ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

LÊ THANH HUYỀN

PHƯƠNG PHÁP LỌC CỘNG TÁC VÀ ỨNG DỤNG TRONG HỆ THÔNG TIN TƯ VẤN

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2015

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

LÊ THANH HUYỀN

PHƯƠNG PHÁP LỌC CỘNG TÁC VÀ ỨNG DỤNG TRONG HỆ THÔNG TIN TƯ VẤN

Chuyên ngành: Khoa học máy tính
Mã số: 60.48.01.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: PGS.TS ĐOÀN QUANG BAN

THÁI NGUYÊN - 2015

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này của tự bản thân tôi tìm hiểu, nghiên cứu dưới sự hướng dẫn của PGS. TS Đoàn Văn Ban. Các chương trình do chính bản thân tôi lập trình, các kết quả là hoàn toàn trung thực. Các tài liệu tham khảo được trích dẫn và chú thích đầy đủ.

Tác giả

Lê Thanh Huyền

LỜI CẢM ƠN

Tôi xin bầy tỏ lời cảm ơn chân thành tới tập thể các Thầy cô Viện Công nghệ thông tin - Viện Hàn Lâm Khoa học và công nghệ Việt Nam, các Thầy cô giáo Đại học Công nghệ thông tin và truyền thông - Đại học Thái Nguyên đã dạy dỗ chúng tôi trong suốt quá trình học tập chương trình cao học tại trường.

Đặc biệt tôi xin bày tỏ lòng biết ơn sâu sắc tới Thầy giáo PGS.TS Đoàn Văn Ban đã quan tâm, định hướng và đưa ra những góp ý, gợi ý, chỉnh sửa quý báu cho tôi trong quá trình làm luận văn tốt nghiệp. Cũng như bạn bè, đồng nghiệp, gia đình và người thân đã quan tâm giúp đỡ, chia sẻ với tôi trong suốt quá trình làm luận văn tốt nghiệp.

Dù đã cố gắng nhưng chắc chắn sẽ không tránh khỏi những thiếu sót vì vậy rất mong nhận được sự đóng góp ý kiến của các Thầy, Cô và các bạn để luận văn được hoàn thiện hơn.

Tôi xin chân thành cảm ơn!

Thái Nguyên, tháng 9 năm 2015

Lê Thanh Huyền

MỤC LỤC

Trang

DANH MỤC CÁC CHỮ VIẾT TẮT

KÝ HIỆU	DIỄN GIẢI
IF	Information Filtering (Lọc thông tin)
IR	Information Retrieval (Truy vấn thông tin)
RS	Recommender System (Hệ thống tư vấn)
u	User (Người dùng)
i	Item (Sản phẩm)

DANH MỤC CÁC BẢNG

Trang

DANH MỤC CÁC HÌNH

Trang

MỞ ĐẦU

Xã hội loài người chứng kiến sự phát triển mạnh mẽ và sôi động của thông tin trong mọi lĩnh vực đặc biệt là sự gia tăng không ngừng lượng thông tin khổng lồ đến từ hàng trăm kênh truyền hình, hàng triệu băng hình, sách, báo, tạp chí, tài liệu thông qua các hệ thống giao dịch điện tử. Vì vậy người dùng sẽ gặp khó khăn trong việc lựa chọn thông tin hữu ích. Nhiều nhà khoa học máy tính trên thế giới nhiệt tình hưởng ứng và quan tâm nghiên cứu phương pháp hạn chế ảnh hưởng của vấn đề quá tải thông tin đối với người dùng, thúc đẩy một lĩnh vực nghiên cứu mới đó là lọc thông tin.

Lọc thông tin *(Infomation Filtering)* [1] là lĩnh vực nghiên cứu quá trình lọc bỏ những thông tin không thích hợp và cung cấp thông tin thích hợp đến với mỗi người dùng. Lọc thông tin được xem là một phương pháp hiệu quả hạn chế tình trạng quá tải thông tin được quan tâm nhiều nhất hiện nay.

Hệ tư vấn *(Recommender System)* [1,2] là hệ thống có khả năng tự động phân tích, phân loại, lựa chọn và cung cấp cho người dùng những thông tin, hàng hóa hay dịch vụ mà họ quan tâm. Hệ tư vấn được xem như một biến thể điển hình có vai trò quan trọng trong lọc thông tin. Nhiều hệ tư vấn đã được thương mại hóa và triển khai thành công, tiêu biểu là hệ tư vấn của các hãng Amazon.com, Netflix.com, Procter & Gamble.

Hệ tư vấn được xây dựng dựa trên hai kỹ thuật lọc thông tin chính: Lọc theo nội dung (*Content-Based Filtering)* và lọc cộng tác *(Collaborative Filterin*g) [1]. Lọc theo nội dung khai thác những khía cạnh liên quan đến nội dung thông tin sản phẩm hoặc người dùng đã từng sử dụng hay truy nhập trong quá khứ để tạo nên tư vấn. Trái lại, lọc cộng tác khai thác những khía cạnh liên quan đến thói quen sở thích của người sử dụng sản phẩm của cộng đồng người dùng có cùng sở thích để tạo nên tư vấn.

So với lọc theo nội dung, lọc cộng tác không phải phân tích, bóc tách, hiểu, đánh chỉ mục cho các đặc trưng nội dung sản phẩm, lọc cộng tác có thể

lọc hiệu quả trên nhiều dạng sản phẩm khác nhau như hàng hóa, sữa, ảnh, tài liệu. Chính vì vậy tác giả đã lựa chọn đề tài *"Phương pháp lọc cộng tác và ứng dụng trong hệ thông tin tư vấn"* để thực hiện trong khuôn khổ luận văn thạc sĩ chuyên ngành khoa học máy tính.

Đối tượng và phạm vi nghiên cứu

- Nghiên cứu phương pháp lọc cộng tác dựa trên bộ nhớ, phương pháp lọc cộng tác dựa trên mô hình và phương pháp lọc cộng tác kết hợp bộ nhớ và mô hình.

- Nghiên cứu lọc cộng tác dựa trên sản phẩm với thuật toán tính độ tương tự, lọc cộng tác dựa trên mô hình đồ thị với thuật toán dựa trên mô hình đồ thị người dùng - sản phẩm nhằm cải thiện độ chính xác của lọc thông tin cho hệ tư vấn và thuật toán lọc bằng phương pháp đồng huấn luyện theo sản phẩm và người dùng. Đặc biệt xây dựng ứng dụng hệ thông tin tư vấn sản phẩm sữa dành cho người tiêu dùng.

Hướng nghiên cứu của đề tài

Tập trung nghiên cứu hai vấn đề chính.

1. Trình bày các phương pháp lọc thông tin, ứng dụng của các phương pháp lọc thông tin, hệ thống thông tin tư vấn với kiến trúc tổng quan của hệ thống lọc thông tin, lọc thông tin và các hệ tư vấn.

2. Nghiên cứu lọc cộng tác dựa trên sản phẩm với thuật toán tính độ tương tự, lọc cộng tác dựa trên mô hình đồ thị với thuật toán dựa trên mô hình đồ thị người dùng - sản phẩm nhằm cải thiện độ chính xác của lọc thông tin cho hệ tư vấn và thuật toán lọc bằng phương pháp đồng huấn luyện theo sản phẩm và người dùng.

Phương pháp nghiên cứu

- Nghiên cứu lý thuyết: Nghiên cứu các khái niệm về lọc thông tin, trong đó đi sâu vào nghiên cứu lọc cộng tác. Nghiên cứu thuật toán tính độ tương tự, phương pháp biểu diễn đồ thị và phương pháp lọc dựa vào lọc đồng huấn luyện.

- Nghiên cứu thực nghiệm: Xây dựng phần mềm ứng dụng hệ thông tin tư vấn sản phẩm sữa dành cho người tiêu dùng.

Ý nghĩa khoa học của đề tài

- Khai thác được thuật toán tính độ tương tự.

- Khai thác phương pháp biểu diễn đồ thị và phương pháp dự đoán trên đồ thị người dùng.

- Khai thác được thuật toán lọc đồng huấn luyện theo sản phẩm và lọc đồng huấn luyện theo người dùng.

Bố cục luận văn

Chương 1: Phương pháp lọc tin.

Trình bày tổng quan về các phương pháp lọc thông tin và hệ thống thông tin tư vấn.

Chương 2: Một số phương pháp lọc cộng tác.

Trình bày thuật toán lọc cộng tác dựa trên sản phẩm, thuật toán dựa trên mô hình đồ thị người dùng - sản phẩm và thuật toán đồng huấn luyện.

Chương 3: Chương trình ứng dụng.

Xây dựng chương trình ứng dụng sản phẩm sữa dành cho người tiêu dùng.

Chương 1
PHƯƠNG PHÁP LỌC TIN

1.1. Các phương pháp lọc thông tin

Lọc thông tin *(Information Filtering)* [1] là lĩnh vực nghiên cứu quá trình lọc bỏ những thông tin không thích hợp và cung cấp thông tin thích hợp đến với mỗi người dùng. Lọc thông tin được xem là một phương pháp hiệu quả hạn chế tình trạng quá tải thông tin được quan tâm nhiều nhất hiện nay. Có 3 phương pháp lọc thông tin.

1.1.1. Phương pháp lọc tin theo nội dung

Lọc theo nội dung là phương pháp thực hiện dựa trên việc so sánh nội dung thông tin hay mô tả hàng hóa, nhằm tìm ra những sản phẩm tương tự với những gì mà người dùng đã từng quan tâm để giới thiệu cho họ những sản phẩm này [3]. Các phương pháp tiếp cận cho lọc theo nội dung có nguồn gốc từ lĩnh vực truy vấn thông tin, trong đó mỗi sản phẩm được biểu diễn bằng một hồ sơ sản phẩm, mỗi người dùng được biểu diễn bằng một hồ sơ người dùng. Phương pháp dự đoán nội dung nguyên bản của sản phẩm thực hiện dựa vào việc xem xét các hồ sơ sản phẩm có mức độ phù hợp cao với hồ sơ người dùng.

1.1.1.1 Bài toán lọc theo nội dung

Bài toán lọc theo nội dung được phát biểu như sau. Cho P = {p_1, p_2,.., p_N} là tập gồm N sản phẩm. Nội dung sản phẩm p∈ P được ký hiệu là *Content(p)* được biểu diễn thông qua tập K đặc trưng nội dung của P. Tập các đặc trưng sản phẩm p được xây dựng bằng các kỹ thuật truy vấn thông tin để thực hiện mục đích dự đoán những sản phẩm khác tương tự với p.

Cho *U* = { u_1, u_2,.., u_M} là tập gồm M người dùng. Với mỗi người dùng *u∈ U,* gọi *ContentBasedProfile(u)* là hồ sơ người dùng u. Hồ sơ của người dùng u thực chất là lịch sử truy cập hoặc đánh giá của người đó đối với các sản phẩm. *ContentBasedProfile(u)* được xây dựng bằng cách phân tích

nội dung các sản phẩm mà người dùng u đã từng truy nhập hoặc đánh giá dựa trên các kỹ thuật truy vấn thông tin.

Bài toán lọc theo nội dung khi đó là dự đoán những sản phẩm mới có nội dung thích hợp với người dùng dựa trên tập hồ sơ sản phẩm *Content(p)* và hồ sơ người dùng *ContendBasedProfile(u)*.

1.1.1.2 Các phương pháp pháp lọc theo nội dung

Lọc theo nội dung được tiếp cận theo hai xu hướng: Lọc dựa trên bộ nhớ và lọc dựa trên mô hình.

Lọc nội dung dựa vào bộ nhớ

Lọc nội dung dựa vào bộ nhớ là phương pháp sử dụng toàn bộ tập hồ sơ sản phẩm và tập hồ sơ người dùng để thực hiện huấn luyện và dự đoán. Trong phương pháp này, các sản phẩm mới được tính toán và so sánh với tất cả hồ sơ người dùng. Những sản phẩm mới có mức độ tương tự cao nhất với hồ sơ người dùng sẽ được dùng để tư vấn cho người dùng này.

Lọc nội dung dựa vào mô hình

Lọc nội dung dựa trên mô hình là phương pháp sử dụng tập hồ sơ sản phẩm và tập hồ sơ người dùng để xây dựng nên mô hình huấn luyện. Mô hình dự đoán sau đó sẽ sử dụng kết quả của mô hình huấn luyện để sinh ra tư vấn cho người dùng. Trong cách tiếp cận này, lọc nội dung có thể sử dụng các kỹ thuật học máy như mạng Bayes, phân cụm, cây quyết định, mạng nơron nhân tạo để tạo nên dự đoán.

Pazzani và Billsus [9] sử dụng bộ phân loại Bayes dựa trên những đánh giá "*thích*" hoặc "*không thích*" của người dùng để phân loại các sản phẩm. Trong đó, phương pháp ước lượng xác suất sản phẩm p_j có thuộc lớp C_i hay không dựa vào tập các đặc trưng nội dung $k_{1j},..,k_{nj}$ của sản phẩm đó.

$$P(C_i| k_{1j}\&k_{2j} \& .. \&k_{nj}) \qquad (1.1)$$

Panzanni và Billsus giả thiết các đặc trưng nội dung xuất hiện độc lập nhau, vì vậy xác suất ở trên tương ứng với:

$$P(C_i) \prod P(k_{xj}| C_i) \qquad (1.2)$$

x: là người dùng chạy từ 1 $\longrightarrow$ n

Vì $P(k_{xj}| C_i)$ và $P(C_i)$ có thể ước lượng dựa vào tập dữ liệu huấn luyện. Do vậy, sản phẩm p_j được xem là thuộc lớp C_i nếu xác suất $P(C_i| k_{1j}\&k_{2\,j} \& .. \&k_{n\,j})$ có giá trị cao nhất thuộc lớp này.

Solombo[5] đề xuất mô hình lọc thích nghi, trong đó chú trọng đến việc quan sát mức phù hợp của tất cả các sản phẩm.

1.1.2. Phương pháp lọc tin theo cộng tác

Không giống như lọc theo nội dung, lọc cộng tác khai thác những khía cạnh liên quan đến thói quen sở thích của người sử dụng sản phẩm để đưa ra dự đoán các sản phẩm mới cho người dùng này. So với lọc theo nội dung, lọc cộng tác không phải phân tích, bóc tách, hiểu, đánh chỉ mục cho các đặc trưng nội dung sản phẩm. Chính vì vậy, lọc cộng tác có thể lọc hiệu quả trên nhiều dạng sản phẩm khác nhau như hàng hóa, sữa, ảnh, tài liệu [4]. Cùng trên một hệ tư vấn, người dùng sẽ được tư vấn nhiều loại mặt hàng khác nhau cho dù các mặt hàng này có thể biểu diễn trên không gian các đặc trưng nội dung khác nhau.

1.1.2.1 Bài toán lọc cộng tác

Ký hiệu $U= \{u_1, u_2,\ldots, u_N\}$ 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 P có thể là hàng hóa, sữa, 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 U đưa ra đánh giá của mình cho một số sản phẩm p_j P bằng một trọng 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} = Ø 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 tư vấn 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 những mặt hàng mà u_a chưa đánh giá (r_{aj} = Ø), trên cơ sở đó tư vấn cho u_a những sản phẩm được đánh giá cao.

Bảng 1.1 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 {Ø, 1, 2, 3, 4, 5}. Giá trị r_{ij} = Ø 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 1.1.Ví dụ về 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	Ø
u_3	3	Ø	2	4
u_4	4	4	Ø	Ø
u_5	4	?	5	5

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.

1.1.2.2 Các phương pháp lọc cộng tác

Cũng giống như lọc theo nội dung, lọc cộng tác 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. Cách tiếp cận cụ thể mỗi phương pháp được thực hiện như sau.

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 các sản phẩm cho người dùng hiện thời. Về thực chất, đây là phương pháp học lười hay học dựa trên ví dụ được sử dụng trong học máy. Phương pháp được thực hiện theo hai bước: Tính toán mức độ tương tự và bước tạo nên dự đoán.

• Tính toán mức 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.

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.

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 tần suất và tần suất xuất hiện ngược, lọc cộng tác sử dụng cosin giữa hai véc tơ biểu diễn đánh giá của người dùng.

Lọc cộng tác dựa vào mô hình

Khác với phương pháp dựa trên bộ nhớ, phương pháp lọc dựa trên mô hình [2] 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. Ưu điểm của của phương pháp này là mô hình huấn luyện có kích thước nhỏ hơn rất nhiều so với ma trận đánh giá và thực hiện dự đoán nhanh. Mô hình chỉ cần cập nhật lại khi có những thay đổi lớn và chỉ thực hiện lại phần xây dựng mô hình.

