Tạp chí Khoa học Đại học Huế: Kỹ thuật và Công nghệ; ISSN 2588–1175

Tập 127, Số 2A, 2018, Tr. 109–120;DOI: 10.26459/hueuni-jtt.v127i2A.5017

HỆ KHUYẾN NGHỊ CỘNG TÁC ĐỒNG TÁC GIẢ

Trần Đình Khang[1], Võ Đức Quang[2], Nguyễn Đăng Tuấn Anh[1]

[1]Trường Đại học Bách khoa Hà Nội, Số 1 Đại Cồ Việt, Hà Nội
[2] Trường Đại học Vinh

Tóm tắt: Mạng đồng tác giả là mạng lưới học thuật giữa các nhà nghiên cứu viết chung bài báo khoa học và mức độ kết hợp đồng tác giả có thể được đặc trưng bởi các độ đo liên kết. Dựa trên các đặc trưng đó, có thể xây dựng nhiều bài toán có ý nghĩa, trong đó có khuyến nghị cộng tác, gợi ý các tác giả có thể kết hợp trong tương lai hoặc tăng cường hợp tác. Bài báo này đề xuất một số độ đo liên kết mới dựa trên cộng đồng tác giả, kịch bản thiết lập bảng ứng viên động theo thời gian và xây dựng hệ khuyến nghị đồng tác giả sử dụng các độ đo đó.

Từ khóa: mạng đồng tác giả, độ đo liên kết, khuyến nghị cộng tác

1 Đặt vấn đề

Trong nghiên cứu khoa học, các nhà khoa học tạo ra sản phẩm là các bài báo khoa học, trong đó thường có nhiều người cùng tham gia và đứng tên đồng tác giả. Một nhà nghiên cứu đóng góp vào nhiều công trình khoa học sẽ có nhiều đồng tác giả khác nhau mà mức độ liên kết giữa họ có thể đo được bằng số các bài báo viết chung hoặc các thông tin khác như sự gắn kết về chuyên môn và nhóm nghiên cứu. Mối quan hệ giữa các tác giả và bài báo là quan hệ nhiều–nhiều, một tác giả có thể tham gia viết nhiều bài báo, một bài báo có thể có một hay nhiều tác giả đứng tên tạo ra một mạng lưới học thuật gọi là mạng đồng tác giả [2, 3, 8] với các nút là các tác giả, các cạnh thể hiện mối liên kết giữa hai tác giả. Theo cách biểu diễn đó, thì có thể coi mạng đồng tác giả là một mạng xã hội đặc biệt kế thừa nhiều đặc trưng của mạng xã hội nói chung như quan hệ lân cận chung và đường dẫn liên kết, nhưng cũng chứa đựng các đặc trưng riêng về chuyên môn, lĩnh vực nghiên cứu, cộng đồng nghiên cứu, v.v…

Với các tính chất như vậy, việc xây dựng mạng đồng tác giả và giải quyết các bài toán đặt ra với mạng đồng tác giả đang thu hút sự quan tâm của nhiều nhóm nghiên cứu. Về các bài toán, có thể biểu diễn mạng đồng tác giả như các cơ sở dữ liệu để thực hiện các truy vấn, tìm kiếm đồng tác giả, nhưng cũng có thể thực hiện các bài toán dẫn xuất thông tin như dự đoán liên kết đồng tác giả hay khuyến nghị liên kết đồng tác giả [1, 4, 7, 11]. Việc dẫn xuất thông tin xem hai nhà khoa học có thể là đồng tác giả trong tương lai hay không là một bài toán có ý nghĩa giúp cho nhà khoa học mở rộng mối quan hệ học thuật của mình và tìm các sự cộng tác

**Liên hệ: khangtd@soict.hust.edu.vn*

Nhận bài: 22–10–2018; Hoàn thành phản biện: 10–11–2018; Ngày nhận đăng: 22–11–2018

phù hợp trong tương lai. Các tính toán như vậy sẽ dựa vào các sự liên kết đồng tác giả trong quá khứ. Người ta thường lượng hóa mức độ liên kết giữa hai tác giả bằng các độ đo liên kết như độ đo lân cận chung và độ đo Jaccard [5, 6, 9]. Ngoài các độ đo thông dụng cho mạng xã hội còn có các nghiên cứu bổ sung các độ đo đặc thù cho mạng đồng tác giả như vị trí tác giả trong bài báo hoặc lĩnh vực chuyên môn [8, 10].

Từ mạng đồng tác giả ở thời điểm hiện tại có thể tính toán được các cặp tác giả tiềm năng liên kết trong tương lai hay còn gọi là ứng viên đồng tác giả. Kèm theo đó là các độ đo liên kết của các cặp ứng viên đó tạo thành bảng ứng viên đồng tác giả. Xét mạng đồng tác giả trong một khoảng thời gian $T1$, thì bảng ứng viên đồng tác giả có các hàng là các ứng viên đồng tác giả xét theo khoảng thời gian $T1$, các cột là các độ đo liên kết tính theo khoảng thời gian $T1$. Nếu $T2$ là khoảng thời gian xảy ra sau $T1$, thì có thể bổ sung thêm cột nhãn, có giá trị là 1 nếu cặp ứng viên thực sự là đồng tác giả trong khoảng $T2$, và có giá trị là –1 nếu cặp ứng viên không là đồng tác giả trong khoảng $T2$. Khi đó, có thể sử dụng bảng ứng viên với các độ đo và cột nhãn như một tập dữ liệu cho học máy để xây dựng mô hình về mối quan hệ giữa nhãn với các độ đo liên kết. Bài toán khuyến nghị cộng tác trở thành bài toán học mô hình và tính toán nhãn liên kết (1/–1) theo mô hình đó. Với mạng đồng tác giả có kích thước lớn thì số liên kết cũng rất lớn, theo bình phương của số nút. Do đó, một đặc tính của bảng ứng viên đồng tác giả là số ứng viên có nhãn –1 vượt trội so với số ứng viên có nhãn 1, tạo ra sự mất cân bằng về nhãn.

