ĐẠI HỌC QUỐC GIA TP.HCM

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

HUỲNH NGỌC TÍN

PHÁT TRIỂN MỘT SỐ PHƯƠNG PHÁP KHUYẾN NGHỊ HỖ TRỢ TÌM KIẾM THÔNG TIN HỌC THUẬT DỰA TRÊN TIẾP CẬN PHÂN TÍCH MẠNG XÃ HỘI

Chuyên ngành: Khoa học Máy tính
Mã số: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

TP. HỒ CHÍ MINH – Năm 2016

Công trình được hoàn thành tại: **Trường Đại học Công nghệ Thông tin – Đại học Quốc gia TpHCM.**

Người hướng dẫn khoa học: GS.TSKH Hoàng Văn Kiếm

Phản biện 1: PGS.TS. Đỗ Phúc
Phản biện 2: PGS.TS. Lê Hoài Bắc
Phản biện 3: PGS.TS. Quản Thành Thơ
Phản biện độc lập 1: PGS.TS. Nguyễn Đình Thúc
Phản biện độc lập 2: PGS.TS. Đỗ Năng Toàn

Luận án đã được bảo vệ trước
Hội đồng chấm luận án cấp Trường tại:
Phòng E 1.1, Trường Đại học Công nghệ Thông tin – ĐHQG TpHCM
Vào lúc 8 giờ 30 ngày 26 tháng 02 năm 2016

Có thể tìm luận án tại:
- Thư viện Quốc gia Việt Nam
- Thư viện Trường Đại học Công nghệ Thông tin – ĐHQG TpHCM.

I. MỞ ĐẦU

I.1 Dẫn nhập

Việc tìm kiếm thông tin khoa học để thực hiện các công việc liên quan đến nghiên cứu là nhu cầu thường xuyên, không thể thiếu đối với những người làm nghiên cứu khoa học, đặc biệt là các nghiên cứu viên (NCV). Các NCV trẻ thì thiếu kinh nghiệm tìm kiếm và xác định các thông tin hữu ích liên quan. Trong khi, các NCV có kinh nghiệm thì phải đương đầu với quá tải thông tin. Để giúp họ dễ dàng hơn trong việc tiếp cận các thông tin học thuật hữu ích liên quan, hệ khuyến nghị trong lĩnh vực học thuật là giải pháp đang được quan tâm nghiên cứu trong những năm gần đây.

Các bài toán khuyến nghị thông tin học thuật phổ biến như: khuyến nghị bài báo, cộng tác, gởi bài, v.v... cũng như các cách tiếp cận truyền thống cho hệ khuyến nghị là lọc dựa trên thông tin lý lịch (Demographic Filtering), lọc dựa trên nội dung CB (Content-Based), lọc cộng tác CF (Collaborative Filtering), lai (Hybrid) phải đương đầu với một số khó khăn, thách thức như: dữ liệu lớn, chưa có dữ liệu chuẩn (benchmark) cho đánh giá thực nghiệm, độ chính xác chưa cao, vấn đề khởi động lạnh (cold-start), chưa có phương pháp phù hợp để đánh giá chất lượng khuyến nghị.

Xu hướng tiếp cận để phát triển các phương pháp mới cho hệ khuyến nghị đó là: phân tích mạng xã hội, khai thác thông tin ngữ cảnh và các phương pháp lai [23]. Trên thực tế, sở thích và quyết định của con người thường chịu ảnh hưởng bởi những người có quan hệ. Các NCV thường cần lời khuyên từ bạn bè, đồng nghiệp, thầy cô để đưa ra những quyết định quan trọng liên quan đến các công việc nghiên cứu. Do đó, luận án chọn tiếp cận phân tích mạng xã hội (có xem xét yếu tố thời gian) kết hợp một số thông tin khác, nhằm giải quyết những hạn chế của một số phương pháp phổ biến, ứng dụng khuyến nghị thông tin học thuật.

I.2 Mục tiêu, nội dung của luận án

– **Mục tiêu chính:** nâng cao kết quả khuyến nghị thông tin học thuật dựa trên tiếp cận phân tích mạng xã hội.

– **Nội dung thực hiện:**

(1) **Xây dựng và làm giàu kho dữ liệu học thuật**.

(2) **Xây dựng mô hình mạng xã hội học thuật**.

(3) **Khai thác mạng xã hội học thuật** → Phát triển một số phương pháp khuyến nghị ứng dụng vào bài toán:

- Khuyến nghị cộng tác.
- Khuyến nghị bài báo khoa học liên quan.

I.3 Các đóng góp chính của luận án

(1) Đề xuất mô hình mạng xã hội học thuật ASN (Academic Social Network) nhận diện từ kho dữ liệu bài báo khoa học. [CT.6]

(2) Bài toán khuyến nghị cộng tác cho NCV

- Đối với NCV có quan hệ đồng tác giả: đề xuất các phương pháp phân tích xu hướng cộng tác trong mạng xã hội học thuật ASN để khuyến nghị các cộng tác viên tiềm năng. Các phương pháp đề xuất bao gồm: MPRS, MPRS+, RSS+ [CT.1, CT.4].
- Đối với NCV chưa có quan hệ đồng tác giả: đề xuất tập đặc trưng để khuyến nghị những mối quan hệ cộng tác tốt, chất lượng [CT.3].
- Đề xuất phương pháp đánh giá chất lượng cộng tác được khuyến nghị [CT.3].

(3) Bài toán khuyến nghị bài báo khoa học: phát triển phương pháp khuyến nghị bài báo khoa học cho NCV dựa trên việc khai thác mạng trích dẫn, quan hệ lòng tin trong mô hình ASN [CT.2, CT.8, CT.11].

(4) Xây dựng kho dữ liệu học thuật hơn 6 triệu bài báo và hệ thống tìm kiếm thông tin khoa học CSPubGuru (www.cspubguru.com) [CT.5, CT.7, CT.9, CT.10, CT.14].

Luận án đã tiến hành triển khai nhiều thử nghiệm trên các tập dữ liệu có kích thước lớn. Kết quả đạt được đã chứng minh được (bằng thực nghiệm) tiếp cận và hiệu quả của các phương pháp cải tiến, đề xuất so với các phương pháp phổ biến hiện nay liên quan đến các bài toán khuyến nghị thông tin học thuật.

I.4 Bố cục của luận án

Luận án bao gồm 153 trang (không tính phần phụ lục), 12 bảng, 29 hình vẽ (không tính bảng và hình vẽ trong phần phụ lục), phần mở đầu và các chương mục: *Phần mở đầu*; *Chương 1*: Hệ khuyến nghị: những phương pháp tiếp cận phổ biến và xu hướng; *Chương 2*: Xác định và mô hình hóa mạng xã hội học thuật; *Chương 3*: Khai thác mạng xã hội học thuật để phát triển các phương pháp khuyến nghị cộng tác; *Chương 4*: Khai thác mạng xã hội học thuật để phát triển các phương pháp khuyến nghị bài báo khoa học; *Kết luận và Hướng phát triển*. Phần tài liệu tham khảo gồm 130 tài liệu (bài báo hội thảo và tạp chí quốc tế). Ngoài ra, Luận án còn có 2 Phụ lục A, B bổ sung các thông tin chi tiết cho phương pháp xây dựng, cấu trúc và nguồn dữ liệu bài báo khoa học đã thu thập.

II. NỘI DUNG LUẬN ÁN

Chương 1 - Hệ khuyến nghị: những phương pháp tiếp cận phổ biến và xu hướng

1.1 Giới thiệu: chương này sẽ tập trung phân tích ưu điểm, hạn chế của các phương pháp khuyến nghị truyền thống. Từ đó dẫn đến tiếp cận của luận án dựa trên phân tích mạng xã hội học thuật để giải quyết các bài toán khuyến nghị trong lĩnh vực học thuật.

1.2 Khái niệm Hệ khuyến nghị

- Hệ khuyến nghị, tiếng anh là Recommender Systems hoặc Recommendation System, là những hệ thống được thiết kế để hướng người dùng đến những đối tượng quan tâm, yêu thích, khi lượng thông tin quá lớn vượt quá khả năng xử lý của người dùng [25, 99].
- Theo Ricci và cộng sự [100], hệ khuyến nghị là những công cụ phần mềm, kỹ thuật cung cấp những đề xuất các đối tượng có thể hữu ích với người dùng. Những đề xuất liên quan đến quyết định của người dùng như: sản phẩm nào nên mua, bài hát nào nên nghe, hay tin tức nào nên đọc.

1.3 Phát biểu bài toán khuyến nghị

Định nghĩa 1.1: *Không gian người dùng* [57]

Không gian người dùng là tập tất cả những người dùng mà hệ thống quan sát được, để thực hiện các phân tích, khuyến nghị. Ký hiệu là U, $U = \{u_1, u_2, u_3, ..., u_n\}$.

Định nghĩa 1.2: *Không gian đối tượng khuyến nghị* [57]

Không gian đối tượng khuyến nghị là tập tất cả những đối tượng sẽ được khuyến nghị cho người dùng. Tùy vào ứng dụng cụ thể, các đối tượng khuyến nghị có thể là sách, báo, phim ảnh, địa điểm, nhà hàng, khách sạn, con người, v.v... Ký hiệu là P, $P = \{p_1, p_2, p_3, ..., p_m\}$.

