ISSN 2354-0575

HỆ THỐNG KHUYẾN NGHỊ CỘNG TÁC DỰA TRÊN PHÂN CỤM BÁN GIÁM SÁT MỜ VÀ ỨNG DỤNG TRONG MẠNG HỢP TÁC KHOA HỌC

Bùi Thế Hồng
Trường Đại học Sư phạm Kỹ thuật Hưng Yên

Ngày tòa soạn nhận được bài báo: 02/07/2018
Ngày phản biện đánh giá và sửa chữa: 02/08/2018
Ngày bài báo được duyệt đăng: 15/08/2018

Tóm tắt:

Bài toán khuyến nghị cộng tác giữa các nhà nghiên cứu có tiềm năng hiện đang rất được chú trọng. Hầu hết các nghiên cứu hiện tại đều giải quyết bài toán khuyến nghị cộng tác dựa trên phương pháp phân lớp nhị phân có cộng tác và không có cộng tác. Tuy nhiên, do mạng hợp tác khoa học rất thưa dẫn đến tập dữ liệu dùng để huấn luyện thường gặp phải vấn đề mất cân bằng dẫn đến hiệu quả phân lớp không cao. Bài báo này đề xuất một hệ thống khuyến nghị cộng tác dựa trên phân cụm bán giám sát mờ để khắc phục nhược điểm của phương pháp phân cụm nhị phân đối với những dữ liệu thưa và không cân bằng. Kết quả thực nghiệm đối với hệ thống khuyến nghị cộng tác đã đề xuất được thực hiện trên một tập dữ liệu thực tế cho thấy trong hầu hết các trường hợp hệ thống khuyến nghị cộng tác dựa trên phân cụm bán giám sát mờ hiệu quả hơn hẳn so với hệ thống khuyến nghị cộng tác dựa trên phân lớp nhị phân.

Từ khóa: *Hệ thống khuyến nghị cộng tác, phân lớp, phân cụm bán giám sát mờ.*

1. Giới thiệu

Ngày nay, với sự phát triển của mạng xã hội liên quan đến thông tin cá nhân của nhiều người, việc gợi ý tự động cho người sử dụng các thông tin cũng như các sản phẩm có thể họ muốn mua hoặc quan tâm, các cá nhân có cùng sở thích hoặc cùng lĩnh vực nghiên cứu với họ là một việc khả thi và đem lại nhiều lợi ích cho con người. Các hệ khuyến nghị đã được quan tâm nghiên cứu và phát triển nhanh chóng, đặc biệt các hệ khuyến nghị trong thương mại điện tử đem lại nhiều lợi nhuận cho các nhà bán sản phẩm. Bên cạnh các hệ khuyến nghị trong thương mại điện tử, các hệ khuyến nghị liên quan đến khoa học kỹ thuật cũng được quan tâm nghiên cứu. Ví dụ, LinkedIn và ResearchGate khuyến nghị các công việc mà ai đó (hoặc người dùng nào đó) có thể ứng tuyển, thông báo các thông tin liên quan đến các nhà khoa học có các công trình nghiên cứu được tham chiếu trong các bài báo khoa học, v.v. Tuy nhiên, việc khuyến nghị các tác giả có các nghiên cứu liên quan đến nhau để hợp tác trong tương lai còn chưa được đưa vào trong các mạng xã hội này. Đây cũng là những khuyến nghị rất có giá trị giúp cho các nhà nghiên cứu tăng cường hợp tác để tạo ra các công trình khoa học mới trong tương lai.

Bài toán khuyến nghị các nhà nghiên cứu có tiềm năng hợp tác còn được gọi với tên là bài toán khuyến nghị cộng tác trong mạng hợp tác khoa học. Trong bài báo này, khái niệm "khuyến nghị cộng tác" ("Collaborations Recommendation" được sử dụng trong [1]) mang hàm ý về sự khuyến nghị hợp tác trong việc xuất bản bài báo khoa học giữa các nhà nghiên cứu (tác giả). Ở đây, mạng hợp tác khoa học là một mạng xã hội có thể biểu diễn dưới dạng một đồ thị vô hướng, trong đó các đỉnh là các nhà khoa học, các cạnh là các mối cộng tác khoa học giữa các nhà khoa học.

