, các mặt hàng mới có thể được gợi ý ngay dựa trên thuộc tính có sẵn.

Tuy nhiên, nhược điểm của gợi ý dựa trên nội dung là khi thông tin mô tả đối tượng có chất lượng kém và bị lỗi. Trong một số trường hợp, những mô tả về nội dung rất khó để so sánh và rút ra gợi ý, chẳng hạn so sánh nội dung của các file video, audio... Việc phân tích nội dung của các đối tượng sản phẩm để đưa ra các sản phẩm tương tự nhau, từ đó đưa ra các gợi ý cho người dùng vẫn chưa phản ánh đúng sở thích của người dùng đó với các sản phẩm.

Thông thường, những hệ thống gợi ý gợi ý những đối tượng tương tự với những đối tượng mà người dùng đã đánh giá trước đó. Tuy nhiên trong một số trường hợp đặc biệt, đối tượng không nên được gợi ý vì chúng có độ tương tự gần như tuyệt đối, nói cách khác là chúng quá tương tự với những thứ người dùng vừa mới xem. Ví dụ như nhiều mục tin tức khác nhau cùng nói về một sự kiện người dùng vừa xem qua ở mục tin tức này, khi đó người dùng sẽ không quan tâm đến những mục tin tức cùng sự kiện kia, hệ thống cũng không nên gợi ý, đôi khi nó còn gợi ý cho người dùng những mặt hàng mà người dùng đã biết hoặc sử dụng trước đó, vì vậy khó có thể tạo ra sự bất ngờ trong gợi ý.

Thêm một bất cập nữa, là người dùng phải có đánh giá cho những đối tượng trước khi hệ thống có thể hiểu được sở thích và gợi ý cho họ những đối tượng khác. Như vậy, hệ thống sẽ gặp vấn đề đối với những người dùng mới, họ chưa cung cấp hoặc cung cấp rất ít những chỉ số dự đoán, hệ thống không đủ dữ liệu ban đầu của người dùng đó để có thể đưa ra những lời gợi ý chính xác dành cho họ.

1.4.2. Phương pháp gợi ý dựa trên lọc cộng tác

Lọc cộng tác là kỹ thuật sử dụng các sở thích cá nhân của người dùng để đưa ra gợi ý. Một hệ thống lọc cộng tác xác định người dùng có sở thích tương tự những người dùng trước và gợi ý các mặt hàng mà họ có thể thích. Bản chất của phương pháp này chính là hình thức gợi ý truyền miệng tự động. Trong phương pháp này, hệ thống sẽ so sánh, tính toán độ tương tự nhau giữa những người dùng hay mặt hàng, từ đó người dùng sẽ được gợi ý những thông tin, mặt hàng được ưa chuộng nhất bởi những người dùng có cùng thị hiếu. Trong phương pháp này, hệ thống thường xây dựng các ma trận đánh giá bởi người dùng lên các mặt hàng, bản tin. Từ đó tính toán độ tương tự giữa họ. Các hệ gợi ý dựa trên lọc cộng tác không yêu cầu quá nặng vào việc tính toán, do đó nó có thể đưa ra những gợi ý có độ chính xác cao và nhanh chóng cho một số lượng lớn người dùng. Hơn nữa, hệ gợi ý này không yêu cầu mô tả nội dung tường minh mà chỉ sử dụng đánh giá của người dùng để ước lượng, do đó những hệ này có khả năng gợi ý phong phú và thường tạo ra

những gợi ý bất ngờ cho người dùng. Với phương pháp này, sở thích của người dùng là đầu vào duy nhất để quyết định kết quả gợi ý.

Nói một cách khác, không giống như phương pháp gợi ý dựa trên nội dung, hệ thống cộng tác dự đoán độ phù hợp của một sản phẩm *i* với người dùng *u* dựa trên độ phù hợp giữa người dùng u_j và *i*, trong đó u_j là người có cùng sở thích với *u*. Ví dụ, để gợi ý một bộ phim cho người dùng *c*, đầu tiên hệ thống cộng tác tìm những người dùng khác có cùng sở thích phim ảnh với *c*. Sau đó, những bộ phim được họ đánh giá cao sẽ được dùng để gợi ý cho *c*.

Đầu vào của bài toán là ma trận thể hiện những hành vi quá khứ, gọi là ma trận Người dùng - Sản phẩm (ma trận User x Item). Hàng là người dùng, cột là sản phẩm, giá trị mỗi ô là đánh giá của người dùng lên sản phẩm đó.

Tùy theo hệ thống mà đánh giá của người dùng được quy ước những giá trị nào. Trong ví dụ này, các đánh giá có giá trị từ 1->5

Bảng 1.1. Ví dụ ma trận Người dùng x Sản phẩm.

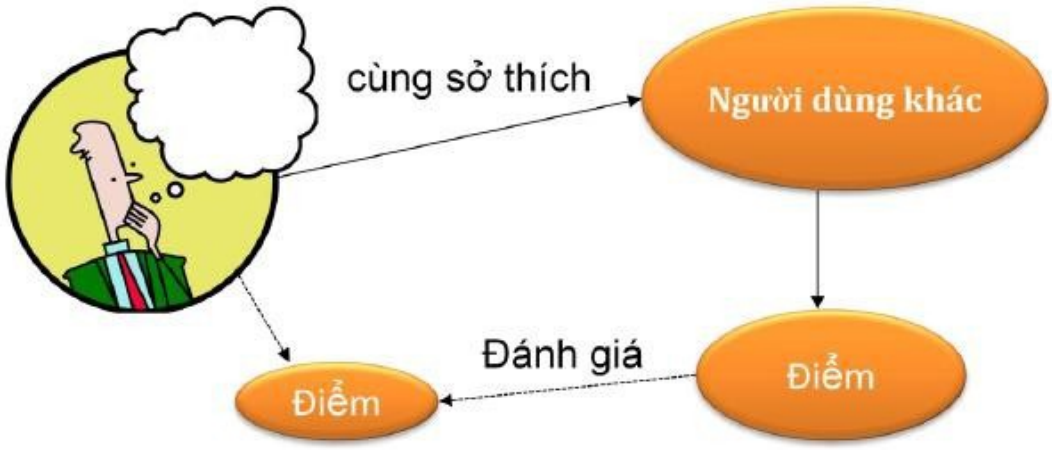
	Sản phẩm 1	Sản phẩm 2	Sản phẩm 3
Người dùng 1	1	0	5
Người dùng 2	4	2	2
Người dùng 3	0	0	0

Ở ma trận này, đánh giá của người dùng 1 đối sản phẩm 1 là 1, sản phẩm 3 là 5, sản phẩm 2 chưa được đánh giá.

Đầu ra của bài toán là: đánh giá của người dùng lên những sản phẩm mà họ chưa đánh giá. Hệ thống gợi ý dựa trên các đánh giá này mà xếp hạng các sản phẩm và gợi ý cho người dùng.

Trong ví dụ này, hệ thống gợi ý phải đưa ra dự đoán, người dùng 1 đánh giá sản phẩm 2 là bao nhiêu. Người dùng 3 đánh giá sản phẩm 1, 2, 3 là bao nhiêu.

Có rất nhiều hệ thống cộng tác đã được phát triển như: Grundy, GroupLens (tin tức), Ringo (âm nhạc), Amazon.com (sách), Phoaks (web)…



Hình 1.3. Phương pháp lọc cộng tác.

Ngược lại với phương pháp tiếp cận dựa trên nội dung thì phương pháp tiếp cận lọc cộng tác lại khắc phục được các giới hạn trên:

- Không giới hạn về loại đối tượng dùng để gợi ý: Phương pháp Lọc cộng tác dựa hoàn toàn vào đánh giá của những người dùng để đưa ra các nhận định về sở thích của người dùng, chính vì thế các tính chất của đối tượng được gợi ý không có ảnh hưởng đến quá trình gợi ý. Ưu điểm này giúp cho phương pháp lọc cộng tác được áp dụng đa dạng trên nhiều hệ thống khác nhau, từ trang thông tin đến âm nhạc, hình ảnh …

- Gợi ý đa dạng: Khắc phục được giới hạn của phương pháp tiếp cận dựa trên nội dung, phương pháp lọc cộng tác có thể đưa ra các đối tượng sản phẩm khuyến nghị hoàn toàn khác so với các sản phẩm mà người dùng *u* đã thích trong quá khứ.

Nhược điểm của lọc cộng tác:

-Vấn đề người dùng mới.
-Vấn đề sản phẩm mới.

1.4.3. Phương pháp gợi ý lai

Hệ gợi ý được dựa trên sự kết hợp của các kỹ thuật đã được đề cập. Một hệ thống lai kết hợp kỹ thuật tiếp cận dựa trên nội dung và lọc cộng tác cố gắng sử

dụng lợi thế của kỹ thuật tiếp cận dựa trên nội dung để sửa chữa những nhược điểm của kỹ thuật lọc cộng tác. Ví dụ, phương pháp lọc cộng tác gặp vấn đề với các mặt hàng mới, nó không có thể gợi ý đối với các mặt hàng mà không có xếp hạng. Điều này lại đơn giản đối với phương pháp tiếp cận dựa trên nội dung khi việc dự đoán cho các mặt hàng mới dựa trên mô tả của người dùng là tính năng có sẵn và khá dễ dàng.

Với hai (hoặc hơn) kỹ thuật gợi ý cơ bản, một số cách đã được đề xuất cho việc kết hợp chúng để tạo ra một hệ thống lai mới.

1.5. MỘT SỐ ỨNG DỤNG

Hệ thống gợi ý đang được thực hiện với sự chú trọng về thực hành với các ứng dụng Thương mại. Như vậy, nghiên cứu hệ gợi ý liên quan đến những khía cạnh thực hành áp dụng đối với việc thực hiện của các hệ thống này. Các khía cạnh đó liên quan đến các giai đoạn khác nhau trong vòng đời của một hệ gợi ý, cụ thể là, thiết kế hệ thống, cài đặt và bảo trì hệ thống trong quá trình hoạt động.

Chúng ta có những ứng dụng cơ bản sau:

Gợi ý sản phẩm: Có lẽ người dùng quan trọng nhất của hệ thống gợi ý là các cửa hàng bán hàng trực tuyến. Chúng ta đã biết đến Amazon hoặc các nhà cung cấp trực tuyến tương tự đang phấn đấu để lôi kéo người dùng quan tâm đến những gợi ý sản phẩm họ có thể mua. Những gợi ý này không phải ngẫu nhiên mà nó dựa trên các quyết định mua hàng được thực hiện trên các khách hàng tương tự hoặc trên các công nghệ khác.

Gợi ý phim ảnh: Netflix cung cấp cho khách hàng các gợi ý về những bộ phim mà họ thích. Những gợi ý này được dựa trên các xếp hạng được cung cấp bởi người sử dụng (biểu diễn dưới dạng ma trận xếp hàng). Các dự đoán xếp hạng chính xác có tầm quan trọng rất lớn, Netflix đã đưa ra giải thưởng một triệu đô la cho người đầu tiên có thuật toán đánh bại 10% hệ thống gợi ý của chính họ. Sau 3 năm nghiên cứu, cuối cùng giải thưởng giành chiến thắng trong năm 2009 thuộc về nhóm các nhà nghiên cứu Bellkor’s Pragmatic Chaos.

Các trang tin tức: Dịch vụ tin tức đã cố gắng xác định sự quan tâm của độc giả dựa trên các bài viết mà họ đã đọc trong quá khứ. Sự giống nhau có thể dựa trên các từ khoá tương tự trong tài liệu hoặc các bài viết đã được đọc từ người đọc có cùng thị hiếu. Nguyên tắc áp dụng để gợi ý là cập nhật thường xuyên nội dung giữa hàng triệu blog có sẵn, video trên YouTube hoặc trên các trang web khác.