Bài báo này có các đóng góp:

- Đề xuất thêm các độ đo về cộng đồng nghiên cứu, kết hợp với các độ đo truyền thống khác. Khảo sát bằng thực nghiệm sự ảnh hưởng của các độ đo với hiệu quả của mô hình để xác định tập độ đo liên kết phù hợp,
- Xây dựng bảng ứng viên theo kịch bản khoảng thời gian động để tận dụng các nhãn liên kết 1 làm cho bảng ứng viên đồng tác giả bớt mất cân bằng hơn,
- Xây dựng hệ khuyến nghị đồng tác giả.

Bài báo được tổ chức như sau: phần tiếp theo trình bày về mạng đồng tác giả, các độ đo liên kết và bảng ứng viên. Phần 3 trình bày về các độ đo liên kết mới, kịch bản cải tiến thiết lập bảng ứng viên và đánh giá ảnh hưởng các độ đo liên kết đến hiệu quả dự báo. Phần 4 giới thiệu về hệ khuyến nghị cộng tác đồng tác giả.

2 Mạng đồng tác giả

2.1 Định nghĩa mạng đồng tác giả

Một mạng đồng tác giả có thể được mô tả bằng hàm $G^T=(V^T,E^T,P^T,T)$, trong đó $T=\{t_1, t_2,\ldots, t_k\}$ là tập các nhãn thời gian; $V^T=\{v_1, v_2,\ldots,\}$ là tập các đỉnh được tạo trong thời gian T, mỗi nút đại diện cho một tác giả trong cộng đồng nghiên cứu; $P^T=\{p_1, p_2,\ldots,\}$ là tập các bài báo trong thời

gian T; E^T= {($v_{i,j}$, p_k, t_h)} là tập các liên kết giữa các tác giả, thể hiện trong thời gian T, hai tác giả (v_i, v_j) có viết chung bài báo p_k tại nhãn thời gian t_h.

Ngoài ra, tập đỉnh V^T còn có thể chứa các thuộc tính của từng nút tương ứng với thông tin cá nhân của các tác giả như quốc tịch, trường Đại học/ Viện Nghiên cứu mà họ công tác, các lĩnh vực chuyên ngành, v.v… Các thuộc tính này được ký hiệu bằng tập A^T= {a_1, a_2,…,$_N$}, trong đó a_i là vector đặc trưng chứa thông tin của tác giả/ đỉnh v_i. Các độ đo sự tương đồng giữa hai tác giả sẽ được xây dựng dựa trên thông tin của các tập E^T và A^T.

Cho trước một khoảng thời gian T thì G^T là mạng đồng tác giả tương ứng với lát cắt thời gian đó. Bài toán khuyến nghị cộng tác sẽ sử dụng các thông tin từ G^T để đưa ra các khuyến nghị cho một tác giả v_i lựa chọn các ứng viên phù hợp để cộng tác đồng tác giả ở thời gian tiếp theo hoặc khuyến nghị cho một cặp tác giả (v_i, v_j) tiếp tục tăng cường cộng tác đồng tác giả.

2.2 Các độ đo liên kết giữa hai tác giả

Mức độ liên kết của một cặp tác giả trong mạng đồng tác giả thường được lượng hóa bởi các độ đo liên kết được trích xuất thông tin từ các tập E^T, A^T. Dưới đây là một số độ đo thông dụng. Các độ đo liên kết này có thể áp dụng trong nhiều loại mạng xã hội khác nhau, không chỉ riêng cho mạng đồng tác giả. Vì tính chất phổ biến của các độ đo này, bài báo sẽ chỉ trình bày sơ lược về tên và nội dung của từng độ đo. Chi tiết về ý tưởng và nguồn gốc của từng độ đo người đọc có thể tham khảo thêm trong các tài liệu liên quan [2, 5].

Với mỗi nút v, ký hiệu $T(v)$ chỉ tập các nút lân cận của v trong mạng đồng tác giả G. Ta có thể chia các độ đo liên kết thành hai nhóm chính: nhóm độ đo dựa trên lân cận và nhóm độ đo dựa trên đường đi.

a/ Nhóm độ đo dựa trên lân cận (neighbour-based metrics)

(i) Độ đo Common Neighbour (CN): Độ đo Common Neighbour giữa hai nút u và v được tính bằng số lượng nút lân cận chung của u và v. Số lượng lân cận chung càng cao thì độ tương đồng CN càng lớn, do đó khả năng (u) có liên kết trong tương lai càng cao.

$$CN(u,v) = |T(u) \cap T(v)| \quad (1)$$

(ii) Độ đo Adamic Adar (AA): Độ đo Adamic-Adar quan sát thêm số lượng nút lân cận của từng lân cận chung. Với z là lân cận chung của cả u và v thì độ đo Adamic-Adar tỷ lệ nghịch với số lượng lân cận của z tính theo logarit.