Định nghĩa 1.3: *Hàm hữu ích* [5]

Hàm hữu ích f là ánh xạ f: $U \times P \rightarrow R$, dùng để ước lượng mức độ hữu ích của $p \in P$ với $u \in U$. Với R là tập có thứ tự các số nguyên hoặc thực trong một khoảng nhất định.

Phát biểu bài toán khuyến nghị

Cho trước,

- $U = \{u_1, u_2, u_3, ..., u_n\}$: không gian người dùng.
- $P = \{p_1, p_2, p_3, ..., p_m\}$: không gian đối tượng khuyến nghị.

Mục đích của hệ khuyến nghị là đi tìm hàm hữu ích f, ước lượng giá trị của $f(u,p)$ (với $u \in U$, $p \in P$). Giá trị của $f(u,p)$ giúp tiên đoán u sẽ thích p nhiều hay ít, hay p hữu ích đối với u như thế nào. Đối với mỗi người dùng $u \in U$, hệ khuyến nghị cần chọn $TopN$ đối tượng $p \in P$ hữu ích nhất đối với người dùng u để khuyến nghị, $P_{TopN} = <p_1, p_2, ..., p_{TopN}>$, (với $TopN << m$). Việc chọn $TopN$ bao nhiêu là tùy thuộc vào nhu cầu thông tin của người dùng, cũng như mục đích cung cấp thông tin của hệ khuyến nghị. Các đối tượng $p \in P_{TopN}$, được chọn thỏa mãn các điều kiện ràng buộc sau:

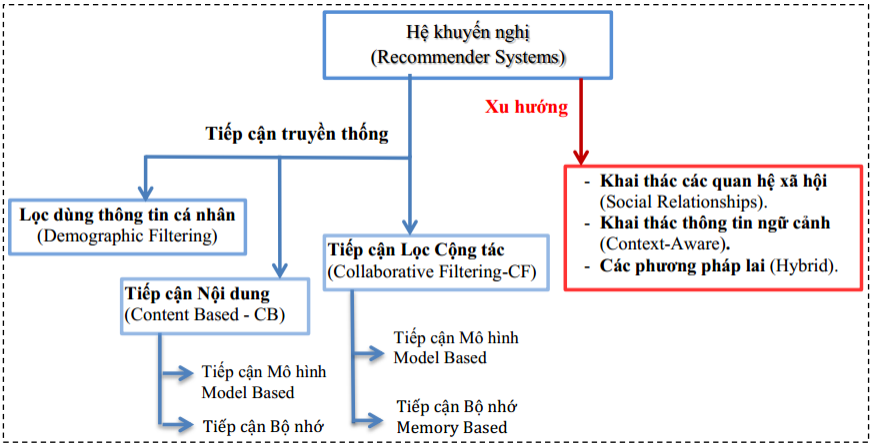
i) $\forall p_k \in P_{TopN}, f(u, p_k) \geq f(u, p_{k+1}), với\ 1 \leq k \leq TopN - 1$. Tức là tập các đối tượng khuyến nghị P_{TopN} là tập có thứ tự. Đối tượng đứng trước có giá trị của hàm hữu ích f lớn hơn hoặc bằng đối

tượng đứng sau, hay đối tượng đứng trước ưu tiên khuyến nghị cho u hơn đối tượng đứng sau.

ii) $\forall p_k \in P_{TopN}, \forall p_i \in P \backslash P_{TopN}, thì\ f(u,p_k) \geq f(u,p_i)$. Tức giá trị hữu ích của các đối tượng được khuyến nghị, được xác định thông qua hàm f, phải lớn hơn hoặc bằng những đối tượng không được khuyến nghị.

Việc xây dựng hàm hữu ích f và ước lượng giá trị hữu ích của các đối tượng khuyến nghị $p \in P$ với những người dùng $u \in U$ có thể thực hiện bằng nhiều phương pháp khác nhau như: dựa vào kinh nghiệm (heuristics), máy học, lý thuyết xấp xĩ, v.v...

1.4 Các cách tiếp cận truyền thống



Hình 1.2: Các cách tiếp cận phổ biến và xu hướng hiện nay cho hệ khuyến nghị.

1.4.1 Tiếp cận nội dung (CB)

Để thực hiện việc ước lượng có hay không người dùng u sẽ thích đối tượng khuyến nghị p, hoặc thích nhiều hay ít. Tức là, xây dựng một hàm hữu ích $f(u,p)$ của các đối tượng khuyến nghị p với người dùng u và ước lượng giá trị hữu ích này. Các phương pháp dựa trên tiếp cận nội dung thông thường sẽ thực hiện các bước sau:

- **Bước 1**: Biểu diễn nội dung đối tượng khuyến nghị $p \in P$, *Content(p)*.
- **Bước 2**: Mô hình hóa sở thích người dùng $u \in U$, gọi tắt là hồ sơ người dùng (User's Profile), ký hiệu *UserProfile(u)*.
- **Bước 3**: Ước lượng giá trị hữu ích dựa trên độ tương tự nội dung của đối tượng khuyến nghị *p* với hồ sơ người dùng *u*. Hệ thống sẽ ưu tiên khuyến nghị những đối tượng *p* có nội dung tương tự cao so với hồ sơ người dùng *u*.

Các phương pháp truyền thống dựa trên nội dung có thể chia thành hai nhóm chính: (1) Một là các phương pháp dựa trên bộ nhớ, thực hiện tính toán độ tương tự giữa *Content(p)* và *UserProfile(u)* dùng các độ đo tương tự Cosine, Euclide; (2) Hai là các phương pháp dựa trên mô hình, với mô hình được học từ dữ liệu dùng các kỹ thuật học máy giám sát để phân các đối tượng khuyến nghị thành những đối tượng người dùng quan tâm (1) hay không quan tâm (0).

Hạn chế của tiếp cận CB:

- Các khó khăn liên quan đến phân tích nội dung.
- Không thể đa dạng trong khuyến nghị (các đối tượng khuyến nghị ngoài lĩnh vực quan sát).
- Người dùng mới (khởi động lạnh).

1.4.2 Tiếp cận lọc cộng tác (CF)

	p_1	p_2	p_3	p_4	p_5	…	p_m
u_1	1	?	5	?	4	?	?
u_2	?	?	4	?	5	?	?
u_3	?	4	?	5	?	?	?
u_4	?	?	?	4	?	?	?
u_5	?	?	?	5	?	?	?
…	?	?	?	?	?	?	?
u_n	?	3	?	?	?	?	5

Hình 1.4: Dấu **?** là những giá trị cần tiên đoán trong ma trận đánh giá.

Tiếp cận CF được xem là tiếp cận thành công nhất để xây dựng các hệ thống khuyến nghị và ứng dụng rộng rãi trong lĩnh vực thương mại điện tử

[110, 57]. Ý tưởng chung của tiếp cận CF là khai thác thông tin, hành vi quá khứ của người dùng dựa trên các đánh giá sẵn có từ ma trận đánh giá (hình 1.4) để tiên đoán, lượng hóa mức độ hữu ích của các đối tượng khuyến nghị mà người dùng chưa biết. Một số các nghiên cứu phổ biến đã thực hiện khảo sát, phân loại, cũng như thực nghiệm, đánh giá các thuật toán CF. Các phương pháp CF nói chung được phân thành hai nhóm chính: (1) CF dựa trên bộ nhớ như các thuật toán tính toán tương tự, lân cận; (2) CF dựa trên mô hình như các thuật toán gom cụm, phân lớp giám sát, thừa số hóa ma trận (Matrix Factorization).

Hạn chế của tiếp cận CF:

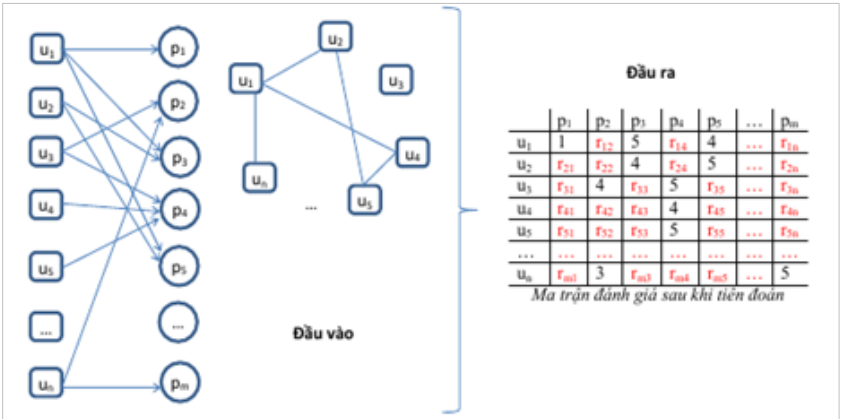
- Ma trận đánh giá thưa.
- Người dùng, đối tượng khuyến nghị mới (khởi động lạnh).

