The Saigon International University

Khóa luận tốt nghiệp

Website: www.siu.edu.vn

Thành phố Hồ Chí Minh - 2024

KHÓA LUẬN TỐT NGHIỆP

Ngành
Khoa học máy tính

Đề tài
Xây dựng hệ thống khuyến nghị cộng tác học giả

Giảng viên hướng dẫn
T.S Huỳnh Ngọc Tín

Sinh viên
Huỳnh Lý Hữu Phúc
MSSV: 91012002066

LỜI CAM ĐOAN

Em là tác giả của báo cáo này, xin cam đoan rằng toàn bộ nội dung và kết quả được trình bày trong báo cáo đồ án là công trình nghiên cứu, phân tích và tổng hợp của em. Em cam đoan rằng không có bất kỳ phần nào trong báo cáo này được sao chép từ công trình của người khác mà không được trích dẫn đúng cách. Mọi thông tin, dữ liệu số, hình ảnh và tài liệu tham khảo đã được trích dẫn rõ ràng và chính xác theo nguồn gốc.

Em cũng cam đoan rằng em đã tuân thủ các quy định, quy tắc và nguyên tắc về nghiên cứu và viết báo cáo, bao gồm việc trích dẫn các tác giả và công trình đã được sử dụng trong quá trình nghiên cứu.

Tp. Hồ Chí Minh, ngày tháng năm 2024

Sinh viên

(Ký và rõ ghi họ tên)

LỜI CẢM ƠN

Em xin chân thành gửi lời cảm ơn đến giảng viên hướng dẫn cũng như các anh chị trong trung tâm SIU AILab đã hỗ trợ em trong quá trình thực hiện đồ án này.

Đầu tiên, em xin chân thành cảm ơn TS. Huỳnh Ngọc Tín, là người thầy đã tận tình hướng dẫn giúp đỡ em vượt qua những khó khăn và hoàn thành khóa luận tốt nghiệp của mình. Thầy đã đưa ra những lời khuyên bổ ích không chỉ áp dụng trong khóa luận mà còn là kim chỉ nam cho hướng phát triển sau này của em.

Em cũng muốn gửi lời cảm ơn tới các anh chị thành viên trong SIU AILab, những người đã hỗ trợ trong lúc hoàn thành khóa luận và nỗ lực của mình để thực hiện đồ án này. Sự cộng tác và tương tác giữa em đã tạo nên một môi trường làm việc tích cực và trao đổi ý tưởng sáng tạo.

Cuối cùng, em muốn tri ân đến gia đình, bạn bè và những người thân yêu đã luôn đứng về phía em, cổ vũ và động viên trong suốt quá trình nghiên cứu và viết báo cáo. Sự động viên và hỗ trợ của họ đã cung cấp cho em sự động lực và niềm tin để vượt qua mọi khó khăn.

Một lần nữa, em xin chân thành cảm ơn tất cả mọi người đã đóng góp và hỗ trợ em trong quá trình thực hiện đồ án này. Sự giúp đỡ của mọi người đã là một phần quan trọng trong thành công của em

Tp. Hồ Chí Minh, ngày tháng năm 2024

Sinh viên

(Ký và rõ ghi họ tên)

