HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Nguyễn Văn Hòa

NGHIÊN CỨU PHƯƠNG PHÁP TƯ VẤN CỘNG TÁC CHO CÁC CỔNG LẬP TRÌNH TRỰC TUYẾN

Chuyên ngành: Hệ thống thông tin
Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SỸ
(Theo định hướng ứng dụng)

Hà Nội - 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **TS. NGUYỄN DUY PHƯƠNG**

Phản biện 1: **PGS.TS. Đỗ Trung Tuấn**

Phản biện 2: **PGS.TS. Phạm Văn Cường**

Luận văn này được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 10 giờ 30 phút ngày 17 tháng 12 năm 2022

Có thể tìm hiểu luận văn này tại:

Thư viện của Học viện Công nghệ Bưu chính Viễn thông

I. LỜI MỞ ĐẦU

Ngày nay chúng ta đang chứng kiến một sự phát triển mạnh mẽ chưa từng có của các cổng lập trình trực tuyến. Các cổng lập trình cung cấp cho người dùng một môi trường lập trình đa ngôn ngữ để có thể lập trình, kết quả lập trình được chấm một cách tự động. Đồng hành cùng với người dùng trên cổng lập trình trực tuyến là các trường đại học, các tập đoàn kinh tế lớn nhằm mục tiêu đào tạo và tuyển dụng nguồn nhân lực chất lượng cao về công nghệ thông tin ví dụ như các tập đoàn lớn về công nghệ Microsoft, Google, Facebook, Apple, Samsung… Ở Việt Nam thì có FPT, Viettel, VNPT, các tập đoàn này không chỉ xây dựng riêng cho mình các cổng lập trình trực tuyến mà còn bảo trợ về tài chính và cung cấp kho nội dung số cho cổng lập trình trực tuyến của các trường đại học khác. Một số trường đại học như MIT, Stanford, Baylor xem cổng lập trình trực tuyến như một công cụ quan trọng trong giảng dạy, rèn luyện và đánh giá kỹ năng lập trình của kỹ sư ngành công nghệ thông tin. Nhận thức được tầm quan trọng và hiệu quả trong công nghệ lập trình trực tuyến, một số trường đại học ở Việt Nam như Đại học Quốc Gia Hà Nội, Đại học Quốc Gia thành phố HCM Đại học Bách Khoa, Học viện công nghệ Bưu Chính Viễn Thông đã xây dựng riêng cho mình các cổng lập trình trực tuyến và đã ứng dụng thành công trong giảng dạy và rèn luyện kỹ năng lập trình của sinh viên.

Có nhiều công nghệ khác nhau để xây dựng nên các cổng lập trình trực tuyến. Ở cấp độ trung học cơ sở hay phổ thông trung học hầu hết các quốc gia chọn công nghệ PC2 hoặc CMS trong giảng dạy, luyện tập và các tổ chức các kỳ thi lập trình quốc gia (NOI) hoặc quốc tế (IOI). Ở các cấp bậc cao hơn, các trường đại học thường lựa chọn các công nghệ Domjudge, Katis hoặc DMOJ trong giảng dạy, luyện tập và tổ chức các kỳ thi lập trình quốc gia hoặc quốc tế theo chuẩn ACM/ICPC. Sự khác biệt giữa các công nghệ này là khá nhỏ và chỉ được phân biệt khi ta triển khai ứng dụng dựa vào quy mô nhỏ hay lớn, nhiều hay ít người dùng, độ lớn dữ liệu của các test hoặc phương pháp đánh giá giải pháp của người lập trình theo mức từng phần hay toàn phần.

Đối với các cổng lập trình trực tuyến, tài nguyên quan trọng nhất là kho nội dung số được nạp bên trong mỗi cổng lập trình. Kho nội dung số được thể hiện dưới dạng tập các bài toán cùng với các bộ dữ liệu kiểm thử tương ứng. Mỗi bài toán cần được xây dựng nhiều bộ dữ liệu kiểm thử, mỗi bộ dữ liệu kiểm thử xác định một tính chất đúng mà giải pháp lập trình cần đạt được. Một giải pháp lập trình được xem là tốt nhất

nếu nó thỏa mãn được tất cả các bộ dữ liệu kiểm thử với thời gian và không gian nhớ xác định. Kho nội dung số được xây dựng bởi các chuyên gia, tổ chức sở hữu cổng lập trình trực tuyến. Kho nội dung số mà càng lớn thì càng thu hút được đông đảo người dùng tham gia.

Một yếu tố rất quan trọng tiếp theo của cổng lập trình trực tuyến là người dùng. Nếu chúng ta có một tập tài nguyên là các bài toán vô cùng đa dạng và phong phú mà lại thiếu mất đi những người tham gia giải quyết kho tài nguyên ấy thì cổng lập trình trực tuyến sẽ đánh mất đi mục đích chính của nó được tạo ra là nhằm phục vụ, nâng cao trình độ, kỹ năng lập trình của lập trình viên hay sinh viên tại các trường đại học. Tập người càng lớn chứng tỏ kho nội dung số của cổng lập trình trực tuyến rất có giá trị. Phần lớn trong số người dùng sử dụng cổng lập trình để học lập trình, một phần trong số họ tham gia sử dụng với mục đích nâng cao kỹ năng lập trình, phần còn lại để chứng tỏ bản thân, khẳng định khả năng của mình trong việc lập trình và cũng dựa vào đó để thu hút các nhà tuyển dụng. Đặc điểm chung nhất của các cổng lập trình trực tuyến là kho nội dung số và số lượng người dùng phong phú và đa dạng. Một số cổng lập trình trực tuyến như CodeForces, TopCoder, ICPC Baylor, CodeLearn, SPOJ,… đã thu hút hàng trăm ngàn lập trình viên trên toàn thế giới tham gia. Chính vì vậy, việc xây dựng một hệ thống gợi ý (Recommendation Systems) là một điều rất cần thiết để có thể tư vấn, gợi ý các bài toán phù hợp với khả năng lập trình của mỗi người dùng. Đây cũng là một điều tối quan trọng trong cổng lập trình trực tuyến vì nó giúp người dùng có thể giải quyết những bài toán phù hợp với khả năng của mình nhằm gây hứng thú với lập trình viên trong quá trình lập trình.

Một trong công nghệ lõi của cổng lập trình trực tuyến là hệ thống tư vấn kho nội dung số đến với người dùng. Một hệ tư vấn tốt có thể tư vấn, gợi ý những bài toán phù hợp với khả năng của lập trình viên với sai số thấp nhất. Trong những năm gần đây có rất nhiều phương pháp được đề xuất sử dụng trong hệ tư vấn nhưng đại đa số lại chỉ tập chung vào hai kiểu tư vấn:

- Tư vấn dựa theo nội dung (Content – Based Systems): Người dùng sẽ được tư vấn dựa theo những sản phẩm tương tự với những sản phẩm người dùng đó đã ưa thích. Có thể hiểu rằng phương pháp này đưa ra gợi ý dựa trên đặc tính của sản phẩm đó.

- Tư vấn cộng tác (Collaborative filtering): Ở phương pháp này hệ thống sẽ gợi ý sản phẩm dựa trên độ tương quan (Similarity) giữa các người dùng với sản phẩm (item) hay người dùng (user) hoặc sản phẩm (item). Có thể hiểu đơn giản ở nhóm này một sản phẩm (item) được gợi ý tới một người dùng (user) dựa trên những người dùng (user) có hành vi tương tự với sản phẩm (item).