Bài toán khuyến nghị cộng tác trong mạng hợp tác khoa học được phát biểu như sau:

Cho thông tin về các tác giả đã từng viết chung bài báo khoa học đến thời điểm t, với một tác giả u nào đó, cần tìm ra một danh sách tác giả có tiềm năng cộng tác (Collaboration) với tác giả u trong tương lai (từ thời điểm t' > t).

Các vấn đề nghiên cứu trong mạng hợp tác khoa học luôn thú vị bởi tính phức tạp chung của bài toán khuyến nghị cộng tác. Mặt khác, việc xây dựng được một hệ thống khuyến nghị cộng tác nghiên cứu sẽ thúc đẩy quá trình giao lưu và hợp tác trong nghiên cứu khoa học.

Bài toán khuyến nghị cộng tác được bắt nguồn từ bài toán dự đoán liên kết trong mạng xã hội, trong đó các độ đo liên kết giữa các cặp tác giả giữ vai trò quan trọng, làm cơ sở để xác định khả năng hình thành liên kết (hợp tác) trong tương lai giữa các cặp tác giả. Hướng tiếp cận phổ biến là chuyển bài toán dự đoán liên kết về bài toán phân lớp nhị phân [2] với hai lớp là có liên kết và không có liên kết. Bài toán dự báo liên kết đã được nhiều nghiên cứu quan tâm [3, 4, 5].

Các nghiên cứu trước đây về khuyến nghị cộng tác thường sử dụng một số độ đo liên kết trọng số như S_{CN}^{pt} [15], S_{AA}^{pt} [15], S_{JC}^{pt} [17],..vv đã được đề xuất trong mạng xã hội thông thường để xây dựng tập đặc trưng. Tuy nhiên, mạng hợp tác khoa học là

một mạng xã hội có nhiều đặc trưng riêng so với các mạng xã hội nói chung. Ví dụ, mức độ cộng tác giữa hai tác giả cùng viết chung các bài báo phụ thuộc vào số lượng bài báo, số lượng tác giả, thứ tự của các tác giả và thời gian công bố của các bài báo mà hai tác giả đã viết chung. Ngoài ra, một nhân tố rất quan trọng có thể ảnh hưởng đến việc cộng tác giữa các tác giả trong tương lai là sự tương đồng về lĩnh vực nghiên cứu. Hai tác giả có thể nghiên cứu nhiều lĩnh vực khác nhau và nếu một số hướng nghiên cứu chính có sự tương đồng cao thì tiềm năng cộng tác trong việc viết chung các bài báo khoa học trong tương lai càng lớn.

Trên thực tế, mỗi nhà nghiên cứu khi công bố các bài báo khoa học ở các tạp chí hoặc hội thảo có thể có cách hành văn khác nhau, trong đó một số từ ngữ đồng nghĩa được sử dụng có thể phản ánh ý nghĩa tương tự nhau hoặc cùng có hàm ý về một số chủ đề nghiên cứu nào đó. Vì vậy, trong nghiên cứu [6] đã đề xuất cách thức xác định mức độ tương đồng giữa các tác giả dựa trên nội dung tóm tắt của bài báo, thông tin về thứ tự của tác giả và thời gian công bố của bài báo.

Hầu hết các nghiên cứu đều tiếp cận giải quyết bài toán khuyến nghị cộng tác dựa trên phân lớp nhị phân, với hai lớp là có cộng tác (nhãn 1) và không cộng tác (nhãn 0). Tuy nhiên, do mạng hợp tác khoa học rất thưa dẫn đến tập dữ liệu dùng để huấn luyện thường gặp phải vấn đề mất cân bằng nhãn, dẫn đến hiệu quả phân lớp không cao. Để giải quyết vấn đề mất cân bằng nhãn, trong bài báo này, chúng tôi đề xuất hệ thống khuyến nghị cộng tác dựa trên hệ thống phân cụm bán giám sát mờ với đặc trưng là các độ đo liên kết trọng số và độ đo liên kết dựa trên nội dung tóm tắt bài báo đã đề xuất trong [6].