Dựa trên các lĩnh vực ứng dụng cụ thể, chúng ta có các lĩnh vực tổng quát cho các ứng dụng phổ biến nhất trong hệ thống gợi ý:

- Giải trí: Gợi ý cho phim ảnh, âm nhạc, và IPTV như MovieLens, EachMovie, Morse, Firefly, Flycasting, Ringo…

- Phân loại nội dung báo chí cho người đọc: Gợi ý tài liệu, gợi ý các trang web, các ứng dụng e-learning và bộ lọc e-mail như Tapestry, GroupLens, Lotus Notes, Anatagonomy…

- Thương mại điện tử: Gợi ý các sản phẩm cho người tiêu dùng mua như sách, máy ảnh, máy tính như Amazon.com, Foxtrot, InfoFinder…

- Dịch vụ: Gợi ý các dịch vụ du lịch như Dietorecs, LifestyleFinder …, các gợi ý của các chuyên gia gợi ý, gợi ý nhà ở hoặc cho thuê, các dịch vụ mai mối… Gợi ý nhà hàng như Adaptive Place Advisor, Polylens, Pocket restaurent finder…

Một vài hệ gợi ý nổi tiếng:

- Phim / TV/ âm nhạc: MovieLens - MovieLens là một trang web giới thiệu phim. Người dùng cho hệ thống biết phim bạn thích và không thích. Hệ thống sử dụng thông tin đó để tạo ra gợi ý cá nhân cho các phim khác mà người dùng có thể sẽ thích hoặc không thích. MovieLens sử dụng công nghệ lọc cộng tác để gợi ý các bộ phim. Nó hoạt động bằng cách kết hợp những người sử dụng có ý kiến tương tự về phim. Mỗi thành viên trong hệ thống có một "vùng lân cận" những người sử dụng tương tự. Đánh giá từ những người láng giềng được sử dụng để tạo ra các gợi ý cá nhân hóa cho người sử dụng.

- Tin tức/báo chí: GroupLens - GroupLens là một phòng thí nghiệm nghiên cứu tại Khoa Khoa học Máy tính và Kỹ thuật tại Đại học Minnesota, tiến hành nghiên cứu trong một số lĩnh vực, bao gồm:

- Hệ thống gợi ý.
- Cộng đồng trực tuyến.
- Công nghệ di động và công nghệ phổ biến.
- Thư viện kỹ thuật số.
- Hệ thống thông tin vùng địa lý.

- Sách/Tài liệu: Amazon.com - Thành lập năm 1994, bắt đầu online vào tháng 7/1995. Từ lĩnh vực kinh doanh ban đầu là sách cho đến nay Amazon đã mở rộng kinh doanh sang nhiều mặt hàng khác như băng đĩa, đồ điện tử, game. Tính đến tháng 7-2005, hãng cung cấp 31 chủng loại hàng tại 7 nước. Hiện Amazon đã cung cấp rất nhiều mặt hàng khác nhau với mục tiêu thực sự trở thành một siêu thị bán lẻ khổng lồ trên Internet theo đúng nghĩa của nó hơn là một cửa hàng bán sách và DVD trực tuyến như trước đây. Amazon.com là một địa chỉ hết sức lôi cuốn mà ngay ngày đầu thành lập đã trở thành địa điểm tham khảo cho bất cứ ai muốn bán mặt hàng của mình. Hiện nay có hơn 900.000 đại lý bán lẻ bên thứ 3 cung cấp mặt hàng của họ lên trang Amazon.

Khi hệ thống gợi ý trở nên phổ biến và được quan tâm hơn, nó sẽ đánh thức tiềm năng lợi thế trong các ứng dụng mới. Các nhà phát triển hệ gợi ý cho một ứng dụng nhất định phải hiểu rõ các các mặt cụ thể của ứng dụng, yêu cầu của nó, thách thức ứng dụng và hạn chế. Chỉ sau khi phân tích những yếu tố này, người ta có thể lựa chọn thuật toán gợi ý tối ưu và thiết kế một sự tương tác giữa con người với máy tính có hiệu quả.

1.6. TIỂU KẾT CHƯƠNG 1

Chương 1 đã trình bày tổng quan về hệ thống gợi ý: Khái niệm chung, chức năng của một hệ thống gợi ý và cơ sở dữ liệu sử dụng trong hệ thống gợi ý. Ngoài ra, trong chương này đã nêu được tầm quan trọng của ứng dụng hệ thống gợi ý trong đời sống.

Chương 2. GIỚI THIỆU KỸ THUẬT LỌC CỘNG TÁC VÀ KỸ THUẬT LÁNG GIỀNG

2.1. PHƯƠNG PHÁP LỌC CỘNG TÁC

2.1.1. Định nghĩa phương pháp lọc cộng tác

Lọc cộng tác (Collaborative Filtering) là một kỹ thuật được áp dụng khá thành công trong các hệ gợi ý, nó được dùng để đánh giá độ quan tâm của người dùng tới một mặt hàng mới. Trong các hệ thống lọc cộng tác, các dự đoán được đưa ra dựa trên tập dữ liệu về sở thích người dùng – mặt hàng có liên quan tới người dùng hoặc mặt hàng. Tuy nhiên, trong trường hợp dữ liệu ít, độ tương tự trực tiếp giữa hai người dùng hoặc hai mặt hàng chỉ cung cấp rất ít thông tin cho ta dự đoán.

Hình 2.1. Hệ thống gợi ý lọc cộng tác của trang web Amazon.com.

Phương pháp gợi ý lọc cộng tác hoàn toàn khác so với phương pháp gợi ý dựa trên nội dung. Thay vì giới thiệu các mặt hàng, vì chúng tương tự như các mặt hàng người dùng đã thích trong quá khứ, cách tiếp cận lọc cộng tác gợi ý các mặt hàng dựa vào ý kiến của những người dùng khác. Thông thường, bằng cách tính toán sự giống nhau của những người sử dụng, một tập hợp láng giềng gần nhất (nearest-neighbor) các người dùng có sở thích tương quan đáng kể với một người dùng nhất

định sẽ được tìm thấy. Như vậy, trong phương pháp này, người dùng chia sẻ sở thích của họ về từng mặt hàng mà họ đã từng tiêu dùng để những người dùng khác của hệ thống có những quyết định tốt hơn đối với những mặt hàng đó. Cách tiếp cận lọc cộng tác là kỹ thuật gợi ý thành công nhất và được chấp nhận rộng rãi cho đến nay.

Bài toán lọc cộng tác:

Ký hiệu $U = \{u_1, u_2, ..., u_N\}$ là tập gồm N người dùng, $P = \{p_1, p_2, ..., p_M\}$ là tập gồm M sản phẩm mà người dùng có thể lựa chọn. Mỗi sản phẩm $p_i \epsilon P$ có thể là hàng hóa, phim, ảnh, tạp chí, tài liệu, sách, báo, dịch vụ hoặc bất kỳ dạng thông tin nào mà người dùng cần đến.

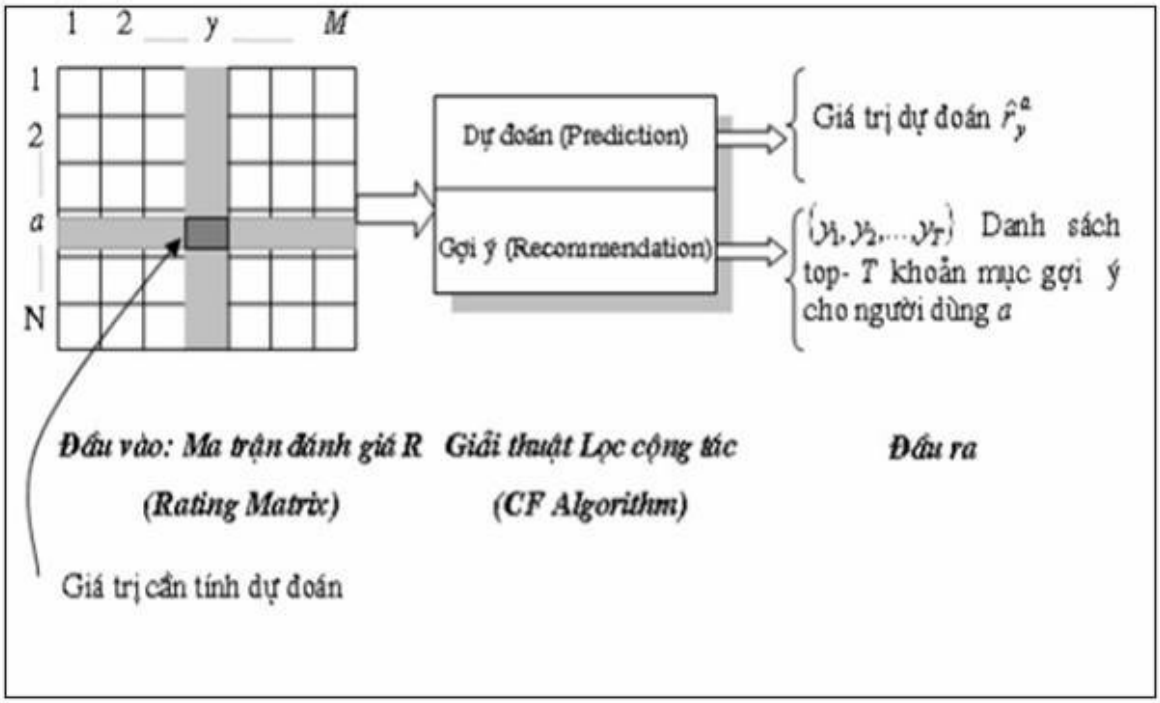
Tiếp theo, ký hiệu R = $\{ r_{ij} \}$, $i = 1..N, j = 1..M$ là ma trận đánh giá, trong đó mỗi người dùng $u_i \epsilon U$ đưa ra đánh giá của mình cho một số sản phẩm $p_i \epsilon P$ bằng một số r_{ij}. Giá trị r_{ij} phản ánh mức độ ưa thích của người dùng u_i đối với sản phẩm p_j, giá trị r_{ij} có thể được thu thập trực tiếp bằng cách hỏi ý kiến người dùng hoặc thu thập gián tiếp thông qua cơ chế phản hồi của người dùng. Giá trị $r_{ij} = \varnothing$ trong trường hợp người dùng u_i chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm p_j.

Với một người dùng cần được gợi ý u_a (Được gọi là người dùng hiện thời, người dùng cần được tư vấn, hay người dùng tích cực), bài toán lọc cộng tác là bài toán dự đoán đánh giá của u_a đối với mặt hàng mà u_a chưa đánh giá ($r_{ij} = \varnothing$), trên cơ sở đó gợi ý cho u_a những sản phẩm được đánh giá cao.

Bảng 2.1 thể hiện một ví dụ với ma trận đánh giá R = (r_{ij}) trong hệ gồm 5 người dùng U = $\{u_1, u_2, u_3, u_4, u_5\}$ và 4 sản phẩm P = $\{p_1, p_2, p_3, p_4\}$. Mỗi người dùng đều đưa ra các đánh giá của mình về các sản phẩm theo thang bậc {∅, 1, 2, 3, 4, 5}. Giá trị $r_{ij} = \varnothing$ được hiểu là người dùng u_i chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm p_j. Các giá trị $r_{5,2} = ?$ là sản phẩm hệ thống cần dự đoán cho người dùng u_5.