Mục đích nghiên cứu của đề tài là nghiên cứu phương pháp tư vấn cộng tác ứng dụng cho các cổng lập trình trực tuyến. Để thực hiện được những mục tiêu trên đề tài cần phải đạt được một số nhiệm vụ nghiên cứu sau:

- Nghiên cứu tổng quan về các cổng nghệ lập trình trực tuyến
- Nghiên cứu phương pháp lọc cộng tác ứng dụng cho hệ tư vấn
- Đề xuất phương pháp tư vấn cộng tác cho cổng lập trình trực tuyến

Đối tượng nghiên cứu của đề tài là các phương pháp tư vấn lọc cộng tác và các công nghệ xây dựng cổng lập trình trực tuyến. Phạm vi nghiên cứu là các phương pháp tư vấn cộng tác cho dữ liệu người dùng và dữ liệu sản phẩm của cổng lập trình trực tuyến, phương pháp nghiên cứu được chia làm 2 hướng:

- Nghiên cứu lý thuyết: Tập trung vào các công nghệ xây dựng cổng lập trình trực tuyến, đặc biệt quan tâm đến việc trích rút dữ liệu người dùng và kết quả lập trình trực tuyến của mỗi lập trình viên. Nghiên cứu các phương pháp tư vấn cộng tác áp dụng trên cổng lập trình trực tuyến dựa trên người dùng hoặc sản phẩm.
- Nghiên cứu thực nghiệm: Xây dựng bộ dữ liệu thử nghiệm riêng cho cổng lập trình trực tuyến của PTIT. Thực hiện thử nghiệm phương pháp đề xuất trên tập dữ liệu đã được xây dựng

Nội dung của luận văn bao gồm 3 chương với cấu trúc như sau:

CHƯƠNG 1: TỔNG QUAN VỀ CỔNG LẬP TRÌNH TRỰC TUYẾN VÀ HỆ TƯ VẤN CỘNG TÁC

Nội dung chính của chương 1 là tập trung nghiên cứu vào những vấn đề cơ bản của cổng lập trình trực tuyến và các phương pháp tư vấn cộng tác. Nội dung của chương bao gồm:

- **Giới thiệu về các công nghệ lập trình trực tuyến**: Ở phần này sẽ trình bày về các công nghệ lập trình trực tuyến, tài nguyên của cổng lập trình trực tuyến và những thách thức, khó khăn của cổng lập trình trực tuyến hiện nay

- **Giới thiệu về hệ tư vấn và một số vấn đề liên quan**: Trình bày về các hệ tư vấn, các phương pháp xây dựng nên hệ tư vấn và một số vấn đề liên quan
- **Phương pháp cộng tác baseline**: Trình bày các phương pháp tư vấn cộng tác cơ sở và được dùng trong việc so sánh, đánh giá với phương pháp đề xuất của luận văn
- **Một số vấn đề của phương pháp lọc cộng tác**: Trình bày một số vấn đề của phương pháp lọc cộng tác và một số nghiên cứu giải pháp
- **Kết luận chương**: Tóm tắt lại những kết quả nghiên cứu của chương

CHƯƠNG 2: PHƯƠNG PHÁP TƯ VẤN CỘNG TÁC CHO CỔNG LẬP TRÌNH TRỰC TUYẾN

Nội dung chính của chương trình bày phương pháp tư vấn cộng tác cho cổng lập trình trực tuyến. Dự kiến nội dung của chương bao gồm:

- **Phát biểu bài toán**: Mô hình hóa bài toán tư vấn cộng tác cho các cổng lập trình trực tuyến.
- **Phương pháp tư vấn cộng cho cổng lập trình trực tuyến**: Trình bày phương pháp tư vấn cộng tác bằng cách biểu diễn mối quan hệ giữa người dùng và các nội dung số trên cổng lập trình trực tuyến như một đồ thị hai phía.
- **Kết luận chương**: Tóm tắt lại những kết quả nghiên cứu của chương.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

Nội dung chính của chương trình bày phương pháp tư thử nghiệm và đánh giá kết quả cài đặt. Dự kiến nội dung của chương bao gồm:

- **Dữ liệu thực nghiệm**: Trình bày phương pháp thu thập dữ liệu từ cổng lập trình trực tuyến của Học viện Công nghệ BCVT.
- **Phương pháp thực nghiệm**: Trình bày các phương pháp đánh giá sai số dự đoán áp dụng cho hệ tư vấn cộng tác.
- **Kết quả thực nghiệm**: Đưa ra các kết quả thực nghiệm và so sánh, đánh giá kết quả so với các phương pháp khác.
- **Kết luận chương**: Tóm tắt lại các kết quả đã đạt được của chương

CHƯƠNG 1: TỔNG QUAN VỀ CỔNG LẬP TRÌNH TRỰC TUYẾN VÀ HỆ TƯ VẤN CỘNG TÁC

1.1. Giới thiệu về công nghệ lập trình trực tuyến

1.1.1. ***PC2***

PC^2 (P-C-Squared) là một hệ thống phần mềm để quản lý các cuộc thi lập trình máy tính được phát triển tại Đại học bang California, Sacramento.

1.1.2. ***CMS***

CMS là phần mềm tổ chức các cuộc thi lập trình tương tự như các cuộc thi nổi tiếng như IOI. Được viết và nhận được sự đóng góp bởi những người tổ chức các cuộc thi tương tự ở cấp quốc gia và quốc tế.CMS là một giải pháp được coi là hoàn chỉnh và đã được thử nghiệm và chứng minh tốt để quản lý các cuộc thi. Tuy nhiên nó chỉ cung cấp các công cụ hạn chế để phát triển dữ liệu, nhiệm vụ thuộc về cuộc thi, Hệ thống được tổ chức theo cách Modular với các dịch vụ khác nhau chạy trên các máy khác nhau và cung cấp khả năng mở rộng thông qua các bản sao dịch vụ trên một số thiết bị.

1.1.3. ***Domjudge***

DOM JUDGE là một hệ thống giám khảo tự động để điều hành các cuộc thi lập trình. Nó có một cơ chế để gửi các giải pháp vấn đề để chúng được đánh giá hoàn toàn tự động và cung cấp giao diện cho các nhóm, giám khảo.

Hệ thống có quy mô khá tốt với số lượng người dùng lớn nhưng lại cần một tài nguyên server đủ mạnh. Nó có đầy đủ các tính năng của một OPJ bao gồm:

- Máy test là một hệ thống chạy độc lập so với giao diện
- Sandbox:
- Hệ dịch:
- Test case:
- Hệ đánh giá:
- Giao diện:

1.1.4. ***Kattis***

Đây là một bộ công cụ khá mới nhưng càng ngày tỏ ra uy tín trong công việc tổ chức các contest chấm thi ACM/ICPC. Website này có đội ngũ chuyên gia đánh giá đề bài rất khắt khe và chuyên nghiệp.

1.1.5. SPOJ

Đây là một website luyện tập trực tuyến rất lớn. Website này cung cấp các API mở cho phép các trường hoặc các tổ chức trên toàn cầu đăng ký sử dụng các site con và công cụ chấm thi trên đó.

1.1.6.DLab

Đây là hệ thống cổng lập trình trực tuyến của khoa Công nghệ thông tin 1 thuộc Học viện Công nghệ Bưu chính Viễn thông với mục đích phục vụ quá trình học tập và luyện tập cho sinh viên trong Học viện.

1.1.7.Tổng kết

Việc xây dựng nên một cổng lập trình trực tuyến hoàn chỉnh bao gồm bộ công cụ kiểm thử, website quản lý, hệ thống tư vấn cho sinh viên luôn là một nhu thiết yếu tại các trường đại học trên thế giới nói chung cũng như Việt Nam nói riêng, nhất là tại các trường chuyên đào tạo nhân lực trong lĩnh vực CNTT.

1.2. Giới thiệu về hệ tư vấn

1.2.1.Phương pháp lọc nội dung (Content Filtering)

Lọc nội dung là kỹ thuật lọc dựa trên sự phân tích về nội dung của dữ liệu chứ không phải là tìm hiểu về nguồn gốc của nó hay các tiêu chí khác. Lọc theo nội dung được thực hiện trên cơ sở so sánh nội dung thông tin hay mô tả hàng hóa để tìm ra những mặt hàng tương tự với những gì mà người dùng đã từng quan tâm để giới thiệu cho họ những mặt hàng này. Lọc dựa trên nội dung thực hiện hiệu quả trên các đối tượng dữ liệu biểu diễn dưới dạng văn bản và được sử dụng rộng rãi trên internet để lọc email và truy cập các trang web.