2. Các nghiên cứu liên quan

Bài toán khuyến nghị truyền thống, chủ yếu tập trung vào ba hướng tiếp cận chính đó là: *(i) hướng tiếp cận dựa trên lọc cộng tác*. Một số thuật toán học máy khác nhau đã được áp dụng trong hướng tiếp cận này, chẳng hạn như Naive Bayes [7] và dựa trên luật [8]. *(ii) hướng tiếp cận dựa trên nội dung* [9, 10], ý tưởng chủ đạo của hướng tiếp cận này là đưa ra khuyến nghị những sản phẩm tương tự (tương đồng) với những sản phẩm mà người dùng đã thích (quan tâm) trong quá khứ sẽ được xem xét. Trong đó, độ tương tự giữa hai sản phẩm được tính toán dựa trên những đặc điểm (đặc trưng) gắn với những sản phẩm được so sánh. *(iii) hướng tiếp cận lai (hybrid)* [11, 12], là một cách kết hợp hai hoặc nhiều phương pháp khuyến nghị nhằm đạt được độ chính xác (hiệu suất) tốt hơn so với khi áp dụng riêng lẻ phương pháp bất kỳ nào đó.

Các nghiên cứu về bài toán khuyến nghị trong mạng xã hội nói chung và mạng hợp tác khoa học nói riêng thường tiếp cận giải quyết bài toán theo hướng học không giám sát. Tức là tính toán độ tương tự giữa một nút (tác giả) v với các nút ứng cử dựa trên thông tin cấu trúc mạng hoặc dựa trên ngữ nghĩa, sau đó lựa chọn ra N nút có mức độ tương tự lớn nhất với nút v. Với cách tiếp cận này, việc đưa ra danh sách khuyến nghị sẽ được thực hiện một cách đơn giản và nhanh chóng.

Tuy nhiên, giữa hai tác giả trong mạng hợp tác khoa học có nhiều đặc trưng, chẳng hạn những đặc trưng dựa trên thông tin cấu trúc mạng (các độ đo liên kết trong mạng), dựa trên sự tương đồng về lĩnh vực nghiên cứu hay dựa trên việc cùng tham gia các sự kiện khoa học (chẳng hạn báo cáo hoặc hội nghị khoa học, …). Việc sử dụng đồng thời nhiều đặc trưng để đưa ra khuyến nghị theo hướng tiếp cận học không giám sát là không dễ và có thể không đạt được kết quả khuyến nghị mong muốn.

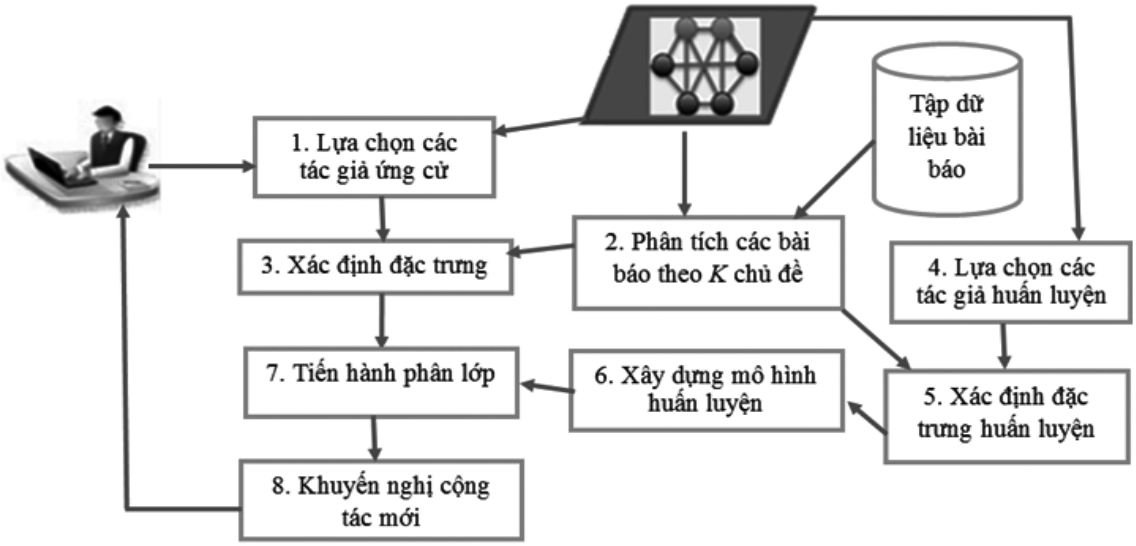
Trong các nghiên cứu về bài toán khuyến nghị cộng tác [2, 7, 8, 10], các tác giả tiếp cận giải bài toán khuyến nghị theo hướng học có giám sát, cụ thể là sử dụng các phương pháp phân lớp. Thông qua kết quả thực nghiệm, phần nào đã khẳng định được tính hiệu quả khi áp dụng phương pháp phân lớp vào bài toán khuyến nghị cộng tác trong mạng hợp tác khoa học. Hình 1 biểu diễn một hệ thống khuyến nghị cộng tác dựa trên phân lớp một cách khái quát thông qua các nghiên cứu [2, 7, 8, 10]. Chi tiết các bước thực hiện khuyến nghị cộng tác được mô tả như sau.