Bảng 2.1. Ví dụ về ma trận ma trận đánh giá của lọc cộng tác.

	p_1	p_2	p_3	p_4
u_1	2	1	3	5
u_2	4	2	1	∅
u_3	3	∅	2	4
u_4	4	4	∅	∅
u_5	4	?	5	5



Hình 2.2. Các thành phần của hệ thống lọc cộng tác.

Ma trận đánh giá R = (r_{ij}) là thông tin đầu vào duy nhất của các phương pháp lọc cộng tác. Dựa trên ma trận đánh giá, các phương pháp lọc cộng tác thực hiện hai tác vụ: Dự đoán quan điểm của người dùng hiện thời (Active user) về các sản phẩm mà họ chưa đánh giá, đồng thời đưa ra một danh sách các sản phẩm có đánh giá cao nhất phân bổ cho người dùng hiện thời. Hình 2.2 mô tả các thành phần của hệ thống lọc cộng tác.

Như vậy, phương pháp lọc cộng tác là phương pháp sử dụng các sở thích của người dùng để tạo ra các gợi ý. Phương pháp lọc cộng tác bao gồm các kỹ thuật như kỹ thuật láng giềng, kỹ thuật mạng Bayesian, kỹ thuật mạng Neural kết hợp SVD (Singular value decomposition), kỹ thuật quy tắc quy nạp.

Lọc cộng tác được tiếp cận theo hai xu hướng chính: Lọc cộng tác dựa trên bộ nhớ và lọc cộng tác dựa trên mô hình. Mỗi phương pháp tiếp cận có những ưu điểm và hạn chế riêng, khai thác các mối liên hệ trên ma trận đánh giá người dùng.

2.1.2. Các phương pháp lọc cộng tác

2.1.2.1. Lọc cộng tác dựa trên bộ nhớ

Các phương pháp lọc dựa trên bộ nhớ sử dụng toàn bộ ma trận đánh giá để sinh ra dự đoán sác sản phẩm cho người dùng hiện thời. Phương pháp thực hiện theo hai bước: Tính toán mức độ tương tự và bước tạo nên dự đoán.

- Tính toán độ tương tự sim (*x,y*): Mô tả khoảng cách, sự liên quan, hay trọng số giữa hai người dùng *x* và *y* (Hoặc giữa hai sản phẩm *x* và *y*).

- Dự đoán: Đưa ra dự đoán cho người dùng cần được tư vấn bằng cách xác định tập láng giềng của người dùng này. Tập láng giềng của người dùng cần tư vấn được xác định dựa trên mức độ tương tự giữa các cặp người dùng hoặc sản phẩm.

Các phương pháp tính toán mức độ tương tự:

Việc tính toán mức độ tương tự giữa hai người dùng *x* và *y* được xem xét dựa vào tập sản phẩm cả hai người dùng đều đánh giá. Tương tự, việc tính toán mức độ tương tự giữa hai sản phẩm *x* và *y* được xem xét dựa vào tập người dùng cùng đánh giá cả hai sản phẩm. Sau đó, sử dụng một độ đo cụ thể để xác định mức độ tương tự giữa hai người dùng hoặc sản phẩm.

Có nhiều phương pháp khác nhau tính toán mức độ tương tự sim (x, y) giữa các cặp người dùng. Hai phương pháp phổ biến nhất được sử dụng là độ tương quan Pearson và giá trị cosin giữa hai véctơ.

- Độ tương quan Pearson giữa hai người dùng x, y (User-Based Similarity) được tính toán theo công thức (2.1).

$$sim(x,y) = \frac{\sum_{p \in P_{x,y}} (r_{x,p} - \bar{r}_x)(r_{y,p} - \bar{r}_y)}{\sqrt{\sum_{p \in P_{x,y}} (r_{x,p} - \bar{r}_x)^2 \sum_{p \in P_{x,y}} (r_{y,p} - \bar{r}_y)^2}} \quad (2.1)$$

Trong đó: + $P_{xy} = \{p \in P \mid r_{x,p} \neq \varphi \wedge r_{y,p} \neq \varphi\}$ là tập tất cả các sản phẩm người dùng *x* và người dùng *y* cùng đánh giá.

+ , là trung bình cộng các đánh giá khác ∅ của người dùng *x* và người dùng *y*.

- Độ tương quan Pearson giữa hai sản phẩm *x*, *y* (Item-Based Similarity) được tính toán theo công thức (2.2).

$$sim(x,y) = \frac{\sum_{u \in U_{x,y}} (r_{x,u} - \bar{r}_x)(r_{y,u} - \bar{r}_y)}{\sqrt{\sum_{u \in U_{x,y}} (r_{x,u} - \bar{r}_x)^2 \sum_{p \in P_{x,y}} (r_{y,u} - \bar{r}_y)^2}} \quad (2.2)$$

Trong đó: + $U_{xy} = \{u \in U \mid r_{u,x} \neq \varphi \wedge r_{u,y} \neq \varphi\}$ là tập tất cả người dùng cùng đánh giá sản phẩm *x* và sản phẩm *y*.

+ Giá trị , là đánh giá trung bình cho sản phẩm *x* và sản phẩm *y*.

- Độ tương tự véctơ giữa hai người dùng *x*, *y* là cosin của hai véctơ *x* và *y* theo công thức (2.3).

$$sim(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}.\vec{y}}{\|\vec{x}\| \times \|\vec{y}\|} = \frac{\sum_{p \in P_{x,y}} r_{x,p} r_{y,p}}{\sqrt{\sum_{p \in P_{x,y}} r_{x,p}^2} \sqrt{\sum_{p \in P_{x,y}} r_{y,p}^2}} \quad (2.3)$$

Trong đó: + Hai người dùng x và y được xem xét như hai véc tơ m chiều.

+ m=|P_{xy}| là số lượng các sản phẩm cả hai người dùng cùng đánh giá.

- Độ tương tự véctơ giữa hai sản phẩm x, y là cosin của hai véctơ x và y theo công thức (2.4).

$$sim(x,y) = \cos(\vec{x},\vec{y}) = \frac{\vec{x}.\vec{y}}{\|\vec{x}\| \times \|\vec{y}\|} = \frac{\sum_{u \in U_{x,y}} r_{x,u} r_{y,u}}{\sqrt{\sum_{u \in U_{x,y}} r_{x,u}^2} \sqrt{\sum_{u \in U_{x,y}} r_{y,u}^2}} \quad (2.4)$$

Trong đó: + Hai sản phẩm x và y được xem xét như hai véctơ cột n chiều.

+ n=|U_{xy}| là số lượng các người dùng cùng đánh giá sản phẩm p.

Chú ý rằng cả hai phương pháp lọc theo nội dung và lọc cộng tác đều sử dụng độ đo cosin giống nhau trên tập các sản phẩm. Tuy nhiên, lọc theo nội dung sử dụng độ tương tự cosin cho các véc tơ của trọng số được tính theo độ đo TF–IDF, lọc cộng tác sử dụng cosin giữa hai véc tơ biểu diễn đánh giá của người dùng. Một số độ tương tự khác cũng được sử dụng trong lọc cộng tác như: Constrained Pearson correlation, Root Mean Square, Spearman rank correlation, Kendall's τ correlation. Về bản chất, những độ đo tương tự này là biến đổi của độ tương quan Pearson.

Các phương pháp dự đoán

Phương pháp dự đoán mức độ thích hợp của sản phẩm p chưa được người dùng u đánh giá được tính toán dựa trên tập những người dùng khác đã đánh giá p. Gọi là tập N người dùng tương tự nhất đối với u. Khi đó, mức độ phù hợp của người dùng u đối với sản phẩm mới p được xác định như một hàm các đánh giá của tập láng giềng. Dưới đây là một số phương pháp thông dụng nhất để dự đoán mức độ phù hợp của sản phẩm p đối với người dùng u.

$$r_{u,p} = \frac{1}{N} \sum_{u' \in \hat{U}} r_{u',p}$$

$$r_{u,p} = k \sum_{u' \in \hat{U}} sim(u,u') \times r_{u',p} \qquad (2.5)$$

$$r_{u,p} = \overline{r_u} + k \sum_{u' \in \hat{U}} sim(u,u') \times (r_{u',p} - \overline{r_{u'}})$$

Trong công thức (2.5), *k* được gọi là nhân tố chuẩn hóa, là trung bình các đánh giá của người dùng *u* được xác định theo (2.6).

$$k = 1 / \sum_{u' \in \hat{U}} \left| sim(u,u') \right|$$

$$\overline{r_u} = \frac{1}{|P_u|} \sum_{p \in P_u} r_{u,p}$$

$$P_u = \{ \mathrm{p} \in \mathrm{P} / \ \mathrm{r}_{u,p} \neq 0 \} \qquad (2.6)$$

2.1.2.2. Lọc cộng tác dựa trên mô hình

Khác với phương pháp dựa trên bộ nhớ, phương pháp lọc dựa trên mô hình sử dụng tập đánh giá để xây dựng mô hình huấn luyện. Kết quả của mô hình huấn luyện được sử dụng để sinh ra dự đoán quan điểm của người dùng về các sản phẩm chưa được họ đánh giá.

Giải thuật lọc cộng tác dựa trên mô hình cung cấp các tư vấn sản phẩm bằng việc phát triển một mô hình đánh giá của người dùng. Giải thuật loại này thuộc phương pháp tính xác suất và xử lý lọc cộng tác như tính toán giá trị kỳ vọng của một dự đoán người dùng, cho đánh giá của người đó với các sản phẩm khác. Xử lý xây dựng mô hình được thực hiện bởi nhiều các giải thuật học máy khác nhau như mạng Bayes, phân cụm, và phương pháp dựa trên luật (rule-based), mô hình hồi quy tuyến tính, mô hình entropy cực đại…

Ưu điểm: Nó có nhiều thuận lợi trong việc cung cấp nhanh và những dự đoán chính xác, giảm thiểu đi tính nhạy cảm trong trường hợp ít dữ liệu.

Nhược điểm: Chúng thường yêu cầu thời gian để nắm bắt mô hình, làm giảm đi hiệu quả trong việc cài đặt trên các ứng dụng trực tuyến - nơi mà dữ liệu thường xuyên được thêm vào.

2.2. KỸ THUẬT LÁNG GIỀNG

2.2.1. Giới thiệu kỹ thuật láng giềng

* Định nghĩa

Để đưa ra một định nghĩa chính thức của nhiệm vụ gợi ý, chúng ta cần quy ước một số ký hiệu. Có thể thống nhất như sau: các thiết lập của người sử dụng trong hệ thống sẽ được ký hiệu là U, các thiết lập của các mặt hàng là I, R là tập xếp hạng ghi nhận trong hệ thống, và S là tập hợp các giá trị có thể cho một đánh giá (ví dụ, S = [1,5] S = {thích, không thích}). Ký hiệu U_i để xác định các tập hợp con của người sử dụng đã đánh giá một mặt hàng i. Tương tự như vậy, I_u đại diện cho các tập hợp con của các mặt hàng đã được đánh giá bởi một người sử dụng u. Cuối cùng, các mặt hàng đã được đánh giá cao nhất bởi hai người dùng u và v, có thể viết là I_{uv}. U được sử dụng để biểu thị các thiết lập của người sử dụng đã đánh giá cả hai mặt hàng i và j.