Mô hình mạng Bayes

Mô hình mạng Bayes [6] biểu diễn mỗi sản phẩm như một đỉnh của đồ thị, trạng thái của đỉnh tương ứng với giá trị đánh giá của người dùng đối với sản phẩm đã được đánh giá. Cấu trúc của mạng được phận biết từ tập dữ liệu huấn luyện. Breese [6] đề xuất phương pháp mạng Bayes đơn giản cho lọc cộng tác, trong đó những đánh giá chưa biết được tính toán theo công thức (1.3). Breese giả thiết các giá trị đánh giá được xem xét như những số nguyên nằm giữa *0* và *n.* Đánh giá chưa biết của người dùng *u* đối với sản phẩm *p* là $r_{u,p}$ được ước lượng thông qua những đánh giá trước đó của người dùng *u*. Gọi $P_u = \{ p' \in P | r_{up'} \neq \emptyset \}$. Khi đó, đánh giá chưa biết của người dùng *u* đối với sản phẩm *p* được tính theo công thức

$$r_{up} = E(r_{up}) = \sum_{i=0}^{n} i \times \Pr\left(r_{up} = i \middle| r_{up'}, p' \in P_u\right) \qquad (1.3)$$

Billsus và Pazzani [9] chuyển đổi dữ liệu có nhiều mức đánh giá thành dữ liệu nhị phân. Khi đó, ma trận đánh giá được chuyển đổi thành ma trận bao gồm đặc trưng nhị phân. Việc chuyển đổi này làm cho việc sử dụng mô hình mạng r trở nên thuận tiện hơn. Tuy nhiên, kết quả phân loại theo các đặc trưng nhị phân không phản ánh đúng các bộ dữ liệu thực

Mô hình phân cụm

Một cụm là tập các đối tượng dữ liệu có các phần tử trong cụm giống nhau nhiều nhất, và khác nhau nhiều nhất đối với các phần tử thuộc các cụm khác. Các phương pháp phân cụm cho lọc cộng tác được sử dụng để phân chia tập người dùng (hoặc tập sản phẩm) thành các cụm người dùng (hoặc sản phẩm) có sở thích tương tự nhau. Khi đó, người dùng (hoặc sản phẩm) thuộc cụm nào sẽ được dự đoán và tư vấn các sản phẩm được đánh giá cao trong cụm đó [7]. Độ đo dùng để ước lượng mức độ giống nhau giữa các đối tượng dữ liệu thường được sử dụng là khoảng cách Minkowski và độ tương quan Pearson.

Cho hai đối tượng dữ liệu $X = (x_1, x_2,..,x_n)$, $Y = (y_1, y_2,..,y_n)$. Khi đó, khoảng cách Minkowski được định nghĩa theo công thức

$$d(X,Y) = \sqrt[q]{\sum_{i=1}^{n} |x_i - y_i|^q}$$

Trong đó, n là số chiều của X và Y; x_i, y_i là giá trị thành phần thứ i của X và Y; q là một số nguyên dương. Nếu q =1, thì d(X,Y) là khoảng cách Minkowski. Nếu q =2, thì d(X,Y)là khoảng cách Euclid.

Sarwar và Herlocker [7] cùng các cộng sự sử dụng các kỹ thuật phân cụm chia tập người dùng thành các cụm. Phương pháp dự đoán sử dụng các thuật toán dựa trên bộ nhớ như độ tương quan Pearson để thực hiện trên mỗi cụm dữ liệu.

Si và Jin [8] đề xuất mô hình phân cụm bằng mô hình pha trộn linh hoạt (Flexible Mixture Model). Phương pháp phân cụm đồng thời cho cả người dùng và sản phẩm và cho phép mỗi người dùng hoặc sản phẩm có thể thuộc nhiều cụm khác nhau, sau đó mô hình hóa các cụm người dùng và các cụm sản phẩm độc lập nhau để thực hiện dự đoán. Kết quả thử nghiệm đã chứng tỏ phương pháp cho lại kết quả tốt hơn so với phương pháp dựa trên độ tương quan Pearson và mô hình định hướng (Aspect Model).

Mô hình ngữ nghĩa ẩn:

Mô hình ngữ nghĩa ẩn cho lọc cộng tác dựa vào các kỹ thuật thống kê, trong đó các tham biến ẩn được thiết lập trong một mô hình hỗn hợp để khám phá ra cộng đồng người dùng phù hợp với mẫu hồ sơ thích hợp.

Si và Jin [8] đề xuất mô hình đa thức (*Multinomial Model*) phân loại tập người dùng với giả thiết chỉ có một kiểu người dùng duy nhất. Marlin [5] đề xuất mô hình pha trộn đa thức (*Multinomial Mixture Model*), kết hợp với mô hình định hướng để tạo nên mô hình hồ sơ đánh giá người dùng(*User Rating Profile*) với giả thiết có nhiều kiểu người dùng và các đánh giá mỗi người dùng độc lập nhau. Marlin khẳng định, hồ sơ đánh giá người dùng thực hiện tốt hơn so với mô hình định hướng và mô hình pha trộn đa thức. Mô hình phân loại và hồi qui: Cho tập gồm N véctơ M chiều {xi}. Mục tiêu

của phân loại hay hồi qui là dự đoán chính xác giá trị đầu ra tương ứng {ci}. Trong trường hợp phân loại, ci nhận một giá trị từ một tập hữu hạn gọi là tập các nhãn. Trong trường hợp hồi qui, ci có thể nhận một giá trị thực. Để áp dụng mô hình phân loại cho lọc cộng tác, mỗi sản phẩm (hoặc người dùng) được xây dựng một bộ phân loại riêng. Bộ phân loại cho sản phẩm y phân loại tập người dùng dựa trên những người dùng khác đã đánh giá sản phẩm y. Các bộ phân loại được tiến hành huấn luyện độc lập nhau trên tập các ví dụ huấn luyện.

1.1.3. Phương pháp lọc tin kết hợp

Lọc kết hợp hay còn gọi là phương pháp lai [1] là phương pháp kết hợp giữa cộng tác và lọc nội dung nhằm tận dụng lợi thế và tránh những hạn chế của mỗi phương pháp. So với các phương pháp khác, lọc kết hợp cho lại kết quả dự đoán tốt và có nhiều triển vọng áp dụng trong các ứng dụng thực tế.

1.1.3.1 Bài toán lọc kết hợp

Ngoài người dùng U, tập sản phẩm P, ma trận lọc cộng tác R như đã được trình bày ở trên, kí hiệu C = $\{c_1,c_2,\ldots, c_k\}$ là tập K đặc trưng biểu diễn nội dung thông tin các sản phẩm $p \in P$ hoặc người dùng $u \in U$. Ví dụ nếu $p \in P$ là một loại sữa , khi đó ta có thể biểu diễn sữa thông qua các đặc trưng c_i *"thể loại", "thành phần", "hãng sản xuất"*, và các đặc trưng khác của sữa; nếu $u \in U$ là một người dùng thì ta có thể xem xét các đặc trưng ci: *"tuổi"*, *" giới tính"*, *"nghề nghiệp"* và các đặc trưng khác phản ánh thông tin người dùng.

Bài toán của lọc kết hợp là dự đoán cho người dùng hiện thời u_a những sản phẩm $p_k \in P$ chưa được u_a đánh giá dựa trên ma trận đánh giá r_{ij} và các đặc trưng nội dung C = $\{c_1,c_2,\ldots, c_k\}$.

1.1.3.2 Các phương pháp lọc kết hợp

Lọc kết hợp được tiếp cận theo 4 xu hướng chính: Kết hợp tuyến tính, kết hợp đặc tính của lọc nội dung vào lọc cộng tác, kết hợp đặc tính của lọc cộng tác vào lọc nội dung và xây dựng mô hình hợp nhất giữa lọc cộng tác và lọc nội dung.

Kết hợp tuyến tính [3] là phương pháp xây dựng hai lược đồ lọc nội dung và lọc cộng tác độc lập nhau. Kết quả dự đoán của toàn bộ mô hình có thể được lựa chọn từ phương pháp cho kết quả tốt hơn. Ưu điểm của phương pháp này là kế thừa được phương pháp biểu diễn và tính toán vốn có của các phương pháp. Nhược điểm lớn nhất của mô hình này là cho kết quả không cao vì chưa có sự kết hợp hiệu quả giữa nội dung và đánh giá người dùng.

Kết hợp đặc tính của lọc nội dung và lọc cộng tác là phương pháp dựa trên các kỹ thuật lọc cộng tác thuần túy nhưng vẫn duy trì hồ sơ người dùng *ContentBasedProfile(u)* như một tham biến tham khảo khi tính toán sự tương tự giữa các cặp người dùng. Phương pháp có thể phát hiện ra những sản phẩm tương tự với hồ sơ người dùng hoặc không tương tự với hồ sơ người dùng. Trong trường hợp dữ liệu thưa hoặc người dùng mới, mức độ tương tự giữa hồ sơ người dùng và sản phẩm sẽ được xem xét đến để tạo nên dự đoán.

Kết hợp đặc tính cuả lọc cộng tác và lọc nội dung là phương pháp xem xét các đánh giá người dùng của lọc cộng tác như một thành phần trong mỗi hồ sơ người dùng. Phương pháp dự đoán thực hiện theo lọc nội dung thuần túy và so sánh với kết quả dựa trên biểu diễn hồ sơ người dùng mở rộng. Phương pháp phổ biến nhất thể hiện theo mô hình này là sử dụng các kỹ thuật giảm số chiều cho hồ sơ người dùng trước khi kết hợp với đánh giá người dùng.

1.1.4. Ứng dụng của các phương pháp lọc tin

Lọc thông tin (IF) là lĩnh vực nghiên cứu các quá trình cung cấp thông tin thích hợp, ngăn ngừa và gỡ bỏ thông tin không thích hợp cho mỗi người dùng. Thông tin được cung cấp (còn được gọi là sản phẩm) có thể là văn bản, trang web, phim, ảnh, dịch vụ hoặc bất kỳ dạng thông tin nào được sản sinh ra từ các phương tiện truyền thông. Phạm vi ứng dụng của lọc thông tin trải rộng trong nhiều ứng dụng thực tế khác nhau của khoa học máy tính. Ứng dụng tiêu biểu nhất của lọc thông tin được kể đến là lọc kết quả tìm kiếm trong các máy tìm kiếm (Search Engine), lọc e-mail dựa trên nội dung thư và hồ sơ người dùng, lọc thông tin văn bản trên các máy chủ để cung cấp thông tin cho tập thể hoặc cá nhân thích hợp, loại bỏ những trang thông tin có ảnh

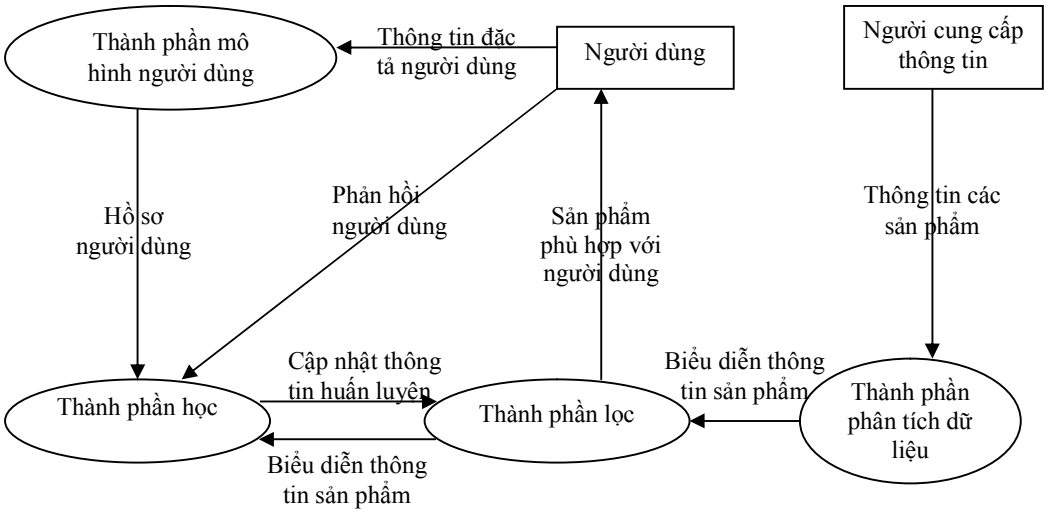
hưởng không tốt đối với người dùng. Đặc biệt, lọc thông tin có vai trò quan trọng cho các hệ thống tư vấn (RS) ứng dụng trong thương mại điện tử.

1.2. Hệ thống thông tin tư vấn

Hệ thống lọc thông tin tư vấn cung cấp cho người dùng những thông tin cần thiết nhất, loại bỏ những thông tin không có giá trị hoặc không thích hợp đối với người dùng. Nguyên lý phổ biến được dùng trong lọc thông tin là nguyên lý dựa vào dữ liệu (Data-Based) và nguyên lý dựa vào tri thức (Knowledge-Based). Các phương pháp lọc có thể được thực hiện dựa vào nội dung thông tin sản phẩm hoặc lọc dựa trên thói quen sở thích người dùng. Các kỹ thuật lọc được phát triển dựa trên nền tảng từ lĩnh vực truy vấn thông tin (Information Retrieval), tách thông tin (Information Extraction), phân loại thông tin (Information Classification). Phạm vi ứng dụng của các hệ thống lọc được áp dụng cho tất cả các mô hình thương mại điện tử thực tế: Khách hàng - Khách hàng (Customer to Customer), Nhà cung cấp - Khách hàng (Business to Customer), Nhà cung cấp - Nhà cung cấp (Business to Business) [6].

1.2.1. Kiến trúc tổng quan của hệ thống lọc thông tin

Một hệ thống lọc thông tin tổng quát bao gồm bốn thành phần cơ bản [6]: Thành phần phân tích dữ liệu (Data Analyser Component), thành phần mô hình người dùng (User Model Component), thành phần học (Learning Component) và thành phần lọc (Filtering Component).



Hình 1.1. Kiến trúc tổng quát của hệ thống lọc thông tin.

• *Thành phần phân tích dữ liệu* có nhiệm vụ thu thập dữ liệu về sản phẩm từ các nhà cung cấp thông tin (ví dụ tài liệu, thư điện tử, sách, báo, tạp chí, sữa, ảnh...). Dữ liệu về sản phẩm được phân tích và biểu diễn theo một khuôn dạng thích hợp, sau đó chuyển đến bộ phận lọc như Hình 1.1.

• *Thành phần mô hình người dùng* có thể "hiện" hoặc "ẩn" dùng để lấy thông tin về người dùng, như giới tính, tuổi, nơi sinh sống và thông tin người dùng đã truy vấn trước đó để tạo nên hồ sơ người dùng. Hồ sơ người dùng sau khi tạo ra được chuyển đến thành phần học để thực hiện nhiệm vụ huấn luyện.

• *Thành phần học* thực hiện huấn luyện trên tập hồ sơ và phản hồi của người dùng theo một thuật toán học máy cụ thể. Thuật toán học lấy dữ liệu từ thành phần mô tả người dùng; lấy dữ liệu về sản phẩm đã được biểu diễn từ thành phần lọc kết hợp với thông tin phản hồi người dùng để thực hiện nhiệm vụ huấn luyện. Kết quả quá trình học được chuyển lại cho bộ phận lọc để thực hiện nhiệm vụ tiếp theo.

• *Thành phần lọc* là thành phần quan trọng nhất của hệ thống, có nhiệm vụ xem xét sự phù hợp giữa hồ sơ người dùng và biểu diễn dữ liệu sản phẩm để đưa ra quyết định phân bổ sản phẩm. Nếu dữ liệu sản phẩm phù hợp với hồ sơ người dùng, sản phẩm sẽ được cung cấp cho người dùng đó. Trong trường hợp ngược lại, hệ thống loại bỏ sản phẩm khỏi danh sách những sản phẩm phân bổ cho người dùng. Người dùng nhận được những sản phẩm thích hợp, xem xét, đánh giá, phản hồi lại cho thành phần học để phục vụ quá trình lọc tiếp theo.

1.2.2. Lọc thông tin và các hệ tư vấn

Hệ tư vấn (RS) là trường hợp riêng của các hệ thống lọc thông tin. Dựa trên thông tin đã có về người dùng, hệ tư vấn xem xét trong số lượng rất lớn hàng hóa hay thông tin và tư vấn cho người dùng một danh sách ngắn gọn nhưng đầy đủ những hàng hóa mà người dùng có khả năng quan tâm.