MỤC LỤC

LỜI CAM ĐOAN i

LỜI CẢM ƠN ii

MỤC LỤC iii

DANH SÁCH HÌNH ẢNH vi

DANH MỤC CÁC KÝ HIỆU, CHỮ VIẾT TẮT vii

CHƯƠNG 1. TỔNG QUAN 1

1.1. Lý do chọn đề tài: 1

1.1.1. Đặt vấn đề 1

1.1.2. Tại sao cần xây dựng hệ thống khuyến nghị cộng tác? 2

1.1.3. Khó khăn khi xây dựng hệ thống khuyến nghị cộng tác học giả 3

1.2. Mục tiêu và phạm vi của đề tài 4

1.2.1. Mục tiêu 4

1.2.2. Phạm vi 4

1.3. Cấu trúc khóa luận 4

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT 6

2.1. Mở đầu 6

2.1.1. Định nghĩa khuyến nghị cộng tác 6

2.1.2. Vai trò của khuyến nghị cộng tác 7

2.2. Các hướng tiếp cận xây dựng hệ thống khuyến nghị cộng tác học giả 8

2.2.1. Tiếp cận nội dung (Content-based filtering). 9

2.2.2. Phương pháp lọc cộng tác (Collaborative filtering). 10

2.2.3. Phương pháp lai (Hybrid). 12

2.3. Các công nghệ sử dụng để xây dựng hệ thống khuyến nghị cộng tác học giả 14

2.3.1. SciBERT. 14

2.3.1.1. Giới thiệu về SciBERT. 14

2.3.1.2. Tại sao sử dụng SciBERT trong hệ thống cộng tác học giả? 17

2.3.2. DBLP. 19

2.3.2.1. Giới thiệu về DBLP. 19

2.3.2.2. Tại sao sử dụng DBLP trong khoá luận? 21

CHƯƠNG 3. THIẾT KẾ HỆ THỐNG 24

3.1. Mở đầu 24

3.1.1. Xử lý dữ liệu từ DBLP. 24

3.1.2. Tạo vector cho các nghiên cứu viên từ dữ liệu DBLP đã qua xử lý bằng SciBERT. 26

3.1.3. Khuyến nghị cộng tác học giả bằng độ tương tự Cosine. 28

3.1.3.1. Độ tương tự Cosine. 28

3.1.3.2. Sử dụng độ tương tự Cosine trong khoá luận. 30

CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ 33

4.1. Mở đầu 33

4.2. Phương pháp đánh giá hệ thống khuyến nghị cộng tác học giả. 33

4.2.1. Tập dữ liệu dùng để đánh giá hệ thống khuyến nghị cộng tác học giả. . 33

4.2.2. MRR. 34

4.2.3. nDCG. 36

4.3. Kết quả thực nghiệm. 40

CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 42

5.1. Kết luận 42

5.2. Hướng phát triển 42

TÀI LIỆU THAM KHẢO 44

DANH SÁCH HÌNH ẢNH

Hình 2.1: Minh hoạ hệ thống khuyến nghị cộng tác học giả 6

Hình 3.1: Tên tác giả bị trùng lắp. 25

Hình 3.2: Quy trình xử lý dữ liệu từ DBLP. 25

Hình 3.3: Tên tác giả đã qua xử lý trùng lắp. 26

Hình 3.4: Sơ đồ quy trình chuyển đổi thành vector từ các bài báo của các tác giả . 27

Hình 3.5: Một vài ví dụ về vector đại diện cho mỗi tác giả..... 28

Hình 3.6: Công thức tính độ tương tự Cosine..... 29

Hình 3.7: Ví dụ về khuyến nghị Top 5 tác giả sử dụng tương tự Cosine. 32

Hình 4.1: Công thức MRR..... 34

Hình 4.2: Công thức của thước đo Cumulative Gain. 36

Hình 4.3: Công thức của thước đo Discounted Cumulative Gain. 37

Hình 4.4: Công thức nDCG. 37

DANH MỤC CÁC KÝ HIỆU, CHỮ VIẾT TẮT

KÍ HIỆU, CHỮ VIẾT TẮT	Ý NGHĨA
BERT	Bidirectional Encoder Representations from Transformers - mô hình biểu diễn từ theo 2 chiều ứng dụng kỹ thuật Transformer
NLP	Natural Language Processing – Xử lý ngôn ngữ tự nhiên
RWR	Random Walk with Restart- là một thuật toán đưa ra sự gần gũi giữa hai nút trong biểu đồ
DBLP	Digital Bibliography & Library Project - là một trang web thư mục khoa học máy tính
LDA	Latent Dirichlet Allocation - mô hình thuộc lớp mô hình sinh xác suất (generative probabilistic model) của một bộ văn bản
CBF	Content-Based Filtering - phương pháp truy xuất thông tin sử dụng các tính năng của mục để chọn và trả về các mục có liên quan đến truy vấn của người dùng
MRR	Mean Reciprocal Rank - là một thang đo phổ biến dùng để đánh giá hiệu suất của các hệ thống khuyến nghị và các hệ thống truy xuất thông tin.
nDCG	Normalized Discounted Cumulative Gain - là một thang đo hiệu quả được sử dụng rộng rãi để đánh giá các hệ thống khuyến nghị.