Trong phương pháp tư vấn dựa trên nội dung, hàm tiện ích $u(c,s)$ của item (sản phẩm) s ứng với người dùng c được đánh giá dựa trên những hàm ước lượng $u(c,s_i)$ được gán bởi người dùng c với những item $s_i \in S$ tương tự với item s.

1.2.2.Phương pháp lọc cộng tác (Collaborative Filtering)

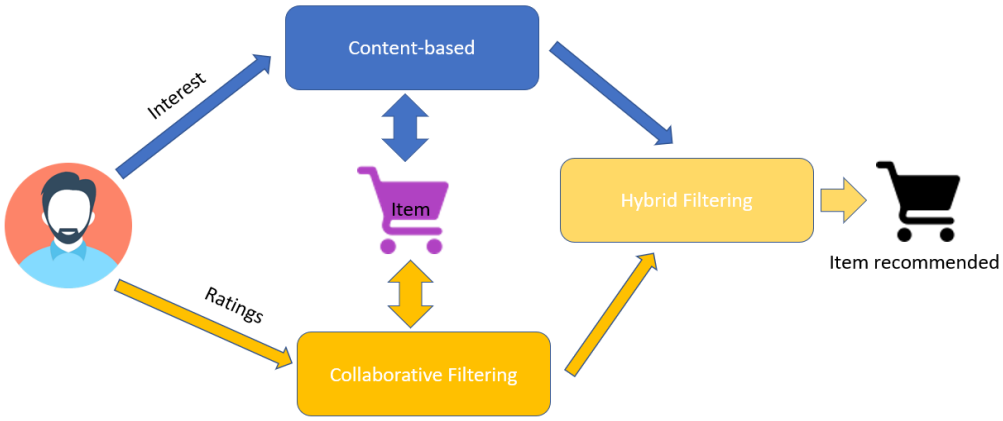
Lọc cộng tác hoạt động bằng cách thu thập phản hồi (Ratings) của người dùng dưới dạng xếp hạng cho các mặt hàng, sản phẩm trong một miền giá trị nhất định có thể từ (-1,1) hay (0,5) tùy thuộc vào bài toán của nhà phát triển. Với giá trị đánh giá này, hệ thống sẽ khai thác các điểm tương đồng về hành vi xếp hạng giữa một số người dùng để xác định cách đề xuất một sản phẩm đến với người dùng.

Các phương pháp lọc cộng tác (CF) có thể được chia nhỏ hơn nữa thành hai phương pháp sau:

- Neighborhood-based (Memory-based):
- Model-based:

1.2.3. Phương pháp lọc kết hợp (Hybrid Filtering)

Để tận dụng thế mạnh của hai phương pháp lọc cộng tác và lọc nội dung đã có một số phương pháp lai được đề xuất kết hợp cả hai. Một cách tiếp cận đơn giản là cho phép cả phương pháp lọc cộng tác và dựa trên nội dung để tạo ra các danh sách đề xuất được xếp hạng riêng biệt và sau đó hợp nhất các kết quả của chúng để tạo ra một danh sách cuối cùng.



Hình 1: *Mô tả phương pháp Hybrid Filtering*

1.3. Phương pháp tư vấn cộng tác based-line

1.3.1. Phương pháp tư vấn cộng tác User-based

Phương pháp này tình toán sự tương tự giữa 2 người dùng *x* và *y* sử dụng công thức độ tương quan Person hoặc dựa trên Vector Cosin sau đó hệ thống sẽ xác định tập các giá trị trọng số của người dùng trên tất cả các item dựa vào độ tương tự giữa các người dùng từ đó lấy k item có trọng số hay đánh giá cao nhất để đưa ra gợi ý,

1.3.2. Phương pháp tư vấn cộng tác Item-based

Phương pháp này tính toán sự tương tự (*similarity*) giữa 2 item *i* và *j*. Sau đó trên cơ sở người dùng cần tư vấn, hệ thống sẽ lấy ra *k* item mà người dùng có khả năng quan tâm nhất. Sau đó hệ thống tư vấn sẽ xếp hạng *k* item này dựa trên trọng số đã tính toán được để đưa ra quyết định kiến nghị cho người dùng. Để tính sự tương đồng giữa 2 item và trọng số đánh giá giữa người dùng *u* trên item *i*, chúng ta sử dụng độ tương quan Pearson hoặc vector tương tự Cosin để tính toán.

1.4. Một số vấn đề của tư vấn cộng tác

Trong phần này, tôi sẽ trình bày một số trở ngại phổ biến trong tư vấn lọc cộng tác cũng như trình bày một số nghiên cứu giải quyết chúng

Sự thưa thớt (*Sparsity***)**:

Cold-start:

Gian lận (Fraud):

1.5. Kết luận

Mặc dù các phương pháp dựa trên nội dung thuần túy tránh được một số cạm bẫy đã thảo luận ở trên, lọc cộng tác vẫn có một số ưu điểm chính so với chúng. Thứ nhất, CF có thể thực hiện trong các miền không có nhiều nội dung được liên kết với các mục hoặc trong đó máy tính khó phân tích nội dung, chẳng hạn như ý tưởng, quan điểm, v.v,.. Thứ hai, hệ thống CF có khả năng cung cấp các tư vấn có tính bất ngờ, tức là nó có thể đề xuất các mục có liên quan đến người dùng, nhưng không chứa nội dung từ hồ sơ của người dùng. Chính vì vậy mà phương pháp lọc cộng tác được đề xuất làm phương pháp trong hệ thống tư vấn của cổng lập trình trực tuyến cũng một phần do tính đặc trưng của cổng lập trình trực tuyến khó có thể gian lận trong đánh giá vấn đề mà lọc cộng tác rất khó giải quyết với những ratings không chính xác. Ở chương 2, luận văn sẽ trình bày về phương pháp lọc cộng tác cho cổng lập trình trực tuyến nhằm nâng cao hiệu quả tư vấn.

CHƯƠNG 2: PHƯƠNG PHÁP TƯ VẤN CỘNG TÁC CHO CỔNG LẬP TRÌNH TRỰC TUYẾN

2.1. Phát biểu bài toán

Tư vấn lọc cộng tác (*collaborative filtering recommendation*) là phương pháp phổ biến trong xây dựng các hệ tư vấn được ứng dụng rộng rãi trong thương mại điện tử, Phương pháp được xây dựng từ tập người dùng $U = \{u_1, u_2, \ldots, u_n\}$ và tập các sản phẩm $P = \{p_1, p_2, \ldots, p_m\}$. Mỗi người dùng $u_i \in U$ đưa ra đánh giá của mình cho một số sản phẩm $p_x \in P$ bằng một số $r_{ix} \in \Omega$ (Ω có thể là tập các số nguyên hoặc tập các số thực). Ma trận đánh giá R là đầu vào duy nhất của các phương pháp tư vấn cộng tác [5]. Để thuận tiện trong trình bày, thay bằng viết $u_i \in U$ và $p_x \in P$ ta viết ngắn gọn thành $i \in U$ và $x \in P$. Dựa vào ma trận đánh giá $R = \{r_{ix}: i=1, 2,.., n; x = 1, 2,.., m\}$, các phương pháp tư vấn lọc cộng tác khai thác những khía cạnh liên quan đến cộng đồng đồng người dùng có cùng chung sở thích để cung cấp cho người dùng này những sản phẩm phù hợp nhất với họ. Tư tưởng chủ đạo của lọc cộng tác là những người dùng có sở thích tương tự nhau trong quá khứ thì họ có thể có chung sở thích trong tương lai. Mỗi người dùng trong hệ tư vấn lọc cộng tác là độc lập với người dùng còn lại.