Sử dụng hệ tư vấn trong các ứng dụng thương mại điện tử sẽ hỗ trợ khách hàng không cần thực hiện các thao tác tìm kiếm sản phẩm, mà chỉ cần lựa chọn hàng hóa hoặc dịch vụ ưa thích do hệ thống cung cấp. Điều này sẽ

làm gia tăng năng lực mua, bán của toàn bộ hệ thống. Chính vì lý do này, hàng loạt các công ty đa quốc gia (Amazon.com, Netflix.com, CDNOW, J.C. Penney, Procter & Gamble..) đã đầu tư và phát triển thành công công nghệ tư vấn để gia tăng hệ thống khách hàng và bán hàng qua mạng [6].

Do là trường hợp riêng của hệ thống lọc tin, hệ tư vấn có nhiều đặc điểm của hệ lọc tin tiêu biểu. Tuy nhiên, do đặc điểm của dữ liệu, người dùng và nội dung, hệ tư vấn cũng như các kỹ thuật được sử dụng có một số khác biệt nhất định. Tùy vào phương pháp lọc tin, các hệ tư vấn được phân loại thành ba loại:

• *Phương pháp tư vấn dựa vào lọc nội dung*: Hệ thống tư vấn cho người dùng những sản phẩm mới có nội dung tương tự với một số sản phẩm họ đã từng mua hoặc từng truy nhập trong quá khứ.

• *Phương pháp tư vấn dựa vào lọc cộng tác*: Người dùng sẽ được tư vấn một số sản phẩm của những người có sở thích giống họ đã từng ưa thích trong quá khứ.

• *Phương pháp tư vấn dựa vào lọc kết hợp*: Hệ thống tư vấn cho người dùng những sản phẩm tương tự với một số sản phẩm họ đã từng mua hoặc từng truy nhập trong quá khứ và sản phẩm của những người có sở thích giống họ đã từng ưa thích trong quá khứ.

Mỗi phương pháp lọc áp dụng cho các hệ tư vấn được phân thành hai hướng tiếp cận: lọc dựa vào bộ nhớ(Memory-Based Filtering) và lọc dựa vào mô hình (Model-Based Filtering).

• *Các phương pháp lọc dựa vào bộ nhớ*: Đây là phương pháp lưu lại toàn bộ các ví dụ huấn luyện. Khi cần dự đoán, hệ thống tìm các ví dụ huấn luyện giống trường hợp cần dự đoán nhất và đưa ra tư vấn dựa trên các ví dụ này. Trường hợp tiêu biểu của lọc dựa vào bộ nhớ là thuật toán K người láng giềng gần nhất. Ưu điểm chính của phương pháp tiếp cận này là đơn giản, dễ cài đặt. Tuy nhiên, phương pháp này có thời gian lọc chậm do việc dự đoán đòi hỏi so sánh và tìm kiếm trên toàn bộ lượng người dùng và sản phẩm.

• *Phương pháp lọc dựa trên mô hình*: Trong phương pháp này, dữ liệu được sử dụng để xây dựng mô hình rút gọn, ví dụ mô hình xác suất hay cây quyết định. Mô hình này sau đó được sử dụng để đưa ra các tư vấn. Phương pháp này cho phép thực hiện việc dự đoán nhanh, do quá trình dự đoán thực hiện trên mô hình đã học trước đó.

1.3. Kết luận

Trong chương này, luận văn đã trình bày khái niệm và các kiến thức cơ sở về các phương pháp lọc thông tin và hệ thông tin tư vấn. Chương 2 tác giả sẽ đi sâu nghiên cứu các phương pháp lọc cộng tác vì phương pháp này có thể lọc hiệu quả trên nhiều dạng sản phẩm khác nhau như hàng hóa, sữa, ảnh, tài liệu.

Chương 2
MỘT SỐ PHƯƠNG PHÁP LỌC CỘNG TÁC

2.1. Lọc cộng tác dựa trên sản phẩm

Giải thuật tư vấn dựa trên sản phẩm nhằm đưa ra các dự đoán cho người dùng bởi đối tượng được xét ở đây là sản phẩm. Quá trình tư vấn bằng phương pháp lọc cộng tác dựa trên sản phẩm sẽ tính toán độ tương tự các sản phẩm, sau đó lựa chọn k sản phẩm tương tự $\{i_1, i_2,\ldots i_k\}$. Khi những sản phẩm có độ tương tự nhất được tìm hết, dự đoán được tính toán dựa trên trung bình của đánh giá người dùng trên những sản phẩm tương tự. Đa số các đề xuất mô tả hai khía cạnh này, cụ thể là việc tính toán độ tương tự và các dự đoán sản phẩm.

Ví dụ minh họa thực tế về một hệ thống lọc cộng tác dựa trên sản phẩm: Giả sử sản phẩm ở đây là sữa, và người dùng là các khách hàng đăng nhập vào 1 hệ thống Webstie để mua sữa. Mỗi người dùng được lưu trữ trên hệ thống với các hồ sơ bao gồm thông tin cá nhân, và các đánh giá của người dùng đó với các loại sữa, đánh giá theo thang điểm từ 1 sao đến 5 sao, với ý nghĩa là đánh giá càng cao thì người dùng càng thích loại sữa đó. Công việc của hệ thống tư vấn là: Khi một người dùng đăng nhập vào hệ thống, hệ thống cần tư vấn những loại sữa cho người dùng đó và những loại sữa được tư vấn đó được dự đoán là người dùng sẽ đánh giá cao. Hệ thống xem xét các loại sữa mà người dùng chưa xem, so sánh độ tương tự giữa loại sữa đó với những sữa khác. Độ tương tự 2 loại sữa được tính dựa trên những người dùng từng đánh giá trên cả 2 loại sữa đó theo 1 thuật toán tính xác suất. Bước cuối cùng của hệ thống tư vấn là dự đoán đánh giá của người dùng với những sữa mà người dùng chưa sử dụng, lựa chọn những sữa được dự đoán có đánh giá cao để đưa vào danh sách tư vấn cho người dùng.

Ví dụ 2.1: Có 9 người dùng đánh giá 9 sản phẩm, với mức độ từ 1 đến 5, nếu người dùng không đánh giá sản phẩm thì giá trị là 0.

Bảng 2.1. Bảng đánh giá người dùng với các sản phẩm

Người dùng	Sản phẩm								
	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
u_1	1	2	1	5	0	0	0	0	0
u_2	2	1	3	0	0	1	0	0	0
u_3	2	3	0	2	3	1	0	0	4
u_4	5	0	2	0	1	0	0	0	2
u_5	0	0	3	5	0	3	4	0	0
u_6	0	1	1	0	3	5	4	1	0
u_7	0	0	0	0	3	4	2	1	0
u_8	0	0	0	0	0	5	1	2	2
u_9	0	0	3	1	0	0	0	2	4

Các bước trong quá trình tư vấn theo phương pháp lọc cộng tác dựa trên sản phẩm:

Bước 1: Tiền xử lý dữ liệu: Dữ liệu được thu thập là những đánh giá sản phẩm của người dùng. Dữ liệu này thường rất lớn tuy nhiên một số đánh giá có thể không có ích trong quá trình tư vấn theo phương pháp lọc cộng tác. Đề xuất được đưa ra để tối ưu dữ liệu đầu vào, một số sản phẩm hoặc người dùng sẽ được loại bỏ nếu người dùng đó đánh giá quá ít sản phẩm, hoặc sản phẩm được quá ít đánh giá.

Bước 2: Xây dựng Ma trận đánh giá: Hàng là các người dùng, Cột là các sản phẩm.

Bước 3: Tính độ tương tự của 2 sản phẩm, xây dựng Ma trận tương tự của các sản phẩm.

Bước 4: Tính dự đoán của người dùng đối với sản phẩm dựa trên những sản phẩm lân cận với sản phẩm dự đoán

2.1.1. Thuật toán tính độ tương tự

Để dự đoán 1 sản phẩm cho 1 người dùng sử dụng phương pháp lọc cộng tác cần xem xét đánh giá của người dùng lên những sản phẩm tương tự

với sản phẩm đó, độ tương tự được xác định dựa vào đánh giá của các người dùng khác đã đánh giá cả 2 sản phẩm. Độ tương tự 2 sản phẩm là 1 xác suất thể hiện 2 sản phẩm đó có tương đồng nhau trên khía cạnh đánh giá của người dùng hay không? Độ tương tự ở đây được hiểu là nếu 2 sản phẩm tương tự nhau thì 1 người dùng thích sản phẩm này sẽ thích sản phẩm kia và ngược lại.

Bước quan trọng trong giải thuật lọc cộng tác dựa trên sản phẩm là tính toán độ tương tự giữa các sản phẩm và sau đó chọn những sản phẩm mà tương đương nhất để sử dụng trong công thức dự đoán. Ý tưởng cơ bản trong tính toán độ tương tự giữa hai sản phẩm i và j là: Chọn các cặp người dùng mà đã đánh giá cả 2 sản phẩm và sau đó áp dụng kỹ thuật tính toán độ tương tự để mô tả độ tương tự $S_{i,j}$.

2.1.1.1 Độ tương tự Cosine

Trong trường hợp này, cả 2 sản phẩm i, j được biểu diễn thông qua 2 véc-tơ cột n chiều, n = $|U_{ij}|$ là số lượng các người dùng cùng đánh giá 2 sản phẩm i và j. Độ tương tự giữa chúng được đo dựa trên tính toán cosine góc giữa 2 véc-tơ đó. Trong ma trận đánh giá m x n, độ tương tự giữa hai sản phẩm i và j, biểu diễn là sim(i,j) được cho bởi công thức:

$$\text{Sim}(i,j) = \text{Cosine}(\vec{i},\vec{j}) = \frac{\vec{i}.\vec{j}}{\left\|\vec{i}\right\|^2 \times \left\|\vec{j}\right\|^2} = \frac{\sum_{u \in Uij} r_{u_i} \times r_{u_j}}{\sqrt{\sum_{u \in U_{ij}} r_{u_i}^2} \sqrt{\sum_{u \in U_{ij}} r_{u_j}^2}} \qquad (2.1)$$

Trong đó:

r_{u_i} : là đánh giá của người dùng u với sản phẩm i

U_{ij}: là tập các người dùng đã đánh giá cả 2 sản phẩm i,j

Mỗi một sản phẩm được đánh giá bởi n người dùng và được xác định như là 1 véc-tơ n chiều trong công thức này, ở đây những người dùng được chọn là những người đã đánh giá cả 2 sản phẩm i và j. Như vậy theo công thức ở trên, kết quả là Cosine của góc hợp giữa 2 véc-tơ đó. Và vì các đánh giá là dương nên, Cosine của 2 véc-tơ bằng 1 thể hiện 2 sản phẩm tương tự nhau hoàn toàn

với những đánh giá của người dùng, cosine của 2 véc-tơ bằng 0, thể hiện 2 sản phẩm này không tương tự nhau.

Dựa vào công thức 2.1 tính độ tương tự và bảng 2.1 đánh giá người dùng với các sản phẩm, ta có:

Bảng 2.2. Bảng tính độ tương tự theo công thức Cosine

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
p_1	1.000	0.891	0.830	0.747	0.646	1.000	0.000	0.000	0.747
p_2	0.891	1.000	0.739	0.824	0.894	0.522	1.000	1.000	1.000
p_3	0.830	0.739	1.000	0.739	0.707	0.659	0.894	0.990	0.992
p_4	0.747	0.824	0.739	1.000	1.000	0.998	1.000	1.000	0.949
p_5	0.646	0.894	0.707	1.000	1.000	0.891	0.949	1.000	0.990
p_6	1.000	0.522	0.659	0.998	0.891	1.000	0.854	0.955	0.614
p_7	0.000	1.000	0.894	1.000	0.949	0.854	1.000	0.713	1.000
p_8	0.000	1.000	0.990	1.000	1.000	0.955	0.713	1.000	0.949
p_9	0.747	1.000	0.992	0.949	0.990	0.614	1.000	0.949	1.000

Bảng 2.2 thể hiện độ tương tự giữa các sản phẩm theo cách tính độ tương tự cosine. Sim(i,j) thể hiện độ tương tự của 2 sản phẩm i, j, với những cặp sản phẩm cùng được đánh giá. Giá trị này dao động từ 0-1, với ý nghĩa 1 là những sản phẩm có giá trị tương tự nhau hoàn toàn, giá trị độ tương tự của 2 sản phẩm càng cao có nghĩa là 2 sản phẩm đó khả năng được đánh giá tương đồng nhau bởi người dùng.

2.1.1.2 Độ tương tự tương quan

Độ tương tự của 2 sản phẩm i, j được cho bởi công thức sau:

$$\text{Sim(i,j)} = \frac{\sum_{u \in Uij} (r_{u_i} - \bar{r_i})(r_{u_j} - \overline{r_j})}{\sqrt{\sum_{u \in U_{ij}} (r_{u_i} - \bar{r_i})^2 \sum_{u \in U_{ij}} (r_{u_j} - \overline{r_j})^2}} \qquad (2.2)$$

Trong đó:

$U_{ij} = \{u \in \underline{U} | \quad r_{u_i} \neq \emptyset \wedge r_{u_j} \neq \emptyset\}|$ là tập tất cả người dùng cùng đánh giá sản phẩm i và sản phẩm j.

r_{u_i} là đánh giá của người dùng u với sản phẩm i.

$\bar{r_i}$ là đánh giá trung bình cho sản phẩm i.

Dựa vào công thức 2.2 tính độ tương tự và bảng 2.1 đánh giá người dùng với các sản phẩm , ta có:

Bảng 2.3. Bảng tính độ tương tự theo công thức tương quan

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
p_1	1.000	0.634	0.526	0.031	0.165	-1.000	0.000	0.000	0.453
p_2	0.634	1.000	-0.140	0.611	0.774	-0.281	1.000	1.000	1.000
p_3	0.526	-0.140	1.000	0.288	-0.669	-0.225	0.486	0.866	0.995
p_4	0.031	0.611	0.288	1.000	1.000	0.497	1.000	-1.000	0.110
p_5	0.165	0.774	-0.669	1.000	1.000	0.584	0.872	1.000	0.954
p_6	-1.000	-0.281	-0.225	0.497	0.584	1.000	0.615	0.855	-0.122
p_7	0.000	1.000	0.486	1.000	0.872	0.615	1.000	0.217	-1.000
p_8	0.000	1.000	0.866	-1.000	1.000	0.855	0.217	1.000	0.857
p_9	0.453	1.000	0.995	0.110	0.954	-0.122	-1.000	0.857	1.000

Bảng 2.3 thể hiện độ tương tự của 2 sản phẩm i, j theo công thức tính độ tương tự tương quan. Khoảng giá trị nằm trong đoạn [-1,1] thể hiện mức độ tương tự theo mức tăng dần. Giá trị độ tương tự càng lớn thể hiện sự tương đồng về mặt đánh giá của 2 sản phẩm i, j. Sự tham gia của giá trị đánh giá trung bình làm tăng tính khách quan đối với các đánh giá lên sản phẩm.

2.1.1.3 Độ tương tự Cosine điều chỉnh

Tính toán độ tương tự sử dụng độ đo Cosine trong trường hợp dựa trên sản phẩm có một sự trở ngại quan trọng: những sự khác nhau trong thang đánh giá giữa các người dùng khác nhau không được đưa vào tài khoản. Độ tương tự Cosine điều chỉnh khắc phục nhược điểm này bằng cách trừ trung bình người dùng tương ứng với mỗi cặp đánh giá. Độ tương tự giữa sản phẩm i và j được cho bởi công thức sau:

$$\text{sim}(i,j)=\frac{\sum_{u\in U_{ij}}(r_{u_i}-\overline{r_u})(r_{u_j}-\overline{r_u})}{\sqrt{\sum_{p\in U_{ij}}(r_{u_i}-\overline{r_u})^2\sum_{p\in P_{xy}}(r_{u_j}-\overline{r_u})^2}} \qquad (2.3)$$

Trong đó:

$U_{ij} = \{u \in U | \quad r_{u_i} \neq Ø \wedge r_{u_j} \neq Ø \quad \}$ là tập tất cả các người dùng đánh giá hai sản phẩm i, j.

$\overline{r_u}$: Là trung bình cộng các đánh giá khác Ø của người dùng u.

r_{u_i} : Là đánh giá của người dùng u với sản phẩm i.