CHƯƠNG 1. TỔNG QUAN

1.1. Lý do chọn đề tài:

1.1.1. Đặt vấn đề

Trong thời đại khoa học công nghệ phát triển nhanh chóng, sự hợp tác giữa các nhà nghiên cứu đóng vai trò then chốt trong việc thúc đẩy tiến bộ khoa học và công nghệ. Sự hợp tác không chỉ giúp tăng cường khả năng tiếp cận tri thức, mà còn mở ra những cơ hội mới để giải quyết các vấn đề phức tạp thông qua sự kết hợp đa dạng các chuyên môn và kỹ năng. Tuy nhiên, việc tìm kiếm và xây dựng mối quan hệ hợp tác phù hợp vẫn là một thách thức lớn đối với nhiều nhà nghiên cứu.

Vấn đề chính nằm ở việc làm thế nào để kết nối các nhà nghiên cứu có chung mối quan tâm hoặc chuyên môn nhưng có thể không biết về sự tồn tại của nhau. Thêm vào đó, sự phát triển nhanh chóng của số lượng tài liệu nghiên cứu và thông tin khoa học càng làm tăng độ khó trong việc xác định các đối tác tiềm năng. Do đó, cần có một giải pháp hiệu quả để giúp các nhà nghiên cứu tìm thấy và tiếp cận lẫn nhau một cách nhanh chóng và chính xác.

Hệ thống khuyến nghị cộng tác học giả là một giải pháp tiềm năng cho vấn đề này. Bằng cách sử dụng các phương pháp phân tích dữ liệu và học máy, hệ thống này có thể đề xuất các đối tác nghiên cứu tiềm năng dựa trên các yếu tố như lĩnh vực nghiên cứu, công trình đã công bố, và mối quan tâm chung. Điều này không chỉ giúp tăng cường khả năng kết nối mà còn mở ra nhiều cơ hội hợp tác đa dạng và phong phú.

1.1.2. Tại sao cần xây dựng hệ thống khuyến nghị cộng tác?

Việc xây dựng hệ thống khuyến nghị cộng tác học giả là cần thiết vì một số lý do quan trọng sau:

- Tăng cường khả năng hợp tác: hợp tác giữa các nhà nghiên cứu đến từ các lĩnh vực khác nhau có thể dẫn đến những phát hiện đột phá và giải quyết những vấn đề phức tạp. Hệ thống khuyến nghị sẽ giúp các nhà nghiên cứu tìm thấy những đối tác có kỹ năng và chuyên môn phù hợp, từ đó tăng cường khả năng hợp tác và sáng tạo.
- Tiết kiệm thời gian và nguồn lực: việc tìm kiếm đối tác nghiên cứu phù hợp thường mất nhiều thời gian và công sức. Hệ thống khuyến nghị sẽ tự động hóa quá trình này, giúp các nhà nghiên cứu tiết kiệm thời gian và tập trung vào công việc nghiên cứu chính của họ.
- Mở rộng mạng lưới khoa học: hệ thống khuyến nghị cộng tác sẽ giúp các nhà nghiên cứu kết nối với nhau trên toàn cầu, mở rộng mạng lưới khoa học và tạo ra cơ hội hợp tác quốc tế. Điều này không chỉ nâng cao chất lượng nghiên cứu mà còn thúc đẩy sự phát triển của khoa học toàn cầu.
- Tối ưu hoá kết quả nghiên cứu: bằng cách kết nối các nhà nghiên cứu có cùng mối quan tâm và chuyên môn, hệ thống khuyến nghị giúp tối ưu hóa kết quả nghiên cứu. Các nhà nghiên cứu có thể chia sẻ dữ liệu, ý tưởng và phương pháp, từ đó nâng cao hiệu quả và chất lượng công trình nghiên cứu.
- Hỗ trợ đào tạo và phát triển: hệ thống khuyến nghị cũng có thể giúp các nhà nghiên cứu trẻ hoặc mới bắt đầu sự nghiệp tìm thấy những người hướng dẫn phù hợp, từ đó hỗ trợ quá trình đào tạo và phát triển chuyên môn.