Bước 1: Từ dữ liệu ban đầu thu thập trong mạng hợp tác khoa học, xác định danh sách các tác giả ứng cử được sử dụng để đưa ra khuyến nghị cộng tác cho một tác giả nào đó. Danh sách các tác giả ứng cử là những tác giả mà chưa từng cộng tác trước đó và có ít nhất một láng giềng chung với tác giả cần được khuyến nghị.

Bước 2: Áp dụng phương pháp phân tích chủ đề (LDA để biểu diễn mỗi bài báo (thông qua tên và nội dung tóm tắt) dưới dạng một véc-tơ trong không gian K chiều, cách thức thực hiện giống như trong [6].

Bước 3: Trong bước này, tập các véc-tơ đặc trưng ứng với mỗi cặp tác giả (cụ thể là giữa tác giả cần được khuyến nghị với các tác giả ứng cử trong bước 1) sẽ được xác định dựa trên các độ đo liên kết (xem trong Bảng 2).

Bước 4: Lựa chọn các tác giả dùng để xây dựng tập đặc trưng huấn luyện. Trong đó, các tác giả này không được trùng với các tác giả ứng cử đã chọn trong bước 1.



Hình 1. *Hệ thống khuyến nghị cộng tác mới dựa trên phân lớp*

Bước 5: Xác định đặc trưng huấn luyện (tương tự như bước 3, nhưng chỉ xét với tập tác giả huấn luyện trong bước 4).

Bước 6: Xây dựng mô hình huấn luyện phân lớp dựa trên tập dữ liệu huấn luyện trong bước 5. Trong bước này, tác giả thử nghiệm với phương pháp phân lớp SVM.

Bước 7: Áp dụng mô hình phân lớp đã thực hiện trong bước 6 với tập các véc-tơ đặc trưng nhận được từ bước 3 để tiến hành phân lớp các cặp tác giả. Kết quả phân lớp sẽ được sử dụng để đưa ra khuyến nghị cộng tác mới.

Bước 8: Từ kết quả phân lớp sẽ xác định được cặp tác giả nào thuộc vào nhãn 1 (có cộng tác). Dựa vào đó sẽ đưa ra khuyến nghị cộng tác mới cho tác giả đã được lựa chọn.

Ngoài hướng tiếp cận học có giám sát, hướng tiếp cận học bán giám sát đã được nhiều nghiên cứu quan tâm. Học có giám sát là quá trình học đi kèm với một tập mẫu huấn luyện. Còn học bán giám sát là học có một số lượng nhỏ các mẫu huấn luyện sử dụng để điều hướng quá trình học. Trong bài báo này, các phương pháp phân cụm mờ [13] và bán giám sát mờ [14] sẽ được sử dụng để xây dựng hệ thống khuyến nghị cộng tác.

Năm 2007, Murata và Moriyasu [15] đã đề xuất ba độ đo liên kết trọng số lấy ý tưởng từ ba độ đo liên kết không trọng số. Sau đó, De Sá và Prudêncio [16] đã thực nghiệm các độ đo liên kết trọng số trên mạng hợp tác khoa học được xây dựng từ DPLP. Gần đây Günes và cộng sự [17] cũng đã thực nghiệm các độ liên kết trọng số trên mạng hợp tác khoa học được xây dựng từ tập các bài báo thuộc lĩnh vực "theoretical high-energy physics" Hep-Th[1].

Để thuận tiện theo dõi, các độ đo liên kết trọng số áp dụng trong mạng hợp tác khoa học được ký hiệu một cách tổng quát là S^{type}_{metric}, trong đó chỉ số trên (type) ký hiệu đại diện cho kiểu trọng số cộng tác ω_{type}, chỉ số dưới (metric) ký hiệu cho độ đo liên kết không trọng số được mở rộng. Như vậy, các độ đo liên kết trọng số được mở rộng từ các độ đo liên kết không trọng số được phân biệt bởi kiểu trọng số cộng tác ω_{type}.