$$AA(u,v) = \sum_{z \in T(u) \cap T(v)} \frac{1}{\log(|T(z)|)} \quad (2)$$

(iii) Độ đo Jaccard Coefficient (JC): Độ đo Jaccard Coefficient giữa hai nút u và v được tính bằng tỉ lệ số lượng lân cận chung trên tổng số lân cận của hai nút.

$$JC(u,v) = \frac{|\,T(u) \cap T(v)\,|}{|\,T(u) \cup T(v)\,|} \quad (3)$$

(iv) Độ đo Preferential Attachment (PA): Độ đo Preferential Attachment thể hiện hai nút càng có nhiều lân cận (bậc càng lớn) thì càng có cơ hội liên kết với nhau trong tương lai.

$$PA\ (u,\ v) = |\ (u)\ | \times |\ (v)\ | \quad (4)$$

(v) Độ đo Resource Allocation (RA): Độ đo Resource Allocation có công thức tương tự như Adamic Adar, chỉ có khác biệt ở phần mẫu số là số lượng lân cận của z.

$$RA(u,v) = \sum_{z\in T(u)\cap T(v)} \frac{1}{|T(z)|} \quad (5)$$

b/ Nhóm độ đo dựa trên đường đi (path-based metrics)

(i) Độ đo ShortestPath: Độ đo ShortestPath được tính bằng nghịch đảo của khoảng cách ngắn nhất giữa hai nút. Trong trường hợp giữa hai nút không có đường đi thì độ đo có giá trị bằng 0.

$$ShortestPath(u,v) = \frac{1}{d(u,v)} \quad (6)$$

(ii) Độ đo Katz: Độ đo Katz được tính dựa trên việc thống kê tất cả đường đi giữa hai nút u và v theo độ dài tăng dần. Các đường đi càng dài thì ảnh hưởng tới độ đo càng giảm do chịu tác động của hàm mũ.

$$Katz(u,v) = \sum_{l=1\to\infty} \beta^l \left|\mathrm{path}^l_{u,v}\right| = \beta A + \beta A^2 + \beta A^3 + \cdots \quad (7)$$

trong đó, $path^l_{u,v}$ là tập các đường đi độ dài l từ u đến v; β là hằng số tùy chọn. Khi β tiến tới 0 thì độ đo trở nên tương tự với độ đo lân cận chung do các đường đi có độ dài lớn đóng góp rất ít vào kết quả cuối cùng.

2.3 Bảng ứng viên đồng tác giả

Từ mạng đồng tác giả ở thời điểm hiện tại, có thể tính toán được các cặp tác giả tiềm năng liên kết trong tương lai, hay còn gọi là ứng viên đồng tác giả. Kèm theo đó là các độ đo liên kết của các cặp ứng viên đó tạo nên bảng ứng viên đồng tác giả. Xét mạng đồng tác giả trong một khoảng thời gian $T1$ thì bảng ứng viên đồng tác giả có các hàng là các ứng viên đồng tác giả xét theo khoảng thời gian $T1$; các cột là các độ đo liên kết tính theo khoảng thời gian $T1$. Nếu $T2$ là khoảng thời gian xảy ra sau $T1$ thì có thể bổ sung thêm cột nhãn, có giá trị là 1 nếu cặp ứng viên thực sự là đồng tác giả trong khoảng $T2$ và có giá trị là –1 nếu cặp ứng viên không là đồng tác giả trong khoảng $T2$.

Bảng 1. Bảng ứng viên đồng tác giả

	Các độ đo liên kết ở khoảng thời gian T_1	Nhãn liên kết =1 (hoặc = –1), nếu là đồng tác giả (hoặc không phải đồng tác giả) trong khoảng thời gian T_2
Các cặp ứng viên đồng tác giả ở khoảng thời gian T_1	Giá trị các độ đo liên kết	Giá trị nhãn

Thủ tục 1: Xây dựng bảng ứng viên đồng tác giả từ mạng đồng tác giả G. Tính các độ đo liên kết trong khoảng thời gian $T1$, và gán nhãn từ mạng đồng tác giả trong khoảng thời gian $T2$ (xảy ra sau $T1$).

– Bước 1:Xác định tập các cặp ứng viên đồng tác giả; (u,v) là một cặp ứng viên nếu $\exists p \in P^{T1}$, $t \in T1$: $(u,v,p,t) \in E^{T1}$,hoặc $\exists z \in V^{T1}$, $p_1,p_2 \in P^{T1}$, $t_1,t_2 \in T1$: (u,z,p_1,t_1), $(z,v,p_2,t_2) \in E^{T1}$.
– Bước 2: Tính các độ đo liên kết của các cặp ứng viên trong khoảng thời gian $T1$.
– Bước 3: Gán nhãn cho các cặp ứng viên; gán nhãn 1 cho cặp (u,v) nếu $\exists p \in P^{T2}$, $t \in T2$: $(u,v,p,t) \in E^{T2}$, ngược lại, gán nhãn –1.

Khi đó, có thể sử dụng bảng ứng viên với các độ đo và cột nhãn như một tập dữ liệu cho học máy để xây dựng mô hình về mối quan hệ giữa nhãn với các độ đo liên kết.