1.1.3. Khó khăn khi xây dựng hệ thống khuyến nghị cộng tác học giả.

Một số khó khăn của các phương pháp khuyến nghị hiện nay có thể kể đến như:

- Khối lượng dữ liệu lớn: Hệ thống cần thu thập và xử lý một lượng lớn dữ liệu từ nhiều nguồn khác nhau, bao gồm các bài báo khoa học, hồ sơ tác giả, cơ sở dữ liệu trích dẫn, và các tài liệu liên quan khác. Điều này đòi hỏi một hạ tầng kỹ thuật mạnh mẽ để xử lý và quản lý dữ liệu hiệu quả. Thêm vào đó, chất lượng dữ liệu không đồng bộ và thiếu chính xác cũng là một thách thức cho việc làm sạch và chuẩn hoá dữ liệu.
- Độ chính xác và chất lượng khuyến nghị: Quan sát thiếu hay không quan sát được một số thông tin về lĩnh vực nghiên cứu yêu thích, cơ quan nghiên cứu khi dữ liệu ban đầu còn ít hoặc thiếu, việc đưa ra các khuyến nghị chính xác gặp nhiều khó khăn. Điều này đặc biệt quan trọng trong môi trường nghiên cứu, nơi mà các lĩnh vực và sở thích có thể rất đa dạng và phức tạp. Khi khuyến nghị cho người dùng mới hay đối tượng khuyến nghị mới, hệ thống cần có khả năng đưa ra các khuyến nghị hợp lý ngay cả khi không có nhiều dữ liệu về người dùng mới hoặc đối tượng khuyến nghị mới.
- Vấn đề khởi động lạnh (cold start): Quan sát thiếu hay không quan sát được một số thông tin về lĩnh vực nghiên cứu yêu thích, cơ quan nghiên cứu. Hoặc làm thế nào để thực hiện khuyến nghị cho những người dùng mới hay đối tượng khuyến nghị mới.
- Các phương pháp đánh giá kết quả khuyến nghị: khó khăn trong việc đo lường sự phù hợp sự phù hợp của một khuyến nghị không phải lúc nào cũng có thể đo lường được bằng các chỉ số định lượng. Phản hồi từ người dùng thường mang tính định tính và có thể khó để tích hợp vào quá trình đánh giá. Thêm vào đó,

các mối quan hệ học thuật có thể phức tạp và đa chiều, không chỉ đơn thuần dựa trên số lượng cộng tác hoặc lĩnh vực nghiên cứu chung mà còn phụ thuộc vào các yếu tố khác như chất lượng công trình nghiên cứu và ảnh hưởng trong cộng đồng khoa học.

1.2. Mục tiêu và phạm vi của đề tài

1.2.1. Mục tiêu

Xây dựng hệ thống khuyến nghị cộng tác học giả sử dụng một mô hình ngôn ngữ tiên tiến dựa trên tập dữ liệu về chuyên ngành khoa học máy tính.

1.2.2. Phạm vi

Lĩnh vực bài báo: chuyên ngành khoa học máy tính từ năm 2010 đến năm 2020.

1.3. Cấu trúc khóa luận

Khoá luận tốt nghiệp này được cấu trúc như sau:

Chương 1: Giới thiệu tổng quan về khóa luận.

Chương 2: Trình bày các cơ sở lý thuyết và các hướng tiếp cận khi xây dựng các hệ thống khuyến nghị cộng tác hiện nay.

Chương 3: Trình bày hướng tiếp cận của khoá luận, đề xuất để xây dựng hệ thống khuyến nghị cộng tác học giả.

Chương 4: Trình bày về phương pháp đánh giá và kết quả đánh giá của hệ thống.

Chương 5: Đưa ra kết luận và hướng phát triển tiềm năng của đề tài.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Mở đầu

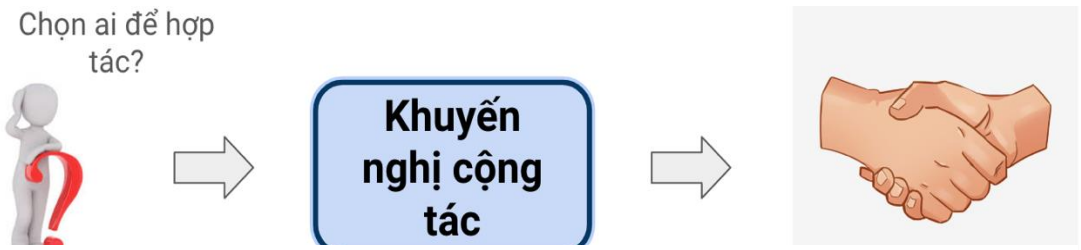
2.1.1. Định nghĩa khuyến nghị cộng tác

Khuyến nghị cộng tác trong nghiên cứu khoa học là bài toán tự động liệt kê những người, nhóm cộng tác tiềm năng ứng với đầu vào là một hay nhóm những nghiên cứu viên. Khuyến nghị cộng tác đóng vai trò quan trọng và gần đây đã bắt đầu thu hút nhiều quan tâm.