Trong nghiên cứu [6], tác giả đã đề xuất độ đo liên kết trọng số dựa trên thứ tự tác giả và thời gian công bố của bài báo. Các độ đo liên kết trọng số lần lượt được ký hiệu là S^{pt}_{CN}, S^{pt}_{AA}, S^{pt}_{JC}, tương ứng với các công thức (1, 2, 3) với trọng số liên kết ω_{pt} được xác định bởi công thức (5) [23].

$$S^{pt}_{CN}(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{\omega_{pt}(u,z)+\omega_{pt}(v,z)}{2} \quad (1)$$

$$S^{pt}_{AA}(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{(\omega_{pt}(u,z)+\omega_{pt}(v,z))}{2Log(\sum_{z'\in\Gamma(z)}\omega_{pt}(z,z'))} \quad (2)$$

$$S^{pt}_{AA}(u,v) = \sum_{z\in\Gamma(u)\cap\Gamma(v)} \frac{\omega_{pt}(u,z)+\omega_{pt}(v,z)}{2Log(1+\sum_{z'\in\Gamma(z)}\omega_{pt}(z,z'))} \quad (3)$$

Trong đó, $\Gamma(u)$ là tập các tác giả đã từng cộng tác với tác giả u; $\omega_{pt}(u,z)$ là số bài báo mà hai tác giả *u, z* đã từng viết chung.

Xét hai tác giả *u, v* trong danh sách các tác giả xuất hiện trong một bài báo và thứ tự tương ứng của hai tác giả là d_u và d_v. Giả sử $d_v > d_u$ và trong mỗi bài báo có nhiều hơn một tác giả. Khi đó, mức

[1] https://arxiv.org/archive/hep-th/

độ liên kết giữa hai tác giả *u, v* (*DCL*(*u, v*)) trong bài báo được tính theo công thức (4).

$$DCL(d_u, d_v) = \begin{cases} \frac{1}{d_u} + \frac{1}{d_v} & if\ 2 \le d_v \le 3 \\ \frac{1}{d_u} + \frac{2}{d_v} & if\ d_v > 3, 1 \le d_u \le 3 \\ \frac{2}{d_u} + \frac{2}{d_v} & if\ d_u > 3 \end{cases} \quad (4)$$

Giả sử hai tác giả *u* và *v* viết chung *P* bài báo. Khi đó trọng số liên kết giữa hai tác giả được tính theo công thức (5).

$$\omega_{pt}(u,v) = \sum_{p=1}^{P} DCL(d_u^p, d_v^p) * k(t_p) \quad (5)$$

Trong đó, d_u^p là thứ tự của tác giả *u* trong bài báo thứ *p*, t_p là thời gian mà bài báo thứ *p* được phản biện hoặc chấp nhận đăng và $k(t_p) = \frac{t_p - t_0}{t_c - t_0}$, với t_0 = thời gian đầu tiên mà hai tác giả này đã cộng tác - 1, t_c là thời gian hiện tại.

Trong [6], nhóm nghiên cứu đã đề xuất một độ đo liên kết dựa trên nội dung tóm tắt của bài báo ($S_{PLC}(u,v)$). Để xác định mức độ tương đồng giữa hai tác giả, có thể kết hợp mức độ tương đồng giữa hai tập bài báo được công bố bởi hai tác giả *u, v* ($S(P_u, P_v)$ có thể xem như là mức độ tương đồng về lĩnh vực nghiên cứu) với mức độ tương tự giữa hai tập bài báo được viết chung bởi hai tác giả (*u, z*) và (*v, z*) ($S(P_{uz}, P_{vz})$) dựa trên ý tưởng của độ đo liên kết trọng số theo láng giềng chung (S_{CN}^{np}).

$$S_{PLC(u,v)} = \frac{1}{e^{1-S(P_u,P_v)}} \times \frac{1}{|\Gamma(u) \cap \Gamma(v)|} \times \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{e^{1-S(P_{uz},P_{vz})}} \quad (6)$$