3 Các độ đo theo cộng đồng tác giả và thiết lập bảng ứng viên đồng tác giả

3.1 Xây dựng các độ đo liên kết dựa trên cộng đồng tác giả

Để so sánh sự tương đồng hay "gần gũi" giữa hai tác giả, ngoài việc sử dụng các đặc trưng liên kết của mạng, chúng ta còn có thể khai thác các thông tin ngữ nghĩa của từng cá nhân tác giả. Một tác giả hay một nhà nghiên cứu được đặc trưng bởi một số thông tin như quốc tịch, nơi làm việc (trường Đại học / Viện nghiên cứu) và lĩnh vực chuyên môn mà họ ưa thích. Các tác giả có chung quốc tịch hoặc nơi làm việc thường có sự gần gũi nhất định về mặt địa lý và ngôn ngữ, do đó khả năng họ có liên kết mới trong tương lai cũng cao hơn so với cặp tác giả không chung thông tin này. Tương tự với cặp tác giả có cùng lĩnh vực chuyên môn ưa thích, sự tương đồng giữa các vấn đề nghiên cứu mà họ quan tâm sẽ dẫn đến xác suất hợp tác lớn hơn.

Ngoài ra, các tác giả có chung quốc tịch, nơi làm việc hoặc lĩnh vực chuyên môn thường có xu hướng hình thành một cộng đồng trong mạng lưới học thuật. Các thành viên trong cộng đồng này thường có mối liên hệ chặt chẽ với nhau và có khả năng chia sẻ thông tin một cách nhanh chóng và dễ dàng hơn. Xuất phát từ mối liên hệ trên, các độ đo liên kết mới sẽ được xây dựng dựa trên thông tin từ nhiều cộng đồng khác nhau, bao gồm cộng đồng tác giả theo quốc gia và cộng đồng tác giả theo lĩnh vực chuyên môn.

a. Độ đo cộng đồng tác giả theo quốc gia

Xét tập tác giả $V = \{v_1, v_2, \ldots, v_N\}$, trong đó tác giả v_i được đặc trưng bởi hai thuộc tính: quốc tịch và nơi công tác (trường Đại học/ Viện nghiên cứu) ký hiệu bằng $affil_{country}(v_i)$ và $affil_{university}(v_i)$.

Ta có hàm so sánh sự giống nhau về nơi công tác và quốc tịch giữa hai hoặc nhiều tác giả:

$$sim\_work(v_1, v_2, \ldots,) = \begin{cases} 2 & if \quad affil_{university}(v_1) = affil_{university}(v_2) = \cdots = affil_{university}(v_n) \\ 1 & if \quad affil_{country}(v_1) = affil_{country}(v_2) = \cdots = affil_{country}(v_n) \\ 0 & if \quad otherwise \end{cases} \tag{8}$$

Độ tương đồng giữa hai tác giả u và v theo cộng đồng quốc gia được tính theo công thức

$$CommCountry\ (u, v) = sim\ \_work\ (u, v) + \sum_{z \in T(u) \cap T(v)} sim\ \_work\ (z, u, v) \quad (9)$$

Có thể thấy độ đo $CommCountry$ sẽ quan sát sự tương đồng về nơi công tác giữa hai tác giả, đồng thời tính đến sự tương đồng của các lân cận chung trong cùng một cộng đồng quốc gia hoặc cộng đồng trường đại học.

b. Độ đo cộng đồng tác giả theo lĩnh vực chuyên môn

Mỗi tác giả trong mạng lưới học thuật còn được đặc trưng bởi các lĩnh vực chuyên môn mà họ quan tâm. Để tìm ra các lĩnh vực chuyên môn này của một tác giả chúng ta có thể lấy thông tin từ nội dung các bài báo được công bố trong quá khứ của họ. Mô hình chủ đề (Topic model) [8] là một trong những phương pháp có thể áp dụng để phân tích các chủ đề từ một tập các bài báo đầu vào. Kết quả của mô hình chủ đề cho ta biết xác suất bài báo p sẽ thiên về chủ đề nào nằm trong số lượng K chủ đề cho trước thể hiện qua vector đặc trưng chủ đề $T = (t_1, t_2, \ldots, t_K)$. Từ kết quả phân tích chủ đề các bài báo, ta có thể xác định danh sách các chủ đề mà một tác giả có khả năng quan tâm theo phương pháp sau.

Gọi $(v_i) = \{p_{i1}, p_{i2}, \ldots, p_{iN}\}$ là danh sách các bài báo mà tác giả v_i đã công bố trong quá khứ. Kết quả phân tích chủ đề của các bài báo này là $paper\_(v_i) = \{T_{i1}, T_{i2}, \ldots, T_{iN}\}$ với T_{iN} là vector gồm K thành phần tương ứng với xác suất bài báo p_{iN} thuộc về một trong số K chủ đề. Từ các thông tin trên, ta có vector đặc trưng về lĩnh vực quan tâm của tác giả v_i được tính theo công thức

$$T_{vi} = \sum_{j=1 \to N} T_{ij} = (t_{i1}, t_{i2}, \ldots, t_{iK}) \quad (10)$$