Hai trong số các vấn đề quan trọng nhất liên quan đến hệ thống gợi ý là *Mặt hàng tốt nhất* và *N gợi ý đầu tiên*. Vấn đề đầu tiên bao gồm tìm kiếm cho một người

dùng cụ thể u, các mặt hàng mới $i \in I \setminus I_u$ mà có khả năng được u quan tâm nhất. Khi xếp hạng, công việc này thường được xác định như là một hồi quy hoặc phân loại vấn đề mà mục tiêu là để tìm hiểu một chức năng $f: U \times I \rightarrow S$ có thể dự đoán đánh giá $f(u, i)$ của một người sử dụng u cho một mặt hàng mới i. Chức năng này sau đó được sử dụng để giới thiệu cho người dùng cần gợi ý u_a một mặt hàng i^* mà ước tính có giá trị đánh giá cao nhất:

$$i^* = \arg\max_{j \in I / I_U} f(u_a, j)$$

* Ưu điểm của kỹ thuật láng giềng

Các ưu điểm chính của kỹ thuật láng giềng là:

Đơn giản: Kỹ thuật láng giềng dựa trên trực quan và thực hiện tương đối đơn giản. Trong đó, chỉ có một tham số (số lượng của láng giềng được sử dụng trong dự đoán) là có thể bị yêu cầu điều chỉnh.

Hợp lý: Kỹ thuật này cũng cung cấp một sự lý giải ngắn gọn và trực quan cho các tính toán dự đoán. Ví dụ: trong gợi ý dựa trên mặt hàng, danh sách các láng giềng cũng như đánh giá được đưa ra bởi người sử dụng cho các mặt hàng này được xem như là một cơ sở, nền tảng cho việc gợi ý. Điều này có thể giúp người sử dụng hiểu rõ hơn về các gợi ý và sự liên quan của nó, được xem như là cơ sở cho một hệ thống tương tác, nơi người dùng có thể lựa chọn những người láng giềng của họ.

Hiệu quả: Một trong những điểm mạnh của các vùng lân cận dựa trên hệ thống là hiệu quả của nó. Không giống như hầu hết các hệ thống dựa trên mô hình, nó yêu cầu các giai đoạn cần phải được thực hiện trong khoảng thời gian thường xuyên đối với các ứng dụng thương mại lớn. Trong khi giai đoạn gợi ý là thường tốn kém hơn so với phương pháp dựa trên mô hình, kỹ thuật láng giềng có thể tính toán trong một bước ẩn, cung cấp các gợi ý gần như ngay lập tức.

Hơn nữa, lưu trữ những láng giềng gần nhất yêu cầu rất ít bộ nhớ, làm cho cách tiếp cận này có khả năng mở rộng cho các ứng dụng có hàng triệu người sử dụng và các mặt hàng.

Ổn định: Một khía cạnh khác hữu ích của hệ thống gợi ý dựa trên cách tiếp cận này là họ ít bị ảnh hưởng bởi việc bổ sung liên tục của người sử dụng, các mặt hàng và đánh giá mà thường gặp trong các ứng dụng thương mại lớn. Ví dụ: một khi mặt hàng tương tự đã được tính toán, một hệ thống dựa trên mặt hàng có thể sẵn sàng đưa ra các gợi ý cho người dùng mới, mà không cần phải cài đặt lại hệ thống. Hơn nữa, một khi đánh giá đã được nhập vào cho một mặt hàng mới, chỉ có tương đồng giữa mặt hàng này và những người đã có trong hệ thống mới cần phải được tính toán.

2.2.2. Phân loại kỹ thuật láng giềng

2.2.2.1. Gợi ý dựa trên người dùng (User-based)

Phương pháp gợi ý dựa trên người dùng dự đoán đánh giá r_{ui} của người dùng u cho một mặt hàng mới i bằng cách sử dụng các đánh giá cho i bởi những người sử dụng tương tự nhất với u, được gọi là láng giềng gần nhất (nearest-neighbors). Giả sử chúng ta cho mỗi người dùng $v \neq u$ một giá trị đại diện cho sở thích giống nhau giữa u và v. k láng giềng gần nhất (k-NN) của u, ký hiệu là N(u), và k người sử dụng v với độ tương tự cao nhất với u. Tuy nhiên, chỉ những người dùng đã đánh giá mặt hàng i mới có thể được sử dụng trong dự đoán của r_{ui}, và thay việc xem xét k người sử dụng tương tự u nhất bằng việc đã đánh giá i.

Giải thuật lọc cộng tác dựa trên người dùng lân cận gần nhất sử dụng độ tương tự Pearson bằng ngôn ngữ giả để dự đoán độ thích cho người dùng u trên sản phẩm i được biểu diễn như sau:

1**: procedure** USERKNN-CF($\bar{r}_u$, r, D^{train})
2: **for** u=1 to N **do**
3: Tính Sim_uu' , sử dụng công thức (CT 1)
4: **end for**

5: **Sort** Sim_uu' // sắp xếp giảm dần độ tương tự
6: **for** *k*=1 to *K* **do**
7: *Ku* ← *k* // Các người dùng k gần nhất của u
8: **end for**
9: **for** *i* = 1 to *M* **do**
10: Tính $\hat{r}_{us}$, sử dụng công thức (CT 2)
11: **end for**
12: end procedure

2.2.2.2. Gợi ý dựa trên các mặt hàng (Item-based)

Trong khi phương pháp dựa trên ý kiến của người sử dụng để dự đoán một đánh giá, phương pháp tiếp cận dựa trên mặt hàng nhìn vào đánh giá cho các mặt hàng tương tự chúng.

Ý tưởng này có thể được thể hiện như sau. Biểu thị bởi $N_u(i)$ các mặt hàng đánh giá bởi người sử dụng tương tự *u* đối với mặt hàng *i*. Các đánh giá dự đoán của *u* cho *i* thu được như là một trọng số trung bình đánh giá được đưa ra bởi *u* đối với các mặt hàng của $N_u(i)$.

2.2.2.3. Đánh giá kỹ thuật gợi ý dựa trên người dùng và mặt hàng

Khi lựa chọn giữa việc hệ thống nên sử dụng kỹ thuật gợi ý dựa trên người dùng hay dựa trên mặt hàng, ta có 5 tiêu chuẩn cần được xem xét:

Độ chính xác: Độ chính xác của phương pháp gợi ý láng giềng phụ thuộc chủ yếu vào tỷ lệ giữa số lượng người sử dụng và các mặt hàng trong hệ thống. Trong trường hợp số lượng người dùng là lớn hơn nhiều hơn số lượng mặt hàng, phương pháp dựa trên các mặt hàng có thể đưa ra các gợi ý chính xác hơn. Tương tự như vậy, hệ thống có người sử dụng ít hơn so với các mặt hàng thì có thể có hiệu quả hơn khi sử dụng phương pháp gợi ý dựa trên người dùng.

Hiệu suất: Bộ nhớ và hiệu quả tính toán của hệ thống gợi ý cũng phụ thuộc vào tỷ lệ giữa số lượng người dùng và các mặt hàng. Vì vậy, khi số lượng người dùng vượt quá số lượng mặt hàng, phương pháp gợi ý dựa trên mặt hàng yêu cầu ít bộ nhớ và thời gian để tính toán độ tương tự hơn phương pháp dựa trên người dùng.

Tuy nhiên, sự phức tạp về thời gian của giai đoạn gợi ý (phụ thuộc vào số lượng các mặt hàng có sẵn và số lượng tối đa các láng giềng) là như nhau cho cả hai phương pháp dựa trên người sử dụng và dựa trên mặt hàng.

Tính ổn định: Sự lựa chọn giữa phương pháp dựa trên người sử dụng và dựa trên mặt hàng phụ thuộc vào tần số và số lượng thay đổi người sử dụng và các mặt hàng của hệ thống. Nếu trong danh sách các mặt hàng có sẵn ít thay đổi so với những người sử dụng của hệ thống, phương pháp dựa trên mặt hàng có thể được ưa thích hơn. Ngược lại, nếu các mặt hàng có sẵn được thay đổi liên tục thì phương pháp dựa trên người dùng sẽ ổn định hơn.

Tính hợp pháp: Một lợi thế của phương pháp dựa trên mặt hàng là nó có thể được sử dụng để giải thích cho một gợi ý. Do đó, danh sách các mặt hàng láng giềng sử dụng trong dự đoán, cũng như độ tương tự của họ, có thể được trình bày cho người sử dụng như là một giải thích về các gợi ý. Tuy nhiên, phương pháp dựa trên người dùng không tuân theo quá trình này bởi vì người dùng cần gợi ý không biết những người sử dụng khác đóng vai trò là các láng giềng trong gợi ý.

Sự ngẫu nhiên: Trong phương pháp dựa trên mặt hàng, đánh giá dự đoán cho một mặt hàng được dựa trên đánh giá cho các mặt hàng tương tự. Do đó, hệ thống gợi ý sử dụng phương pháp này sẽ có xu hướng giới thiệu loại mặt hàng cho một người sử dụng có liên quan đến những người thường đánh giá cao chính các loại mặt hàng đó. Điều này có thể dẫn đến các gợi ý không phong phú, đa dạng. Mặt khác, sử dụng phương pháp gợi ý dựa trên người dùng có nhiều khả năng đưa ra các gợi ý tình cờ, ngẫu nhiên hơn.

2.2.3. Các bước kỹ thuật láng giềng

Có 3 bước quan trọng trong việc cài đặt một hệ thống gợi ý với kỹ thuật láng giềng là: 1) Chuẩn hóa đánh giá, 2) Tính toán độ tương tự, và 3) Lựa chọn các láng giềng.

2.2.3.1. Chuẩn hóa đánh giá

Khi nói đến việc đánh giá cho một mặt hàng, mỗi người dùng đều có ý kiến cá nhân riêng của mình. Ngay cả khi thống nhất một khuôn mẫu rõ ràng cho việc đánh giá (ví dụ: 1 = "hoàn toàn không đồng ý", 2 = "không đồng ý", 3 = "trung lập", …), một số người dùng có thể miễn cưỡng cho điểm cao/thấp cho mặt hàng mà họ thích/không thích. Hai trong số các phương pháp chuẩn hóa đánh giá phổ biến nhất được đề xuất để chuyển đổi xếp hạng cá nhân đến một quy mô thống nhất hơn là phương pháp Điểm trung bình và Điểm số Z.

a. Phương pháp điểm trung bình

Ý tưởng của phương pháp là để xác định xem một đánh giá là tích cực hay tiêu cực bằng cách so sánh nó với đánh giá trung bình. Trong gợi ý dựa trên người dùng, chuẩn hóa đánh giá bằng cách trừ đi r_{ui} cho các giá trị trung bình r_u của các đánh giá được đưa ra bởi người sử dụng u cho các mặt hàng trong I_u:

$$hr_{ui} = r_{ui} - r_u$$

Tương tự như vậy, phương pháp chuẩn hóa đánh giá Điểm trung bình đối với các mặt hàng trong đánh giá r_{ui} được cho bởi:

$$hr_{ui} = r_{ui} - r_i$$

b. Phương pháp điểm số Z

Xem xét hai người dùng A và B mà cả hai đều có một đánh giá trung bình là 3. Giả sử, xếp hạng của A nằm trong khoảng từ 1 đến 5, trong khi người B luôn 3. Một đánh giá 5 cho một mặt hàng bởi B là đặc biệt hơn so với đánh giá tương tự bởi A, và do đó, phản ánh sự đánh giá cao hơn cho mặt hàng này. Phương pháp chuẩn hóa đánh giá điểm số Z xem xét sự lây lan trong thang đánh giá cá nhân.