Bài toán tư vấn nội dung số cho mỗi người dùng trên cổng lập trình trực tuyến có thể được xem xét như bài toán tư vấn lọc cộng tác điển hình. Cụ thể luận văn này xây dựng hệ tư vấn bài toán cho người dùng trên cổng lập trình trực tuyến Dlab sử dụng phương pháp lọc cộng tác. Gọi tập người dùng là $U = \{u_1, u_2,.., u_n\}$, tập sản phẩm là $P = \{p_1, p_2,.., p_m\}$. Tập U và P lần lượt là tập người dùng và bài tập là dữ liệu thực được thu thập trên cổng lập trực tuyến trình Dlab. Kết quả lập trình của mỗi người dùng $i \in U$ giải quyết bài toán $x \in P$ được hệ thống ghi nhận một cách tự động bằng một số r_{ix}. Trong đó, r_{ix} được ghi nhận giá trị *1* nếu giải pháp lập trình của người dùng $i \in U$ thỏa mãn tất cả các bộ dữ liệu kiểm thử đối với bài toán $x \in P$, r_{ix} được ghi nhận giá trị *-1* nếu giải pháp lập trình của người dùng $i \in U$ chưa thỏa mãn đầy đủ các bộ dữ liệu kiểm thử đối với bài toán $x \in P$, r_{ix} được ghi nhận giá trị *0* nếu người dùng $i \in U$ chưa giải quyết bài toán $x \in P$. Nhiệm vụ của phương pháp tư vấn lọc cộng tác là cung cấp

tập các bài toán phù hợp với khả năng lập trình của mỗi người dùng trên cổng lập trình trực tuyến Dlab.

Có nhiều đề xuất khác nhau để giải quyết bài toán tư vấn lọc cộng tác, tuy vậy ta có thể phân loại các phương pháp thành hai cách tiếp cận chính: tư vấn lọc cộng tác dựa vào bộ nhớ (Memory-Based) và tư vấn lọc cộng tác dựa vào mô hình(Model-Based). Trong luận văn này tập trung vào phương pháp tư vấn dựa vào bộ nhớ. Tư vấn lọc cộng tác dựa trên bộ nhớ được tiếp cận theo hai phương pháp chính: Phương pháp tư vấn lọc cộng tác dựa vào người dùng (UserBased) và tư vấn lọc cộng tác dựa vào sản phẩm (ItemBased). Mỗi phương pháp đều có những ưu điểm riêng khai thác khía cạnh liên quan đến người dùng hoặc sản phẩm. Đặc điểm chung của cả hai phương pháp này là sử dụng toàn bộ tập dữ liệu đánh giá để dự đoán quan điểm của người dùng cần được tư vấn về các sản phẩm mà họ chưa hề biết đến. Về bản chất, đây là phương pháp học lười hay học dựa trên ví dụ được sử dụng trong học máy. Ngoài hai phương pháp tư vấn lọc cộng tác dựa vào người dùng (UserBased) và tư vấn lọc cộng tác dựa vào sản phẩm (ItemBased) đã được trình bày ở chương 1 thì trong luận văn sẽ trình bày thêm phương pháp tư vấn lọc cộng được thực hiện bằng cách biểu diễn mối quan hệ giữa người dùng và các nội dung số trên cổng lập trình trực tuyến như một đồ thị hai phía (GraphBased). Phương pháp này sẽ được trình bày chi tiết ở mục 2.2.

2.2. Phương pháp tư vấn cộng tác cho cổng lập trình trực tuyến

Để nâng cao chất lượng tư vấn, trong mục này luận văn sẽ trình bày đề xuất một thuật toán tư vấn lọc cộng tác cho cổng lập trình trực tuyến. Phương pháp được thực hiện bằng cách biểu diễn mối quan hệ giữa người dùng và các nội dung số trên cổng lập trình trực tuyến như một đồ thị hai phía. Lợi dụng tính chất này, luận văn xem xét bài toán cần giải quyết như một vấn đề tìm kiếm trên đồ thị. Phương pháp sẽ được tiến hành như dưới đây.

2.2.1. Phương pháp ước lượng mức độ phù hợp của người dùng đối với sản phẩm

Giả sử ta có hệ n người dùng $U = \{u_1, u_2,.., u_n\}$, m sản phẩm $P = \{p_1, p_2,,., p_m\}$. Trong đó, tập người dùng U được thu thập một cách tự động ngay từ khi người dùng đăng ký tham gia cổng lập trình trực tuyến. Tập sản phẩm P là tập các bài toán cùng với các bộ dữ liệu kiểm thử đã được nạp trong cổng lập trình trực tuyến. Ma trận đánh

giá $R = \{r_{ix}: i=1, 2,.., n; x = 1, 2,,.m\}$ được ghi nhận tự động trong cổng lập trình trực tuyến theo công thức (17).

$$r_{ix} = \begin{cases} 1 & \text{nếu người dùng } i \in U \text{ submit đúng bài toán } x \in P \\ 0 & \text{nếu người dùng } i \in U \text{ chưa submit bài toán } x \in P \\ -1 & \text{nếu người dùng } i \in U \text{ submit sai bài toán } x \in P \end{cases} \quad (17)$$

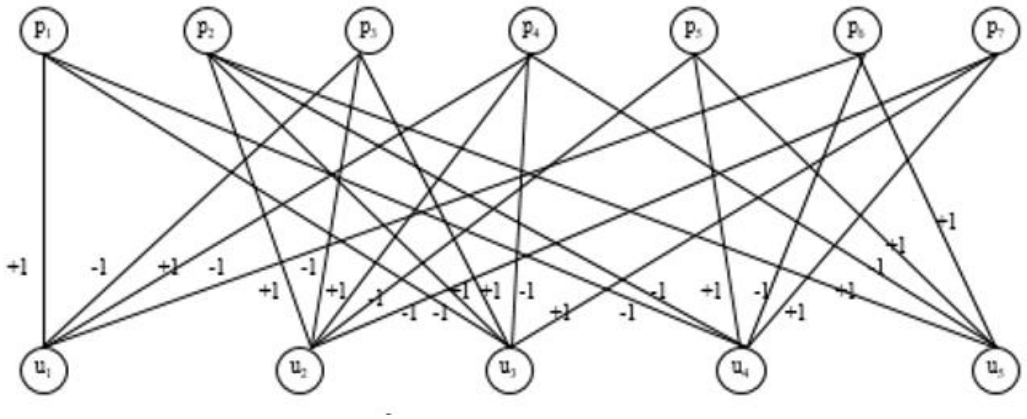
Dễ dàng nhận thấy, ma trận đánh giá R được xác định theo (17) được biểu diễn như một đồ thị hai phía (bipart graph). Một phía là tập người người dùng U, phía còn lại là tập bài toán P trong cổng lập trình trực tuyến. Tập cạnh của đồ thị chỉ bao gồm các cạnh có dạng $e = (i, x)$, trong đó $i \in U$ và $x \in P$. Không tồn tại các cạnh nối các đỉnh ở cùng một phía. Nói cách khác, không tồn tại các cạnh nối giữa đỉnh người dùng và đỉnh người dùng và không tồn tại các cạnh nối giữa đỉnh sản phẩm và đỉnh sản phẩm. Ví dụ với ma trận đánh giá của hệ gồm 5 người dùng và 7 sản phẩm như trong Bảng 3 sẽ cho ta biểu diễn đồ thị hai phía tương ứng trong Hình 2. Trong đó, giá trị $r_{ix} = 1$ thể hiện người dùng i lập trình đúng bài toán x, $r_{ix} = -1$ thể hiện người dùng i lập trình chưa đúng bài toán x, $r_{ix} = 0$ thể hiện người dùng i chưa lập trình bài toán x.