Vector T_{vi} gồm K thành phần thể hiện sự quan tâm của tác giả v_i đến một số lĩnh vực (chủ đề) nhất định trong danh sách K lĩnh vực chuyên môn. Bằng việc chọn một ngưỡng θ thích hợp, ta có thể lọc ra danh sách các lĩnh vực được tác giả v_i quan tâm nhất:

$$Topics\ (v_i) = \{j \mid j \in [1. .K] \wedge t_{ij} > \theta\} \quad (11)$$

Mặt khác, các phần tử của tập (v_i) sẽ thể hiện các cộng đồng chuyên môn mà tác giả v_i là một thành viên. Từ thông tin của các cộng đồng này, ta sẽ xây dựng độ đo liên kết giữa hai tác giả (u, v) dựa trên cộng đồng tác giả theo lĩnh vực chuyên môn như sau:

$$CommTopic(u, v) = |Topics(u) \cap Topics(v)| + \sum_{z \in T(u) \cap T(v)} |Topics(z) \cap Topics(u) \cap Topics(v)| \quad (12)$$

Có thể thấy với độ đo $CommTopic$, hai tác giả có càng nhiều lĩnh vực chung thì càng có khả năng liên kết với nhau trong tương lai. Hơn nữa, số lượng các lân cận chung nằm trong cùng cộng đồng chuyên môn cũng làm tăng khả năng liên kết giữa hai người.

3.2 Kịch bản thiết lập bảng ứng viên

Để thiết lập bảng ứng viên, có thể chia các khoảng thời gian và tính toán các ứng viên, độ đo và gán nhãn như trình bày ở Thủ tục 1. Đặc trưng của bảng ứng viên là số lượng các cặp ứng

viên có nhãn –1 lớn hơn rất nhiều so với số lượng cặp ứng viên có nhãn 1. Thực tế là một cặp ứng viên (u,v) từ khoảng thời gian $T1$ có thể trở thành đồng tác giả thực sự sau này, nhưng nếu chỉ gán nhãn trong khoảng thời gian $T2$ thì vẫn lấy nhãn –1 do chưa phải là đồng tác giả ở $T2$. Điều này có thể làm mất đi nhiều mẫu có nhãn 1 nếu xét theo các khoảng thời gian cố định.

Bài báo đề xuất một kịch bản cải tiến mới phù hợp hơn, trong đó các liên kết mới xuất hiện ở thời điểm t được gán độ đo từ thông tin của mạng đồng tác giả trong cả khoảng thời gian trước đó $[0, t-1]$ hay mốc thời gian phân chia giai đoạn thay đổi theo thời điểm quan sát. Cách tiếp cận này có ưu điểm là tận dụng được toàn bộ thông tin về liên kết giữa các tác giả trong quá khứ, đồng thời không bỏ sót liên kết mới nào để thiết lập bảng ứng viên. Hơn nữa, kịch bản này cũng mô phỏng chính xác hơn quá trình xuất hiện các liên kết mới trong thực tế được kỳ vọng là sẽ giúp tăng hiệu quả khuyến nghị.

Thủ tục 2: Xây dựng bảng ứng viên đồng tác giả từ mạng đồng tác giả G trong khoảng thời gian $T= \{t_1,t_2,\ldots,t_k\}$.

- Bước 1: Xác định tập các cặp ứng viên đồng tác giả; (u,v) là một cặp ứng viên, nếu $\exists p \in P^T$, $t \in T$: $(u,v,p,t) \in E^T$, hoặc $\exists z \in V^T$, $p_1,p_2 \in P^T$, $t_1,t_2 \in T$: (u,z,p_1,t_1), $(z,v,p_2,t_2) \in E^T$.
- Bước 2: Xét các nhãn thời gian t_i, bắt đầu từ t_k đến t_1.
- Với mỗi $(u,v,p,t_i) \in E^T$ thì tính các độ đo cho (u,v) trong khoảng thời gian $[t_1,t_{i-1}]$, gán nhãn 1 cho (u,v), và từ bây giờ không tính lại với cặp (u,v) này nữa.
- Bước 3: Với các cặp ứng viên chưa được gán nhãn thì đều gán nhãn –1.

Kịch bản cải tiến có ưu điểm là tận dụng được các nhãn 1. Sau đây là thực nghiệm với các dữ liệu thu thập từ thư viện khoa học trực tuyến ScienceDirect (sciencedirect.com) gồm các bài báo và tác giả thuộc ba tạp chí: *Chemical Physics Letters*, *Journal of Molecular Biology* và *Biochemical and Biophysical Research Communications* [12, 13, 14]. Các bài báo được lấy nằm trong khoảng thời gian từ năm 2000 cho đến hết năm 2017. Thông tin về số bài, số tác giả có trong Bảng 2.

Bảng 2. Thông tin về dữ liệu thử nghiệm

Tên tạp chí (tên bộ dữ liệu)	ISSN	Số bài báo	Số tác giả	Số bài báo trung bình trong 1 năm	Số quốc gia có bài báo được xuất bản
Chemical Physics Letters (***chem_letter***)	00092614	18 931	41 806	1 113	114
Journal of Molecular Biology (***mole_bio***)	00222836	10 806	35 217	635	97
Biochemical and Biophysical Research Comm. (***biophy_chem***)	0006291X	34 848	134 448	2 049	128

Tổng cộng		64 585	211 471	3 797	176

Bảng 3 trình bày thông tin về số lượng nhãn dương trong mỗi bộ dữ liệu kiểm tra tương ứng với các kịch bản truyền thống và kịch bản cải tiến. Có thể thấy kịch bản cải tiến giúp tận dụng được nhiều nhãn dương hơn trong các bộ dữ liệu.