Trong phương pháp dựa trên người dùng (User-based), chuẩn hóa đánh giá *r* chia Điểm trung bình của người sử dụng bởi độ lệch chuẩn σ_u của các đánh giá được đưa ra bởi người sử dụng *u*:

$$hr_{ui} = \frac{r_{ui} - r_u}{\sigma_u}$$

Tương tự như vậy, việc chuẩn hóa điểm số Z của đánh giá trong phương pháp dựa trên việc chia nghĩa định tâm bởi độ lệch chuẩn đánh giá cho mặt hàng *i*:

$$hr_{ui} = \frac{r_{ui} - r_i}{\sigma_i}$$

Trong hai phương pháp, phương pháp Điểm số Z được cho là tốt hơn so với Điểm trung bình. Vì Điểm số Z sử dụng giá trị độ lệch chuẩn sẽ làm rõ sự khác biệt của các giá trị đánh giá. Phương pháp này nhạy cảm và tập trung chú ý sự thay đổi thường xuyên của các giá trị đánh giá, loại bỏ các giá trị nằm ngoài thang đánh giá để đưa ra giá trị đánh giá chính xác nhất.

2.2.3.2. Tính toán độ tương tự

Độ tương tự đóng một vai trò kép trong phương pháp gợi ý láng giềng. Thứ nhất, nó cho phép lựa chọn láng giềng đáng tin cậy được sử dụng trong dự đoán, và thứ hai, nó cung cấp giá trị để cho biết tầm quan trọng nhiều hay ít của những người láng giềng trong dự đoán. Việc tính toán độ tương tự là một trong những khía cạnh quan trọng nhất của việc xây dựng một hệ thống gợi ý, vì nó có thể có một tác động đáng kể trên cả tính chính xác và hiệu quả của nó.

Ta có nhiều cách để tính toán độ tương tự giữa hai người dùng như: sử dụng Hệ số tương quan Pearson, tính Khoảng cách Euclide, sử dụng Hệ số tương quan Pearson hạn chế, Hệ số tương quan thứ hạng Spearman, tính Độ tương tự theo Cosine, tính sự Khác biệt trung bình bình phương. Bên cạnh đó, ta cũng đã đưa ra

kết luận nên sử dụng Hệ số tương quan Pearson để đạt được hiệu suất tốt nhất về sự cân bằng giữa độ chính xác của dự báo và các mặt hàng có thể được dự đoán. Vậy nên, với việc tính toán độ tương tự giữa hai người dùng, ta nên sử dụng cách tính Hệ số tương quan Pearson để có kết quả tốt nhất.

2.2.3.3. Lựa chọn láng giềng

Việc lựa chọn số lượng láng giềng gần nhất và tiêu chuẩn sử dụng cho việc lựa chọn này có thể cũng có tác động đến kết quả của hệ thống gợi ý.

Việc lựa chọn những người láng giềng sử dụng trong việc giới thiệu các mặt hàng thường được thực hiện theo hai bước: 1) Trước khi tiến hành tính toán dự đoán, và 2) Trong quá trình tính toán dự đoán.

a. Trước khi tiến hành tính toán dự đoán

Trong hệ thống gợi ý lớn, có thể có hàng triệu người sử dụng và các mặt hàng,

nó thường không thể lưu trữ các điểm tương đồng (khác 0) giữa mỗi cặp người dùng hoặc mặt hàng do hạn chế về bộ nhớ. Hơn nữa, làm như vậy sẽ rất lãng phí vì chỉ các giá trị được sử dụng trong các dự đoán mới quan trọng nhất. Quá trình lọc các láng giềng là một bước cần thiết cho việc tiếp cận vùng lân cận bằng cách giảm số lượng độ tương tự để lưu trữ, và hạn chế số lượng láng giềng để xem xét trong các dự đoán. Có một số cách để làm điều này:

Lọc top N: Cho mỗi người dùng hoặc mặt hàng, chỉ có một danh sách N-láng giềng gần nhất và độ tương tự tương ứng của họ được lưu lại. Để tránh sai sót một cách hiệu quả và chính xác, N nên được lựa chọn cẩn thận. Vì vậy, nếu N là quá lớn, đồng nghĩa với việc cần nhiều không gian bộ nhớ để lưu trữ các danh sách vùng lân cận và dự đoán đánh giá sẽ bị chậm. Mặt khác, lựa chọn một giá trị N quá nhỏ có thể giảm phạm vi của phương pháp gợi ý, gây ra việc một số mặt hàng không bao giờ được gợi ý.

Lọc định mức: Thay vì giữ một số cố định các láng giềng gần nhất, cách tiếp cận này sẽ giúp tất cả những láng giềng có độ tương tự có độ lớn lớn hơn một

ngưỡng nhất định. Trong khi điều này là linh hoạt hơn so với kỹ thuật lọc trước, như chỉ có những người láng giềng quan trọng nhất được lưu giữ, thì giá trị lại khó để xác định.

Lọc đánh giá âm: Nói chung, đánh giá âm là ít đáng tin cậy hơn đánh giá dương. Điều này là do mối tương quan dương mạnh mẽ giữa hai người sử dụng là một hỉ số tốt thuộc về một nhóm phổ biến (ví dụ, nhóm thanh thiếu niên, người hâm mộ khoa học viễn tưởng, vv.) Tuy nhiên, mặc dù tương quan âm có thể chỉ ra thành viên các nhóm khác nhau, nó không nói các nhóm này khác nhau như thế nào, hay các nhóm này là tương thích cho các thể loại khác của các mặt hàng. Mối tương quan âm không cải thiện nhiều độ chính xác dự đoán, cho dù mối tương quan như vậy có thể được loại bỏ phụ thuộc vào dữ liệu. Ba cách tiếp cận trên không loại trừ lẫn nhau và có thể được kết hợp để phù hợp với nhu cầu của hệ thống gợi ý. Ví dụ, người ta có thể loại bỏ tất cả các điểm tương quan âm cũng như những người có độ tương tự thấp hơn so với ngưỡng cụ thể.

b. Trong quá trình tính toán dự đoán

Khi một danh sách các láng giềng đã đợc tính toán cho mỗi người dùng hoặc mặt hàng, dự đoán đánh giá mới được thực hiện với k-láng giềng gần nhất, có nghĩa là, k láng giềng có độ tương tự gần nhau nhất. Vấn đề quan trọng ở đây là có giá trị hay khoảng giá trị cụ thể nào để sử dụng cho k.

Thực tế, không nên đưa tất cả các láng giềng vào tính toán. Nếu chúng ta bao gồm tất cả láng giềng, điều này sẽ không chỉ ảnh hưởng xấu đến thời gian tính toán, mà còn ảnh hưởng đến tính chính xác của các gợi ý. Trong hầu hết các tình huống thực tế, một vùng lân cận từ 20 đến 50 láng giềng là hợp lý.

Khi số lượng các láng giềng hạn chế do sử dụng một k nhỏ (ví dụ: $k < 20$), độ

chính xác dự báo là thấp. Khi k tăng, láng giềng nhiều hơn góp phần vào việc dự đoán và giới thiệu các mặt hàng cho người dùng. Kết quả, độ chính xác dự đoán được cải thiện. Cuối cùng, độ chính xác thường giảm khi có quá nhiều người hàng xóm được sử dụng trong dự đoán (ví dụ: $k > 50$), do số lượng láng giềng k là quá

cao, quá nhiều láng giềng làm loãng các dự đoán. Mặc dù một số láng giềng thường từ 20 đến 50, giá trị tối ưu của k nên được xác định cụ thể.

2.3. TIỂU KẾT CHƯƠNG 2

Trong chương 2 đã tập trung giới thiệu về phương pháp lọc cộng tác và phương pháp láng giềng.

Trong chương tiếp theo sẽ trình bày về ứng dụng phương pháp lọc cộng tác trong hệ thống bán hàng trực tuyến.

Chương 3. XÂY DỰNG HỆ THỐNG BÁN SÁCH TRỰC TUYẾN

3.1. GIỚI THIỆU HỆ THỐNG GỢI Ý

Hệ thống Website được xây dựng với mục đích tìm hiểu, nghiên cứu hoạt động của hệ gợi ý Recommender Systems.

Hệ thống giới thiệu và bán sách trực tuyến cho phép khách hàng bất kỳ có thể tìm kiếm và xem các tác phẩm của tất cả các tác giả mà người dùng ưa thích. Hệ thống sẽ hiển thị các cuốn sách được ưa thích nhất và bán chạy nhất.

Hệ thống cho phép người dùng đăng ký thành viên và đăng nhập vào hệ thống để tìm kiếm thông tin các cuốn sách cũng như đánh giá cho các cuốn sách đó. Thông tin đăng ký sẽ bao gồm Tên đăng nhập, Mật khẩu, Tên khách hàng, Ngày sinh, Số điện thoại, và Địa chỉ. Sau khi đăng ký thành công, hệ thống tự động đăng nhập và những thông tin đăng ký của người dùng sẽ được lưu vào cơ sở dữ liệu.

Sau khi đăng ký và đăng nhập thành công, hệ thống cho phép người dùng đánh giá, bình chọn những cuốn sách mà mình yêu thích với điểm số dao động từ 1-5 tương ứng với mức độ hài lòng của mỗi cá nhân. Đặc biệt hệ thống sẽ gợi ý cho khách hàng các loại sách trong quá trình chọn sản phẩm sử dụng kỹ thuật lọc cộng tác và hiển thị các cuốn sách tương tự với cuốn sách mà khách hàng đang xem sử dụng các thuộc tính tương tự về thể loại của sách đó. Chỉ khi nào người dùng đăng ký và đăng nhập thành công thì mới có thể đánh giá cho các cuốn sách của hệ thống đưa ra và những cuốn sách mà người dùng đó yêu thích. Những thông tin đánh giá của người dùng sẽ được lưu vào cơ sở dữ liệu nhằm sử dụng cho việc tính toán sau này. Nếu là khách hàng mới thì hệ thống sẽ dựa vào thông tin về thể loại yêu thích của khách hàng để tư vấn sách theo thông tin vừa thu thập.

Sau khi đăng ký, đăng nhập và đánh giá hoặc đăng nhập thành hiện đang đăng nhập dựa trên chính thông tin đánh giá mà người dùng đó cung cấp. Đồng thời sẽ tiếp tục đưa ra danh sách các cuốn sách khác để người dùng tìm kiếm thông tin và đánh giá.

Người dùng có thể thay đổi điểm số đánh giá đối với các cuốn sách của mình.

3.2. PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

Tác nhân “Khách hàng” sử dụng hệ thống để đặt hàng. Các trường hợp sử dụng ở dạng tổng quát này là “xem sản phẩm”, “mua hàng”, “đặt hàng” và “đăng ký thành viên”. Trường hợp sử dụng “xem sản phẩm” có thể được sử dụng bởi khách hàng chỉ khi khách hàng chỉ muốn tìm và xem sản phẩm. Trường hợp sử dụng này cũng có thể được sử dụng như là một phần của trường hợp sử dụng “mua hàng”. Trường hợp sử dụng “đăng ký thành viên” cho phép khách hàng đăng ký trên hệ thống.

Trường hợp sử dụng “xem sản phẩm” được mở rộng thành một vài trường hợp sử dụng tuỳ chọn - khách hàng có thể tìm sản phẩm, xem chi tiết sản phẩm, xem những sản phẩm tương tự với sản phẩm, đánh giá và chấm điểm cho sản phẩm, xem những sản phẩm gợi ý cho khách hàng và thêm sản phẩm vào giỏ hàng.