Người dùng	*Sản phẩm*						
	p_1	p_2	p_3	p_4	p_5	p_6	p_7
u_1	1	0	-1	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

Bảng 1: *Ví dụ với ma trận đánh giá của hệ gồm 5 người dùng và 7 sản phẩm*

Theo tính chất của đồ thị hai phía, đường đi từ đỉnh người dùng $i \in U$ đến đỉnh sản phẩm $x \in P$ luôn có độ dài L lẻ ($L=1, 3, 5, 7...$). Ví dụ các đường đi u_1-p_4-u_3-p_2, u_1-p_3-u_2-p_7, u_1-p_1-u_2-p_2 là những đường đi độ dài 3, các đường đi u_1-p_4-u_2-p_2-u_3-p_7, u_2-p_3-u_1-p_6-u_4-p_1, u_1-p_1-u_3-p_7-u_4-p_2 là những đường đi độ dài 5. Tập tất cả các đường đi từ đỉnh người dùng $i \in U$ đến đỉnh sản phẩm $x \in P$ được chia thành 3 loại: loại 1 bao gồm tập các đường đi chỉ đi qua các cạnh có trọng số dương (+1), loại 2 bao gồm tập các đường đi chỉ đi qua các cạnh có trọng số âm (-1), loại 3 bao gồm tập các đường đi đi qua các cạnh có trọng số hoặc (+1) hoặc(-1). Ví dụ: u_1-p_4-u_3-p_2, u_1-p_4-u_2-p_2-u_3-p_7 là các

đường đi loại 1; u_1-p_3-u_2-p_7, u_2-p_3-u_1-p_6-u_4-p_1 là các đường đi loại 2; u_1-p_1-u_2-p_2, u_1-p_1-u_3-p_7-u_4-p_2 là các đường đi loại 3.



Hình 2: *Đồ thị hai phía tương ứng với ma trận đánh giá của hệ gồm 5 người dùng và 7 sản phẩm*

Bằng trực quan ta dễ dàng quan sát được, nếu số lượng đường đi độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$ thuộc loại 1 chiếm đa số trong cả ba loại đường đi thì việc dự đoán sản phẩm x phù hợp với người dùng i có khả năng đúng cao nhất. Trong trường hợp này, ta dự đoán quan điểm của người dùng i đối với sản phẩm mới x có giá trị r_{ix} = 1. Nếu số lượng đường đi độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$ thuộc loại 2 chiếm đa số trong cả ba loại đường đi thì việc dự đoán sản phẩm x không phù hợp với người dùng i cũng có khả năng đúng cao nhất. Trong trường hợp này, ta dự đoán quan điểm của người dùng i đối với sản phẩm mới x có giá trị r_{ix} = -1. Trường hợp cuối cùng, nếu số lượng đường đi độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$ thuộc loại 3 chiếm đa số trong cả ba loại đường đi thì chúng ta không dự đoán được sản phẩm x có phù hợp hay không đối với người dùng i. Trong trường hợp này, ta dự đoán quan điểm của người dùng i đối với sản phẩm mới x có giá trị r_{ix} = 0. Dựa trên nhận xét này, chúng tôi đề xuất phương pháp tính toán mức độ phù hợp của người dùng đối với các sản phẩm mới như sau.

Gọi $W = \{w_{ix}: i = 1,2,..,n; x = 1,2,,.,m\}$ là ma trận liền kề biểu diễn đồ thị hai phía của ma trận đánh giá R được xác định theo công thức(2), $W^1 = \{w_{ix}^1: i = 1,2,,.,n; x = 1,2,,.,m\}$ là ma trận liền kề biểu diễn đồ thị hai phía cho các giá trị đánh giá dương được xác định theo công thức(3), $W^2 = \{w_{ix}^2: i = 1,2,,.,n; x = 1,2,,.,m\}$ là ma trận liền kề biểu diễn đồ thị hai phía cho các giá trị đánh giá âm được xác định theo công thức (4). Nói cách khác ta thực hiện

tách đồ thị hai phía biểu diễn ma trận đánh giá R thành hai đồ thị con, đồ thị con W^1 chỉ bao gồm các cạnh có trọng số dương (+1), đồ thị con W^2 chỉ bao gồm các cạnh có trọng số âm (-1).

$$W = \begin{cases} 1 & nếu\ r_{ix} \neq 0 \\ 0 & otherwise \end{cases} \quad (18)$$

$$W^1 = \begin{cases} 1 & nếu\ r_{ix} > 0 \\ 0 & otherwise \end{cases} \quad (19)$$

$$W^2 = \begin{cases} 1 & nếu\ r_{ix} < 0 \\ 0 & otherwise \end{cases} \quad (20)$$

Khi đó, số lượng tất cả các đường đi có độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$ trên toàn bộ ma trận đánh giá R được xác định theo công thức (21). Đây cũng chính là số lượng của cả ba loại đường đi từ đỉnh $i \in U$ đến đỉnh $x \in P$. Số lượng đường đi loại 1(đường đi qua các cạnh có trọng số 1) từ đỉnh $i \in U$ đến đỉnh $x \in P$ được xác định theo công thức (22) . Số lượng đường đi loại 2(đường đi qua các cạnh có trọng số -1) từ đỉnh $i \in U$ đến đỉnh $x \in P$ được xác định theo công thức (23) . Số lượng đường đi loại 3(đường đi qua các cạnh có trọng số 1 hoặc -1) từ đỉnh $i \in U$ đến đỉnh $x \in P$ được xác định theo công thức (24). Trong đó, W^T là ma trận chuyển vị của W, $(W^1)^T$ là ma trận chuyển vị của W^1, $(W^2)^T$ là ma trận chuyển vị của W^2.

$$W^L = \begin{cases} W & nếu\ \ L = 1 \\ W.W^T.W^{L-2} & nếu\ L = 3, 5, \ldots \end{cases} \quad (21)$$

$$(W^1)^L = \begin{cases} W^1 & nếu\ \ L = 1 \\ W^1.(W^1)^T.(W^1)^{L-2} & nếu\ L = 3, 5, \ldots \end{cases} \quad (22)$$

$$(W^2)^L = \begin{cases} W^2 & nếu\ \ L = 1 \\ W^1.(W^2)^T.(W^1)^{L-2} & nếu\ L = 3, 5, \ldots \end{cases} \quad (23)$$

$$(W^3)^L = W^L - (W^1)^L - (W^2)^L \quad (24)$$

Như vậy ta đã xác định được số lượng của từng loại đường đi có độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$, Gọi MAX là số lượng đường đi có độ dài L lớn nhất từ đỉnh $i \in U$ đến đỉnh $x \in P$ được xác định theo (25). Khi đó, phương pháp dự đoán mức độ phù hợp của người dùng $i \in U$ đối với các sản phẩm mới $x \in P$ được xác định theo công thức (26).

$$MAX = max\{w_{ix}^L : i = 1, 2, .., n; x = 1, 2, .., m\} \quad (25)$$

$$r_{ix} = \begin{cases} 1 & nếu\ \frac{(w^1)^L_{ix}}{MAX} > 0.5 \\ 0 & nếu\ \frac{(w^3)^L_{ix}}{MAX} > 0.5 \\ -1 & nếu\ \frac{(w^2)^L_{ix}}{MAX} > 0.5 \end{cases} \quad (26)$$

Trong công thức dự đoán, để hạn chế các cặp (i, x) có số lượng đường đi độ dài L nhỏ nhưng có số lượng đường đi độ dài L thuộc mỗi loại vẫn chiếm đại đa số so với w^L_{ix}, Chính vì vậy, chúng tôi so sánh với giá trị lớn nhất trong ma trận W^L để tiến hành so sánh. Giá trị dự đoán quan điểm của người dùng i đối với sản phẩm mới x là r_{ix}=1 khi tỉ số $\frac{(w^1)^L_{ix}}{MAX} > 0.5$. Điều này có nghĩa số lượng đường đi độ dài L từ đỉnh i đến đỉnh x phải đạt giá trị đủ lớn và số lượng đường đi độ dài L đi qua các cạnh có trọng số dương chiếm đại đa số. Tương tự như vậy, giá trị dự đoán r_{ix} =-1 khi số lượng đường đi độ dài L từ đỉnh i đến đỉnh x phải đạt giá trị đủ lớn và số lượng đường đi độ dài L đi qua các cạnh có trọng số âm chiếm đại đa số. Giá trị dự đoán r_{ix} =0 khi số lượng đường đi độ dài L từ đỉnh i đến đỉnh x phải đạt giá trị đủ lớn và số lượng dường đi độ dài L đi qua các cạnh có trọng số cả âm lẫn dương chiếm đại đa số. Dựa vào phương pháp dự đoán đã được xây dựng theo (26) luận văn đề xuất thuật toán tư vấn cộng tác cho cổng lập trình trực tuyến như trong Mục 2.2.2.