Bảng 3. Số mẫu dữ liệu có nhãn dương theo các kịch bản

Bộ dữ liệu \Kịch bản	Truyền thống	Cải tiến
chem_letter	1250	1460
mole_bio	780	910
biophy_chem	1780	2110

3.3 Đánh giá sự ảnh hưởng của các độ đo

Phần này sẽ xem xét sự ảnh hưởng của các độ đo thông qua thực nghiệmvới dữ liệu và các kịch bản thiết lập bảng ứng viên như mô tả ở phần trên. Bảng ứng viên được đưa vào một thủ tục phân lớp dựa vào các độ đo liên kết để phân lớp nhãn. Chia bảng dữ liệu thành bộ dữ liệu huấn luyện và bộ dữ liệu kiểm tra. Đánh giá hiệu quả phân lớp thông qua các tiêu chí AUC (Area Under Curve) và F1-score. Độ đo AUC đặc trưng cho xác suất chọn ngẫu nhiên hai cặp tác giả thì xác suất dự báo (predict probability) của cặp tác giả có liên kết sẽ lớn hơn cặp tác giả không có liên kết. Nếu AUC = 1 tương ứng với việc dự báo là tốt nhất, trong khi với phương pháp dự báo ngẫu nhiên thì AUC = 0,5. Độ đo F1-score = 2 ×Precision/ (Precision + Recall).

Các độ đo liên kết được thử nghiệm bao gồm các độ đo truyền thống được trình bày ở Phần 2.2 và hai độ đo cộng đồng mới được trình bày ở Phần 3.1 là Community country và Community topics. Các thử nghiệm so sánh hiệu quả phân lớp các tổ hợp độ đo theo kịch bản cải tiến thiết lập bảng ứng viên. Kết quả ở Bảng 4 và Bảng 5 theo các độ đo AUC và F1-Score.

Bảng 4. Kết quả AUC của các tổ hợp độ đo cộng đồng mới + truyền thống

Bộ dữ liệu	**Độ đo**						
	Comm Country	*Comm Topic*	*CN+JC +AA+PA+ Katz*	*Comm Country + Comm Topic*	CN+Katz+Comm Country	CN+Katz +Comm Topic	CN+Katz+Comm Topic+ Comm Country
chem_letter	0,7345	0,7540	0,7356	0,8341	0,8130	0,8033	**0,8651**
mole_bio	0,713	0,6929	0,6929	0,7625	0,7477	0,7012	**0,7780**
biophy_chem	0,8916	0,6845	0,8192	0,9279	0,9117	0,8363	**0,9387**

Bảng 5. Kết quả F1-Score của các tổ hợp độ đo cộng đồng mới + truyền thống

Bộ dữ liệu	**Độ đo**						
	Comm Country	*Comm Topic*	*CN+JC +AA+PA+ Katz*	*Comm Country + Comm Topic*	*CN+Katz+ Comm Country*	*CN+Katz +Comm Topic*	*CN+Katz+ Comm Topic+ Comm Country*
chem_letter	0,6911	0,6800	0,6761	0,7623	0,7455	0,7294	**0,8116**
mole_bio	0,7009	0,6573	0,6221	0,7128	0,7112	0,6854	**0,7223**
biophy_chem	0,8244	0,6456	0,7742	0,8503	0,8335	0,7820	**0,8710**

Kết quả cho thấy sự cải thiện đáng kể về hiệu quả dự báo khi sử dụng kết hợp các độ đo cộng đồng mới với các độ đo truyền thống. Tỉ lệ cải thiện trung bình là 15%. Các thử nghiệm trên định hướng cho việc lựa chọn tổ hợp các độ đo liên kết đồng tác giả khi thiết lập bảng ứng viên đồng tác giả cho tính toán các khuyến nghị.

4 Xây dựng hệ khuyến nghị cộng tác đồng tác giả

Việc xây dựng hệ khuyến nghị cộng tác bao gồm bagiai đoạn:

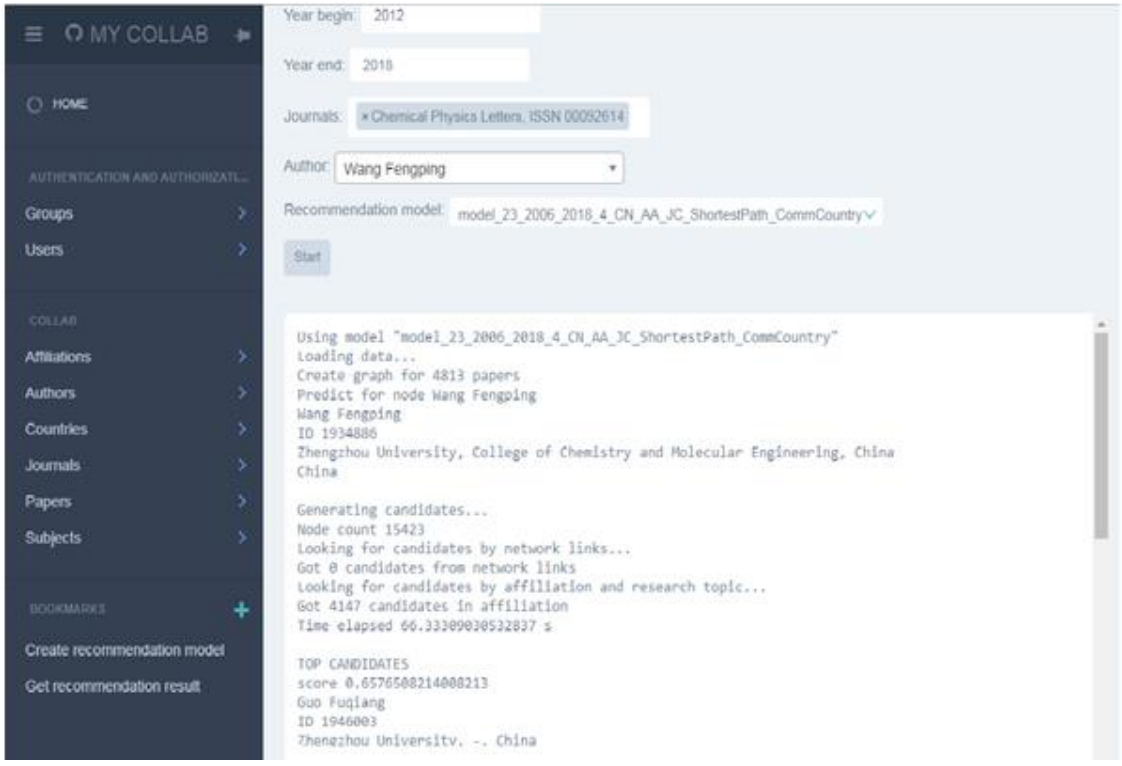
- Thu thập dữ liệu, phân tích, tổ chức dữ liệu,
- Tính toán các độ đo liên kết và thiết lập các bản ứng viên,
- Xây dựng mô hình khuyến nghị.