1.4.3 Tiếp cận lai

Những phương pháp khác nhau đều có những điểm mạnh, cũng như điểm yếu của nó (bảng 1.2). Để tận dụng những điểm mạnh và hạn chế điểm yếu của những tiếp cận khác nhau, nhiều nghiên cứu đã tập trung phát triển các hệ khuyến nghị dựa trên việc kết hợp các tiếp cận khác nhau, được gọi là tiếp cận lai (Hybrid Approach) hay hệ khuyến nghị lai (Hybrid Recommender System). Robin Burke đã khảo sát các phương pháp lai cho hệ khuyến nghị và trình bày tóm tắt 7 nhóm phương pháp tiếp cận lai phổ biến: Lai có trọng số (Weighted Hybrid); Lai chuyển đổi (Switching Hybrid); Lai trộn (Mixed Hybrid); Lai kết hợp đặc trưng (Feature Combination Hybrid); Lai theo đợt (Cascade Hybrid); Lai tăng cường đặc trưng (Feature Augmentation Hybrid); Lai meta (Meta-Level Hybrid) [25].

1.4.4 Tiếp cận phân tích mạng xã hội

Bên cạnh việc khai thác thông tin sở thích của người dùng dựa trên dữ liệu quá khứ như tiếp cận CB, CF thì tiếp cận phân tích mạng xã hội thực hiện khuyến nghị dựa trên việc xem xét ảnh hưởng, chi phối hành vi sở thích của người dùng thông qua các mối quan hệ xã hội (Hình 1.7)



Hình 1.8: Minh họa khuyến nghị xã hội

1.4.5 Xu hướng mới cho hệ khuyến nghị

- Kết hợp sử dụng thông tin ngữ cảnh để nâng cao hiệu quả khuyến nghị [3, 6]. Xem xét sự ảnh hưởng của thời gian, xu hướng đến kết quả khuyến nghị như thế nào [22, 109].
- Tìm cách kết hợp thông tin xã hội rõ ràng, tìm ẩn vào các phương pháp truyền thống [22].
- Tiếp cận lai nhằm giải quyết những hạn chế của mỗi phương pháp khác nhau [5, 22, 25].
- Lưu vết, thu thập thông tin tiềm ẩn về hành vi của người dùng từ Internet để xác định sở thích của họ.

Ưu điểm, hạn chế của các cách tiếp cận truyền thống và xu hướng cho hệ khuyến nghị có thể tóm tắt trong bảng 1.2.

Bảng 1.2: Ưu, nhược điểm các cách tiếp cận phổ biến và xu hướng nghiên cứu.

Ưu điểm & Hạn chế	Tiếp cận truyền thống và xu hướng				
	Truyền thống			Xu hướng	
	Nội dung (CB)	Lọc Cộng tác (CF)	CB kết hợp CF	Phân tích mạng xã hội	Khai thác thông tin ngữ cảnh
Phù hợp văn bản	Có	Có	Có	Có	Có

Đa dạng đối tượng khuyến nghị	Không	Có	Có	Có	Có
Hạn chế về phân tích nội dung	Có	Không	Không	Không	Không
Có thể đa dạng hóa khuyến nghị.	Không	Có	Có	Có	Có
Người dùng mới (khởi động lạnh)	Có	Có	Có	Có	Có
Đối tượng mới (khởi động lạnh)	Không	Có	Có	Có	Có
Vấn đề ma trận thưa	Không	Có	Có	Có	Có
Có thể giải quyết ma trận thưa, khởi động lạnh	Không	Không	Có	Có	Có

Khó khăn chung:

- Dữ liệu lớn.
- Độ chính xác, chất lượng khuyến nghị chưa cao.
- Dữ liệu đánh giá thưa.
- Chưa có phương pháp tốt để đánh giá kết quả, chất lượng khuyến nghị.
- Vấn đề khởi động lạnh.

Trong lĩnh vực học thuật, các NCV thường dựa trên ý kiến đề xuất của giáo sư, đồng nghiệp, những người có kinh nghiệm để đưa ra những quyết định liên quan đến công việc nghiên cứu khoa học như: chọn hội thảo gởi bài, chọn người hợp tác, chọn bài báo để đọc, v.v… Để thực hiện được việc khai thác các mối quan hệ xã hội trong học thuật, chương tiếp theo sẽ trình bày việc rút trích, mô hình hóa các mạng xã hội học thuật từ kho dữ liệu bài báo khoa học.

Chương 2 - Xác định và mô hình hoá mạng xã hội học thuật

2.1 Giới thiệu

Với mục tiêu phát triển các phương pháp khuyến nghị trong lĩnh vực học thuật dựa trên tiếp cận phân tích mạng xã hội, luận án cần xem xét: (1)

Chuẩn bị kho dữ liệu học thuật đủ lớn và đủ phong phú; (2) Xác định và mô hình các mối quan hệ xã hội học thuật; (3) Khai thác các mối quan hệ học thuật để phát triển các phương pháp khuyến nghị.

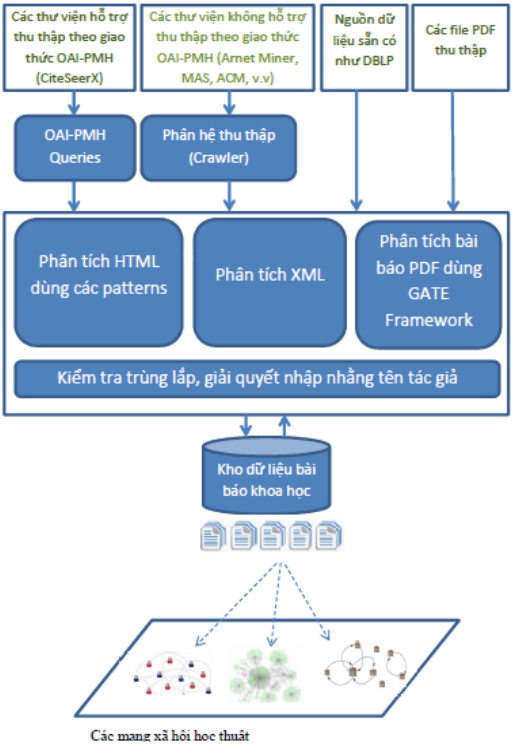
Về các kho dữ liệu học thuật thì các nghiên cứu phổ biến hiện nay thực hiện trên nhiều tập dữ liệu khác nhau được rút trích từ nhiều nguồn khác nhau. Chẳng hạn, Chen và cộng sự [27, 28, 29], S. D. Gollapalli và cộng sự [48], thì tiến hành thử nghiệm trên dữ liệu được trích xuất từ CiteSeerX[1]. Trong khi đó, Tang và cộng sự [117], Sugiyama và cộng sự [111, 112, 113], Luong và cộng sự [75, 76], tiến hành thực nghiệm trên tập dữ liệu bài báo khoa học được trích xuất từ các hội thảo chuyên ngành và gán nhãn thủ công. Một số nghiên cứu phổ biến khác thì trích xuất từ kho dữ liệu khoa học DBLP[2] để xây dựng tập dữ liệu thực nghiệm. Nói chung, theo hiểu biết của chúng tôi thì hiện nay chưa có những tập dữ liệu chuẩn (benchmark) đối với các bài toán khuyến nghị trong lĩnh vực học thuật. Bên cạnh đó, cho đến nay thì những thông tin có được từ các tập dữ liệu phổ biến cho download như DBLP, CiteSeerX vẫn còn khá hạn chế, thiếu nhiều thông tin cần thiết (bảng 2.1). Vì vậy, việc xây dựng và làm giàu một kho dữ liệu khoa học đủ lớn và đủ phong phú và công bố rộng rãi cho cộng đồng tham khảo để tiến hành các đánh giá thực nghiệm là cần thiết.

Chương này sẽ tập trung trình bày 2 phần chính: (1) Giải pháp, kết quả của việc xây dựng và làm giàu kho dữ liệu học thuật; (2) Mô hình các mạng xã hội học thuật ASN, cũng như các phương pháp lượng hóa trên các mạng xã hội học thuật ASN. Kết quả liên quan đã được công bố trong các công trình: [CT.5, CT.6, CT.7, CT.9, CT.10, CT.14].

2.2 Xây dựng và làm giàu kho dữ liệu học thuật

Quá trình xây dựng và làm giàu kho dữ liệu học thuật có thể minh họa tóm tắt thông qua hình vẽ 2.1.

[1] http://csxstatic.ist.psu.edu/about/data
[2] http://dblp.uni-trier.de/xml/



Hình 2.1: Tích hợp dữ liệu bài báo khoa học từ nhiều nguồn không đồng nhất

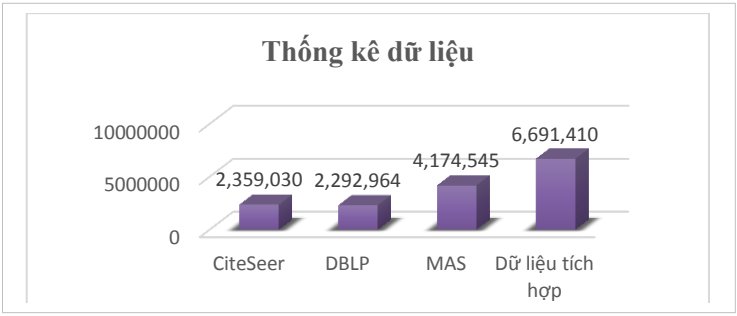
Kết quả kho dữ liệu đã xây dựng (CSPubGuru)

Tính đến tháng 03/2013, luận án đã thu thập được hơn 6 triệu bài báo chuyên ngành khoa học máy tính và thông tin liên quan. Tập dữ liệu đã thu thập, tích hợp đặt tên là CSPubGuru. Kích thước và thông tin lưu trữ của CSPubGuru được trình bày trong bảng 2.4 và hình 2.4. Hiện nay, CSPubGuru và các tập dữ liệu thực nghiệm liên quan được công bố tại: https://sites.google.com/site/tinhuynhuit/dataset.