Dựa vào công thức 2.3 tính độ tương tự và bảng 2.1 đánh giá người dùng với các sản phẩm, ta có:

Bảng 2.4. Bảng tính độ tương tự theo công thức Cosine điều chỉnh

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
p_1	1.000	-0.098	0.121	-0.847	-0.992	0.600	0.000	0.000	-0.496
p_2	-0.098	1.000	0.413	-0.600	-0.447	-0.747	-1.000	1.000	1.000
p_3	0.121	0.413	1.000	-0.986	0.000	-0.726	-0.956	0.800	0.894
p_4	-0.847	-0.600	-0.986	1.000	-1.000	-0.083	1.000	1.000	-0.894
p_5	-0.992	-0.447	0.000	-1.000	1.000	0.440	0.447	-1.000	0.600
p_6	0.600	-0.747	-0.726	-0.083	0.440	1.000	-0.109	-0.866	-0.759
p_7	0.000	-1.000	-0.956	1.000	0.447	-0.109	1.000	-0.158	1.000
p_8	0.000	1.000	0.800	1.000	-1.000	-0.866	-0.158	1.000	-0.447
p_9	-0.496	1.000	0.894	-0.894	0.600	-0.759	-1.000	-0.447	1.000

Bảng 2.4 thể hiện độ tương tự của 2 sản phẩm i, j theo công thức tính độ tương tự cosine điều chỉnh. Khoảng giá trị nằm trong đoạn [-1, 1] thể hiện mức độ tương tự theo mức tăng dần. Giá trị độ tương tự càng lớn thể hiện sự tương đồng về mặt đánh giá của 2 sản phẩm i, j. Sự thay đổi của công thức tính độ tương tự này so với công thức tính độ tương tự Cosine là sự tham gia của giá trị đánh giá trung bình $\overline{r_u}$, $\overline{r_u}$ đánh giá trung bình của các người dùng đối với các sản phẩm mà người dùng u đã đánh giá. Xem xét giá trị $\overline{r_u}$ này, giả sử 1 người sử dụng u đánh giá 1 sản phẩm với giá trị đánh giá $[1, 5]$, với người u này cho đánh giá 1 với các sản phẩm họ không thích và 3 với cách

sản phẩm họ rất thích, một người dùng u' đánh giá các sản phẩm họ không thích là 3 và những sản phẩm họ thích là 5, giá trị $\overline{r_u}$ sẽ trở thành giá trị phân biệt giữa thích và không thích, tạo ra sự cân đối hơn với các giá trị r tham gia đánh giá trong công thức tính độ tương tự.

2.1.2. Tính toán dự đoán và tư vấn

Bước quan trọng nhất của hệ thống lọc cộng tác là đưa ra kết quả dự đoán. Phần trên đã đưa ra những sản phẩm tương tự nhất dựa trên độ tương tự, bước tiếp theo là nghiên cứu kỹ mục tiêu xếp hạng của người dùng và sử dụng kỹ thuật để thu được dự đoán.

Dự đoán đánh giá của một người dùng lên một sản phẩm được suy ra từ các đánh giá của người dùng đó trên các sản phẩm lân cận.

2.1.2.1 Công thức dự đoán dựa trên trung bình đánh giá sản phẩm lân cận

Dựa vào công thức đơn giản nhất để dự đoán của 1 người dùng u lên 1 sản phẩm i là dựa vào những người dùng lân cận của u mà đã đánh giá sản phẩm i.

Trong đó:

$$P_{a_i} = \frac{1}{N} \sum_{i' \in T_i \cap S_a} r_{a_{i'}} \qquad (2.4)$$

N: là tổng các sản phẩm lân cận của i đã được a đánh giá.

T_i: là tập hợp các sản phẩm i' lân cận với i mà u đã đánh giá

Vấn đề là chọn ra các sản phẩm i' với các tiêu chí như thế nào, trong bài toán này, kết quả phụ thuộc vào tiêu chí chọn ra các lận cận của sản phẩm i.

Ví dụ: Giả sử với ví dụ được cho ở bảng 2.1. Xét với các i' là những sản phẩm mà u đã đánh giá và i' là lân cận với i nếu $Sim(i, i') \neq 0$

Trường hợp Sim(i, j) được tính theo công thức tính độ tương tự Adjusted Cosine. Áp dụng công thức ta có thể dự đoán như sau cho người dùng u_1 với những sản phẩm mà họ chưa đánh giá như sau:

Bảng 2.5. Bảng dự đoán và tư vấn theo phương pháp tính trung bình dự đoán

u_1	Sản phẩm chưa đánh giá	Dự đoán	Tư vấn(N=3)
1	7	2.667	7,8,5
	8	2.667	
	5	2.667	
	9	2.25	
	6	2.25	

2.1.2.2 Công thức dự đoán dựa trên tổng trọng số

Dự đoán đánh giá của người dùng a với sản phẩm i được cho bởi công thức sau:

$$P_{a_i} = \frac{\sum_{j \in S_a \cap T_i} Sim(i,j) \times r_{a_j}}{\sum_{j \in S_a \cap T_i} |Sim(i,j)|} \qquad (2.5)$$

Trong đó:

S_a: Các sản phẩm mà người dùng a đã đánh giá.

r_{a_j} : Là đánh giá sản phẩm j của người dùng a.

Sim(i,j): Là độ tương tự của 2 sản phẩm i, j.

Ti: Tập các sản phẩm lân cận của sản phẩm i.

Công thức tính toán đánh giá của người dùng a lên sản phẩm i dựa vào những đánh giá của người dùng a lên các sản phẩm tương tự với i. Giá trị dự đoán sẽ nằm trong khoảng [1, 5]. Trường hợp dự đoán sẽ cho kết quả cao (người dùng a thích sản phẩm i) khi những sản phẩm lân cận với i (có độ tương tự cao) được người dùng a đánh giá cao.

Trường hợp Sim(i, j) được tính theo công thức tính độ tương tự Adjusted Cosine. Với số liệu được cho ở Bảng 2.1, cho các lân cận của sản phẩm i là

tập các sản phẩm mà có độ tương tự khác 0. Ta có thể dự đoán như sau cho người dùng u_1 với những sản phẩm mà họ chưa đánh giá như sau:

Bảng 2.6. Bảng dự đoán và tư vấn theo phương pháp Weigth Sum

u_1	Sản phẩm chưa đánh giá	Dự đoán	Tư vấn(N=3)
1	7	3.669	7,8,6
	8	2.786	
	6	2.304	
	5	1.677	
	9	1.237	

2.1.2.3 Công thức dự đoán dựa trên tổng trọng số với đánh giá trung bình của người dùng

$$P_{a_i} = \overline{R_a} + \frac{\sum_{j \in S_a \cap T_i} Sim(i,j) \times (r_{a_j} - \overline{R_a})}{\sum_{j \in S_a \cap T_i} |Sim(i,j)|} \qquad (2.6)$$

Trong đó:

S_a: Các sản phẩm mà người dùng a đã đánh giá.

r_{a_j} : Là đánh giá sản phẩm j của người dùng a.

Sim(i, j): Là độ tương tự của 2 sản phẩm i, j.

Ti: Tập các sản phẩm lân cận của của sản phẩm i.

$\overline{R_a}$: Là đánh giá trung bình của người dùng a.

Công thức tính toán đánh giá của người dùng a lên sản phẩm i dựa vào những đánh giá của người dùng a lên các sản phẩm tương tự với i. Giá trị đánh giá trung bình của người dùng a là R nhằm làm tăng tính cân đối của người dùng a lên sản phẩm. Giá trị R của mỗi người dùng là là một mức để xác định người dùng đó đánh giá cao hay không cao một sản phẩm. Giá trị dự đoán sẽ nằm trong khoảng [1, 5]. Trường hợp dự đoán sẽ cho kết quả cao

(người dùng a thích sản phẩm i) khi những sản phẩm lân cận với i (có độ tương tự cao) được người dùng a đánh giá cao.

Trường hợp sim(i, j) được tính theo công thức tính độ tương tự Adjusted cosine. Với số liệu được cho ở Bảng 2.2, cho các lân cận của sản phẩm i là tập các sản phẩm mà có độ tương tự khác 0. Ta có thể dự đoán như sau cho người dùng u_1 với những sản phẩm mà họ chưa đánh giá như sau:

Bảng 2.7. Bảng dự đoán và tư vấn theo phương pháp tổng trọng số với đánh giá trung bình của người dùng và sử dụng độ tương tự Ajusted Cosine

u_1	Sản phẩm chưa đánh giá	Dự đoán	Tư vấn(N=3)
1	7	3.669	7,8,6
	8	2.786	
	6	2.304	
	5	1.677	
	9	1.237	

2.1.2.4 Công thức dự đoán dựa trên tổng trọng số với trung bình đánh giá lên sản phẩm

$$P_{a_i} = \overline{R_i} + \frac{\sum_{j \in S_a \cap T_i} Sim(i,j) \times (r_{a_j} - \overline{R_j})}{\sum_{j \in S_a \cap T_i} |Sim(i,j)|} \qquad (2.7)$$

Trong đó:

S_a: Là các sản phẩm mà người dùng a đã đánh giá.

r_{a_j} : Là đánh giá sản phẩm j của người dùng a.

Sim(i, j): Là độ tương tự của 2 sản phẩm i, j.

T_i: Tập các sản phẩm lân cận của của sản phẩm i.

$\overline{R_i}$: Là đánh giá trung bình sản phẩm i.

Công thức tính toán đánh giá của người dùng a lên sản phẩm i dựa vào những đánh giá của người dùng a lên các sản phẩm tương tự với i. $\overline{R_i}$ là giá trị đánh giá trung bình sản phẩm i của các người dùng. Giá trị $\overline{R_j}$ của mỗi sản

phẩm là một mức để xác định khi 1 người dùng đánh giá sản phẩm i , thì so với tất cả các người dùng, đánh giá của người đó cho sản phẩm i đó là cao hay không cao. Giá trị dự đoán sẽ nằm trong khoảng [1, 5]. Trường hợp dự đoán sẽ cho kết quả cao (người dùng a thích sản phẩm i) khi những sản phẩm lân cận với i (có độ tương tự cao) được người dùng a đánh giá cao.

Trường hợp sim (i, j) được tính theo công thức tính độ tương tự Adjusted cosine. Với số liệu được cho ở Bảng 2.2, cho các lân cận của sản phẩm i là tập các sản phẩm mà có độ tương tự khác 0. Ta có thể dự đoán như sau cho người dùng u_1 với những sản phẩm mà họ chưa đánh giá như sau:

Bảng 2.8. Bảng dự đoán và tư vấn theo phương pháp tổng trọng số với đánh giá trung bình sản phẩm và sử dụng độ tương tự Ajusted Cosine.

u_1	Sản phẩm chưa đánh giá	Dự đoán	Tư vấn(N=3)
1	7	3.635	7,6,9
	6	2.988	
	9	2.509	
	5	2.347	
	8	1.881	

2.1.3. Thuật toán lọc cộng tác dựa trên sản phẩm

2.1.3.1 Độ tương tự Cosine

```
Function Cosine(a[], b[])
Begin
        ab, a2, b2 = 0;
        For (i=1…n) do
        If (a[i]<>0) and (b[i]<>0) then
        Begin
                ab = a[i] * b[i] + ab;
                a2 = a[i] * a[i] + a2;
                b2 = b[i] * b[i] + b2;
        end;
        return ab/(sqrt(a2)*sqrt(b2));
End;
```

Độ phức tạp thuật toán độ đo Cosine: O(n) với n là số lượng phần tử của mảng a và b

2.1.3.2 Độ tương tự Cosine điều chỉnh

```
Function CosineDieuChinh(a[], b[])
Begin
    ab, a2, b2 = 0;
    tb_a, tb_b = 0, num_a=0, num_b = 0;
    for (i=1..n) do
    begin
        if (a[i]>0) then
        begin
            tb_a = tb_a + a[i];
            num_a = num_a + 1;
        end;
        if (b[i]>0) then
        begin
            tb_b = tb_b + b[i];
            num_b = num_b + 1;
        end;
    end;
    tb_a = tb_a / num_a;
    tb_b = tb_b/num_b;
    For (i=1...n) do
    If (a[i]<>0) and (b[i]<>0) then
    Begin
        ab = (a[i] - tb_a) * (b[i] - tb_b) + ab;
        a2 = (a[i] - tb_a) * (a[i] - tb_a) + a2;
        b2 = (b[i] - tb_b) * (b[i] - tb_b) + b2;
    end;
    return ab/(sqrt(a2)*sqrt(b2));
End;
```

Độ phức tạp thuật toán độ đo Cosine: O(n) với n là số lượng phần tử của mảng a và b

2.1.3.3 Dự đoán dựa trên trung bình đánh giá sản phẩm lân cận

```
Function DuDoan(user, item, lancan)
Begin
    Tong=0; dem=0;
    For (i=1..n) do
        If (R[user,i]>0) and (Cosine[i,item]==lancan) then
        Begin
            Tong = tong + R[user,i];
            Dem = dem+1;
        End;
    return Tong / Dem;
End;
```

Độ phức tạp thuật toán dự đoán dựa trên trung bình đánh giá: O(n) với n là số lượng sản phẩm

2.1.3.4 Dự đoán dựa trên tổng trọng số

```
Function Weighted_Sum(user, item, lancan)
Begin
    S = 0; n = 0;
    For (i=1..n) do
        If (R[user,i]>=0) and (sim[item,i] == lancan) then
        Begin
            S = R[user,i] * Sim[item,i] + S;
            n = Abs(sim[item,i]) + n;
        end;
    return s/n;
End;
```

Độ phức tạp thuật toán dự đoán dựa trên Weighted Sum: O(n) với n là số lượng sản phẩm

2.1.3.5 Dự đoán dựa trên tổng trọng số với trung bình đánh giá lên người dùng

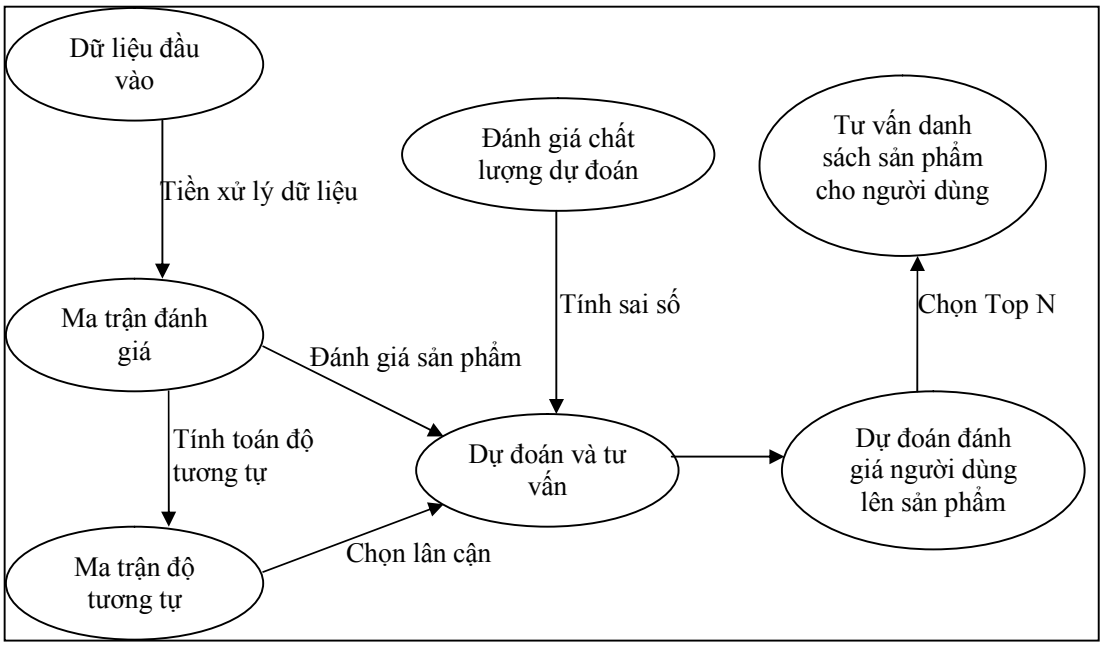
```
Function DuDoan3(user, item, lancan)
Begin
    S = 0; n = 0;
```

```
tb = 0, num = 0;
for (i=1..n) do
        if (R[user,i]>0) then
        begin
                tb = tb + R[user, i];
                num = num + 1;
        end;
tb = tb / num;
For (i=1..n) do
        If (R[user,i]>=0) and (sim[item,i] == lancan) then
        Begin
                S = (R[user,i]  - tb) * Sim[item,i] + S;
                n = Abs(sim[item,i]) + n;
        end;
return tb + s/n;
End;
```

Độ phức tạp thuật toán dự đoán dựa trên tổng trọng số với đánh giá trung bình người dùng: O(n) với n là số lượng sản phẩm.

2.1.4. Đánh giá các yếu tố ảnh hưởng đến độ chính xác kết quả tư vấn

Một hệ tự vấn tốt khi đề xuất được những sản phẩm mà người dùng đó sẽ thích và chọn lựa.