Trong đó,

$$S(P_u, P_v) = \frac{\overline{x_u} . \overline{x_v}}{\|\overline{x_u}\| \times \|\overline{x_v}\|} \quad (7)$$

$$\overline{x_u}(j) = \frac{1}{m} \sum_{i=1}^{m} x_i^u(j),\ \ j = \overline{1{:}K} \quad (8)$$

$$S(P_{uz}, P_{vz}) = \frac{\overline{x_{uz}} . \overline{x_{vz}}}{\|\overline{x_{uz}}\| \times \|\overline{x_{vz}}\|} \quad (9)$$

$$\overline{x_{uz}}(j) = \frac{1}{k} \sum_{i=1}^{k} x_i^{uz}(j),\ \ j = \overline{1{:}K} \quad (10)$$

$X_u = \{x_1^u, x_2^u, ...x_m^u\}$, $X_v = \{x_1^v, x_2^v, ...x_n^v\}$, $X_{uz} = \{x_1^{uz}, x_2^{uz}, ...x_k^{uz}\}$ lần lượt là tập các véc-tơ trong không gian *K* chiều, biểu diễn các bài báo trong P_u, P_v và P_{vz} tương ứng; $\overline{x_u}$ là véc-tơ trung bình từ tập các bài báo của tác giả *u; m, n* lần lượt là số lượng bài báo được công bố bởi tác giả *u, v; k, q* lần lượt là số bài báo được viết chung bởi tác giả *u* và *z*, và *v* và *z*.

Để đánh giá sự hiệu quả của bài toán khuyến nghị cộng tác, có thể sử dụng tiêu chí đánh giá độ bao phủ (Recall) và F1-measure.

3. Hệ thống khuyến nghị cộng tác dựa trên phân cụm bán giám sát mờ

Hệ thống khuyến nghị cộng tác trong mạng hợp tác khoa học cần lựa chọn ra một tập các tác giả mà chưa từng cộng tác với một tác giả nào đó trong quá khứ có tiềm năng năng cộng tác với họ trong tương lai. Trên thực tế, với một tác giả bất kỳ trong mạng hợp tác khoa học thì số lượng tác giả mà chưa từng có cộng tác với tác giả đó là rất nhiều bởi đồ thị biểu diễn mạng hợp tác khoa học là rất thưa. Do vậy, để hạn chế được tập các tác giả ứng cử nghiên chỉ xét những cặp tác giả có ít nhất một láng giềng chung. Chi tiết hệ thống khuyến nghị cộng tác mới dựa trên phân cụm bán giám sát mờ (SSSFC[19]) được trình bày trong Hình 2.

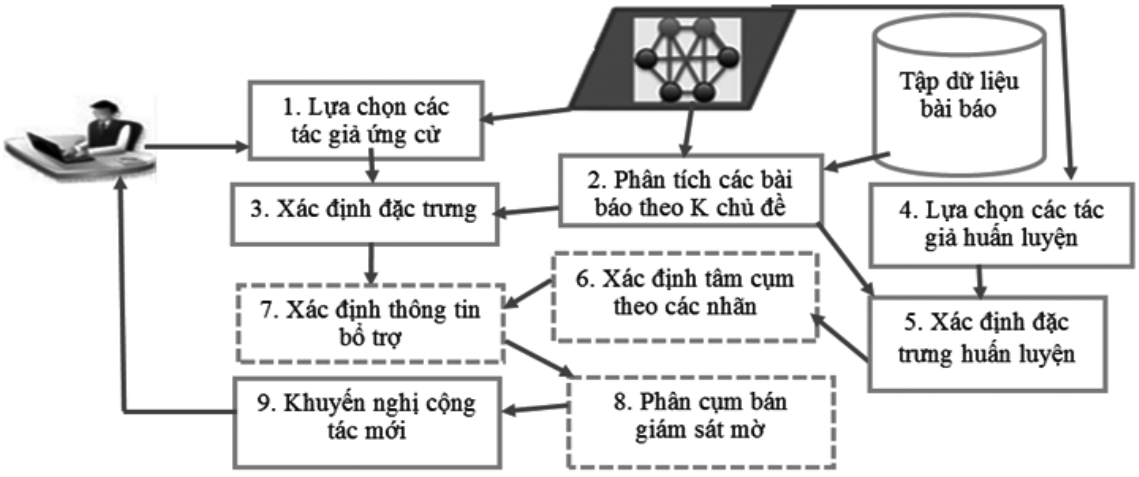
Sự khác biệt chính đối với khuyến nghị cộng tác dựa trên phân cụm bán giám sát mờ so với dựa trên phân lớp đó là sử dụng dữ liệu tập huấn luyện để điều hướng quá trình phân cụm (các bước 6, 7, 8), bằng việc xác định tâm cụm khởi tạo từ tập huấn luyện. Nhằm tạo ra sự hiệu quả phân cụm và nâng cao chất lượng khuyến nghị cộng tác hơn so với khuyến nghị cộng tác dựa trên phân lớp thường hay nhạy cảm với sự mất cân bằng nhãn trong tập huấn luyện.