Bảng 2.4: Thông tin bài báo từ DBLP, CiteSeerX, CSPubGuru

Thông Tin bài báo	DBLP	CiteSeer	CSPubGuru
Tiêu đề	✓	✓	✓
Tác giả	✓	✓	✓
Cơ quan			✓
Tóm tắt		✓	✓
Nơi công bố	✓	✓	✓

Năm	✓	✓	✓
Từ khóa		✓	✓



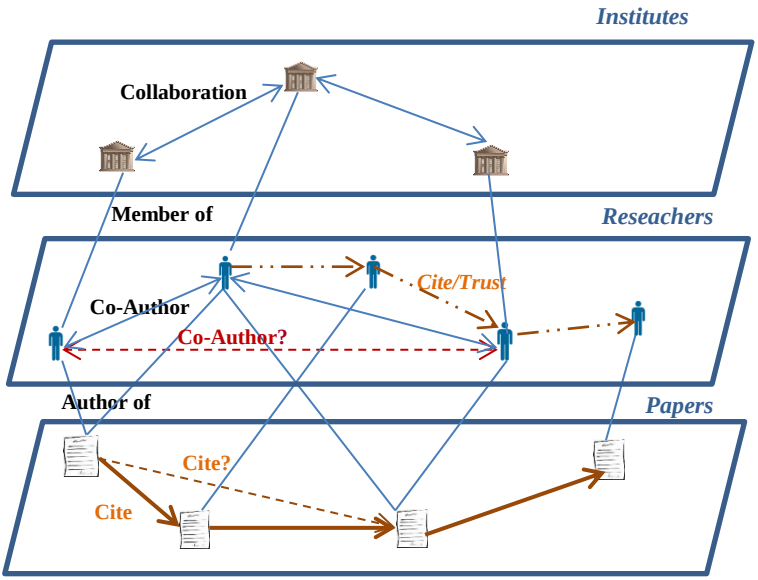
Hình 2.4: Kích thước kho dữ liệu tích hợp tính đến 03/2013.

2.3 Xác định và mô hình mạng xã hội học thuật (ASN)

Từ kho dữ liệu học thuật thu thập được, chúng ta có thể nhận diện ra một số đối tượng nghiên cứu như: nghiên cứu viên, bài báo khoa học, các trường, các viện hay cơ quan công tác của các tác giả. Hình 2.5 minh họa các mạng xã hội có thể quan sát được từ kho dữ liệu học thuật.

ASN = (CoNet, CiNet_Author, CiNet_Paper, AffNet, M)

- **CoNet<R, E_1>**: Mạng cộng tác đồng tác giả.
- **CiNet_Author <R,E_2>**: Mạng trích dẫn của các tác giả.
- **CiNet_Paper <P, E_3>**: Mạng trích dẫn của các bài báo khoa học.
- **AffNet <Aff, E_4>**: Mạng cộng tác giữa các viện, trường.
- **M**: Các phương pháp tính toán trên ASN. Các phương pháp tính toán mới được đề xuất trong thành phần M:
  - **Mô hình hồ sơ của NCV**
    - Sở thích dựa trên xu hướng [CT.02]
    - Uy tín của nghiên cứu viên [CT.03]
    - Mức độ năng động của nghiên cứu viên [CT.03]
  - **Mô hình các mối quan hệ dựa trên xu hướng**
    - Xu hướng cộng tác giữa các nghiên cứu viên: RSS+(r_i,r_j), MPRS+(r_i,r_j) [CT.01, CT.04]
    - Quan hệ giữa các cơ quan (Org_RSS(o_i, o_j)) [CT.03]
    - Quan hệ lòng tin (đồng tác giả và trích dẫn) [CT.02]



Hình 2.5: Các cấu trúc xã hội từ kho dữ liệu bài báo khoa học.

Chương 3 - Khai thác mạng xã hội học thuật để phát triển các phương pháp khuyến nghị cộng tác

3.1 Giới thiệu

Cộng tác là hành động hay quá trình hai hay nhiều cá nhân, tổ chức làm việc cùng nhau để thực hiện một mục đích chung[3]. Trong nghiên cứu khoa học, có thể quan niệm cộng tác nghiên cứu là quá trình làm việc cùng nhau của những NCV để đạt được một mục đích chung trong việc tìm ra các tri thức khoa học mới [61]. Cộng tác nghiên cứu giúp các NCV có cơ hội để trao đổi kiến thức, kinh nghiệm. Những NCV càng có nhiều quan hệ công tác tốt thì càng có khả năng tạo ra nhiều tri thức mới trong khoa học [61, 74].

Có thể nói đối tác hay người cộng tác là một trong những yếu tố then chốt quyết định chất lượng, kết quả đạt được của quá trình cộng tác. Câu hỏi đặt ra là làm thế nào có thể tìm được những người cộng tác phù hợp? Mục đích của chương này là trình bày, phát biểu bài toán khuyến nghị cộng

[3] http://oxforddictionaries.com/definition/english/collaboration

tác trong nghiên cứu khoa học và phát triển các phương pháp mới dựa trên tiếp cận khai thác các mối quan hệ xã hội học thuật từ mô hình ASN (đã đề cập trong chương trước) để giải quyết bài toán này cho từng nhóm NCV khác nhau.

3.2 Bài toán khuyến nghị cộng tác

Định nghĩa 3.1: *NCV có đồng tác giả (un-isolated researcher)*

NCV có đồng tác giả là các NCV mà tồn tại ít nhất một bài báo đã công bố trong quá khứ có đồng tác giả với một NCV khác.

Định nghĩa 3.2: *NCV chưa có đồng tác giả (isolated researcher)*

NCV chưa có đồng tác giả là các NCV mà trong quá khứ, tính tới thời điểm hiện tại chưa có bài báo công bố nào có đồng tác giả với một NCV khác.

Trong phạm vi luận án này, chúng tôi xem xét giải quyết bài toán khuyến nghị cộng tác với đầu vào là một NCV, 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:

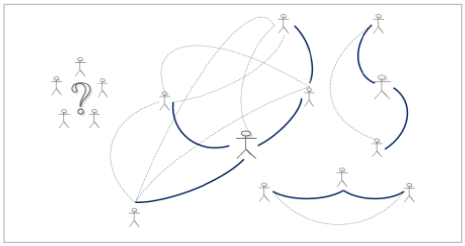
- **Đầu vào**:
  - $R=\{r\}$: tập tất cả các nghiên cứu viên.
  - $P=\{p\}$: tập tất cả các bài báo trong kho dữ liệu.
  - $O=\{o\}$: danh sách các cơ quan nơi các NCV đang làm việc.
- **Đầu ra**:
  - Xác định hàm $f(r_i,r_j)$ để ước lượng tiềm năng quan hệ cộng tác của $r_i \in R$ với $r_j \in R$, $r_i \neq r_j$.
  - $\forall r \in R$, dựa trên hàm f chọn $TopN$ các NCV tiềm năng nhất, $R_{TopN} \subset R$, $R_{TopN} = <r_1, r_2, ..., r_{TopN}>$, (với TopN << $|R|$, $r_i \in R_{Top\text{-}N}$, $r_i \neq r$) để khuyến nghị cho r.

3.3 Trường hợp các NCV có đồng tác giả

3.3.1 Tiếp cận phổ biến

Hầu hết các nghiên cứu phổ biến nhất hiện nay tập trung phân tích, khai thác các mối quan hệ học thuật và sử dụng các độ đo tương tự đỉnh

cục bộ và toàn cục như: Cosine, Jaccard, AdamicAdar, RSS để thực hiện khuyến nghị cộng tác (Chen và cộng sự [27, 28, 29], Lopes và cộng sự [72], Brandao và cộng sự [23]) (hình 3.1).



Hình 3.1: Những phương pháp dựa trên phân tích mạng đồng tác giả có thể khuyến nghị cho các NCV có đồng tác giả (nét đức trong hình), nhưng không thực hiện được đối với các NCV chưa có đồng tác giả (quanh dấu chấm hỏi)

3.3.2 Các phương pháp đề xuất

Đóng góp của luận án: Đề xuất phương pháp khuyến nghị dựa trên phân tích xu hướng quan hệ giữa các nghiên cứu viên: phương pháp RSS+, MPRS+ thuộc thành phần M trong mô hình ASN [CT.1, CT.4].

Tóm tắt phương pháp RSS+ và MPRS+

Đầu vào: $R = \{r\}$: tập tất cả các NCV có đồng tác giả (un-isolated)

$CoNet = (R, E_1)$: mạng đồng tác giả giữa các NCV trong R

Đầu ra:

- Xác định hàm $f(r_i,r_j)$ để ước lượng mức độ tiềm năng cho quan hệ cộng tác của $r_j \in R$ với $r_i \in R$, $r_i \neq r_j$.
- $\forall r_i \in R$, chọn *TopN* các NCV $r_j \in R$, $r_j \neq r_i$ để khuyến nghị cho r_i dựa trên giá trị hàm $f(r_i,r_j)$

- **Bước 1**: Tính trọng số theo xu hướng cho cạnh nối giữa 2 đỉnh u, v bất kỳ trong CoNet theo công thức:

$$Direct\_Sim\ (u, v,\ t_0) = \begin{cases} \dfrac{f_{Trend}(u,v,t_0)}{\sum_{\forall c \in N_u} f_{Trend}(u,c,t_0)}, & Nếu\ tồn\ tại\ cạnh\ giữa\ u,\ v\ trong\ E_1 \\ 0, & ngược\ lại \end{cases}$$

Với, $f_{Trend}(u,v,t_0)$ là hàm phụ thuộc yếu tố xu hướng cộng tác:

$$f_{Trend}(u,v,t_0) = \sum_{t_i=t_0}^{t_c} n(u,v,t_i) * \frac{1}{e^{(t_c-t_i)}})$$

Trong đó:

- N_u *là tập các đồng tác giả của u.*
- $n(u,v,t_i)$*: số bài báo u và v cộng tác viết tại thời điểm* t_i.