2.2.2. Thuật toán tư vấn cộng tác cho cổng lập trình trực tuyến

Thuật toán tư vấn cộng tác cho cổng lập trình trực tuyến (ký hiệu là GraphBased) được thực hiện thông qua 4 bước như trong Hình 6. Tại bước 1 của thuật toán ta tiến hành xây dựng các đồi thị hai phía. Đồ thị W dùng để xác định số lượng các đường đi có độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$. Đồ thị W^1 dùng để xác định số lượng các đường đi có độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$ chỉ đi qua các cạnh có trọng số dương. Đồ thị W^2 dùng để xác định số lượng các đường đi có độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$ chỉ đi qua các cạnh có trọng số âm. Trong đó, giá trị L được xác định thông qua thử nghiệm. Trong bài báo này, chúng thử nghiệm và lấy L=7 đã cho kết quả tốt nhất.

Tại bước 2 của thuật toán, chúng ta tiến hành tìm số lượng đường đi có độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$ trên đồ thị W. Kết quả nhận được là ma trận W^L ghi nhận số lượng tất cả các loại đường đi có độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$. Tiếp đến ta tìm được số lượng đường đi độ dài L đi qua các cạnh có trọng số dương trong ma trận

$(W^1)^L$ và số lượng đường đi độ dài *L* đi qua các cạnh có trọng số dương trong ma trận $(W^2)^L$. Cuối cùng ta xác định được ma trận $(W^4)^L$ ghi lại số lượng đường đi độ dài *L* đi qua các cạnh có trọng số cả âm cả dương.

Tại bước 3 của thuật toán chúng ta tiến hành điền các giá trị dự đoán r_{ix} theo công thức (20). Trong đó, một số nhãn lúc đầu có $r_{ix} = 0$ được thay thế bằng giá trị 1, một số khác được thay thế bằng giá trị -1, phần còn lại có giá trị 0 tương ứng với việc ta chưa thể đưa ra dự đoán. Thuật toán chi tiết được thể hiện trong Hình 6.

Thuật toán GraphBased:

Đầu vào:

- Ma trận đánh giá *R* được xác định theo công thức (17).

Đầu ra:

- Danh sách *k* sản phẩm mới $x \in P$ phù hợp nhất đối với người dùng $i \in U$.

Các bước tiến hành:

Bước 1. Xây dựng các đồ thị hai phía từ ma trận đánh giá *R*:

1.1. Xây dựng ma trận kề biểu diễn đồ thị hai phía trên toàn bộ *R* theo công thức (18):

$$W = \begin{cases} 1 & nếu\ r_{ix} \neq 0 \\ 0 & otherwise \end{cases}$$

1.2. Xây dựng ma trận kề biểu diễn đồ thị hai phía cho các đánh giá có giá trị 1 theo công thức (19):

$$W^1 = \begin{cases} 1 & nếu\ r_{ix} > 0 \\ 0 & otherwise \end{cases}$$

1.3. Xây dựng ma trận kề biểu diễn đồ thị hai phía cho các đánh giá có giá trị -1 theo công thức (20):

$$W^2 = \begin{cases} 1 & nếu\ r_{ix} < 0 \\ 0 & otherwise \end{cases}$$

Bước 2. Tìm số lượng đường đi độ dài *L* của mỗi loại:

2.1. Tìm số lượng đường đi độ dài *L* từ đỉnh $i \in U$ đến đỉnh $x \in P$ trên toàn bộ *R* theo công thức (21):

$$W^L = \begin{cases} W & nếu\ \ L = 1 \\ W.W^T.W^{L-2} & nếu\ L = \ 3,5,\ldots \end{cases}$$

2.2. Tìm số lượng đường đi loại 1 có độ dài L từ đỉnh i∈U đến đỉnh x∈P theo công thức (22):

$$(W^1)^L = \begin{cases} W^1 & nếu\ \ L = 1 \\ W^1.(W^1)^T.(W^1)^{L-2} & nếu\ L = \ 3,5,\ldots \end{cases}$$

2.3. Tìm số lượng đường đi loại 2 có độ dài L từ đỉnh i∈U đến đỉnh x∈P theo công thức (23):

$$(W^2)^L = \begin{cases} W^2 & nếu\ \ L = 1 \\ W^2.(W^2)^T.(W^2)^{L-2} & nếu\ L = \ 3,5,\ldots \end{cases}$$

2.4. Tìm số lượng đường đi loại 3 có độ dài L từ đỉnh i∈U đến đỉnh x∈P theo công thức (24):

$$(W^3)^L = W^L - (W^1)^L - (W^2)^L$$

Bước 3. Dự đoán quan điểm của $i \in U$ đối với các sản phẩm mới $x \in P$:

3.1. Tìm số lượng đường đi độ dài L từ đỉnh $i \in U$ đến đỉnh $x \in P$ trên toàn bộ R theo công thức (25):

$$MAX = max\{w_{ix}^L : i = 1,2,..,n; x = 1,2,..,m\}$$

3.2. Sinh ra dự đoán quan điểm $i \in U$ đối với các sản phẩm mới $x \in P$ theo công thức (26):

$$r_{ix} = \begin{cases} 1 & nếu \ \frac{(w^1)_{ix}^L}{MAX} > 0.5 \\ 0 & nếu \ \frac{(w^3)_{ix}^L}{MAX} > 0.5 \\ -1 & nếu \ \frac{(w^2)_{ix}^L}{MAX} > 0.5 \end{cases}$$

Bước 4. Tạo nên tư vấn cho người dùng $i \in U$ các sản phẩm mới $x \in P$**:**

4.1. Sắp xếp r_{ix} theo thứ tự tăng dần của trọng số.

4.2. Chọn k sản phẩm mới đầu tiên có r_{ix}=1 để tư vấn cho người dùng i.

Hình 2: *Thuật toán GraphBased.*

Ví dụ: Xét ma trận trọng số sau:

	U_1	U_2	U_3	U_4	U_5	U_6
P_1	1	0	-1	-1	0	1
P_2	0	1	1	-1	1	0
P_3	-1	-1	1	0	0	1
P_4	1	1	-1	0	-1	-1
P_5	0	-1	0	1	1	0
P_6	-1	0	0	-1	1	1

Bảng 2: *Ma trận đánh giá R*

Bước 1: Xây dựng các đồ thị hai phía

- Ma trận kề biểu diễn đồ thị hai phía trên toàn bộ R:

$$\begin{vmatrix} 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \end{vmatrix}$$

- Ma trận kề biểu diễn cho các đánh giá có giá trị 1:

$$\begin{vmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{vmatrix}$$

$$\begin{vmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{vmatrix}$$

- Ma trận kề biểu diễn cho các đánh giá có giá trị -1:

$$\begin{vmatrix} 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{vmatrix}$$

Bước 2: Xác định số lượng đường đi độ dài L của mỗi loại, ở đây chúng ta xét L = 5.