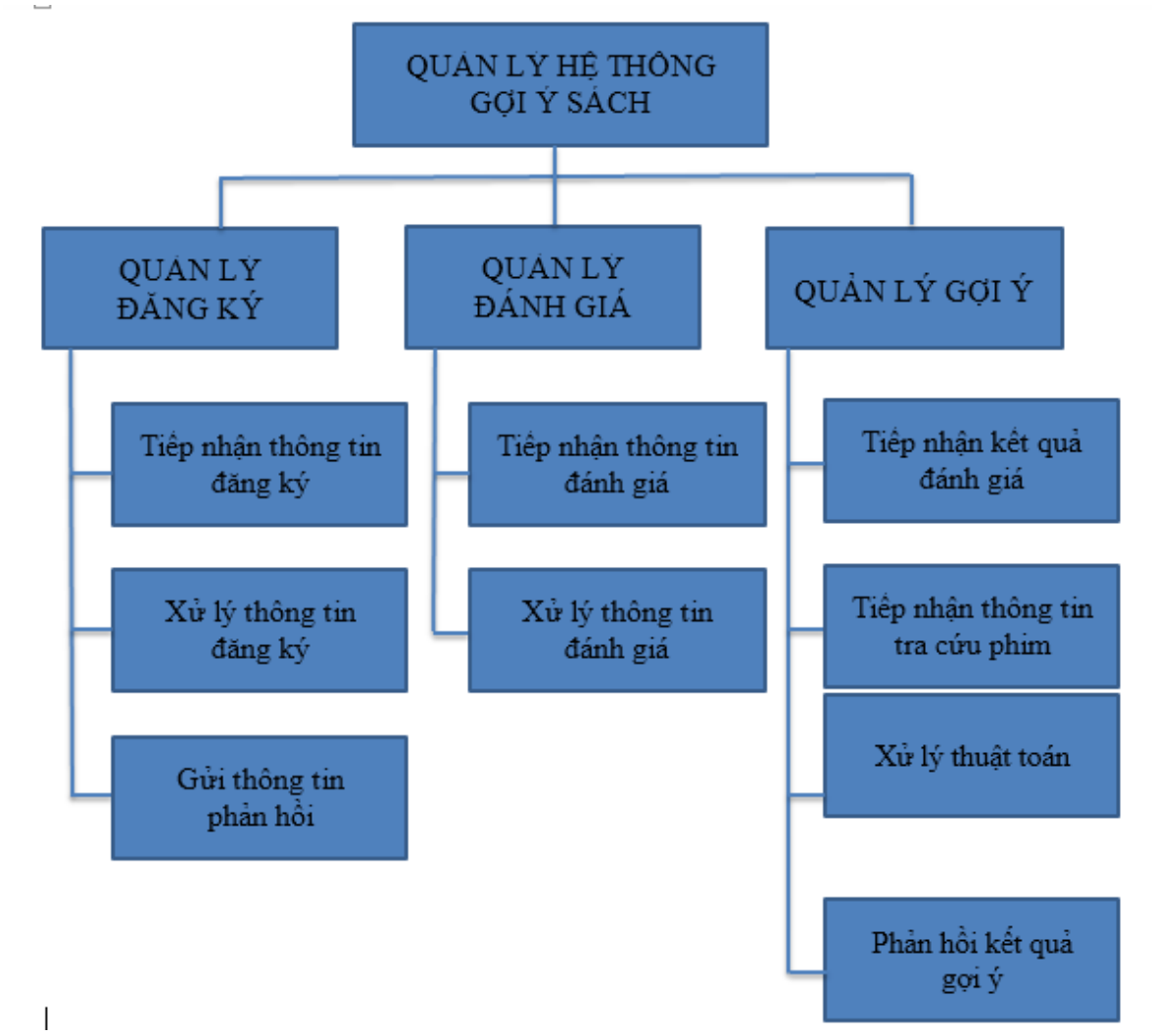
Trường hợp sử dụng “đăng nhập hệ thống” được bao gồm trong trường hợp sử dụng “đánh giá sản phẩm”, “chấm điểm sản phẩm”, “xem sản phẩm gợi ý” và "thêm sản phẩm vào giỏ hàng” bởi vì các thành phần này yêu cầu khách hàng phải chứng thực tài khoản.

Trường hợp sử dụng “đặt hàng” bao gồm một vài trường hợp sử dụng cần thiết như “xem, cập nhật số lượng hàng và xoá đơn hàng trong giỏ hàng”, “tính toán tổng tiền”. Khách hàng phải chứng thực tài khoản. Điều này có thể được thực hiện thông qua đăng nhập khách hàng (login page).

3.2.1. Sơ đồ chức năng kinh doanh (BFD)

- Quản lý đăng ký: Người dùng truy cập website có thể tạo tài khoản và đăng nhập vào hệ thống. Hệ thống sau khi nhận được những thông tin đăng ký của người dùng sẽ xử lý thông tin và phản hồi lại cho người dùng.
- Quản lý đánh giá: Người dùng sau khi đăng ký và đăng nhập vào hệ thống sẽ được chuyển đến trang chủ để xem các sản phẩm và đánh giá sản phẩm. Hệ thống sẽ tiếp nhận và xử lý thông tin đánh giá sau khi người dùng đánh giá xong.

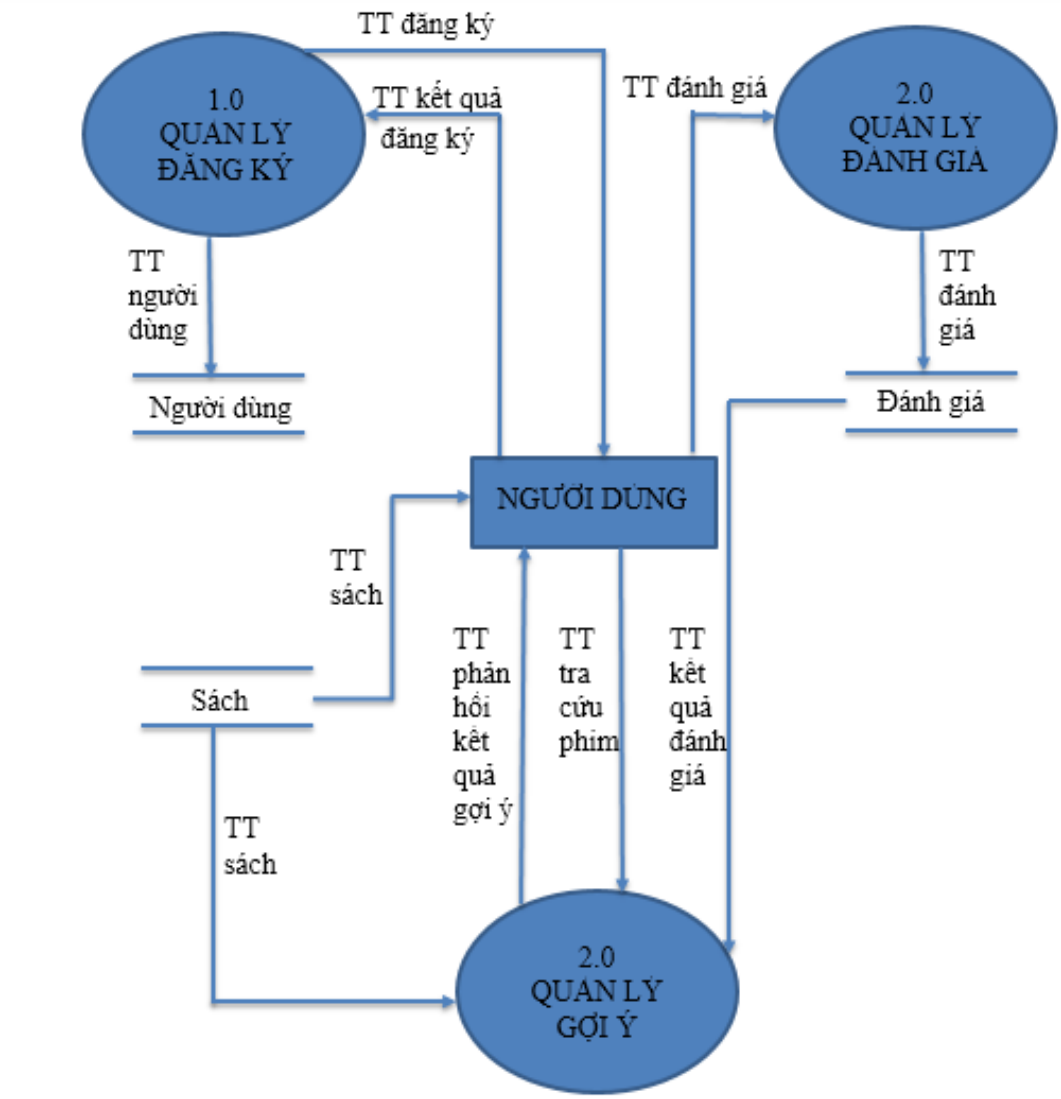
- Quản lý gợi ý: Sau khi người dùng đánh giá xong, hệ thống xử lý và tiếp nhận kết quả đánh giá. Lúc này người dùng có thể tra cứu chi tiết các cuốn sách và hệ thống sẽ đưa ra kết quả gợi ý cho người dùng.



Hình 3.1. Sơ đồ chức năng kinh doanh BFD.

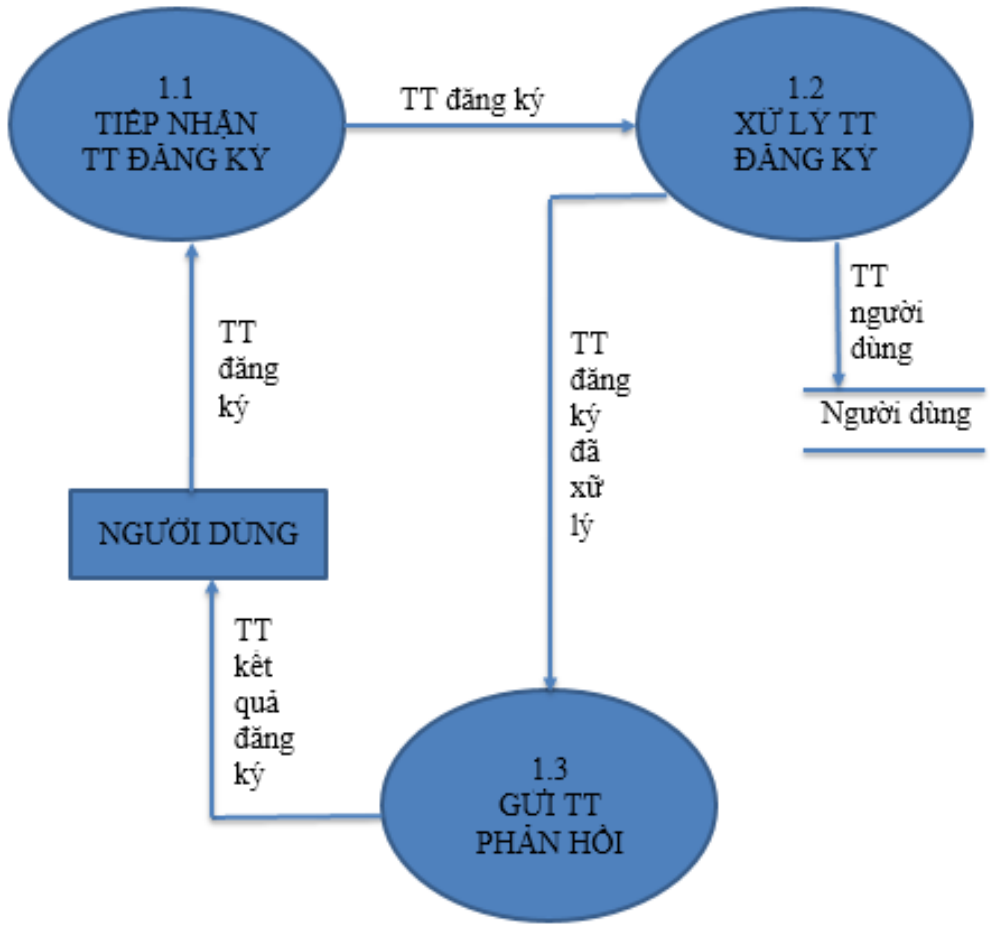
3.2.2. Sơ đồ luồng dữ liệu (DFD)