- t_0: *năm bắt đầu xem xét xu hướng cộng tác*
- t_c: *năm hiện tại*

• **Bước 2**: Tìm tất cả các đường đi đơn $p \in P_{u,v}$ có độ dài nhỏ hơn 4 giữa 2 đỉnh u, v bất kỳ trong CoNet.

$\forall u \in R$:

Duyệt theo chiều sâu từ đỉnh u, qua k đỉnh (z_1, z_2,…, z_k) (z_1 là u, z_k là v, với $\forall v \in R$, $v \neq u$), với $k < 5$

Thêm p= (z_1, z_2,…, z_k) vào tập $P_{u,v}$

• **Bước 3**: Tính trọng số theo xu hướng cho tất cả các đường đi đơn $p \in P_{u,v}$.

$\forall u \in R$, $\forall v \in R$, $u \neq v$:

$\forall p \in P_{u,v}$, tính:

$$WeightOf\_DirectPath_p(u, v, t_0) = \prod_{i=1}^{k-1} Direct\_Sim(z_i, z_{i+1}, t_0)$$

• **Bước 4**: Tính mức độ quan hệ giữa 2 đỉnh u, v trong CoNet:

Theo RSS$^+$:

$$Indirect\_Sim(u, v, t_0) = Indirect\_Sim_{RSS^+} = \sum_{p_i \in P_{u,v}} WeightOf\_DirectPath_{p_i}(u, v, t_0)$$

Theo MPRS$^+$:

$$Indirect\_Sim(u, v, t_0) = Indirect\_Sim_{MPRS^+} = \max_{p_i \in P_{u,v}} \left(WeightOf\_DirectPath_{p_i}(u, v, t_0)\right)$$

• **Bước 5**: Thực hiện khuyến nghị

$\forall r_i, r_j \in R$, $r_i \neq r_j$:

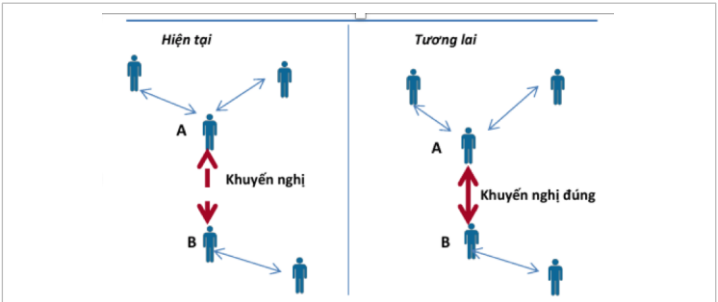
- $f(r_i, r_j) = Indirect\_Sim(r_i, r_j, t_0)$
- Chọn TopN các r_j có $f(r_i, r_j)$ lớn nhất để khuyến nghị.

***Độ phức tạp tính toán**: O(|R|^2 d^3). (d: bậc trung bình của một NCV = 2|E|/|R|)*

3.3.3 Thực nghiệm đánh giá

Hiện nay chưa có tập dữ liệu chuẩn để đánh giá cho bài toán khuyến nghị cộng tác. Hầu hết các nhóm nghiên cứu đều tiến hành thực nghiệm trên tập dữ liệu do họ thu thập và xây dựng. Với tính phổ biến của DBLP, NCS đã chọn thực nghiệm trên tập DBLP và tập CSPubGuru tự xây dựng.

Về phương pháp đánh giá cho hệ khuyến nghị, đây là một vấn đề vẫn đang được nghiên cứu. Những nghiên cứu phổ biến dùng kết quả tiên đoán liên kết đồng tác giả để đánh giá hiệu năng của các phương pháp khuyến nghị cộng tác [27, 28, 29, 117]. Chẳng hạn, hệ thống khuyến nghị A cộng tác với B. Sau đó, A có cộng tác với B thì đó là một khuyến nghị đúng, ngược lại là sai (hình 3.3). Luận án cũng dùng kết quả tiên đoán liên kết đồng tác giả để so sánh hiệu năng các phương pháp đề xuất với một số phương pháp phổ biến khác.

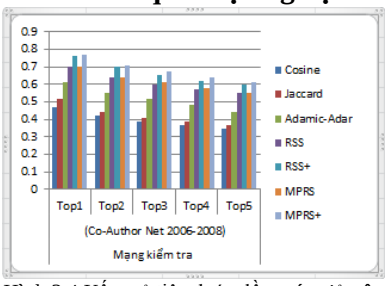


Hình 3.3. Minh họa đánh giá độ chính xác khuyến nghị cộng tác

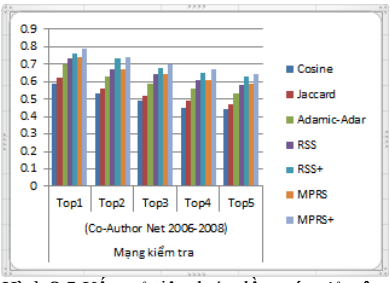
3.3.3.1 Thiết lập thực nghiệm cho DBLP và CSPubGuru

- **Huấn luyện**: Co-Author Net [2001-2005]
- **Đánh giá (GroundTruth)**: Co-Author Net [2006-2008]
- **Dữ liệu đầu vào:** phân các NCV đầu vào theo nhóm bậc: Thấp, Trung Bình, Cao. Chọn ngẫu nhiên 300 NCV, từ 3 nhóm bậc Thấp, Trung Bình, Cao.

3.3.3.2 Kết quả thực nghiệm



Hình 3.4 Kết quả tiên đoán đồng tác giả trên tập DBLP



Hình 3.5 Kết quả tiên đoán đồng tác giả trên tập CSPubGuru

Bảng 3.2: Kết quả tiên đoán đồng tác giả trên tập DBLP

Phương pháp	Mạng kiểm tra (Co-Author Net 2006-2008)				
	Top1	**Top2**	**Top3**	**Top4**	**Top5**
Cosine	0.47	0.42	0.39	0.37	0.35
Jaccard	0.52	0.44	0.41	0.39	0.37
Adamic-Adar	0.61	0.55	0.52	0.48	0.44
RSS	0.70	0.64	0.60	0.57	0.55
MPRS	0.70	0.64	0.61	0.58	0.55
RSS+	0.76	0.70	0.65	0.62	0.60
MPRS+	0.77	0.71	0.67	0.64	0.61

Bảng 3.3: Kết quả tiên đoán đồng tác giả trên tập CSPubGuru

Phương pháp	Mạng kiểm tra (Co-Author Net 2006-2008)				
	Top1	**Top2**	**Top3**	**Top4**	**Top5**
Cosine	0.59	0.53	0.49	0.45	0.44
Jaccard	0.62	0.56	0.52	0.49	0.47
Adamic-Adar	0.70	0.63	0.59	0.56	0.53
RSS	0.73	0.67	0.64	0.61	0.58
MPRS	0.74	0.67	0.64	0.61	0.59
RSS+	**0.76**	**0.73**	**0.68**	**0.65**	**0.63**
MPRS+	**0.79**	**0.74**	**0.70**	**0.67**	**0.64**

3.3.3.3 Nhận định

- Phương pháp đề xuất (phân tích quan hệ dựa trên xu hướng) cải tiến độ chính xác khuyến nghị cộng tác cho các NCV có liên kết đồng tác giả so với các phương pháp tương tự đỉnh phổ biến hiện nay.

3.4 Trường hợp các NCV chưa có đồng tác giả

3.4.1 Tiếp cận của luận án

Không có các thông tin đồng tác giả, quá trình cộng tác các phương pháp phân tích mạng đồng tác giả phổ biến hiện nay không thể thực hiện được (hình 3.1). Để giải quyết vấn đề này, luận án đã đề xuất dùng các thông tin hỗ trợ khác: tương tự sở thích nghiên cứu, quan hệ của các cơ quan, mức độ quan trọng, và tích cực của các nghiên cứu viên. Các thông tin hỗ trợ này được dùng như tập đặc trưng để học mô hình tiên đoán liên kết đồng tác giả dựa trên học máy giám sát [CT.3].

3.4.1.1 Tương tự nội dung nghiên cứu

Độ tương tự nội dung nghiên cứu của r và r' được tính như sau:

$$ContentSim(r, r') = \frac{(w_r . w_{r'})}{\|w_r\| . \|w_{r'}\|}$$

Trong đó, w_r: vector biểu diễn sở thích nghiên cứu của r.