- Tìm số lượng đường di độ dài L = 5 trên toàn bộ R:

$W^5 = W.W^T.W.W^T.W$

$$= \begin{vmatrix} 196 & 171 & 191 & 161 & 169 & 196 \\ 175 & 175 & 180 & 162 & 175 & 075 \\ 199 & 178 & 197 & 159 & 171 & 199 \\ 241 & 222 & 240 & 201 & 217 & 241 \\ 125 & 129 & 129 & 122 & 132 & 125 \\ 188 & 169 & 183 & 163 & 172 & 188 \end{vmatrix}$$

- Tìm số lượng đường đi độ dài L = 5 có giá trị 1:

$(W^1)^5 = W^1.W^{1T}.W^1.W^{1T}.W^1$

$$= \begin{vmatrix} 11 & 7 & 8 & 1 & 9 & 17 \\ 8 & 18 & 19 & 6 & 26 & 16 \\ 7 & 8 & 13 & 2 & 14 & 18 \\ 10 & 11 & 7 & 1 & 8 & 8 \\ 2 & 7 & 8 & 6 & 17 & 8 \\ 8 & 9 & 13 & 5 & 20 & 19 \end{vmatrix}$$

- Tìm số lượng đường đi độ dài L = 5 có giá trị 1:

$(W^2)^5 = W^2.W^{2T}.W^2.W^{2T}.W^2$

$$\begin{vmatrix} 6 & 1 & 12 & 16 & 5 & 5 \\ 5 & 1 & 5 & 11 & 1 & 1 \end{vmatrix}$$

$$= \begin{vmatrix} 10 & 9 & 1 & 6 & 0 & 0 \\ 1 & 0 & 15 & 7 & 10 & 10 \\ 4 & 5 & 0 & 1 & 0 & 0 \\ 11 & 5 & 6 & 16 & 1 & 1 \end{vmatrix}$$

- Tìm số lượng đường đi loại 3 độ dài L = 5:

$$(W^3)^5 = W^5 - (W^1)^5 - (W^2)^5$$

$$= \begin{vmatrix} 179 & 163 & 171 & 144 & 155 & 174 \\ 162 & 156 & 156 & 145 & 148 & 158 \\ 182 & 161 & 183 & 151 & 157 & 181 \\ 230 & 211 & 218 & 193 & 199 & 223 \\ 119 & 117 & 121 & 115 & 115 & 117 \\ 170 & 155 & 164 & 142 & 151 & 168 \end{vmatrix}$$

Bước 3: Sinh dự đoán

Ví dụ ta sinh dự đoán của U_2 lên bài toán P_1

$$r_{1,2} = 0 \text{ vì } \frac{W^3{}_{1,2}}{Max} = \frac{163}{171} > 0.5$$

Điều đó có nghĩa là chưa thể dự đoán được U_2 có thể làm được bài toán P_1 hay không.

2.3. Kết luận

Chương này tôi đã đề xuất 3 phương pháp cho cổng lập trinh trực tuyến, ở hai phương pháp *User-based* và *Item-based,* đây là hai phương pháp phổ biến và chúng khắc phục những nhược điểm của nhau, ở phương pháp *User-based* sẽ không có tính hiệu quả, thực tế cho thấy nếu cổng lập trình trực tuyến có một lượng người dùng lớn hơn rất nhiều so với số lượng bài toán thì lúc này hệ thống sẽ gặp rất nhiều khó khăn trong khâu tính toán ma trận tương tự giữa các cặp người dùng vì lúc này ma trận tương quan là rất lớn còn về phương pháp *Item-based* cho thấy hiệu quả hơn ở phương diện người dùng và bài toán bởi trên thực tế số lượng người dùng đa phần sẽ lớn hơn rất nhiều so với bài toán cũng chính vì điều này sẽ làm ma trận đánh giá của *Item-based* sẽ được đầy đủ và có được những đánh giá chất lượng hơn từ nhiều người dùng không những thế với số lượng bài toán ít hơn rất nhiều người dùng sẽ giúp hệ thống tính toán và đưa ra gợi ý nhanh hơn so với phương pháp *User-based.* Tuy vậy cả hai phương pháp lại có nhược điểm là nếu dữ liệu thưa thớt sẽ rất khó để đưa ra phán đoán chính xác bởi dữ liệu thưa thớt sẽ ảnh hưởng rất lớn đến khâu tính toán độ tương quan

như tôi đã trình bày hai phương pháp khi sử dụng công thức tương quan cosin, để tăng độ chính xác phải cần một lượng dữ liệu lớn điều này rất khó với hệ thống mới chạy giai đoạn đầu chính vì thế mà luận văn đã đề xuất phương pháp lọc cộng tác dựa trên mô hình đồ thị, phương pháp này mạnh mẽ ở điểm có thể đưa ra dự đoán ngay khi dữ liệu không đủ lớn bằng cách tính toán số đường theo các loại mà bài toán đưa ra từ số lượng đường đi tính được ta có thể phán đoán quan điểm của người dùng và bài toán theo tỷ lệ số loại đường đi mà cặp người dùng và bài toán có thể đi được. Ở chương sau, luận văn sẽ thực nghiệm trên cổng lập trình trực tuyến của Học Viện Công Nghệ Bưu Chính Viễn Thông để thấy được sự ưu việt của phương pháp lọc cộng tác theo mô hình đồ thị hai phía so với hai phương pháp còn lại.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Phương pháp thực nghiệm

Thuật toán dựa trên đồ thị (GraphBased) đề xuất được tiến hành thử nghiệm trên tập dữ liệu được thu thập từ cổng lập trình trực tuyến Dlab của Học viện Công nghệ Bưu chính Viễn thông. Phương pháp xây dựng bộ dữ liệu và kết quả thử nghiệm được đánh giá và so sánh với các phương pháp khác theo thủ tục được mô tả như dưới đây.

Trước tiên, toàn bộ tập người dùng được chia thành hai phần, một phần U_{tr} được sử dụng làm dữ liệu huấn luyện, phần còn lại U_{te} được sử dụng để kiểm tra. Tập dữ liệu huấn luyện dùng để xây dựng mô hình theo các thuật toán lọc được sử dụng. Với mỗi người dùng $u \in U_{te}$, các đánh giá $r_{u,p} \neq \varnothing$ được chia thành hai phần O_u và P_u. O_u được coi là đã biết, trong khi đó P_u là đánh giá cần dự đoán từ dữ liệu huấn luyện và O_u. Giả sử phương pháp lọc đưa ra dự đoán cho người dùng trong tập P_u là P'_u, Khi đó, sai số dự đoán được thực hiện bằng cách so sánh các đánh giá trong hai tập P_u là P'_u,

Độ đo trung bình giá trị tuyệt đối lỗi

Đánh giá sai số phân loại trung bình giá trị tuyệt đối lỗi (MAE) được Breese đề xuất năm 1998 được xem là phương pháp tiêu chuẩn cho lọc cộng tác. Sai số dự đoán MAE_u cho mỗi người dùng u thuộc tập dữ liệu kiểm tra được tính bằng trung bình giá trị tuyệt đối giữa hiệu số giá trị dự đoán và giá trị thực đối với tất cả mặt hàng thuộc tập P_u.

$$MAE_u = \frac{1}{|P_u|} \sum_{y \in P_u} \left| \hat{r}_y^{\,u} - r_y^{\,u} \right| \qquad (27)$$

Sai số dự đoán trên toàn tập dữ liệu kiểm tra được tính bằng trung bình cộng sai số dự đoán cho mỗi khách hàng thuộc U_{te} được tính toán theo công thức (27). Giá trị MAE càng nhỏ, phương pháp dự đoán càng chính xác.

$$MAE = \frac{\sum_{u \in U_{te}} MAE_u}{|U_{te}|} \qquad (28)$$

3.2. Dữ liệu thực nghiệm

Phương pháp tư vấn cộng tác do nhóm nghiên thu thập trực tiếp từ cổng lập trình trực tuyến của Học viện Công nghệ Bưu chính Viễn thông, Tập dữ liệu thu thập từ tháng 8/2020 đến tháng 6 năm 2021 được 6435 người dùng đăng ký tham gia cổng lập trình trực tuyến. Kho nội dung số được xây dựng trong vòng 1 năm với 1245 bài toán để người học có thể

lập trình và chấm bài tự động. Tổng số lượt giải bài (submission) của 6435 người dùng ghi nhận trong cổng lập trình trực tuyến đến ngày 20/6/2021 là 1,151,000. Trong đó, người dùng *i* lập trình đúng bài toán *x* cổng lập trình ghi lại giá trị 1, người dùng *i* lập trình chưa đúng bài toán *x* cổng lập trình ghi lại giá trị -1, người dùng *i* chưa giải bài toán *x* cổng lập trình ghi lại giá trị 0. Mỗi người lập trình có thể submit một bài nhiều lần và hệ thống chỉ ghi nhận giá trị 1, -1 cho kết quả cuối cùng. Trong số 6435 người dùng, luận văn lọc ra được 6120 người dùng đã tham gia lập trình ít nhất 20 bài dù đúng hoặc sai để tiến hành thử nghiệm. Dữ liệu đầu vào được minh họa ở hình 3.