Trong phạm vi khoá luận này, bài toán khuyến nghị cộng tác được giải quyết với đầu vào là một nghiên cứu viên, hệ thống có nhiệm vụ sinh ra danh sách xếp hạng những người cộng tác tiềm năng. Bài toán có thể được định nghĩa một cách hình thức như sau:

• Đầu vào: R = {r}: tập tất cả các nghiên cứu viên. P = {p}: tập tất cả các bài báo trong kho dữ liệu. O = {o}: danh sách các cơ quan nơi các nghiên cứu viên đang làm việc.

• Đầu ra: Xác định hàm f(ri , rj) để ước lượng tiềm năng quan hệ cộng tác của ri ∈ R với rj ∈ R, ri ≠ rj . ∀r ∈ R, dựa trên hàm f chọn TopN các tác giả tiềm năng nhất, R TopN ⊂ R, R TopN =< r1, r2, ..., rT opN >, (với T opN << |R|, ri ∈ RTopN , ri ≠ r) để khuyến nghị cho r.



Hình 2.1: Minh hoạ hệ thống khuyến nghị cộng tác học giả

2.1.2. Vai trò của khuyến nghị cộng tác

Hệ thống khuyến nghị cộng tác học giả đóng vai trò quan trọng trong việc hỗ trợ và thúc đẩy các hoạt động nghiên cứu khoa học như:

- Kết nối các nhà nghiên cứu: hệ thống khuyến nghị cộng tác giúp các nhà nghiên cứu tìm thấy và kết nối với những người có chung mối quan tâm, chuyên môn hoặc lĩnh vực nghiên cứu, tạo điều kiện thuận lợi cho sự hợp tác liên ngành và đa ngành.
- Thúc đẩy hợp tác khoa học: bằng cách gợi ý các đối tác nghiên cứu tiềm năng, hệ thống khuyến nghị cộng tác thúc đẩy sự hợp tác giữa các nhà khoa học, từ đó tạo ra những công trình nghiên cứu chất lượng cao và mang tính đột phá.
- Tiết kiệm thời gian và nguồn lực: việc tìm kiếm đối tác nghiên cứu phù hợp có thể tốn nhiều thời gian và công sức. Hệ thống khuyến nghị cộng tác tự động hóa quá trình này, giúp các nhà nghiên cứu tiết kiệm thời gian và nguồn lực để tập trung vào công việc nghiên cứu chính.
- Nâng cao chất lượng nghiên cứu: sự hợp tác giữa các nhà nghiên cứu có thể nâng cao chất lượng của các công trình nghiên cứu thông qua việc chia sẻ kiến thức, kỹ năng và tài nguyên. Hệ thống khuyến nghị cộng tác giúp tìm ra những đối tác phù hợp, tăng khả năng thành công của các dự án nghiên cứu.
- Hỗ trợ phát triển chuyên môn: đối với các nhà nghiên cứu trẻ hoặc những người mới vào nghề, hệ thống khuyến nghị cộng tác có thể giúp họ tìm thấy những người hướng dẫn hoặc cố vấn phù hợp, hỗ trợ quá trình học tập và phát triển chuyên môn.
- Tăng cường tính ứng dụng của thông tin: bằng cách phân tích hành vi và sở thích của người dùng, hệ thống khuyến nghị cộng tác giúp tăng cường khả năng

tiếp cận thông tin quan trọng và hữu ích, giúp các nhà nghiên cứu nhanh chóng nắm bắt những phát hiện mới và xu hướng nghiên cứu hiện tại.

- Thúc đẩy đổi mới và sáng tạo: sự kết hợp giữa các nhà nghiên cứu từ các lĩnh vực khác nhau có thể dẫn đến những ý tưởng mới mẻ và sáng tạo. Hệ thống khuyến nghị cộng tác tạo điều kiện thuận lợi để các nhà khoa học chia sẻ và phát triển các ý tưởng này.