3.4.1.2 Quan hệ giữa các cơ quan

Giả thuyết: những mối quan hệ mới tiềm năng thường xuất phát từ các cơ quan có quan hệ cộng tác mạnh.

$$w(o_i,\ o_{i+1}) = \frac{Coll\_Num(o_i, o_{i+1})}{Total\_Coll\_Num(o_i)}$$

$$Path\_Weight_p\ (o, o') = \prod_{i=1}^{k} w(o_i, o_{i+1})$$

$$OrgRS(o, o') = \sum_{i=1}^{m} Path\_Weight_{p_i}\ (o, o')$$

3.4.1.3 Uy tín của NCV

Giả thuyết: uy tín của NCV càng cao khi họ có nhiều trích dẫn của những NCV uy tín khác. Luận án dùng CiNet_Author<R, E_2> trong mô hình ASN đề xuất để tính uy tín của một NCV.

$$I.Rate(r_i) = \frac{1-d}{N} + d * (\sum_{r_j}^{LinkTo\ r_i} \frac{I.Rate(r_j)}{|OutLink(r_j)|} + \sum_{r_j has\ no}^{out-links} \frac{I.Rate(r_j)}{N})$$

Trong đó,

- N: Tổng số các NCV trong mạng trích dẫn (CiNet_Author)
- |OutLink(r)|: số lượng các out-link của r
- d: nhân tố thẩm thấu (damping factor) trong Random Walk with Restart (RWR) (**H. Tong và cộng sự [121]**).

3.4.1.4 Độ năng động của nghiên cứu

Giả thuyết: NCV năng động nếu ngày càng cho ra nhiều bài báo.

$f_{active}(r, t_0) = \sum_{i=0}^{c} N(r, t_i) * \frac{1}{e^{(t_c - t_i)}}$, trong đó,

- t_c: năm hiện tại
- t_0: năm bắt đầu xét mức độ năng động
- N(r, t_i): số lượng bài báo của NCV r tại thời điểm t_i

3.4.2 Phương pháp đánh giá

3.4.2.1 Độ chính xác tiên đoán liên kết

Tương tự với các nghiên cứu [28, 117], để lượng hóa độ chính xác tiên đoán liên kết cho các NCV chưa có đồng tác giả cần được khuyến nghị với các NCV khác, luận án dùng các độ đo phổ biến trong truy vấn thông tin như độ chính xác (Precision), độ bao phủ (Recall), độ đo F, độ chính xác trung bình AP (Average Precision) [9]. Nếu hệ thống tiên đoán một cặp (một NCV chưa có đồng tác giả và một NCV khác) sẽ là một cộng tác

đồng tác giả và mối quan hệ đồng tác giả này xảy ra trong tương lai thì xem như đây là một tiên đoán đúng, ngược lại là sai (hình 3.3).

3.4.2.2 Đề xuất phương pháp đánh giá chất lượng cộng tác

Luận án đưa ra giả thuyết: "Một quan hệ cộng tác tốt hơn những quan hệ cộng tác khác nếu nó tạo ra nhiều bài báo hơn". Khi đó, chất lượng của *TopN* những người cộng tác tiềm năng được khuyến nghị có thể lượng hóa như sau:

$$Collaboration\_Quality\_TopN(r,\{r_i\}) = \sum_{i=1}^{TopN} \frac{1}{e^i} * Coll\_Num(r,r_i)$$

$\{r_i\}$: là sanh sách xếp hạng các NCV khuyến nghị cho r.

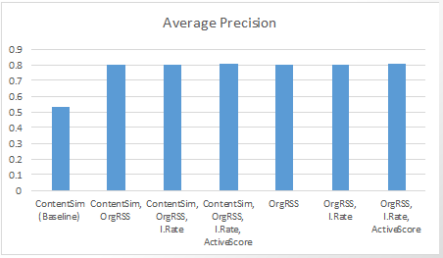
Coll_Num(r,r$_i$): số lần đồng tác giả của r với r_i.

3.4.3 Thực nghiệm

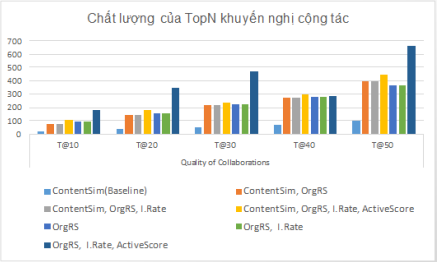
3.4.3.1 Tập dữ liệu thực nghiệm

- Rút trích từ CSPubGuru trong khoảng 2001 - 2011
- Researchers: 807.005
- Publications: 1.266.790
- G_0: [2001, 2005] (Chọn NCV chưa có đồng tác giả)
- G_1: [2006, 2011] (Chọn dữ liệu huấn luyện và kiểm tra)
- NCV chưa có đồng tác giả: 1491

3.4.3.2 Kết quả thực nghiệm



Hình 3.8: Độ chính xác tiên đoán đồng tác giả AP khi thêm các đặc trưng mới.



Hình 3.9: Chất lượng tiên đoán đồng tác giả khi thêm các đặc trưng mới.

Nhận định:

- Tương tự sở thích không ảnh hưởng đến quyết định cộng tác.
- Quan hệ giữa các cơ quan (OrgRS) là yếu tố đóng vai trò quyết định.

- Độ năng động của NCV là yếu tố quan trọng quyết định chất lượng cộng tác.

Chương 4 - Khai thác mạng xã hội học thuật để phát triển các phương pháp khuyến nghị bài báo khoa học

4.1 Giới thiệu

Trong phạm vi luận án, khuyến nghị bài báo khoa học cho NCV là bài toán với đầu vào là một hay nhiều NCV và tập các bài báo khoa học quan sát được. Hệ thống sẽ trả về danh sách xếp hạng các bài báo khoa học tiềm năng, ứng với quan tâm nghiên cứu của mỗi NCV.

4.2 Bài toán khuyến nghị bài báo khoa học

Cho trước,

- $R = \{r\}$: tập tất cả các NCV
- $P = \{p\}$: tập tất cả các bài báo đã quan sát.
- $R_p \subseteq R$: tập những nghiên cứu viên $r \in R$ đã thể hiện đánh giá, quan tâm với các bài báo khoa học $p \in P$.
- $P_r \subseteq P$: tập những bài báo được NCV r đánh giá, thể hiện sự quan tâm thông qua việc trích dẫn.
- Existed_Rating = $\{v(r', p')\}$, thể hiện mức độ liên quan của bài báo $p' \in P_r$ với NCV $r' \in R_p$.

Mục đích của hệ khuyến nghị bài báo khoa học là xây dựng hàm hữu ích $f(r,p)$ và ước lượng giá trị của hàm f để tiên đoán xem r sẽ quan tâm đến p nhiều hay ít, hay p tiềm năng và hữu ích đối với r như thế nào. Đối với mỗi NCV r_i, hệ khuyến nghị cần chọn $TopN$ bài báo khoa học, $P_{TopN} = <p_1, p_2, ..., p_{TopN}>$, tiềm năng và hữu ích nhất đối với NCV r_i để khuyến nghị. Các bài báo $P_{TopN} = <p_1, p_2, ..., p_{TopN}>$ được chọn thỏa mãn các điều kiện sau:

i) $\forall p_k \in P_{TopN}$, $v(r_i, p_k) \notin Existed\_Rating$. Tức phải khuyến nghị những bài báo p_k mà NCV r_i chưa biết.

ii) $\forall p_k \in P_{TopN}$, $f(r_i,p_k) \geq f(r_i,p_{k+1})$, *với* $1 \leq k \leq n-1$. Tức tập các bài báo khuyến nghị P_{TopN} là tập có thứ tự. Bài báo đứng trước có giá trị

hàm hữu ích f lớn hơn hoặc bằng bài báo đứng sau và ưu tiên khuyến nghị cho r_i hơn.

iii) $\forall p_k \in P_{TopN}$, $\forall p_{no\_rec} \in P \backslash P_{TopN}$, *thì* $f(r_i, p_k) \geq f(r_i, p_{no\_rec})$. Tức giá trị hữu ích của các bài báo được khuyến nghị, được xác định thông qua hàm f, phải lớn hơn hoặc bằng những bài báo không được khuyến nghị.

4.3 Khó khăn, thách thức

Tương tự các hệ khuyến nghị khác, hệ khuyến nghị bài báo khoa học cũng có những khó khăn, thách thức như:

- Dữ liệu lớn. Không gian NCV *R* và bài báo *P* là rất lớn.
- Ma trận đánh giá thưa. Ma trận thể hiện sự đánh giá, quan tâm của các NCV đối với các bài báo là rất thưa.
- Vấn đề khởi động lạnh. Quan sát thiếu hay không thể quan sát được các thông tin về NCV, cũng như bài báo khoa học.
- Chưa có tập dữ liệu chuẩn cho thực nghiệm, đánh giá.
- Độ chính xác khuyến nghị chưa cao.
- Chưa có phương pháp phù hợp để đánh giá kết quả bài báo khuyến nghị.

4.4 Phương pháp phổ biến và đề xuất

Luận án đề xuất khái niệm lòng tin và phương pháp lượng hóa lòng tin trong lĩnh vực học thuật. Tiếp cận của luận án dựa trên khai thác mạng xã hội học thuật ASN (mạng trích dẫn & mạng đồng tác giả). Kết hợp xu hướng sở thích và quan hệ lòng tin của NCV để thực hiện khuyến nghị bài báo tiềm năng có liên quan [CT.2, CT.9].