Hình 2.1. Mô hình hệ thống lọc cộng tác dựa trên sản phẩm

2.1.4.1 Đánh giá chất lượng của hệ thống tư vấn

Hệ thống tư vấn sử dụng Độ chính xác (Precision), Độ nhạy(Recall) và F-Measure để đánh giá chất lượng hệ thống tư vấn

Độ chính xác = Số sản phẩm tư vấn chính xác/ Tổng số sản phẩm tư vấn.

Độ nhạy = Số sản phẩm tư vấn chính xác/ Tổng số sản phẩm.

F-Measure = 2(*Độ chính xác * Độ nhạy*)/(*Độ chính xác + Độ nhạy*)

2.1.4.2 Các yếu tố ảnh hưởng đến độ chính xác tư vấn

- Ảnh hưởng của dữ liệu đầu vào: Trong nhiều hệ thống tư vấn, số những đánh giá thu được thường rất nhỏ so với số những đánh giá cần có cho dự đoán. Sự thành công của hệ thống tư vấn lọc cộng tác phụ thuộc vào giá trị của đại đa số những người dùng chính.

- Ảnh hưởng của thuật toán tính độ tương tự: Để xây dựng ma trận tương tự, các giá trị được tính toán theo một công thức được đề xuất ở trên hoặc một công thức khác. Các công thức tính toán khác nhau sẽ cho ra các ma trận đánh giá khác nhau, dẫn đến kết quả tư vấn không đồng nhất. Nhiều chuyên gia đánh giá, với thuật toán có sự tham gia của đánh giá trung bình sẽ có kết quả tốt hơn (Độ tương tự tương quan và Độ tương tự Cosin điều chỉnh)

- Ảnh hưởng của số lượng lân cận tham gia vào dự đoán: Thông qua quá trình tính ma trận tương tự, hệ thống tư vấn sử dụng những sản phẩm lân cận nhất với sản phẩm đang xét để đưa vào dự đoán, số lượng lân cận ảnh hưởng đến chất lượng của kết quả tư vấn.

2.2. Lọc cộng tác dựa trên mô hình đồ thị

Lọc cộng tác có thể xem xét như bài toán tìm kiếm trên đồ thị dựa trên biểu diễn mối quan hệ đánh giá của người dùng đối với các sản phẩm. Mục này trình bày một mô hình đồ thị cho lọc cộng tác.

2.2.1. Phương pháp biểu diễn đồ thị

Mô hình đồ thị cho lọc cộng tác có thể mô tả như sau. Cho ma trận đánh giá đầu vào của lọc cộng tác $R = (r_{ij})$. Gọi $X = (x_{ij})$ là ma trận cấp $N \times M$ có các phần tử được xác định theo công thức(2.8). Trong đó, $x_{ij} = 1$ tương ứng với trạng thái người dùng u_i đã đánh giá sản phẩm p_j, $x_{ij} = 0$ tương ứng với trạng thái người dùng chưa đánh giá sản phẩm p_j.

$$x_{ij} = \begin{cases} 1 & \text{if } r_{ij} \neq \theta \\ 0 & \text{otherwise} \end{cases} \qquad (2.8)$$

Đồ thị biểu diễn đánh giá của người dùng đối với các sản phẩm (Gọi tắt là Người dùng - Sản phẩm) G = (V, E) được biểu diễn theo ma trận X, trong đỉnh $V = U \cup P$ (U là tập người dùng, P là tập sản phẩm); tập cạnh E bao gồm tập các cạnh biểu diễn đánh giá của người dùng đối với sản phẩm. Cạnh nối giữa đỉnh $u_i \in U$ và đỉnh $p_j \in P$ được thiết lập nếu người dùng ui đã đánh giá sản phẩm p_j ($x_{ij} = 1$). Trọng số của mỗi cạnh được lấy tương ứng là r_{ij}. Như vậy, trong biểu diễn này đồ thị Người dùng- Sản phẩm có hai loại cạnh: Cạnh có trọng số dương $r_{ij} = +1$ biểu diễn người dùng u_i"thích" sản phẩm p_j, cạnh có trọng số âm $r_{ij} = -1$ biểu diễn người dùng u_i "không thích" sản phẩm p_j.

Ví dụ 2.2. Hệ gồm 5 người dùng U = {u_1, u_2, u_3, u_4, u_5}, 7 loại sữa P ={p_1, p_2, p_3, p_4, p_5, p_6, p_7,}. Ma trận đánh giá r_{ij} được cho trong bảng 2.9. Giả sử p_1, p_2, p_4, p_5, p_6 có đặc trưng sữa dành cho trẻ c1 "suy dinh dưỡng"; p_3, p_4, p_5, p_7 có đặc trưng c2 "cao to khỏe".

Bảng 2.9. Ma trận đánh giá R

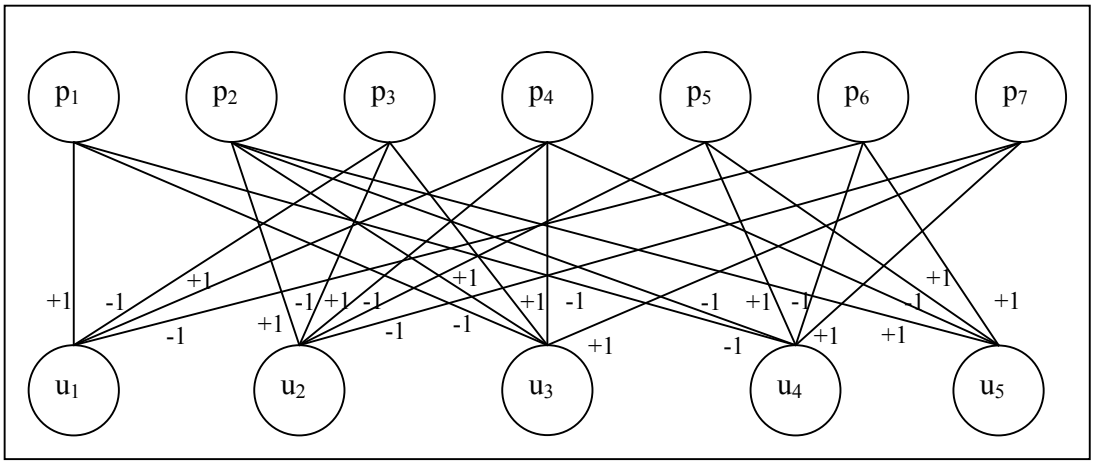
Người dùng	Sản phẩm						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	1	Ø	Ø	1	Ø	1	Ø
u_2	Ø	1	1	1	1	Ø	1
u_3	1	1	1	1	Ø	Ø	1
u_4	1	1	Ø	Ø	1	1	1
u_5	?	1	?	1	1	1	?

với ma trận đánh giá R được cho trong Bảng 2.9 thì ma trận X được thể hiện như Bảng 2.10.

Bảng 2.10. Ma trận X biểu diễn đánh đồ thị Người dùng- Sản phẩm

Người dùng	Sản phẩm						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	1	0	0	1	0	1	0
u_2	0	1	1	1	1	0	1
u_3	1	1	1	1	0	0	1
u_4	1	1	0	0	1	1	1
u_5	0	1	0	1	1	1	0

Đồ thị Người dùng- Sản phẩm có hai loại cạnh: Cạnh có trọng số dương r_{ij} = +1 biểu diễn người dùng u_i"thích" sản phẩm p_j, cạnh có trọng số âm r_{ij} = -1 biểu diễn người dùng u_i "không thích" sản phẩm p_j. Khi đó, đồ thị được biểu diễn như Hình 2.2.



Hình 2.2. Đồ thị người dùng - sản phẩm

2.2.2. Phương pháp dự đoán trên đồ thị người dùng - sản phẩm

Các phương pháp lọc cộng tác dựa trên độ tương quan thực hiện bằng cách xác định những người dùng tương tự nhất với người dùng hiện thời để tạo nên tư vấn. Trong ví dụ trên dễ dàng nhận thấy u_5 tương tự nhất với u_2, u_3 và u_4 vì u_5, u_2, u_3 cùng "thích" p_2 và u_5, u_4 cùng "thích" p_5. Dựa trên mức độ tương tự này, các sản phẩm p_3, p_4 và p_7 sẽ được tư vấn cho người dùng u_5.

Cách làm trên có thể được thực hiện dễ dàng trên mô hình đồ thị bằng cách xem xét các đường đi độ dài 3 từ đỉnh người dùng đến đỉnh sản phẩm, những sản phẩm nào có nhiều số đường đi độ dài 3 từ đỉnh người dùng hiện thời đến đỉnh sản phẩm sẽ được dùng để tạo nên tư vấn. Ví dụ ta cần phân bổ sản phẩm cho người dùng u_5, các đường đi u_5-p_5-u_4-p_7, u_5-p_2-u_2-p_4, u_5-p_2-u_3-p_3, u_5-p_2-u_3-p_7 được xem xét đến trong khi dự đoán các sản phẩm cho u_5. Những sản phẩm có nhiều đường đi nhất đến u5 sẽ được dùng để tư vấn. Ví dụ p_7 có nhiều đường đi độ dài 3 hơn so với $p_3$và p_4(u_5-p_5-u_4-p_7, u_5-p_2-u_3-p_7) sẽ được tư vấn cho u_5.

Hơn thế nữa, phương pháp lọc dựa trên độ tương quan sẽ không bao giờ được xem xét đến p_1 trong các khả năng tư vấn vì u_5 và u_1 được xác định là không tương tự nhau. Điều này không đúng trong trường hợp dữ liệu thưa của lọc cộng tác, u_5 và u_1 không tương tự nhau vì chúng có quá ít dữ liệu đánh giá để thực hiện tính toán. Nhược điểm này có thể khắc phục trên mô hình đồ

thị bằng cách mở rộng phương pháp dự đoán đến các đường đi độ dài lẻ lớn hơn 3 (5, 7, 9...). Những sản phẩm có nhiều đường đi nhất đến nó được dùng để tư vấn cho người dùng hiện thời. Với cách làm này, p_1 cũng được xem xét đến vì có đường đi độ dài 5 (u_5-p_2-u_2-p_4-u_1-p_1). Phương pháp dự đoán trên đồ thị Người dùng - Sản phẩm có thể được thực hiện thông qua các bước sau:

2.2.2.1. Tách đồ thị Người dùng-Sản phẩm thành các đồ thị con

Trong số các đường đi từ ui đến p_j, ta xem xét đến hai loại đường đi: Đường đi theo các cạnh có trọng số dương (ví dụ đường đi u_5-p_2-u_3-p_3) và đường đi theo các cạnh có trọng số âm (ví dụ đường đi u_5-p_4-u_3-p_1). Để tính toán hiệu quả cho mỗi loại đường đi, ta tách đồ thị Người dùng-Sản phẩm thành hai đồ thị con: Đồ thị con chỉ bao gồm các cạnh có trọng số dương và đồ thị con chỉ bao gồm các cạnh có trọng số âm.

Cho đồ thị Người dùng - Sản phẩm G =(V, E) được biểu diễn theo ma trận $X= (x_{ij})$ cấp N×M. Ký hiệu $X^+ = (x^+_{ij})$ ma trận cấp N×M được xác định theo công thức (2.9).Ký hiệu $X^- = (x^-_{ij})$ là ma trận cấp N×M được xác định theo công thức (2.10).

$$x^+_{ij} = \begin{cases} 1 & \text{if } r_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2.9)$$

$$x^+_{ij} = \begin{cases} 1 & \text{if } r_{ij} < 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2.10)$$

Đồ thị $G^+ = (V, E^+)$ được biểu diễn theo ma trận X^+ có tập đỉnh đúng bằng tập đỉnh của G, có tập cạnh E^+ bao gồm các cạnh có trọng số dương của G.

$$E^+ = \{e = (ui, pj) \in E \mid r_{ij} = 1 \qquad (2.11)$$

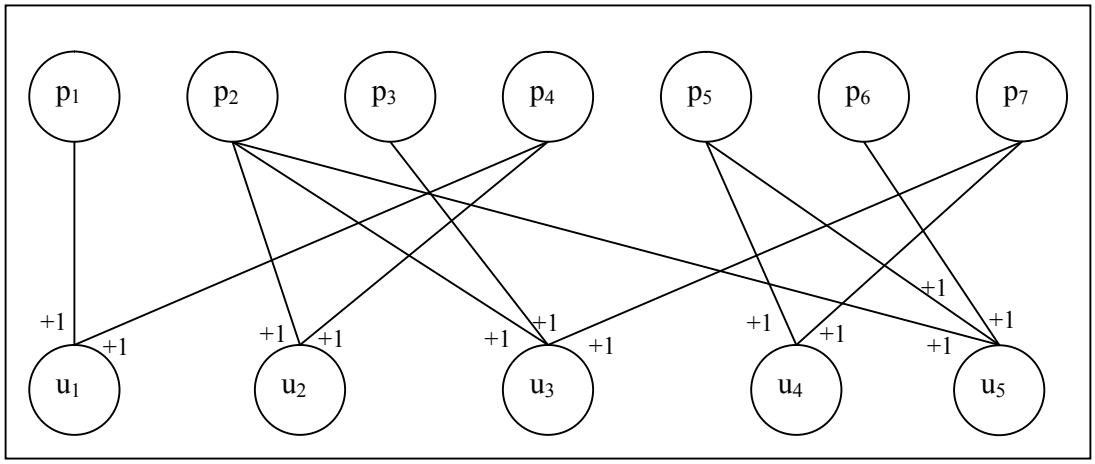
Đồ thị $G^- = (V, E^-)$ được biểu diễn theo ma trận X có tập đỉnh đúng bằng tập đỉnh của G, có tập cạnh E bao gồm các cạnh có trọng số âm của G.

$$E^- = \{e = (ui, pj) \in E \mid r_{ij} = -1 \qquad (2.12)$$

Ví dụ 2.3: Với ma trận đánh giá R được cho trong Bảng 2.9, đồ thị G được biểu diễn theo ma trận X trong Bảng 2.10 thì ma trận X^+, X^- được thể hiện trong Bảng 2.11 và Bảng 2.12. Đồ thị G^+, G tương ứng được biểu diễn theo Hình 2.3 và Hình 2.4.

Bảng 2.12. Ma trận X^+ biểu diễn các đánh giá thích hợp Sản phẩm Người dùng

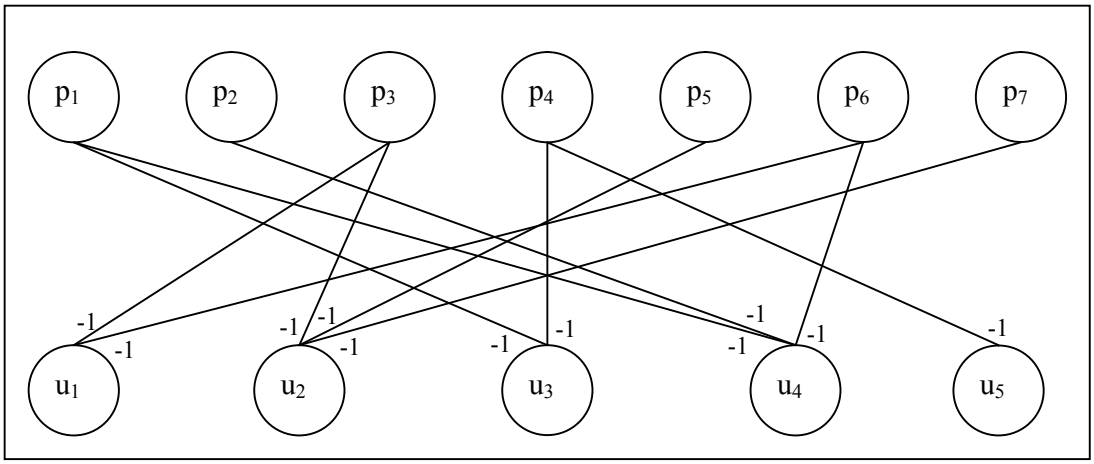
Người dùng	Sản phẩm						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	1	0	0	1	0	0	0
u_2	0	1	0	1	0	0	0
u_3	0	1	1	0	0	0	1
u_4	0	0	0	0	1	0	1
u_5	0	1	0	0	1	1	0



Hình 2.3. Đồ thị G biểu diễn cách đánh giá thích hợp

Bảng 2.12. Ma trận X^- biểu diễn các đánh giá không thích hợp

Người dùng	Sản phẩm						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	0	0	1	0	0	1	0
u_2	0	0	1	0	1	0	1
u_3	1	0	0	1	1	0	0
u_4	1	1	0	0	1	1	0
u_5	0	0	0	1	1	0	0



Hình 2.4. Đồ thị G biểu diễn cách đánh giá không thích hợp

2.2.2.2. Phương pháp dự đoán trên đồ thị có trọng số dương G^+

Trọng số đường đi từ đỉnh người dùng ui đến đỉnh sản phẩm p_j theo các cạnh có trọng số dương được ghi nhận là một số dương phản ánh mức độ "thích"của sản phẩm đối với người dùng. Những đường đi có độ dài lớn sẽ được đánh trọng số thấp, những đường đi có độ dài nhỏ được đánh trọng số cao. Những sản phẩm nào có trọng số cao sẽ được dùng để tư vấn cho người dùng hiện thời.