Nhìn chung, hệ thống khuyến nghị cộng tác học giả đóng góp quan trọng vào việc phát triển khoa học và công nghệ, tối ưu hóa nguồn lực và thời gian, nâng cao chất lượng nghiên cứu, và thúc đẩy sự hợp tác và sáng tạo trong cộng đồng nghiên cứu.

2.2. Các hướng tiếp cận xây dựng hệ thống khuyến nghị cộng tác học giả

Tương tự như các lĩnh vực khuyến nghị học thuật khác, nghiên cứu về phương pháp phát triển khuyến nghị cộng tác có thể được phân loại thành các loại sau: phương pháp lọc cộng tác (Collaborative filtering), tiếp cận nội dung (Content-Based filtering) và phương pháp lai (Hybrid). Trong phần này, khoá luận giới thiệu các phương pháp được sử dụng rộng rãi trong từng loại khuyến nghị. Ngoài ra, khoá luận còn cung cấp cái nhìn tổng quan về các khía cạnh và kỹ thuật quan trọng nhất được sử dụng trong các lĩnh vực này.

Tương tự như các lĩnh vực khuyến nghị học thuật khác, nghiên cứu về phương pháp phát triển khuyến nghị cộng tác có thể được phân loại thành các loại sau: phương pháp lọc cộng tác (Collaborative filtering), tiếp cận nội dung (Content-Based filtering) và phương pháp lai (Hybrid). Trong phần này, khoá luận giới thiệu các phương pháp được sử dụng rộng rãi trong từng loại khuyến nghị. Ngoài ra, khoá luận còn cung cấp

cái nhìn tổng quan về các khía cạnh và kỹ thuật quan trọng nhất được sử dụng trong các lĩnh vực này.

2.2.1. Tiếp cận nội dung (Content-based filtering).

CBF tập trung vào sự tương đồng ngữ nghĩa giữa các đặc điểm cá nhân của nhà nghiên cứu, chẳng hạn như hồ sơ cá nhân, lĩnh vực chuyên môn và sở thích nghiên cứu. Các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) được sử dụng để trích xuất từ khóa từ các tài liệu liên quan, nhằm xác định lĩnh vực chuyên môn và mối quan tâm của các nhà nghiên cứu.

Mô hình không gian vectơ (VSM) được sử dụng rộng rãi trong các phương pháp đề xuất dựa trên nội dung. Bằng cách biểu diễn các truy vấn và tài liệu dưới dạng vectơ trong không gian đa chiều, các vectơ này có thể tính toán mức độ liên quan hoặc tương tự. Yukawa và cộng sự [10] đã đề xuất một hệ thống khuyến nghị chuyên gia sử dụng mô hình không gian vectơ mở rộng để tính toán vectơ tài liệu cho mỗi tài liệu mục tiêu của tác giả hoặc tổ chức, cung cấp danh sách theo thứ tự mức độ liên quan giữa các chủ đề học thuật và các nhà nghiên cứu.

Các mô hình phân cụm chủ đề sử dụng VSM đã được áp dụng rộng rãi để lập hồ sơ các lĩnh vực nghiên cứu của các nhà khoa học bằng danh sách từ khóa có lược đồ trọng số. Bằng cách sử dụng mô hình tính trọng số từ khóa, Afzal và Maurer [10] đã triển khai một phương pháp tự động để đo lường hồ sơ chuyên môn trong giới học thuật, kết hợp nhiều số liệu để đo lường mức độ chuyên môn tổng thể. Gollapalli và cộng sự [11]đã đề xuất một hệ thống khuyến nghị dựa trên nội dung học thuật bằng cách tính toán sự tương đồng giữa các nhà nghiên cứu dựa trên hồ sơ cá nhân trích xuất từ các ấn phẩm và trang chủ học thuật của họ.