4.4.1 Xu hướng sở thích của NCV (CB-Recent): trên thực tế quan tâm nghiên cứu của NCV sẽ thay đổi theo thời gian và bị chi phối bởi nội dung của những bài báo gần đây nhiều hơn so với những bài đã công bố quá lâu trong quá khứ. Suyigama và đồng nghiệp đã khai thác yếu tố thời gian, đề xuất phương pháp khuyến nghị bài báo dựa trên mô hình quan tâm nghiên

gần đây của NCV, gọi tắt là CB-Recent [111]. Phương pháp CB-Recent có thể tóm tắt như sau:

Đầu vào: R = {r}, tập các nhà nghiên cứu quan sát được
P = {p}, tập bài báo của các nhà nghiên cứu.

Đầu ra: ∀r∈R, trả về TopN những p∈P dựa trên giá trị hữu ích tiên đoán.

Bước 1, 2: $\forall p \in P$.

- Rút trích phần tiêu đề và tóm tắt. Loại bỏ stopwords và stemming.
- Xây dựng vector biểu diễn nội dung bài báo p, là $\overrightarrow{f_p}$, dùng phương pháp gán trọng số TFIDF.

Bước 3: Xây dựng vector Profile cho các NCV $r \in R$, $\overrightarrow{P_r}$.

∀r∈R: xây dựng vector profile $\overrightarrow{P_r}$ cho mỗi nhà nghiên cứu ***r***.

$$\overrightarrow{P_r} = \sum_{i=1}^{n} e^{\gamma*(t_c - t(p_i))} * \overrightarrow{f_{p_i}}$$, Trong đó

γ: hệ số xu hướng. (γ ∈[0,1]. Trường hợp đơn giản α = 1)

t_c: năm hiện tại thực hiện khuyến nghị.

$t(p_i)$: năm công bố của bài báo p_i.

n: Tổng số bài báo mà *r* công bố trong quá khứ.

Bước 4: Thực hiện khuyến nghị

$\forall r \in R, \forall p \in P$

- $f(r,p) = Sim_{CB}(r,p) = Cosine(\overrightarrow{P_r}, \overrightarrow{f_p})$
- Chọn *TopN* những p∈P có f(r,p) lớn nhất khuyến nghị cho r∈R.

Độ phức tạp: O(|R||P|) *(|R|: số lượng NCV, |P|: số lượng bài báo)*

4.4.2 Xu hướng quan hệ lòng tin và sở thích (CB-TrendTrust)

Bên cạnh quan tâm nghiên cứu, các NCV thường đặt lòng tin vào một số chuyên gia trong lĩnh vực, cũng như hành vi lần theo các bài báo tham khảo và trích dẫn để chọn bài phù hợp liên quan đến quan tâm nghiên cứu của họ. Do đó, luận án đã đề xuất phương pháp lượng hóa xu hướng lòng tin kết hợp xu hướng nghiên cứu của NCV để phát triển phương pháp khuyến nghị bài báo khoa học liên quan cho NCV.

Đầu vào: R = {r}, tập các nhà nghiên cứu quan sát được
P = {p} tập bài báo của các nhà nghiên cứu.

Đầu ra: ∀r∈R, trả về TopN những p∈P dựa trên giá trị hữu ích tiên đoán.

Bước 1: Xây dựng mạng trích dẫn CiNet_Author, CoNet giữa các NCV

Bước 2: Mô hình hóa quan tâm nghiên cứu của NCV

$\forall p \in P$, tiền xử lý, vector hóa bài báo p dùng TFIDF, $\overrightarrow{f_p}$

$\forall r \in R$: xây dựng vector profile $\overrightarrow{P_r}$ cho mỗi nhà nghiên cứu ***r***.

$$\overrightarrow{P_r} = \sum_{i=1}^{n} e^{\gamma*(t_c - t(p_i))} * \overrightarrow{f_{p_i}}$$, (n: tổng số bài báo của r đã công bố)

Bước 3: Lượng hóa quan hệ lòng tin của r_i và r_j tính từ thời điểm t_0, $w_{trust}(r_i, r_j, t_0)$ dựa trên quan hệ trích dẫn.

$\forall r_i \in R$, $\forall r_j \in R$, $r_i \neq r_j$:

$$w_{coAuthor}(r_i, r_j, t_0) = f_{Trend}(r_i, r_j, t_0) = \frac{\sum_{t_i=t_0}^{t_c} NumColl(r_i, r_j, t_i)}{e^{\gamma*(t_c - t_i)} * TotalColl(r_i, t_0)}$$

$$w_{cite}(r_i, r_j, t_0) = \frac{\sum_{t_i=t_0}^{t_c} NumCitation(r_i, r_j, t_i)}{e^{\gamma*(t_c - t_i)} * TotalCitation(r_i, t_0)}$$

$$w_{trust}(r_i, r_j, t_0) = w_{cite}(r_i, r_j, t_0) + \frac{\sum_{r_u \in CoAuthor(r_i)} w_{coauthor}(r_i, r_u, t_0) * w_{cite}(r_u, r_j, t_0)}{|CoAuthor(r_i)|}$$

- *NumCitation(r_i,r_j,t_i)*: số lần r_i trích dẫn r_j trong năm t_i.
- *TotalCitation(r_i, t_0)*: Tổng số trích dẫn của r_i tính từ t_0 đến hiện tại (t_c).
- *NumColl (r_i, r_j, t_i)*: số lần mà r_i đồng tác giả với r_j trong năm t_i.
- *TotalColl(r_i, t_0)*: tổng số cộng tác của r_i tính từ năm t_0.

Bước 4: Tính mức độ lòng tin của r_i với bài báo p_j từ thời điểm t_0, $w_{trust}(r_i, p_j, t_0)$

$\forall r_i \in R$, $\forall p_j \in P$

$$w_{trust}(r_i, p_j, t_0) = \text{MAX}_{r_k \in Authors(p_j)}(w_{trust}(r_i, r_k, t_0))$$

(với $a_j \in A$: tập các tác giả của bài báo p_j)

Bước 5: Kết hợp lòng tin với xu hướng nghiên cứu của NCV.

Lặp $\forall r_i \in R$, $\forall p_j \in P$

- $f(r_i,p_j,t_0) = \alpha * w_{trust}(r_i, p_j, t_0) + (1 - \alpha) * Sim_{CB}(r_i, p_j)$
- Chọn TopN các $p_j \in P$ có $f(r_i,p_j,t_0)$ lớn nhất để khuyến nghị.

Độ phức tạp: O(|R||P|l) *(l: số tác giả trung bình của một bài báo)*

4.5 Thực nghiệm, đánh giá

4.5.1 Tập dữ liệu và thiết lập thực nghiệm

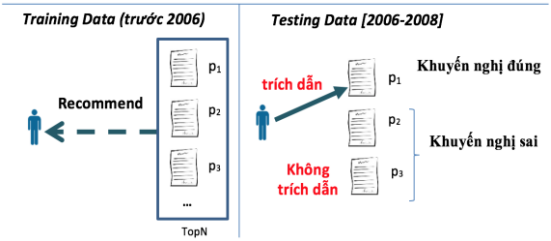
Sử dụng tập dữ liệu CSPubGuru để tiến hành thực nghiệm (https://sites.google.com/site/tinhuynhuit/dataset). Tương tự các nghiên cứu liên quan khác, luận án chia dữ liệu theo các khoảng thời gian là quá khứ (trước 2006) và tương lai [2006-2008]. Dữ liệu quá khứ để phân tích,

khuyến nghị. Dữ liệu tương lai làm GroundTruth để đánh giá độ chính xác khuyến nghị.

- 1000 NCV và bài báo của họ trước 2006 → dữ liệu đầu vào.
- GroundTruth: các bài báo 1000 NCV này trích dẫn từ 2006 đến 2008 (52.254 bài).

4.5.2 Phương pháp đánh giá kết quả khuyến nghị

TopN những đối tượng tiềm năng trả về từ hệ thống sẽ được dùng để đánh giá độ chính xác của phương pháp khuyến nghị. Nếu trong tương lai, NCV có trích dẫn bài báo được hệ thống khuyến nghị cho họ thì xem như khuyến nghị đúng, ngược lại là sai (hình 4.1). Các độ đo đánh giá được dùng phổ biến trong các nghiên cứu hiện nay đều có nguồn gốc từ lĩnh vực truy vấn thông tin (IR). Tương tự các nghiên cứu của Sugiyama và Kan [111, 112, 113], chúng tôi tập trung phân tích kết quả thực nghiệm với độ đo NDCG [58], MRR [123].



Hình 4.1 Minh họa cách tính độ chính xác khuyến nghị bài báo

4.5.3 Kết quả thực nghiệm

Bảng 4.1 Tóm tắt so sánh, đánh giá các phương pháp khuyến nghị bài báo

Phương pháp Khuyến nghị	**Độ đo đánh giá**		
	NDCG@5	NDCG@10	MRR
(CF-kNN, k=40)	0.0357	0.0330	0.0934
CB	0.2945	0.2334	0.5128
CB+R+C, Th_j = 0.8	0.2877	0.2282	0.4985
CB-Recent	**0.3577**	**0.2735**	**0.6142**
CBTrendTrust	**0.3610**	**0.2778**	**0.6164**

4.5.4 Nhận định

- Tiếp cận CF cho thấy không phải là tiếp cận phù hợp cho bài toán khuyến nghị bài báo liên quan, trong khi tiếp cận CB là tiếp cận phù hợp mà các nghiên cứu hiện nay đang dùng cho bài toán này.
- Khai thác yếu tố xu hướng để mô hình hoá sở thích NCV đã cải tiến đáng kể độ chính xác khuyến nghị.
- Kết hợp nội dung và quan hệ lòng tin góp phần cải tiến độ chính xác khuyến nghị bài báo, nhưng chưa đáng kể. (Tiếp tục nghiên cứu).