Hệ thống khuyến nghị đồng tác giả được xây dựng với mục đích giúp các nhà nghiên cứu có thể tìm được danh sách những người có thể cộng tác hiệu quả trong tương lai một cách nhanh chóng và thuận tiện nhất.

Dữ liệu thử nghiệm của hệ thống là các thông tin về bài báo và tác giả (tiêu đề bài báo, tóm tắt nội dung, từ khóa, thông tin tác giả, v.v…) từ 3 tạp chí *Chemical Physics Letters*, *Journal of Molecular Biology*, và*Biochemical and Biophysical Research Communications* của Sciencedirect trong khoảng thời gian 2000–2017 thông qua **API** của **Sciencedirect**. Các thông tin khoa học được thiết kế tổ chức lại thành các cơ sở dữ liệu quan hệ. Cụ thể các bảng dữ liệu: *Journal*, *Country*, *Subject*, *Institute*, *Author*, *Paper*, *PaperAuthor*, *CoAuthorship*.Các bảng *Country*, *Subject*, *Institute* bổ sung thông tin cho *Author*, bảng *Journal* bổ sung thông tin cho*Paper*, bảng *AuthorPaper* cho biết tác giả của các bài báo cụ thể. Từ đó tính được *CoAuthorship* chứa các cặp đồng tác giả.

Với CSDL đã có, tiến hành xây dựng hoàn thiện bảng ứng viên với kịch bản thiết lập đã trình bày (Phần 3.2) sử dụng các phương pháp tính toán các độ đo liên kết (Phần 2.2 và 3.1). Sử

dụng phương pháp Tf-Idf để vector hóa nội dung gồm tiêu đề và tóm tắt của các bài báo; sau đó sử dụng phương pháp NMF (Non-Negative Matrix Factorization) để xác định vector đặc trưng chủ đề. Các tham số được sử dụng gồm số *topic n_topics* = 40 và độ dài vector *Tf-Idf n_length* = 600. Từ đó, tính toán được các độ đo liên kết cho các cặp ứng viên. Chức năng khuyến nghị được xây dựng dựa trên mô hình phân lớp Support Vector Machine (SVM) với dữ liệu đã gán nhãn của bảng ứng viên bao gồm học mô hình từ dữ liệu huấn luyện là bảng ứng viên đã gán nhãn, lưu trữ mô hình và sử dụng mô hình để tính toán khuyến nghị đồng tác giả.



Hình 1. Giao diện của hệ thống

Về công nghệ, hệ khuyến nghị đồng tác giả được thiết kế theo mô hình MVC (Model-View-Controller) sử dụng hệ quản trị CSDL MySQL, ngôn ngữ lập trình Python, thư viện Django Web Framework và thư viện ScikitLearn (Python) để cài đặt các thành phần chức năng và giao diện của hệ thống. Các kết quả thử nghiệm ở Phần 3.3 được chương trình cài đặt và thực thi trên máy tính chạy hệ điều hành Ubuntu 64bit, cấu hình i5 4200U@2.5Ghz, 8GB RAM.

Hệ khuyến nghị đồng tác giả cho phép đưa ra top-N ứng viên theo thứ tự có khả năng cộng tác phù hợp nhất đối với một tác giả bất kỳ. Ngoài ra, hệ thống còn xây dựng các chức năng bổ sung như tìm kiếm, truy vấn và cập nhật thông tin tác giả, bài báo, tạp chí, quốc gia, và cho phép hiện thị trực quan mạng đồng tác giả.

Ví dụ với tác giả *Wang Fengping*, hệ thống khuyến nghị top-5 các ứng viên tiềm năng: *Guo Fuqiang*, *Guan Xinxin*, *Wang Yanan*, *Liu Pu*, *Huang Qiuying*. Theo giao diện như ở Hình 1, người dùng cung cấp thông tin về khoảng thời gian, lựa chọn các độ đo liên kết và tên tác giả cần khuyến nghị. Hệ thống sẽ thực hiện phân lớp theo mô hình đã được huấn luyện để chọn ra các cặp ứng viên nhãn 1, trong đó có một thành phần là tác giả đó. Top-N đồng tác giả tiềm năng được lấy ra từ thành phần còn lại trong các cặp vừa được tính.