Phương pháp dự đoán trên đồ thị G^+ được Huang đề xuất dựa trên việc tính toán trọng số các đường đi từ đỉnh người dùng đến đỉnh sản phẩm [7]. Những sản phẩm nào có trọng số cao nhất sẽ được dùng để tư vấn cho người dùng hiện thời.

Để ý rằng, đồ thị G, G^+, G^- đều là những đồ thị hai phía, một phía là các đỉnh người dùng, phía còn lại là các đỉnh sản phẩm. Do vậy, các đường đi từ đỉnh người dùng đến đỉnh sản phẩm luôn có độ dài lẻ.

Đối với đồ thị hai phía, số các đường đi độ dài L xuất phát từ một đỉnh bất kỳ thuộc phía người dùng đến đỉnh bất kỳ thuộc phía sản phẩm được xác định theo công thức 2.13, trong đó X là ma trận biểu diễn đồ thị hai phía, X^T là ma trận chuyển vị của X, L là độ dài đường đi.

if L=1

if L=3,5,7...

$$X = \begin{cases} X \\ X.X^T X_\alpha^{L-2} \end{cases} \quad (2.13)$$

Để ghi nhận trọng số của các đường đi từ đỉnh sản phẩm đến đỉnh người dùng trên đồ thị G^+ sao cho những đường đi dài có trọng số thấp, những đường đi ngắn có trọng số cao, ta sử dụng hằng khử nhiễu $\alpha(0<\alpha\leq 1)$ theo công thức (2.14), trong đó X^+ là ma trận biểu diễn đồ thị G^+,$(X^+)^T$ là ma trận chuyển vị của X^+, L là độ dài đường đi. Thuật toán dự đoán trên đồ thị G^+ được thể hiện trong:

$$(X^+)_\alpha^L = \begin{cases} \alpha.X^+ & if\ \ L=1 \\ \alpha^2.X^+.(X^+)^T(X^+)_\alpha^{L-2} & if\ \ L=3,5,7... \end{cases} \quad (2.14)$$

Thuật toán dự đoán trên đồ thị G^+

Đầu vào:

- *Ma trận X^+ là biểu diễn của đồ thị G^+;*

Đầu ra:

- *K sản phẩm có trọng số cao nhất chưa được người dùng đánh giá*

Các bước thực hiện:

Bước 1. Tìm trọng số các đường đi độ dài lẻ L trên đồ thị G^+ sao cho các đường đi độ có dài nhỏ được đánh trọng số cao, các đường đi có độ dài lớn được đánh trọng số thấp.

$$(X^+)_\alpha^L = \begin{cases} \alpha.X^+ & if\ \ L=1 \\ \alpha^2.X^+.(X^+)^T(X^+)_\alpha^{L-2} & if\ \ L=3,5,7... \end{cases}$$

Bước 2. Sắp xếp các sản phẩm theo thứ tự giảm dần của trọng số $(X^+)_\alpha^L$

Bước 3. Chọn K sản phẩm có trọng số cao nhất chưa được đánh giá để tư vấn cho người dùng hiện thời.

Độ phức tạp thuật toán dự đoán trên đồ thị G^+ là $O(L.N^{2.376})$. Trong đó, L là độ dài đường đi từ đỉnh người dùng đến đỉnh sản phẩm, N là số lượng người dùng.

Ví dụ 2.4:

Ví dụ với ma trận X^+ biểu diễn đồ thị G^+ trong Bảng 2.12, lấy $\alpha = 0.5$, L = 5. Giả sử ta cần tư vấn K = 2 sản phẩm cho người dùng u5, khi đó thuật toán thực hiện như sau:

Bước 1: số các đường đi độ dài 5 từ đỉnh người dùng đến đỉnh sản phẩm được xác định theo công thức (2.14). Khi đó,

$$(X^+)^3_{0.5} = \begin{pmatrix} 0.250 & 0.125 & 0.000 & 0.375 & 0.000 & 0.000 & 0.000 \\ 0.125 & 0.500 & 0.125 & 0.375 & 0.125 & 0.125 & 0.125 \\ 0.000 & 0.625 & 0.375 & 0.125 & 0.250 & 0.125 & 0.500 \\ 0.000 & 0.250 & 0.125 & 0.000 & 0.375 & 0.125 & 0.375 \\ 0.000 & 0.625 & 0.125 & 0.125 & 0.500 & 0.375 & 0.250 \end{pmatrix}$$

$$(X^+)^5_{0.5} = \begin{pmatrix} 0.15625 & 0.18750 & 0.03125 & 0.28125 & 0.03125 & 0.03125 & 0.03125 \\ 0.12500 & 0.59375 & 0.18750 & 0.34375 & 0.25000 & 0.18750 & 0.25000 \\ 0.03125 & 0.81250 & 0.37500 & 0.21875 & 0.43750 & 0.25000 & 0.56250 \\ 0.00000 & 0.43750 & 0.18750 & 0.06250 & 0.37500 & 0.18750 & 0.37500 \\ 0.03125 & 0.81250 & 0.25000 & 0.21875 & 0.56250 & 0.37500 & 0.43750 \end{pmatrix}$$

Bước 2: Sắp xếp các sản phẩm theo thứ tự giảm dần của trọng số cho người dùng u_5 ta nhận được: p_2, p_5, p_7, p_6, p_3, p_4, p_1.

Bước 3: Chọn K=2 sản phẩm chưa được người dùng đánh giá có trọng số cao để tư vấn cho u^5 ta nhận được: p_3, p_7.

2.2.2.3. Phương pháp dự đoán trên đồ thị có trọng số âm G^-

Trọng số đường đi từ đỉnh người dùng ui đến đỉnh sản phẩm theo các cạnh có trọng số âm được ghi nhận là một số âm phản ánh mức độ "không thích" của người dùng đối với sản phẩm. Những đường đi có độ dài lớn sẽ được đánh trọng số cao, những đường đi có độ dài nhỏ được đánh trọng số thấp. Những sản phẩm nào có trọng số thấp được loại bỏ ra khỏi danh sách các sản phẩm cần tư vấn cho người dùng hiện hiện thời.

Để xem xét ảnh hưởng các đánh giá "không thích" vào quá trình dự đoán, ta có thể ước lượng mức độ đóng góp của các đánh giá này trên đồ thị G bằng cách phủ định lại phương pháp dự đoán trên đồ thị G^+.

Cụ thể phương pháp thay thế việc dự đoán trên đồ thị G^+ bằng đồ thị G^-

Thay việc ước lượng trọng số đường đi từ đỉnh người dùng đến đỉnh sản phẩm dài sẽ có trọng số thấp, đường đi ngắn có trọng số cao bằng việc ước lượng trọng số các đường đi dài có trọng số cao, đường đi ngắn có trọng số thấp. Thay việc sử dụng hằng số khử nhiễu +α bằng hằng số khử nhiễu -α để trọng số các đường đi luôn âm và tăng dần theo độ dài đường đi. Thay việc sắp xếp các sản phẩm theo thứ tự giảm dần của trọng số bằng việc sắp xếp các sản phẩm theo thứ tự tăng dần của trọng số.Thay quá trình phân bổ các sản phẩm có trọng số cao cho người dùng hiện thời bằng việc loại bỏ các sản phẩm có trọng số thấp.

Thuật toán dự đoán trên đồ thị G^-.

Đầu vào:

- *Ma trận X^- là biểu diễn của đồ thị G^-;*

Đầu ra:

- *K sản phẩm có trọng số nhỏ nhất chưa được người dùng đánh giá*

Các bước thực hiện:

Bước 1. Tìm trọng số các đường đi độ dài lẻ L trên đồ thị G^- sao cho các đường đi độ có dài nhỏ được đánh trọng số thấp, các đường đi có độ dài lớn được đánh trọng số cao. $(X^-)_\alpha^L$

$$(X^+)_\alpha^L = \begin{cases} \alpha.X^+ & if \;\; L=1 \\ \alpha^2.X^+.(X^+)^T(X^+)_\alpha^{L-2} & if \;\; L=3,5,7... \end{cases}$$

Bước 2. Sắp xếp các sản phẩm theo thứ tự tăng dần của trọng số

Bước 3. Loại bỏ K sản phẩm có trọng số $(X^-)_\alpha^L$ *thấp nhất chưa được đánh giá để tư vấn cho người dùng hiện thời.*

Độ phức tạp thuật toán dự đoán trên đồ thị G là $O(L.N^{2.376})$. Trong đó, L là độ dài đường đi từ đỉnh người dùng đến đỉnh sản phẩm, N là số lượng người dùng.

Ví dụ 2.5:

Ví dụ với ma trận X^- trong Bảng 2.13, lấy L= 5 và α=0.5. Giả sử ta cần gỡ bỏ K=2 các sản phẩm cho người dùng u_5. Khi đó,

Bước 1: Tính được: $(X)^5_{0.5}$

$$(X^+)^5_{0.5} = \begin{pmatrix} -0.18750 & -0.15625 & -0.34375 & -0.03125 & -0.15625 & -0.34375 & -0.15625 \\ -0.03125 & -0.03125 & 0.46875 & -0.00000 & -0.31250 & -0.18750 & -0.31250 \\ -0.34375 & -0.15625 & -0.03125 & -0.28125 & -0.00000 & -0.18750 & -0.00000 \\ -0.50000 & -0.34375 & -0.18750 & -0.18750 & -0.03125 & -0.50000 & -0.03125 \\ -0.12500 & -0.03125 & -0.00000 & -0.15625 & -0.00000 & -0.03125 & -0.00000 \end{pmatrix}$$

*Bước 2: S*ắp xếp các sản phẩm theo thứ tự tăng dần của trọng số, ta nhận được: p_4, p_1, p_2, p_6, p_3, p_5, p_7.

Bước 3: Chọn các sản phẩm có trọng số nhỏ nhất chưa được u_5 đánh giá đưa ra khỏi danh sách các sản phẩm cần tưvấn cho u5, ta nhận được: p_1, p_3.

2.2.2.4. Phương pháp dự đoán theo tất cả đánh giá

Một sản phẩm người dùng "thích" vẫn có thể xuất hiện trong danh sách các sản phẩm loại bỏ khỏi quá trình tư vấn, một sản phẩm người dùng "không thích" vẫn có thể xuất hiện trong danh sách các sản phẩm cần tư vấn. Để ngăn ngừa tình trạng này, luận văn đề xuất phương pháp dự đoán trên tất cả đánh giá.

Phương pháp dự đoán trên đồ thị G^+ chỉ được thực hiện trên những đánh giá "thích" của người dùng đối với sản phẩm, phương pháp dự đoán trên đồ thị G chỉ được thực hiện trên những đánh giá "không thích" của người dùng đối với sản phẩm. Việc bỏ qua những đánh giá "không thích" của người dùng đối với sản phẩm có những ảnh hưởng không nhỏ đến chất lượng dự đoán, vì đánh giá "thích" hay "không thích" đều phản ánh thói quen và sở thích sử dụng sản phẩm của người dùng.

Trong ví dụ trên, nếu thực hiện dự đoán trên đồ thị G^+ thì p_3 được xem là phương án dùng để tư vấn cho u_5. Nếu thực hiện dự đoán trên đồ thị G thì p3 được xem là phương án loại bỏ ra khỏi danh sách các sản phẩm dùng để tư vấn cho u5. Để khắc phục mâu thuẫn này, ta có thể mở rộng phương pháp dự đoán cho tất cả các đánh giá "thích" và "không thích" của người dùng.

Các bước cụ thể của phương pháp được tiến hành.

Đầu vào:

- *Ma trận X^+, X^- là biểu diễn của đồ thịG^+, G^-*

Đầu ra:

- *K sản phẩm có trọng số cao nhất chưa được người dùng đánh giá*

Các bước thực hiện:

Bước 1. *Tính toán ma trận trọng số* $(X^+)^L_\alpha$ *của các đường đi độ dài lẻ L trên ma trận* X^+ *sao cho các đường đi có độ dài nhỏ được đánh trọng số cao, các đường đi có độ dài lớn được đánh trọng số thấp.*

$$(X^+)^L_\alpha = \begin{cases} \alpha.X^+ & if \;\; L=1 \\ \alpha^2.X^+.(X^+)^T(X^+)^{L-2}_\alpha & if \;\; L=3,5,7... \end{cases}$$

Bước 2. *Tính toán ma trận trọng số* $(X^+)^L_\alpha$ *của các đường đi độ dài lẻ L trên ma trận* X^- *sao cho các đường đi có độ dài nhỏ được đánh trọng số thấp, các đường đi có độ dài lớn được đánh trọng số cao.*

$$(X^-)^L_\alpha = \begin{cases} \alpha.X^- & if \;\; L=1 \\ \alpha^2.X^-.(X^-)^T(X^-)^{L-2}_\alpha & if \;\; L=3,5,7... \end{cases}$$

Bước 3 .*Kết hợp ma trận trọng số* $(X)^L_\alpha = (X^-)^L_\alpha + (X^-)^L_\alpha$

Bước 4. *Sắp xếp các sản phẩm theo thứ tự tăng dần của trọng số* x^L_a

Bước 5. *Chọn K sản phẩm có trọng số* x^L_a *cao nhất chưa được đánh giá để tư vấn cho người dùng hiện thời.*

Độ phức tạp thuật toán dự đoán trên tất cả đánh giá là $O(L.N^{2.376})$. Trong đó, L là độ dài đường đi từ đỉnh người dùng đến đỉnh sản phẩm, N là số lượng người dùng.

Ví dụ 2.6:

Với ma trận X^+ trong Bảng 2.12, X^- trong Bảng 2.13, lấy L= 5 và α=0.5. Giả sử ta cần tư vấn K=2 các sản phẩm cho người dùng u_5. Khi đó,

Bước 1:

$$(X^{+})^{5}_{0.5} = \begin{pmatrix} 0.15625 & 0.18750 & 0.03125 & 0.28125 & 0.03125 & 0.03125 & 0.03125 \\ 0.12500 & 0.59375 & 0.18750 & 0.34375 & 0.25000 & 0.18750 & 0.25000 \\ 0.03125 & 0.81250 & 0.37500 & 0.21875 & 0.43750 & 0.25000 & 0.56250 \\ 0.00000 & 0.43750 & 0.18750 & 0.06250 & 0.37500 & 0.18750 & 0.37500 \\ 0.03125 & 0.81250 & 0.25000 & 0.21875 & 0.56250 & 0.37500 & 0.43750 \end{pmatrix}$$

Bước 2:

$$(X^{-})^{5}_{0.5} = \begin{pmatrix} -0.18750 & -0.15625 & -0.34375 & -0.03125 & -0.15625 & -0.34375 & -0.15625 \\ -0.03125 & -0.03125 & 0.46875 & -0.00000 & -0.31250 & -0.18750 & -0.31250 \\ -0.34375 & -0.15625 & -0.03125 & -0.28125 & -0.00000 & -0.18750 & -0.00000 \\ -0.50000 & -0.34375 & -0.18750 & -0.18750 & -0.03125 & -0.50000 & -0.03125 \\ -0.12500 & -0.03125 & -0.00000 & -0.15625 & -0.00000 & -0.03125 & -0.00000 \end{pmatrix}$$

Bước 3:

$$(X)^{5}_{0.5} = \begin{pmatrix} -0.03125 & +0.03125 & -0.03125 & +0.25000 & -0.12500 & -0.32150 & -0.12500 \\ +0.09375 & +0.56250 & -0.28125 & +0.34375 & -0.62550 & +0.00000 & -0.00625 \\ -0.31250 & +0.65625 & +0.34375 & -0.06250 & +0.43750 & +0.62500 & +0.56250 \\ -0.50000 & -0.09375 & +0.00000 & -0.12500 & +0.34375 & -0.31250 & +0.34375 \\ -0.09375 & -0.78125 & +0.25000 & +0.06250 & +0.56250 & +0.37500 & +0.43750 \end{pmatrix}$$

Bước 4: Sắp xếp được: p_2, p_5, p_7, p_6, p_3, p_4, p_1.

Bước cuối cùng của thuật toán ta chọn p_7 và p_3 tư vấn cho u5.