Các mô hình dựa trên chủ đề cũng đã được áp dụng rộng rãi trong xử lý tài liệu. Mô hình dựa trên chủ đề giới thiệu một lớp chủ đề giữa nhà nghiên cứu và tài liệu được trích xuất. Ví dụ, theo phương pháp phân bổ Dirichlet tiềm ẩn (LDA), mỗi tài liệu được coi là sự kết hợp của các chủ đề và mỗi từ trong tài liệu được coi là được chọn ngẫu nhiên từ các chủ đề của tài liệu đó. Yang và cộng sự [12] đã đề xuất một phương pháp đề xuất cộng tác viên bổ sung để truy xuất các chuyên gia cho cộng tác nghiên cứu, sử dụng thuật toán tham lam heuristic nâng cao với phân kỳ Kullback–Leibler đối xứng dựa trên mô hình chủ đề xác suất. Kong và cộng sự [13] đã áp dụng hệ thống đề xuất cộng tác viên bằng cách tạo ra danh sách đề xuất dựa trên các vectơ học giả được học từ mối quan tâm nghiên cứu của các nhà nghiên cứu trích xuất từ các tài liệu dựa trên mô hình chủ đề.

Tuy nhiên, các phương pháp dựa trên nội dung thường có chi phí tính toán cao do số lượng lớn tài liệu được phân tích và không gian vectơ. Để giảm thiểu chi phí này và tối đa hóa sự ưa thích, Kong và cộng sự [14] đã trình bày một phương pháp khuyến nghị cộng tác viên học thuật dựa trên lý thuyết đối sánh, áp dụng nhiều chỉ số trích xuất từ các tài liệu liên quan để tích hợp ma trận ưu tiên giữa các nhà nghiên cứu. Một số nhà nghiên cứu cũng đã sửa đổi các đặc trưng có trọng số và phương pháp trích xuất chủ đề kết hợp với các yếu tố khác để đạt được độ chính xác cao hơn. Ví dụ, Sun và cộng sự [15] đã thiết kế một mô hình đề xuất cộng tác viên học thuật nhận thức được độ tuổi nghề nghiệp, bao gồm trích xuất quyền tác giả từ các thư viện kỹ thuật số, trích xuất chủ đề dựa trên các tóm tắt đã xuất bản và bước đi ngẫu nhiên nhận thức được độ tuổi nghề nghiệp để đo lường mức độ tương đồng của học giả.

2.2.2. Phương pháp lọc cộng tác (Collaborative filtering).

Các đề xuất dựa trên lọc cộng tác truyền thống nhằm mục đích tìm người hàng xóm gần nhất trong bối cảnh xã hội tương tự như bối cảnh của người dùng mục tiêu. Nó chọn những người hàng xóm gần nhất dựa trên sự tương đồng về đánh giá của người dùng. Khi người dùng xếp hạng một nhóm mặt hàng theo cách tương tự như cách của người dùng mục tiêu, hệ thống đề xuất sẽ xác định những người hàng xóm gần nhất này là các nhóm có cùng sở thích và đề xuất các mặt hàng được các nhóm này ưa thích nhưng người dùng mục tiêu không phát hiện ra. Để áp dụng phương pháp này vào đề xuất cộng tác viên, hệ thống sẽ đề xuất những người đã từng làm việc với đồng nghiệp của tác giả mục tiêu chứ không phải với chính tác giả mục tiêu. Tương tự, hệ thống coi mỗi tác giả là một mục cần được xếp hạng và các hoạt động học thuật như viết một bài báo cùng nhau là một hoạt động xếp hạng, tuân theo phương pháp khuyến nghị dựa trên lọc cộng tác truyền thống. Hoạt động xuất bản của các nhà nghiên cứu được chuyển thành hành động xếp hạng và tần suất các bài báo có đồng tác giả được coi là giá trị xếp hạng.

Dựa trên mạng đồng tác giả này được chuyển đổi từ các hoạt động xuất bản của các nhà nghiên cứu, một số phương pháp dự đoán liên kết và trọng số cạnh đã được sử dụng. Benchettara và cộng sự [16] đã giải quyết vấn đề dự đoán liên kết trong các mạng đồng tác giả bằng cách sử dụng phương pháp học máy được giám sát theo cấu trúc liên kết đôi. Koh và Dobbie [17] đã đề xuất một phương pháp đề xuất cộng tác viên học thuật sử dụng mạng đồng tác giả với cách tiếp cận quy tắc kết hợp có trọng số sử dụng cơ chế trọng số được gọi là tính xã hội. Các phương pháp đề xuất dựa trên mạng đồng tác giả này được chuyển đổi từ các hoạt động xuất bản, trong đó tất cả các nút có cùng chức năng, được gọi là các phương pháp đề xuất dựa trên mạng đồng nhất.