- **Sơ đồ luồng dữ liệu mức 0 (DFD mức 0)**



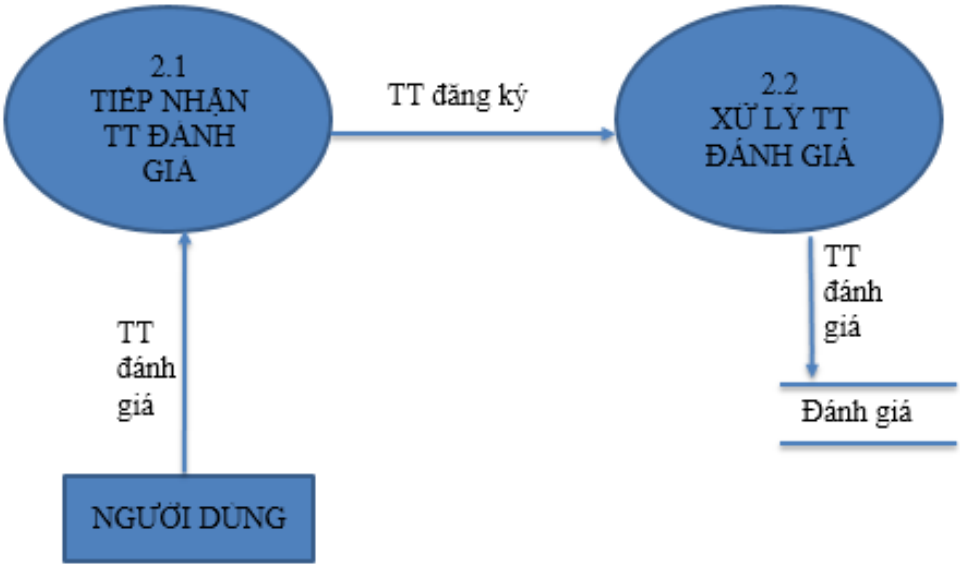
Hình 3.2. Sơ đồ DFD mức 0.

- **Sơ đồ luồng dữ liệu mức 1 (DFD mức 1)**



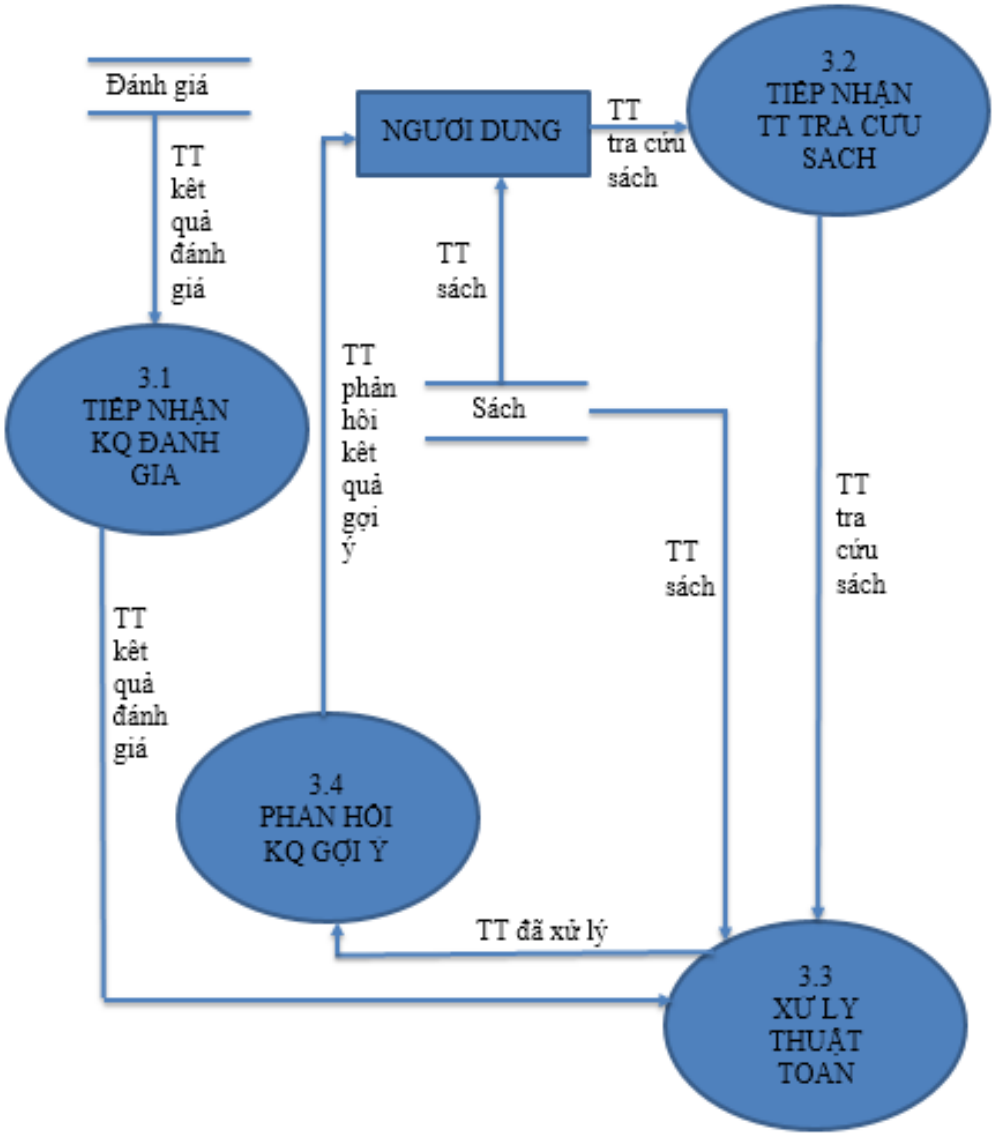
Hình 3.3. Sơ đồ DFD mức 1.

- **Sơ đồ luồng dữ liệu mức 2 (DFD mức 2)**



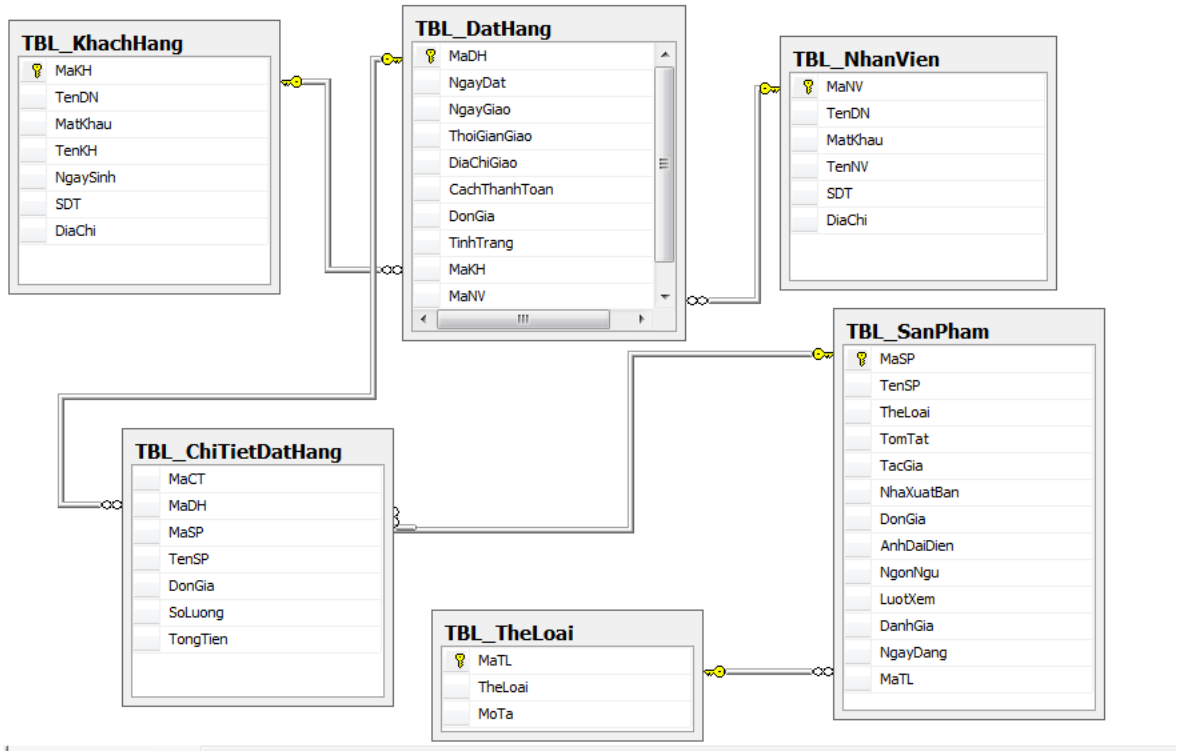
Hình 3.4. Sơ đồ DFD mức 2.

- **Sơ đồ luồng dữ liệu mức 3 (DFD mức 3)**



Hình 3.5. Sơ đồ DFD mức 3.

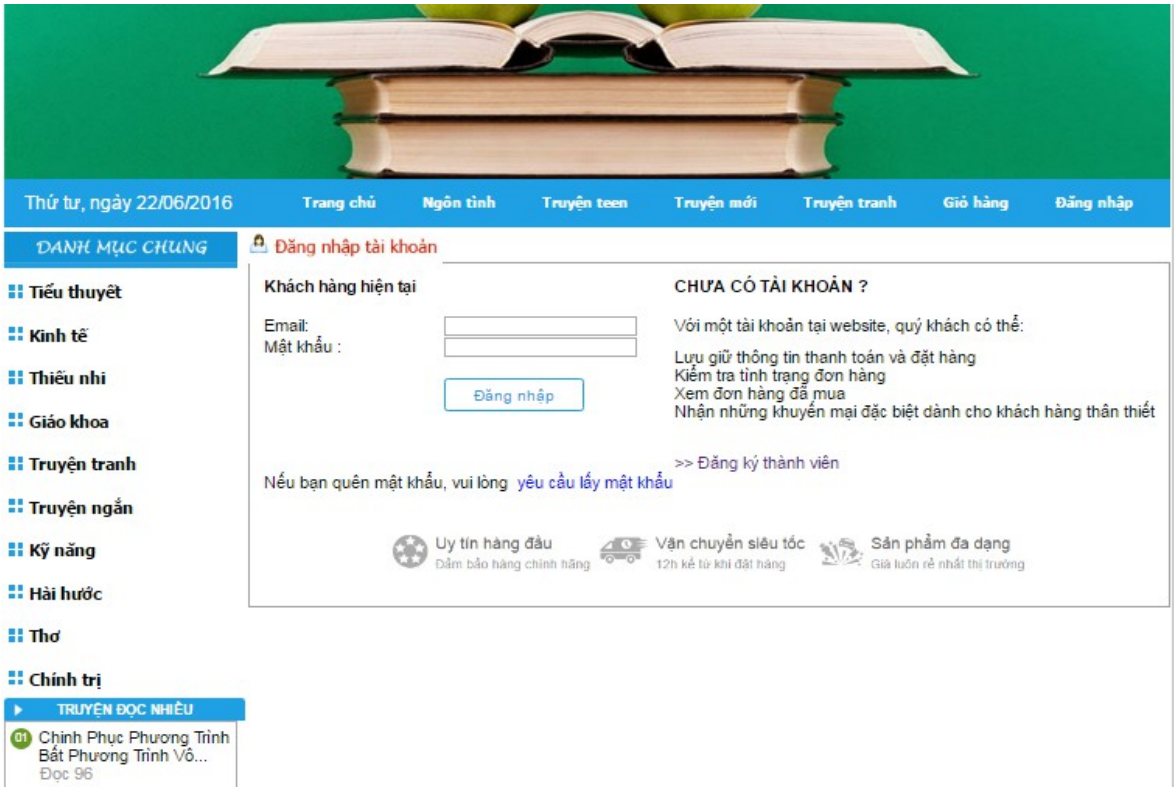
3.2.3. Sơ đồ mô hình cơ sở dữ liệu



Hình 3.6. Sơ đồ mô hình cơ sở dữ liệu.