Lọc cộng tác trong trường hợp dữ liệu thưa thường dựa vào phương pháp giảm số chiều ma trận đánh giá. Hạn chế lớn nhất của phương pháp này là có thể mất thông tin trong khi giảm số chiều ma trận. Hạn chế này cũng có thể khắc phục dựa trên việc xem xét và mở rộng độ dài đường đi trên mô hình đồ thị trên.

2.3. Lọc cộng tác dựa vào lọc đồng huấn luyện

Học nửa giám sát đã thu hút nhiều sự chú ý từ các nhà nghiên cứu bởi một số lượng lớn các ví dụ không có nhãn có thể làm tăng hiệu suất cho thuật toán học khi chỉ có một số ví dụ nhỏ hơn là có nhãn. Blum và Mitchell là

những người đầu tiên xem xét việc thiết định bài toán mà tập đặc trưng của mỗi ví dụ có thể được chia thành 2 khung nhìn khác biệt. Xem xét bài toán lọc cộng tác theo cách tiếp cận đồng huấn luyện, thì 2 khung nhìn được xác định ở đây là khung nhìn theo người dùng và khung nhìn theo sản phẩm. Tập các nhãn được xác định có thể là những giá trị rõ ràng (các giá trị nằm trong đoạn [1,5]). Và cặp người dùng - sản phẩm mà người dùng chưa đánh giá sản phẩm là những mẫu huấn luyện cần được xác định nhãn.

2.3.1. Mô tả thuật toán đồng huấn luyện

Thuật toán đồng huấn luyện áp dụng khi tập dữ liệu có sự phân chia đặc trưng tự nhiên. Quá trình đồng huấn luyện được mô tả hình thức như sau: Quá trình đồng huấn luyện được thực hiện như sau. Cho không gian mẫu X = X_1 x X_2 trong đó: X_1, X_2 tương ứng là 2 khung nhìn khác nhau của một mẫu. Mỗi mẫu x đã cho là một cặp (x_1, x_2). Giả sử rằng mỗi khung nhìn là đầy đủ để phân loại đúng. Cho D là một phân phối trên X, và cho C1, C2 lần lượt là các lớp khái niệm được định nghĩa tương ứng trên X_1, X_2.

Giả sử rằng tất cả các nhãn của các mẫu có xác suất khác 0 dưới D là phù hợp với hàm mục đích f1 ∋ C1 và cũng phù hợp với hàm mục đích f2 ∋ C2. Hay nói cách khác, nếu f biểu thị cho khái niệm mục đích kết hợp trên toàn bộ mẫu, thì với bất kỳ mẫu x = (x_1, x_2) được quan sát với nhãn ℓ, chúng ta có f(x) = f(x1) = f(x2) = ℓ. Trong thực tế, thì điều này có nghĩa là D gán xác suất bằng 0 cho bất kỳ mẫu nào mà f(x1) ≠ f(x2).

2.3.2. Thuật toán lọc cộng tác bằng phương pháp đồng huấn luyện theo người dùng

Phương pháp lọc cộng tác bằng đồng huấn luyện theo người dùng được thực hiện thông qua các vòng lặp t. Tại bước khởi tạo t=0, ma trận dự đoán $R^{(0)}$ = ($r^{(0)}_{ij}$) được lấy bằng chính ma trận đánh giá ban đầu R=(r_{ij}).

Các bước cụ thể của phương pháp được tiến hành.

Đầu vào:

- *Khởi tạo ma trận đánh giá* $R^{(0)} = (r^{(0)}_{ij}) = (r_{ij})$

Đầu ra:

- *Ma trận dự đoán* $R^{(t)} = (r^{(t)}_{ij})$

Thuật toán

Bước 1: Khởi tạo số bước lặp ban đầu: t$\leftarrow$0;

Bước 2: Lặp

2.1. Huấn luyện theo người dùng:

a) Tìm tập các người dùng cùng đánh giá cho sản phẩm $S_i^{(t)}$

$$S_i = \{ j \in u : |P_i \cap P_j| \geq \gamma \} \qquad (2.15)$$

γ: Hằng số người dùng cùng đánh giá sản phẩm

Sử dụng công thức độ tương tự tương quan để tính tập tất cả các người dùng cùng đánh giá sản phẩm i và j (u_{ij})

$$u_{ij} = \begin{cases} 0 & if \quad j \notin S_i \\ \dfrac{\sum_{x \in P_i \cap Pj} (r_{ix} - \bar{r}_i)(r_{jx} - \bar{r}_j)}{\sqrt{\sum_{x \in P_i \cap P_j} (r_{ix} - r_i)^2} \sqrt{\sum_{x \in P_i \cap P_j} (r_{jx} - \bar{r}_j)^2}}, & otherwise \end{cases} \qquad (2.16)$$

b) Tìm K_i là người dùng đánh giá sản phẩm cao nhất

$$K_i = \{ j \in S_i : u_{ij} \rightarrow \max \} \qquad (2.17)$$

c) Dự đoán người dùng x với sản phẩm i

$$r_{ix} = \bar{r}_i + \frac{\sum_{j \in K_i} (r_{jx} - \bar{r}_j) u_{ij}}{\sum_{j \in K_i} |u_{ij}|} \qquad (2.18)$$

$\bar{r}_i$: là đánh giá trung bình cộng cho sản phẩm i.

2.2. Huấn luyện theo sản phẩm:

a) Tìm tập các sản phẩm được người dùng đánh giá $C_x^{(t)}$.

$$C_x = \{ y \in P : |u_x \cap u_y \geq \gamma| \} \qquad (2.19)$$

Sử dụng công thức Cosin điều chỉnh để tính độ tương tự giữa hai sản phẩm

$$P_{xy} = \begin{cases} 0 & if \quad y \notin C_x \\ \dfrac{\sum_{i \in Px \cap Py} (r_{ix} - \bar{r}_x)(r_{jy} - \bar{r}_y)}{\sqrt{\sum_{i \in P_x \cap P_y} (r_{ix} - r_x)^2} \sqrt{\sum_{x \in P_i \cap P_j} (r_{jy} - \bar{r}_y)^2}}, & otherwise \end{cases} \qquad (2.20)$$

b) Tìm K_i là sản phẩm mà người dùng đánh giá cao nhất

$$K_i=\{j \in S_i : u_{ij} \to \max\} \qquad (2.21)$$

c) Dự đoán người dùng x với sản phẩm i

$$r_{ix}=\bar{r_i} + \frac{\sum_{j \in K_i}(r_{jx} - \bar{r_j})u_{ij}}{\sum_{j \in K_i}|u_{ij}|} \qquad (2.22)$$

$\bar{r_i}$: là đánh giá trung bình cộng cho sản phẩm i.

2.3. Tăng bước lặp: t←t+1;

Until Converges: không có nhãn phân loại nào được bổ sung vào ma trận dự đoán

2.3.3 Lọc cộng tác bằng phương pháp đồng huấn luyện theo sản phẩm

Gần giống với lọc cộng tác bằng phương pháp đồng huấn luyện theo người dùng, phương pháp này chỉ có một điểm khác trong quá trình huấn luyện đó là thứ tự thực hiện huấn luyện, quá trình huấn luyện theo sản phẩm sẽ được thực hiện trước quá trình huấn luyện theo người dùng.

Thuật toán lọc cộng tác bằng phương pháp đồng huấn luyện theo sản phẩm.

Đầu vào:

- *Khởi tạo ma trận đánh giá* $R^{(0)} = (r^{(0)}_{ij}) = (r_{ij})$

Đầu ra:

- *Ma trận dự đoán* $R^{(t)} = (r^{(t)}_{ij})$

Thuật toán

Bước 1: Khởi tạo số bước lặp ban đầu: t←0;

Bước 2: Lặp.

2.1. Huấn luyện theo sản phẩm:

a) Tìm tập các sản phẩm được người dùng đánh giá $C_x^{(t)}$.

$$C_x=\{y \in P : |U_x \cap U_y \geq \gamma|\} \qquad (2.19)$$

γ : Hằng số sản phẩm được người dùng đánh giá

Sử dụng công thức cosin điều chỉnh để tính tập các sản phẩm được người dùng đánh giá

$$P_{xy} = \begin{cases} 0 & if \quad y \notin C_x \\ \dfrac{\sum_{i \in Px \cap Py} \left(r_{ix} - \overline{r_x}\right)\left(r_{jy} - \overline{r_y}\right)}{\sqrt{\sum_{i \in P_x \cap P_y} \left(r_{ix} - r_x\right)^2} \sqrt{\sum_{x \in P_i \cap P_j} \left(r_{jy} - \overline{r_y}\right)^2}}, & otherwise \end{cases} \qquad (2.20)$$

b) Tìm K_i là sản phẩm được người dùng đánh giá cao nhất.

$$K_i = \left\{ j \in S_i : u_{ij} \to \max \right\} \qquad (2.21)$$

c) Dự đoán sản phẩm i với người dùng x

$$r_{ix} = \overline{r_i} + \frac{\sum_{j \in K_i} \left(r_{jx} - \overline{r_j}\right) u_{ij}}{\sum_{j \in K_i} \left| u_{ij} \right|} \qquad (2.22)$$

2.2. Huấn luyện theo người dùng:

a) Tìm tập các người dùng cùng đánh giá cho sản phẩm $S_i^{(t)}$

$$S_i = \left\{ j \in U : \left| P_i \cap P_j \right| \geq \gamma \right\} \qquad (2.15)$$

γ : Hằng số người dùng cùng đánh giá sản phẩm

Sử dụng công thức độ tương tự tương quan để tính tập tất cả các người dùng cùng đánh giá sản phẩm i và j (u_{ij})

$$u_{ij} = \begin{cases} 0 & if \quad j \notin S_i \\ \dfrac{\sum_{x \in P_i \cap Pj} \left(r_{ix} - \overline{r_i}\right)\left(r_{jx} - \overline{r_j}\right)}{\sqrt{\sum_{x \in P_i \cap P_j} \left(r_{ix} - r_i\right)^2} \sqrt{\sum_{x \in P_i \cap P_j} \left(r_{jx} - \overline{r_j}\right)^2}}, & otherwise \end{cases} \qquad (2.16)$$

b) Tìm K_i là người dùng đánh giá sản phẩm lớn nhất

$$K_i = \left\{ j \in S_i : u_{ij} \to \max \right\} \qquad (2.17)$$

c) Dự đoán người dùng x với sản phẩm i

$$r_{ix} = \overline{r_i} + \frac{\sum_{j \in K_i} \left(r_{jx} - \overline{r_j}\right) u_{ij}}{\sum_{j \in K_i} \left| u_{ij} \right|} \qquad (2.18)$$

$\overline{r_i}$: là đánh giá trung bình cộng cho sản phẩm i.

2.3. Tăng bước lặp: t←t+1;

Until Converges: không có nhãn phân loại nào được bổ sung vào ma trận dự đoán.

Ví dụ 2.6:

Xét bài toán lọc cộng tác 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 7 sản phẩm $P = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$. Mỗi người dùng đều đưa ra các đánh giá của mình về các sản phẩm theo thang bậc {Ø, 1, 2, 3, 4, 5}. Giá trị r_{ij} = Ø được hiểu là người dùng u_i chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm p_j. Các giá trị $r_{5,1}$ = ? là sản phẩm hệ thống cần dự đoán cho người dùng u_5.

Bảng 2.13: Người dùng và sản phẩm

	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	4	2	5	Ø	3	Ø	3
u_2	5	Ø	5	5	4	Ø	Ø
u_3	4	Ø	Ø	4	3	4	3
u_4	Ø	3	5	5	Ø	5	Ø
u_5	?	5	?	?	Ø	4	4

Lọc cộng tác bằng phương pháp đồng huấn luyện theo người dùng:

Chọn γ = 3 với người dùng u_1 của bảng (2.13), theo công thức (2.15) thì: S1 = $\{u_2, u_3\}$, S2 = $\{u_1\}$, S3 = $\{u_1, u_2\}$, S4 = $\{\theta\}$, S5 = $\{\theta\}$. Khi đó mức độ tương tự giữa hai người dùng được xác định theo công thức (2.16).

Các nhãn phân loại chắc chắn chỉ được dự đoán từ những người dùng j $\in$ Si theo công thức (2.18).

Với tập người dùng đã cho trong bảng (2.13), tìm K_j theo (2.17) ta được K1 = $\{u_3\}$, K2 = $\{u_1\}$, K3 = $\{u_1\}$.

Bảng 2.14: Bảng giá trị đánh giá theo người dùng.

	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	4	2	5	**4**	3	**4**	3
u_2	5	**2**	5	5	4	Ø	**3**
u_3	4	**2**	**5**	4	3	4	3
u_4	Ø	3	5	5	Ø	5	Ø
u_5	?	5	?	?	Ø	4	4

Lọc cộng tác bằng phương pháp đồng huấn luyện theo sản phẩm:

Chọn $\gamma = 3$ với sản phẩm p_1 của bảng (2.13), theo công thức (2.19) thì:

C1= {p_5},C2= {Ø},C3= {Ø},C4= {Ø},C5= {Ø}.

Tuy vậy việc quan sát theo sản phẩm được thực hiện sau quá trình quan sát theo người dùng ta sẽ xác định được:

C1= {p_2, p_3, p_4, p_5, p_7}, C2= {p_1, p_3, p_4, p_5, p_7}, C3= {p_1, p_2, p_4, p_5, p_7},

C4= {p_1, p_2, p_3, p_5, p_6, p_7}, C5= {p_1, p_2, p_3, p_4, p_7}, C6= {p_2, p_3, p_4, p_7},

C7= {p_1, p_2, p_3, p_4, p_5, p_6}.

Mức độ tương tự giữa hai sản phẩm được xác định theo công thức (2.20)

Các nhãn phân loại chắc chắn chỉ được dự đoán từ các sản phẩm y $\in$ Cx theo công thức (2.21) và (2.22)

Dựa theo kết quả quan sát theo người dùng ta tìm được:

K1={p_4}, K2={p_7}, K3={p_4}, K4={p_1}, K5={p_7}, K6={p_4}, K7={p_5}.

Bảng 2.15: Bảng giá trị đánh giá theo sản phẩm

	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	4	2	5	**4**	3	**4**	3
u_2	5	**2**	5	5	4	***5***	**3**
u_3	4	**2**	**5**	4	3	4	3
u_4	***5***	3	5	5	Ø	5	Ø
u_5	?	5	?	?	***4***	4	4

Như vậy, quá trình lọc cộng tác bằng đồng huấn luyện với 1 bước lặp đã bổ sung được các giá trị đánh giá còn thiếu vào tập dữ liệu huấn luyện.

2.5. Kết luận

Chương này tác giả đã trình bày phương pháp lọc cộng tác dựa trên sản phẩm với các thuật toán tính độ tương tự và dự đoán, phương pháp lọc cộng tác dựa trên mô hình đồ thị với thuật toán dựa trên mô hình người dùng - sản phẩm đề xuất phương pháp dự đoán trên tất cả các đánh giá và thuật toán đồng huấn luyện dựa trên người dùng và sản phẩm. Với mỗi thuật toán đều có ví dụ minh họa quá trình xử lý để đưa ra các kết quả tư vấn.

Chương 3

XÂY DỰNG HỆ THÔNG TIN TƯ VẤN SẢN PHẨM SỮA DÀNH CHO NGƯỜI TIÊU DÙNG

3.1. Phát biểu bài toán

Hiện nay trên thị trường có rất nhiều các loại sữa, người tiêu dùng gặp nhiều khó khăn cho việc lựa chọn sản phẩm sữa phù hợp cho bản thân và gia đình, mỗi người dùng lại có một nhu cầu và sở thích khác nhau. Vấn đề đặt ra là cần lựa chọn những sản phẩm sữa phù hợp cho từng đối tượng người tiêu dùng, đồng thời cần sự đánh giá của khách hàng về sản phẩm.

Dựa trên cơ sở nghiên cứu về phương pháp lọc cộng tác và ứng dụng trong lọc thông tin tư vấn tác giả muốn xây dựng phần mềm thử nghiệm hệ thông tin tư vấn sản phẩm sữa dành cho người tiêu dùng.