5 Kết luận

Bài báo đã trình bày và phân tích về các độ đo liên kết trong mạng đồng tác giả, từ đó phát triển thêm các độ đo bổ sung về cộng đồng nghiên cứu. Bài báo cũng cải tiến xây dựng bảng ứng viên theo kịch bản khoảng thời gian động để tận dụng các nhãn liên kết dương, làm cho bảng ứng viên đồng tác giả bớt mất cân bằng hơn. Các khảo sát thực nghiệm cho thấy việc phối hợp các độ đo cơ bản và độ đo cộng đồng mới, kết hợp với việc sử dụng kịch bản xây dựng bảng ứng viên cải tiến đã cho hiệu quả khuyến nghị chính xác và hiệu quả hơn. Dựa trên cơ sở dữ liệu về thông tin bài báo học thuật thu thập được, các tác giả đã xây dựng một hệ thống khuyến nghị cộng tác khá hoàn chỉnh về chức năng, đáp ứng nhu cầu tra cứu, tham khảo và có nhiều tiềm năng phát triển mở rộng.

Tài liệu tham khảo

1. Zervas P, Tsitmidelli A, Sampson DG, Chen NS, Kinshuk (2014), Studying research collaboration patterns via co-authorship analysis in the field of Tel: The case of educational technology & society journal, *Educ Technol Soc* 17(4), pp 1–16
2. M. A. Brandão, M. M. Moro, G. R. Lopes, and J. P. M. Oliveira (2013), *Using link semantics to recommend collaborations in academic social networks*, in Proc.22nd Int. Conf. World Wide Web Companion (WWW Companion), pp.833–840
3. W. Glänzel and A. Schubert (2005), *Analysing scientific networks through co-authorship*, in Handbook of Quantitative Science and Technology Research.New York, NY, USA: Springer-Verlag, pp. 257–276.
4. S. Lee and B. Bozeman (2005), The impact of research collaboration on scientific productivity, *Soc. Stud. Sci.*, vol. 35, no. 5, pp. 673–702.
5. D. Liben-Nowell and J. Kleinberg (2007), The link-prediction problem for social networks, *J. Amer. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031.
6. R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla (2010), *New perspectives and methods in link prediction*, in Proc. 16th ACM SIGKDD Int. Conf. Knowl.Discovery Data Mining (KDD), Washington, DC, USA, pp. 243–252.
7. Milen Pavlov, Ryutaro Ichise (2007), *Finding Experts by Link Prediction in Co-authorship Networks*, Proceeding of 2nd International ExpertFinder Workshop (FEWS2007), pp. 42–55
8. Pham Minh Chuan, Le Hoang Son, Mumtaz Ali, Tran Dinh Khang, Le Thanh Huong, Nilanjan Dey (2018), Link Prediction in Co-authorship Networks based on Hybrid Content Similarity Metric, *Applied Intelligence*, 48(8), ISSN: 0924-669X. Doi: 10.1007/s10489-017-1086-x, pp. 2470–2486

9. Phạm Minh Chuẩn, Trịnh Khắc Linh, Trần Đình Khang, Lê Hoàng Sơn (2017), *Phân tích sự ảnh hưởng của một số độ đo liên kết áp dụng vào bài toán dự đoán liên kết trong mạng đồng tác giả*, Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR) – Đà Nẵng, 17–18/8/2017. ISBN: 978–604– 913–614–6, trang 760–767.
10. [Pham Minh Chuan, Cu Nguyen Giap, Le Hoang Son, Chintan Bhatt, Tran Dinh Khang (2017), *Enhance Link Prediction in Online Social Networks Using Similarity Metrics, Sampling, and Classification*, Proceeding of the 4th International Conference on Information System Design and Intelligent Applications (INDIA–2017), 15–17 June 2017, Danang, Vietnam, DOI: 10.1007/978-981-10-7512-4_81, pp. 823 – 833
11. [Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008), *ArnetMiner: extraction and mining of academic social networks,* In Proceeding of the 14thACM SIGKDD International conference on Knowledge discovery and datamining, KDD '08, pages 990–998, New York, NY, USA. ACM
12. https://www.sciencedirect.com/journal/chemical-physics-letters/, truy cập tháng 6/2017
13. https://www.sciencedirect.com/journal/journal-of-molecular-biology, truy cập tháng 6/2017
14. https://www.sciencedirect.com/journal/biochemical-and-biophysical-research-communications, truy cập tháng 6/2017

CO-AUTHORSHIP RECOMMENDATION SYSTEMS

Tran Dinh Khang[1], Vo Duc Quang[2], Nguyen Dang Tuan Anh[1]

[1] Hanoi University of Science and Technology, No. 1 Đai Co Viet Street, Hà Nội
[2] Vinh University

Abstract. A co-authorship network is an academic network amongresearchers who could write a joint scientific paper, where the degree of co-authorship can be characterized by linking measures. On the basis of these characteristics, many meaningful problems can be created, including recommendations for collaboration, suggestions for future collaborators or increased collaboration. This article proposes some new linking measures based on the research community, a new time-dependent candidates set-up scenario, and the development of a co-authorship recommendation system using those measures.

Keywords: co-authorship network, linking measure, co-authorship recommendation