3.2.4. Một số giao diện minh họa của hệ thống

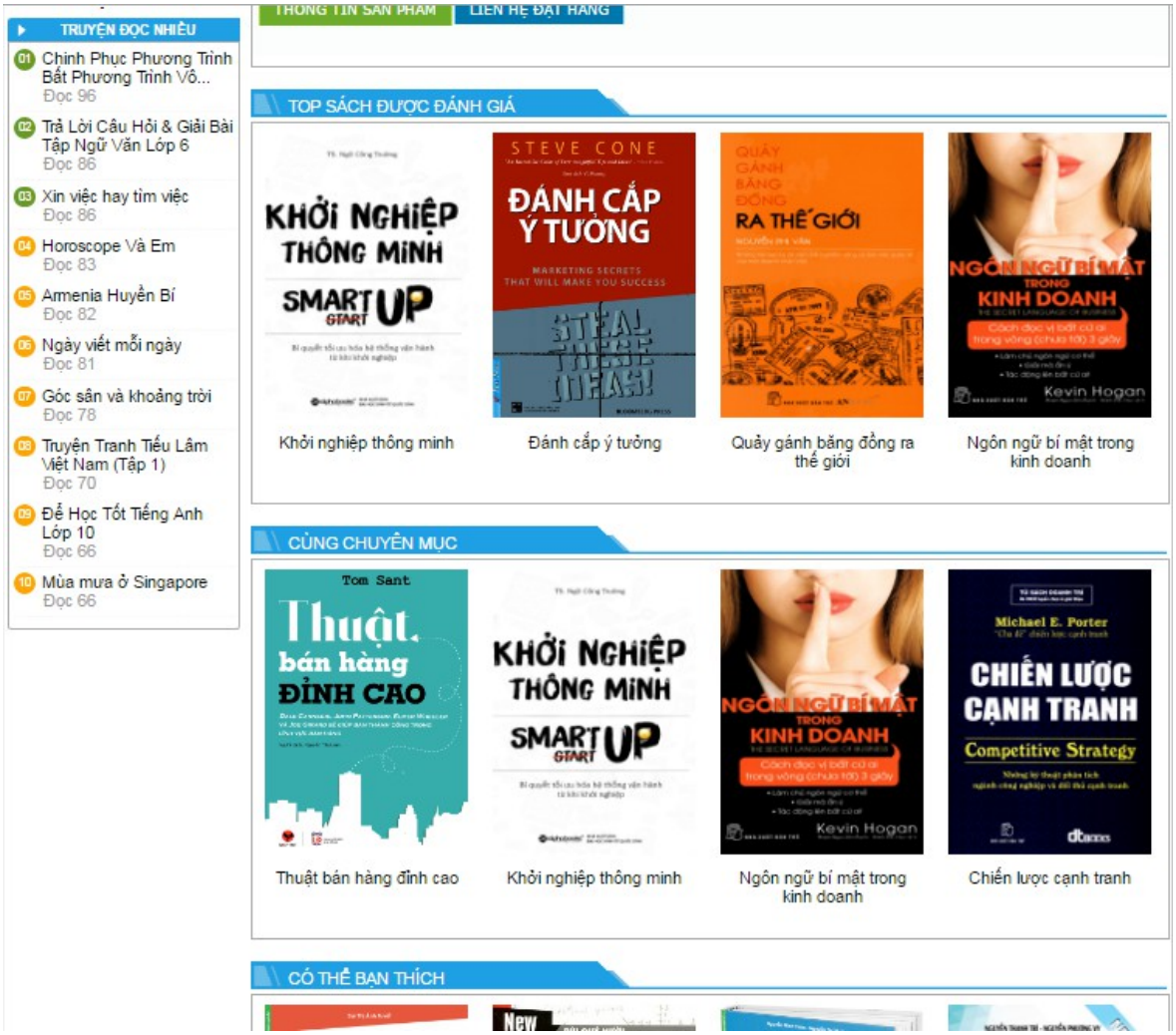
Hình 3.7. Giao diện chính của website.



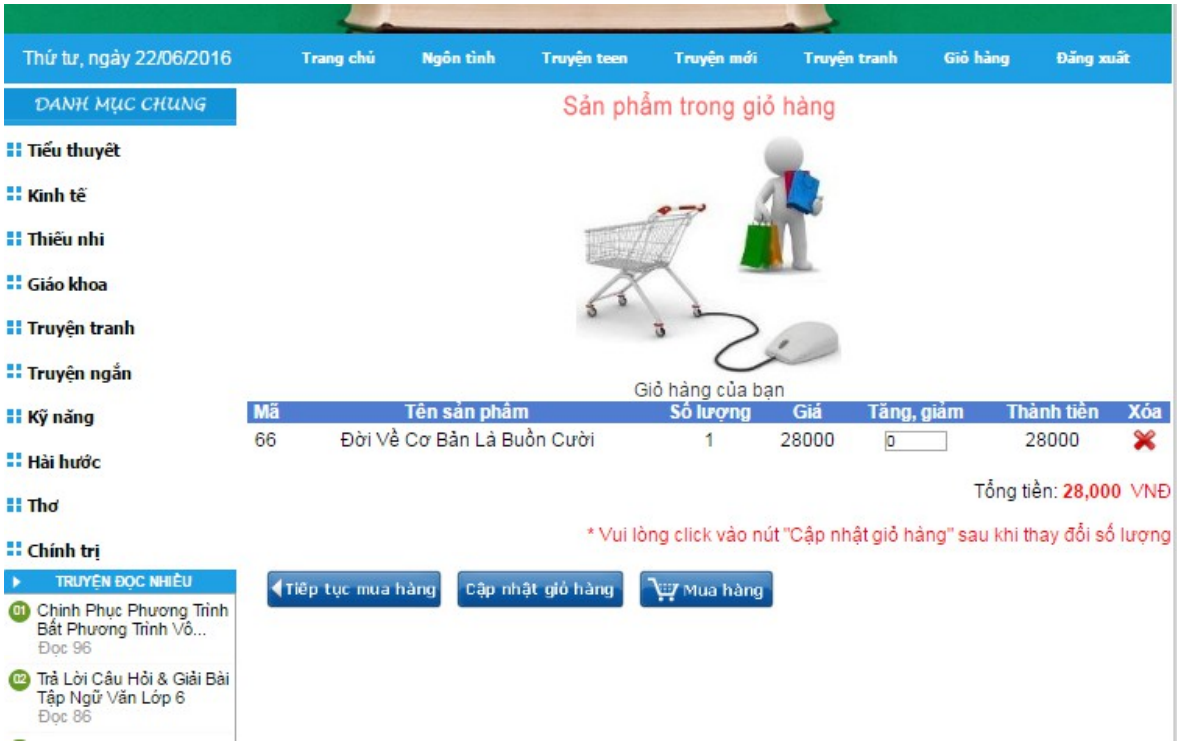
Hình 3.8. Giao diện khi khách hàng đăng nhập hệ thống.

Hình 3.9. Giao diện khi khách hàng đăng ký thành viên mới.

Hình 3.10. Giao diện thông tin sản phẩm.



Hình 3.11. Giao diện sản phẩm gợi ý cho người dung.



Hình 3.12. Giao diện khi khách hàng thực hiện tác vụ mua hàng.

3.3. TIỂU KẾT CHƯƠNG 3

Trong chương 3 luận văn tập trung cài đặt ứng dụng của hệ thống gợi ý bán sách trực tuyến. Một số các kết quả chạy mô phỏng với mục đích cho chúng ta thấy được những ưu điểm của hệ thống gợi ý theo lọc cộng tác. Hệ thống gợi ý bán sách

trực tuyến đã được xây dựng hoàn chỉnh và thực nghiệm từ phản hồi từ người dùng cho thấy hệ thống đã đưa ra những lời gợi ý khá phù hợp.

KẾT LUẬN

Trong bài luận văn này, tôi đã trình bày mô hình láng giềng trong lọc cộng tác – mô hình tư vấn dựa trên độ tương tự trực tiếp giữa hai người dùng hoặc sản phẩm. Trong mô hình này, tôi tính toán độ tương tự giữa hai người dùng, từ đó đưa ra dự đoán đánh giá của người dùng với sản phẩm mới. Đối với những người dùng mới thì chúng tôi sẽ lọc dựa trên một số thuộc tính thu thập từ người dùng để tư vấn những sản phẩm.

Bên cạnh đó, tôi đã cài đặt, xây dựng hệ thống bán hàng trực tuyến hoàn chỉnh có tích hợp kỹ thuật lọc cộng tác để gợi ý sản phẩm cho khách hàng. Qua đó, giúp người đọc có thể nắm được một quy trình xây dựng hệ thống gợi ý trong thực tế, đây là công việc vẫn chưa thấy đề cập đến trong các nghiên cứu liên quan

- **Hướng phát triển:**

Hướng nghiên cứu trong tương lai của tôi là kiểm nghiệm lại thuật toán này dựa trên bộ dữ liệu của hệ thống bán hàng trực tuyến (sau một thời gian vận hành) và đưa ra một số cải tiến cho giải thuật.

TÀI LIỆU THAM KHẢO

1. Tiếng Việt:

[1]. Nguyễn Hùng Dũng và Nguyễn Thái Nghe (2014), *Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc cộng tác*. Tạp chí Khoa học Trường Đại học Cần Thơ, số 31a, trang 36-51.

2. Tiếng Anh:

[2]. Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl (2001), *Item-based Collaborative Filtering Recommendation Algorithms*, *WWW '01: Proceedings of the 10th International Conference on World Wide Web, Seite 285--295. New York, NY, USA, ACM.*

[3]. Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan (2011), *Collaborative Filtering Recommender Systems*. Foundations and Trends® in Human–Computer Interaction: Vol. 4: No. 2, pp 81-173.

[4]. Prem Melville and Vikas Sindhwani (2010), *Recommender Systems*, IBMT.J.Watson Research Center, Yorktowwn Heights, NY 10598.

[5]. J. Ben Schafer, Dan Freankowski, Jon Herlocker, and Shilad Sen (2007), *Collaborative Filtering Recommender Systems*, pp. 291-324, Springer-Verlag Berlin, Heidelberg, Vol. 4321.

[6]. Xiongcai Cai, Michael Bain, Alfred Krzywicki, Wayne Wobcke, Yang Sok Kim, Paul Compton and Ashesh Mahidadia (2010), *Learning Collaborative Filtering and Its Application to People to People Recommendation in Social Networks*. Sydney, Ustralia Dec. 13, ISBN: 978-0-7695-4256-0, pp: 743-748.

[7]. Prem Melville and Vikas Sindhwani (2010), *Recommender Systems,* IBMT.J.Watson Research Center, Yorktowwn Heights, NY 10598.