3.2. Phân tích thiết kế hệ thống tư vấn sản phẩm sữa

3.2.1 Xác định bài toán

- Dữ liệu đầu vào: Bảng đánh giá các sản phẩm của người dùng
- Dữ liệu đầu ra: Tư vấn các sản phẩm mà người dùng chưa đánh giá

3.2.2. Phân tích các yêu cầu

Người dùng là các khách hàng đăng nhập vào 1 hệ thống Webstie để mua sữa. Mỗi người dùng được lưu trữ trên hệ thống với các hồ sơ bao gồm thông tin cá nhân, và các đánh giá của người dùng đó với các loại sữa. Đánh giá theo thang điểm từ 0 đến 5, với ý nghĩa là đánh giá càng cao thì người dùng càng thích loại sữa đó, điểm 0 dành cho loại sữa mà người dùng chưa đánh giá hoặc chưa biết về loại sữa đó. Hệ thống cần phải dự đoán cho khách hàng các sản phẩm sữa mà khách hàng chưa đánh giá. Tuy nhiên có rất nhiều loại sữa đã được đánh giá chỉ bởi một vài người và những sữa này khả năng được tư vấn là rất ít, thậm chí ngay cả khi trong số đó có những người dùng đưa ra đánh giá rất cao về chúng. Cũng như vậy, đối với những người dùng mà thị hiếu của họ khác thường so với một số đông người khác thì sẽ không có người dùng nào được tư vấn về những thị hiếu giống họ, dẫn đến việc tư vấn

nghèo nàn. Một phương pháp vượt qua tính thưa thớt trong đánh giá là sử dụng thông tin cá nhân của người dùng khi tính toán sự tương đồng giữa những người dùng. Hai người dùng được xem là giống nhau không khi được đánh giá có sở thích về các loại sữa là giống nhau mà chúng còn phải thuộc cùng một đối tượng.

Một vấn đề trước khi xây dựng ma trận đánh giá, với những sản phẩm ít được đánh giá, hoặc những người dùng ít đánh giá sản phẩm, những người dùng và sản phẩm này sẽ không hữu ích trong quá trình tư vấn. Vấn đề cần chọn lọc ra những sản phẩm và người dùng để tham gia trong quá trình tư vấn. Rõ ràng những sản phẩm mới hoặc người dùng mới không thể tham gia trong quá trình dự đoán, hoặc những sản phẩm hay người dùng có đánh giá ít hơn 1 ngưỡng nào đó cũng được loại ra tư vấn cho họ trong những sản phẩm mà họ chưa đánh giá thì sản phẩm nào là phù hợp nhất dựa trên những người có sở thích giống họ.

Công việc của hệ thông tin tư vấn:

Người dùng mới sẽ đăng kí thông tin cá nhân của mình để tạo nên một bộ hồ sơ người dùng được lưu trữ trong cơ sở dữ liệu

Khi một người dùng đăng nhập vào hệ thống, hệ thống có nhiệm vụ tư vấn những loại sữa mà người dùng đó chưa từng biết đến và những loại sữa tư vấn đó được dự đoán là người dùng sẽ đánh giá cao.

Các bước được thực hiện như sau:

Bước 1: Hệ thống sẽ xem xét các loại sữa mà người dùng chưa đánh giá, so sánh độ tương tự giữa loại sữa đó với những sữa khác, độ tương tự 2 loại sữa được tính dựa trên những người dùng từng đánh giá trên cả 2 loại sữa đó theo một thuật toán tính xác suất.

Bước 2: Hệ thống tư vấn sẽ dự đoán đánh giá của người dùng với những sữa mà người dùng chưa sử dụng, lựa chọn những sữa được dự đoán có đánh giá cao để đưa vào danh sách tư vấn cho người dùng

3.2.3. Thiết kế hệ thống tư vấn sản phẩm sữa

Thiết kế cơ sở dữ liệu

Sử dụng phần mềm Acsess tạo cơ sở dữ liệu "Lọc cộng tác" với bảng "Người dùng" để lưu trữ thông tin của khách hàng.

NguoiDung : Table

Field Name	Data Type
ID	AutoNumber
CMTND	Text
Ten	Memo
HoDem	Memo
NgaySinh	Date/Time
GioiTinh	Yes/No
DiaChi	Memo
TenDangNhap	Text
MatKhau	Text
Sua1	Text
Sua2	Text
Sua3	Text
Sua4	Text
Sua5	Text
Sua6	Text
Sua7	Text
Sua8	Text
Sua9	Text

Hình 3.1: Bảng Người dùng ở chế độ thiết kế

NguoiDung : Table

ID	CMTND	Ten	HoDem	NgaySinh	GioiTinh	DiaChi	TenDangNhap	MatKhau	Sua1	Sua2	Sua3	Sua4	Sua5	Sua6	Sua7	Sua8	Sua9
1	111	Quang	Vũ Đức	8/30/1981	☑	Quang Trung	quangvd	1234	1	2	1	5	0	0	0	0	0
2	101	Trang	Phùng T Thu	0/19/1981	☐	Đống Quang	trangptt	1234	2	1	3	0	0	1	0	0	0
3	102	Huyền	Lê Thanh	1/1/1983	☑	Quang Trung	huyenlt	1234	2	3	0	2	3	1	0	0	4
4	103	An	Nguyễn Văn	3/1/1980	☑	Đống Quang	annv	1234	5	0	2	0	1	0	0	0	2
5	104	An	Nguyễn Thị	2/1/1988	☐	Độc Lập	annt	1234	0	0	3	5	0	3	4	0	0
6	105	Anh	Đặng Hải	1/20/1980	☐	Quang Trung	anhdh	1234	0	1	1	0	3	5	4	1	0
7	106	Anh	Trần Lan	1/28/1982	☑	Đống Quang	anhtl	1234	0	0	0	0	3	4	2	1	0
8	107	Anh	Lê Quang	1/14/1982	☐	Độc Lập	anhlq	1234	0	0	0	0	0	5	1	2	2
9	108	Dung	Vũ Thùy	5/6/1988	☐	Độc Lập	dungvt	1234	0	0	3	1	0	0	0	2	4
ber)					☐				0	0	0	0	0	0	0	0	0

Hình 3.2: Bảng Người dùng ở chế độ trang dữ liệu

Thiết kế chức năng.

Các chức năng chính của chương trình:

Đăng ký: Người dùng mới đăng kí thông tin khách hàng và đánh giá một số loại sữa mà họ đã sử dụng để tạo nên bộ hồ sơ người dùng

Đăng nhập: Người dùng đăng nhập vào hệ thống với tên truy nhập và mật khẩu của riêng mình, nếu muốn người dùng có thể thay đổi các đánh giá sản phẩm trước đó.

- Hệ thống cho phép khách hàng lựa chọn lọc trên bộ nhớ hoặc đồ thị.

- Người dùng yêu cầu tư vấn sản phẩm sữa chưa đánh giá hệ thống sẽ tính toán đưa ra kết quả.

- Bước cuối cùng hệ thống sẽ tư vấn cho khách hàng sản phẩm mà khách hàng sẽ đánh giá cao

3.3. Xây dựng chương trình ứng dụng

Chương trình được chạy trên bộ visual studio 2012, ngôn ngữ lập C# thực hiện cài đặt thuật toán lọc cộng tác dựa trên sản phẩm và lọc cộng tác dựa trên mô hình đồ thị để tư vấn sản phẩm sữa cho người tiêu dùng.

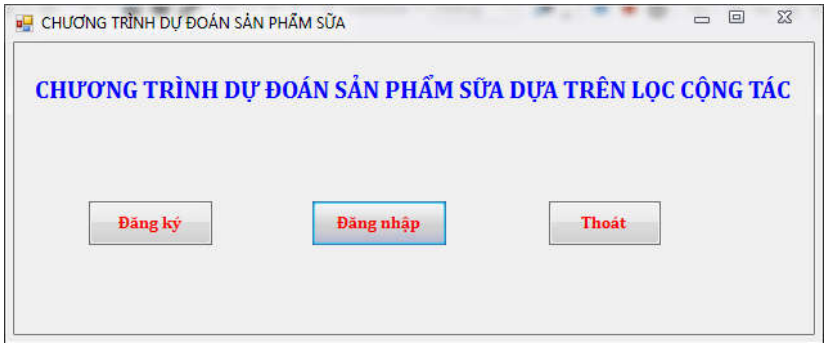
Các thuật toán sử dụng để cài đặt

- Thuật toán tính độ tương tự cosine điều chỉnh, khoảng giá trị luôn nằm trong đoạn [-1,1] thể hiện mức độ tương tự theo mức tăng dần giá trị độ tương tự, giá trị độ tương tự càng lớn thể hiện sự tương đồng về mặt đánh giá của 2 sản phẩm. Tác giả sử dụng thuật toán này bởi công thức cosin điều chỉnh có thêm thêm trung bình cộng các đánh giá khác rỗng của người dùng.

- Thuật dự đoán dựa trên tổng trọng số với việc đánh giá của người dùng lên sản phẩm dựa vào những đánh giá của người dùng đó lên các sản phẩm tương tự.

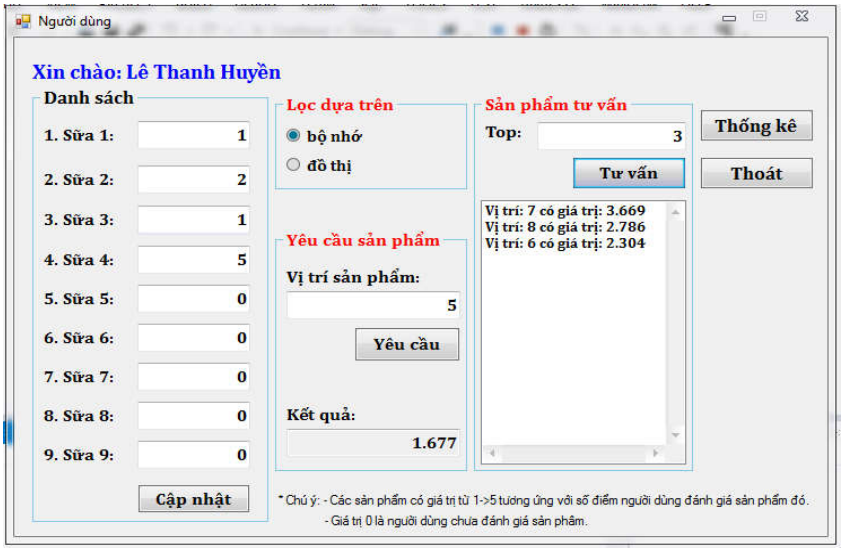
- Thuật toán dựa trên mô hình đồ thị với phương pháp dự đoán theo tất cả đánh giá(đồ thị có trọng số dương và đồ thị có trọng số âm).

Demo chương trình

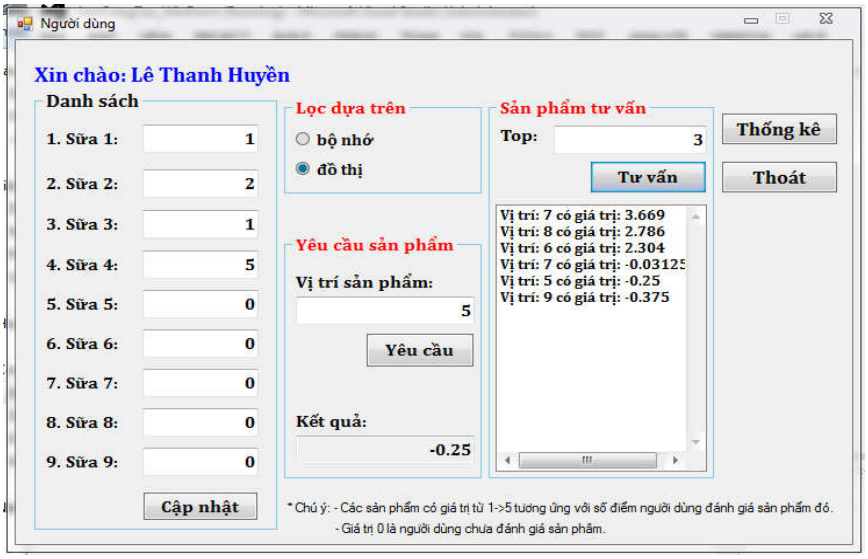


Hình 3.3: Giao diện chương trình dự đoán sản phẩm sữa cho người tiêu dùng.

Hình 3.4: Người dùng đăng nhập vào hệ thống.



Hình 3.5: Hệ thống lọc cộng tác dựa vào bộ nhớ



Hình 3.6: Hệ thống lọc cộng tác dựa vào đồ thị

3.4. Kết luận

Chương 3 tác giả xây dựng ứng dụng sản phẩm sữa cho người tiêu dùng sử dụng lọc cộng tác dựa vào bộ nhớ và lọc cộng tác dựa vào đồ thị. Do sản phẩm sữa chưa có bộ dữ liệu chuẩn nên tác giả hướng theo cách tiếp cận mới là làm thực nghiệm bằng cách phát phiếu thăm dò ý kiến cho 9 người dùng với 9 sản phẩm sữa nên tác giả chưa thể tiến hành đánh giá sản phẩm sữa theo Precision, Recall và F-Measure được.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận

Luận văn tiến hành nghiên cứu một số phương pháp lọc cộng tác và đã đạt được những yêu cầu sau:

- Nghiên cứu lọc cộng tác dựa trên sản phẩm với thuật toán tính độ tương tự và tính toán dự đoán tư vấn.

- Nghiên cứu lọc cộng tác dựa trên mô hình đồ thị với thuật toán dựa trên mô hình đồ thị người dùng - sản phẩm.

- Nghiên cứu lọc cộng tác bằng phương pháp đồng huấn luyện theo người dùng và theo sản phẩm.

- Xây dựng ứng dụng hệ thông tin tư vấn sản phẩm sữa dành cho người tiêu dùng. Ứng dụng cho phép người dùng đăng nhập để đánh giá đồng thời nhận được gợi ý những sản phẩm hợp với sở thích của mỗi người dùng.

Hướng phát triển

Luận văn mới chỉ nghiên cứu được một phương pháp lọc thông tin cho hệ tư vấn đó là phương pháp lọc cộng tác, phương pháp này còn nhiều hạn chế về vấn đề dữ liệu thưa, người dùng và sản phẩm mới. Bởi vậy, trong tương lai phương hướng phát triển tiếp theo của tác giả sẽ nghiên cứu thêm các phương pháp lọc thông tin cho hệ tư vấn khác để khắc phục các hạn chế trên đồng thời xây dựng chương trình ứng dụng thông tin tư vấn được tốt hơn với những đánh giá cụ thể.

Do thời gian và kinh nghiệm nghiên cứu còn thiếu, kiến thức còn hạn chế, mặc dù đã nỗ lực cố gắng, tuy nhiên luận văn không tránh khỏi những thiếu sót. Rất mong nhận được những chỉ bảo của các thầy cô, sự đóng góp của các bạn đồng nghiệp để tác giả có thể hoàn thành công trình nghiên cứu này tốt hơn.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. Nguyễn Duy Phương, Từ Minh Phương (2009), ”Lọc cộng tác và lọc theo nội dung dựa trên mô hình đồ thị”, *Chuyên san các công trình nghiên cứu, phát triển và ứng dụng Công nghệ thông tin và Truyền thông*, Tập V-1 số1, trang: 4-12.

[2]. Nguyễn Duy Phương, Từ Minh Phương (2008), “Một thuật toán lọc cộng tác cho trường hợp ít dữ liệu”, *Tạp chí Tin học và Điều khiển học*, tập 24, trang: 62-74.

[3]. Nguyễn Duy Phương, Phạm Văn Cường, Từ Minh Phương (2008), “Một số giải pháp lọc thư rác tiếng Việt”, *Chuyên san các công trình nghiên cứu, phát triển và ứng dụng Công nghệ thông tin và Truyền thông*, số19, trang: 102-112.

[4]. Nguyễn Duy Phương, Lê Quang Thắng, Từ Minh Phương (2008), “Kết hợp lọc cộng tác và lọc theo nội dung sử dụng đồ thị”, *Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông,* trang: 155-166.

Tiếng Anh

[5]. G. Somlo and A. Howe (2001), “Adaptive Lightweight Text Filtering”, *Proc. Fourth Int’l Symp. Intelligent Data Analysis.*

[6]. J. S. Breese, D. Heckerman, and C. Kadie (1998), “Empirical analysis of

[7]. Predictive Algorithms for Collaborative Filtering”, *In Proc. of 14th Conf. on Uncertainty in Artificial Intelligence,* pp. 43-52.

[8]. J.L. Herlocker, J.A. Konstan, and J. Riedl (2000), “Explaining Collaborative Filtering Recommendations”, *Proc. ACM Conf Computer Supported Cooperative Work.*

[9]. L. Si and R. Jin (2003), “Flexible Mixture Model for Collaborative Filtering”, *Proc. 20th Int’l Conf. Machine Learning.*

[10]. M. Pazzani and D. Billsus (1997), “Learning and Revising User Profiles: The Identification of Interesting Web Sites”, *Machine Learning*, vol. 27, pp. 313-331.