III. KẾT LUẬN

Các kết quả đạt được

Nhằm hỗ trợ các NCV dễ dàng hơn trong việc tìm kiếm, khai thác các thông tin học thuật, luận án đã tập trung nghiên cứu và phát triển các phương pháp khuyến nghị dựa trên tiếp cận phân tích mạng xã hội cho hai bài toán chính: (1) Khuyến nghị cộng tác; (2) Khuyến nghị bài báo khoa học. Sau quá trình nghiên cứu thực hiện, luận án đã đạt được một số kết quả có ý nghĩa khoa học như sau:

(1) Khảo sát, phân tích, đánh giá các cách tiếp cận cho hệ khuyến nghị và các nghiên cứu liên quan đến khuyến nghị thông tin học thuật.

(2) Đề xuất mô hình hóa các mạng xã hội học thuật nhận diện được từ kho dữ liệu học thuật, mô hình ASN [CT.6].

(3) Bài toán khuyến nghị cộng tác cho NCV:

- Đối với NCVcó quan hệ đồng tác giả: đề xuất, cải tiến các phương pháp phân tích xu hướng cộng tác trong mạng xã hội học thuật ASN để khuyến nghị các cộng tác viên tiềm năng. Các phương pháp đề xuất bao gồm: MPRS, MPRS+, RSS+ [CT.1, CT.4].
- Đối với NCV chưa có quan hệ đồng tác giả: đề xuất tập đặc trưng để khuyến nghị những mối quan hệ cộng tác tốt, chất lượng [CT.3].
- Đề xuất phương pháp đánh giá chất lượng cộng tác [CT.3].

(4) Bài toán khuyến nghị bài báo khoa học: phát triển phương pháp khuyến nghị bài báo khoa học cho NCV dựa trên việc khai thác mạng trích dẫn, quan hệ lòng tin trong mô hình ASN [CT.2, CT.8].

(5) Xây dựng kho dữ liệu hơn 6 triệu bài báo khoa học và triển khai thử nghiệm hệ thống tìm kiếm thông tin khoa học CSPubGuru (www.cspubguru.com) [CT.5, CT.7, CT.9, CT.10, CT.14].

Giá trị thực tiễn của luận án

- Ứng dụng các phương pháp khai thác mạng xã hội học thuật ASN vào các bài toán khuyến nghị trong lĩnh vực học thuật, hỗ trợ cộng đồng làm nghiên cứu khoa học. Một số bài toán ứng dụng đã được thử nghiệm như: khuyến nghị cộng tác, khuyến nghị bài báo liên quan, khuyến nghị hội thảo, tạp chí gởi bài.
- Kết quả nghiên cứu của luận án về hệ khuyến nghị có thể áp dụng cho nhiều lĩnh vực khác nhau.

Việc nghiên cứu, phát triển các phương pháp, hệ khuyến nghị, giải pháp thông minh giúp người dùng dễ dàng hơn trong việc tìm kiếm thông tin liên quan là một vấn đề lớn, còn nhiều khó khăn, thách thức đã và đang thu hút nhiều nghiên cứu của cộng đồng khoa học trên khắp thế giới. Những kết quả đạt được bước đầu giúp nghiên cứu sinh có được nền tảng tri thức để bước vào lĩnh vực nghiên cứu tiềm năng này.

Trong quá trình thực hiện luận án, nghiên cứu sinh cũng tham gia một số hoạt động khoa học khác như:

- Hợp tác với nhóm nghiên cứu Đại học Arkansas, USA để phát triển các phương pháp khuyến nghị gởi bài dựa trên tiếp cận phân tích mạng xã hội. [CT.12, CT.13].
- Chủ trì các đề tài nghiên cứu khoa học: đề tài cấp ĐHQG TpHCM loại C, 2011 (nghiệm thu loại tốt); Đề tài cơ sở, 2013 (nghiệm thu loại khá); Đề tài cấp ĐHQG loại C, 2014-2015 (đã hoàn thành báo cáo giữa kỳ); Đề tài cơ sở 2015 (đang thực hiện).

CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ

Tạp chí chuyên ngành

[CT. 1] Tin Huynh, Kiem Hoang. ***New Methods for Calculating Trend - Based Vertex Similarity for Collaboration Recommendation***. Journal of Computer Science and Cybernetics (Tạp chí Tin học và Điều khiển học – Viện KH&CN Việt Nam), vol.29, No.4, pages 338-350, 2013.

[CT. 2] Huỳnh Ngọc Tín, Hoàng Kiếm. ***Khai thác xu hướng sở thích và quan hệ lòng tin để phát triển phương pháp khuyến nghị bài báo khoa học***. Chuyên san "Các công trình nghiên cứu, phát triển và ứng dụng Công nghệ thông tin và Truyền thông" - Tạp chí Công nghệ Thông tin và Truyền thông, Tập V-1, Số 13 (33), trang 67-78, 06-2015.

Hội thảo chuyên ngành

[CT. 3] Tin Huynh, Atsuhiro Takasu, Tomonari Masada, Kiem Hoang. ***Collaborator Recommendation for Isolated Researchers***. In Proceedings of the 28th IEEE International Conference on Advanced Information Networking and Applications (AINA-2014), pages 639-644, Victoria, Canada, May 13-16, 2014.

[CT. 4] Tin Huynh, Kiem Hoang, Dao Lam. ***Trend Based Vertex Similarity for Academic Collaboration Recommendation***. In Proceedings of 5th International Conference on Computational Collective Intelligence Technologies and Applications (ICCCI 2013), pages 11-20, Craiova, Romani, September 11-13, 2013.

[CT. 5] Tin Huynh, Kiem Hoang, Tien Do, Duc Huynh. ***Vietnamese Author Name Disambiguation for Integrating Publications from Heterogeneous Sources***. In Proceedings of the 5th Asian conference on Intelligent Information and Database Systems (ACIIDS 2013), pages 226-235, Kuala Lumpur, Malaysia, March 18-20, 2013.

[CT. 6] Tin Huynh, Kiem Hoang. ***Modeling Collaborative Knowledge of Publishing Activities for Research Recommendation***. In Proceedings of the 4th International Conference on Computational Collective Intelligent Technologies and Applications (ICCCI 2012), pages 41-50, Ho Chi Minh City, Vietnam, November 28-30, 2012.

[CT. 7] Tin Huynh, Hiep Luong, and Kiem Hoang. ***Integrating bibliographical data of computer science publications from online digital libraries***. In Proceedings of the 4th Asian conference on Intelligent Information and Database Systems (ACIIDS'12), pages 226-235, Kaohsiung, Taiwan, March 19-21, 2012.

[CT. 8] Tin Huynh, Kiem Hoang, Loc Do, Huong Tran, Hiep Luong, Susan Gauch. ***Scientific Publication Recommendations Based on Collaborative Citation Networks***. In Proceedings of the 2012 International Conference on Collaboration Technologies and Systems (CTS 2012), pages 316-321, Denver, Colorado, USA, May 21-25, 2012.

[CT. 9] Tin Huynh, Kiem Hoang. ***GATE framework based metadata extraction from scientific papers***. In Proceedings of the International Conference on Education and Management Technology (ICEMT 2010), page 188 – 191, Cairo, Egypt, November 02-04, 2010.

[CT. 10] Hung Nghiep Tran, Tin Huynh, Tien Do. ***Author Name Disambiguation by Using Deep Neural Network***. In Proceedings of the 6th Asian conference on Intelligent Information and Database Systems (ACIIDS'14), pages 123-132, Bangkok, Thailand, April 7-9, 2014.

[CT. 11] Hung Nghiep Tran, Tin Huynh, Kiem Hoang. ***A Potential Approach to Overcome Data Limitation in Scientific Publication Recommendation***. In Proceedings of the seventh international conference on knowledge and systems engineering (KSE-2015), TpHCM, Vietnam, Oct 8-10, 2015.

[CT. 12] Hiep Luong, Tin Huynh, Susan Gauch, Loc Do, and Kiem Hoang. ***Publication venue recommendation using author network's publication history***. In Proceedings of the 4th Asian conference on Intelligent Information and Database Systems, Kaohsiung, Taiwan, March 2012 (ACIIDS'12), pages 426-435, Kaohsiung, Taiwan, March 19-21, 2012.

[CT. 13] Hiep Luong, Tin Huynh, Susan Gauch, Kiem Hoang. ***Exploiting Social Networks for Publication Venue Recommendations***. In Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2012), pages 239- 245, Barcelona, Spain, October 4-7, 2012.

[CT. 14] Tien Do, Dao Lam, Tin Huynh. ***A Framework for integrating bibliographical data of computer science publications***. 2014 International Conference on Computing, Management and Telecommunications (ComManTel 2014), pages 245-250, Da Nang, Vietnam, April 27-29, 2014.