1	1	1
1	2	1
1	3	1
1	9	1
1	10	1
1	11	1
1	17	1
1	20	1
1	21	1
1	27	1
1	33	1
1	34	1
1	45	-1
2	1	1
2	2	1
2	3	1
2	8	-1
2	9	-1
2	11	1
2	12	-1
2	17	1
2	21	1
2	33	1
2	34	-1
2	45	-1
2	67	1

Hình 3: *Ảnh minh họa dữ liệu đầu vào*

Ở hình 3, cột đầu tiên là thể hiện id của người dùng, cột thứ 2 là id của bài tập, cột thứ 3 là trạng thái giải bài của người dùng đối với bài tập.

Chọn ngẫu nhiên 2000, 3000, và 4000 trong tập 6120 người dùng làm dữ liệu huấn luyện. Chọn ngẫu nhiên 400, 600, và 800 người trong số còn lại làm tập dữ liệu kiểm tra. Để thử nghiệm khả năng của phương pháp mới đề xuất so với những phương pháp khác trong trường hợp có ít dữ liệu, chúng tôi thay đổi số lượng bài lập trình của mỗi người dùng trong

tập kiểm tra sao cho số lượng bài đã lập trình lần lượt là 5, 10 và 20, phần còn lại là những đánh giá cần dự đoán.

3.3. Kết quả thực nghiệm

Phương pháp GraphBased đề xuất trong Mục 2.2 được thử nghiệm và so sánh với những phương pháp sau:

- Phương pháp Userbased sử dụng độ tương quan Person, đây là phương pháp lọc cộng tác dựa trên người dùng được trình bày ở mục 1.3.1
- Phương pháp Itembased sử dụng độ tương quan Person, đây là phương pháp lọc cộng tác dựa trên sản phẩm được trình bày ở mục 1.3.2

Giá trị MAE trong Bảng 10 được ước lượng từ trung bình của 10 lần thử nghiệm ngẫu nhiên. Kết quả thử nghiệm cho thấy phương pháp đề xuất đều cho giá trị MAE nhỏ hơn phương pháp UseBassed và ItemBased trong tất cả các trường hợp dữ liệu biết trước là 5, 10 hay 20 đánh giá. Cụ thể, trong trường hợp dữ liệu rất thưa với số lượng đánh giá biết trước trong tập dữ liệu kiểm tra là 5 thì giá trị MAE của phương pháp Usebased, ItemBased, GraphBased lần lượt là 0.2158, 0.2172, 0.208 trên tập dữ liệu huấn luyện 2000 người dùng. Giá trị MAE của các phương pháp Userbased, ItemBased, GraphBased có xu hướng nhỏ dần khi kích cỡ tập dữ liệu huấn luyện tăng lên 3000 và 4000 người dùng. Tuy nhiên, phương pháp GraphBased vẫn có giá trị MAE nhỏ hơn các phương pháp còn lại. Điều này chứng tỏ phương pháp đề xuất phát huy được hiệu quả ngay cả trường hợp dữ liệu thưa của lọc cộng tác.

Trong trường hợp có tương đối đầy đủ dữ liệu, cụ thể với số lượng đánh giá biết trước là 20 thì giá trị MAE của phương pháp GraphBased có giá trị nhỏ hơn hẳn các phương pháp còn lại. Giá trị MAE của phương pháp GraphBased là 0.1812, 0.1792, 0.1712, hai phương pháp Userbased-Graph, Itembased-Graph cũng cho 1 kết quả tương đối tốt không kém Graph-Based với MAE trong khoảng [0.17,0.19] trong khi đó các phương pháp UserBased và ItemBased đều cho kết quả MAE >1.820. Điều này chỉ có thể lý giải các phương pháp tư vấn dựa trên độ tương quan thực hiện ước lượng mức độ giữa các cặp người dùng hoặc sản phẩm trực tiếp trên tập giao các đánh giá về sản phẩm hay người dùng. Tuy nhiên, với phương pháp dựa vào đồ thị, chúng ta có thể suy diễn mức độ trên tập các đường đi có trọng số dương và tập các đường đi có trọng số âm đồng thời không đưa ra dự đoán với các đường đi có cả trọng số âm lẫn dương. Điều này cho phép ta tận dụng được các mối liên hệ gián tiếp vào kết quả dự đoán.

Mặc dù kết quả của các phương pháp Graph cho ra 1 kết quả tốt hơn hẳn những phương pháp thông thường nhưng lại gặp 1 vấn đề về sự liên thông giữa các đỉnh trong các đồ thị tức sẽ có 1 cặp đỉnh không có cạnh nối và hoàn toàn bị tách ra khỏi đồ thị nếu chúng ta có 1 tập người dùng và bài toán khổng lồ nhưng dữ liệu lại cực kì thưa thớt điều này sẽ dẫn đến khi ta tách mô hình đồ thị tổng quát thành các đồ thị con thể hiện các đường âm và dương sẽ dẫn đến đồ thị này sẽ không thể liên thông, khi đồ thị không liên thông sẽ ảnh hưởng rất cao đến kết quả khi ta tăng độ dài $L = 5,7,9 \ldots$

Để khắc phục nhược điểm này tôi đã tăng cường lọc dữ liệu cho tập training, có thể thấy rằng vấn đề đồ thị không liên thông là do sự thưa thớt dữ liệu giữa các đỉnh với nhau chính vì vậy mà tôi đã chọn ra 6000 người dùng đã làm ít nhất 20 bài toán và mỗi bài toán phải tối thiểu có 20 người đã làm điều này không thể chắc chắn sẽ cho chúng ta một đồ thị liên thông nhưng nó có thể hạn chế bớt sự không liên thông giữa các đỉnh trong đồ thị. Có thể tăng dữ liệu lên 30, 50 để khắc phục vấn đề này. Một hướng khác là chúng ta có thể thấy tôi đã thử nghiệm mỗi phương pháp 10 lần để lấy ra kết quả trung bình, điều này cũng nhằm hạn chế trong 1 vài trường hợp có thể xuất hiện đồ thị không liên thông nên lấy trung bình kết quả cũng là 1 cách có thể giảm thiểu sự sai số khi gặp đồ thị không liên thông. Để mà có thể chắc chắn khi tách thành các đồ thị con thể hiện đường đi âm và dương là 1 đồ thị liên thông thì yếu tố dữ liệu rất quan trọng, với 2 phương pháp trên chỉ 1 phần nào hạn chế được sự không liên thông xảy ra chứ không hoàn toàn khắc phục triệt để vấn đề này.

Kích thước tập dữ liệu huấn luyện	Phương pháp	Số lượng đánh giá biết trước		
		5	10	20
2000 người dùng	UserBased	0.2158	0.2117	0.2042
	ItemBased	0.2172	0.2146	0.1992
	GraphBased	**0.2068**	**0.1984**	**0.1872**
3000 người dùng	UserBased	0.2147	0.2012	0.1924
	ItemBased	0.2145	0.2116	0.1967
	GraphBased	**0.2043**	**0.1877**	**0.1792**
4000 người dùng	UserBased	0.2037	0.2117	0.1942
	ItemBased	0.2012	0.2146	0.1820
	GraphBased	**0.1987**	**0.1784**	**0.1712**

Bảng 3: *Giá trị MAE của các phương pháp*

3.4. Kết luận

Nghiên cứu đã đề xuất một phương pháp tư vấn lọc cộng tác trên cổng lập trình trực tuyến của Học viện Công nghệ Bưu chính Viễn thông nhằm cung cấp cho tập bài toán phù hợp với khả năng lập trình cho mỗi người dùng. Mô hình đề xuất phát huy hiệu quả ngay cả trong trường hợp dữ liệu thưa mà các mô hình tư vấn lọc cộng tác khác thường gặp phải khó khăn. Phương pháp tư vấn cộng tác đề xuất đã đem lại hiệu quả thiết thực cho người học lập trình trực tuyến của Học viện Công nghệ Bưu chính Viễn thông.