Bước 6: Từ mỗi loại nhãn của tập dữ liệu huấn luyện, xác định tâm các cụm cho từng nhãn thông qua véc-tơ trung bình chung của các véc-tơ mang nhãn tương ứng trong tập huấn luyện. Các tâm cụm được xác định trong quá trình huấn luyện sẽ kết hợp với dữ liệu kiểm tra để xác định ma trận độ thuộc bổ trợ và ma trận này là khoảng cách Euclid từ các cặp tác giả đến tâm cụm của nhãn trên tổng số khoảng cách Euclid từ các cặp tác giả đó đến tâm các cụm của nhãn.

Bước 7: Xác định thông tin bổ trợ, cụ thể ở đây là xác định ma trận độ thuộc dựa trên phương pháp phân cụm mờ (FCM [18]).

Cụ thể, dựa trên tập dữ liệu kiểm tra, sử dụng phân cụm mờ (FCM) với tâm cụm khởi tạo được lấy từ Bước 6. Từ đó, sẽ xác định được ma trận độ thuộc bổ trợ và sử dụng trong phân cụm bán giám sát SSSFC [19] trong Bước 7.

Bước 8: Thuật toán phân cụm bán giám sát chuẩn SSSFC [19] với thông tin bổ trợ được xác định ở bước 7 được thực hiện với tập đặc trưng xác định trong bước 3 với số cụm bằng 2. Khi đó, phân cụm SSSFC xác định được ma trận độ thuộc của các cặp tác giả vào các cụm.



Hình 2. *Hệ thống khuyến nghị cộng tác mới dựa trên phân cụm bán giám sát mờ*

4. Kết quả thực nghiệm

Để so sánh hệ thống khuyến nghị cộng tác mới dựa trên SSSFC với hệ thống dựa trên phân lớp. Nghiên cứu sẽ tiến hành thực nghiệm trên mạng hợp tác khoa học được xây dựng dựa trên tập các bài báo được công bố trên tạp chí BJ (Biophysical Journal) từ năm 2006 đến 2017 và trên môi trường Matlab. Để kiểm chứng hệ thống khuyến nghị cộng tác mới, trong phần này chúng tôi sẽ chia dữ liệu các bài báo thành bốn tập bài báo khác nhau (D1, D2, D3 và D4) ứng với các khoảng thời gian 8 năm liên tiếp sau:

- D1: với khoảng thời gian T1 từ năm 2006 đến năm 2013,
- D2: với khoảng thời gian T2 từ năm 2007 đến năm 2014,
- D3: với khoảng thời gian T3 từ năm 2008 đến năm 2015 và
- D4: với khoảng thời gian T4 từ năm 2009 đến năm 2016.

Trong mỗi tập dữ liệu bài báo Dk (k = 1, 2, 3, 4) sử dụng tập các bài báo xuất hiện trong 6 năm đầu để xây dựng mạng hợp tác khoa học và sử dụng hai năm cuối để gán nhãn cho các cặp tác giả ứng cử đã công bố bài báo trong 6 năm đầu. Tập huấn luyện và kiểm tra được xây dựng theo cách sau ứng với mỗi tập Dk.

Bảng 1. *Thống kê các tập dữ liệu*

Tập dữ liệu	Khoảng thời gian	Số bài báo trong 6 năm đầu tiên	Số tác giả dùng để huấn luyện	Số tác giả dùng để kiểm tra
D1	2006 - 2013	2254	43	28
D2	2007 - 2014	1530	37	43
D3	2008 - 2015	1345	40	18
D4	2009 - 2016	1204	18	12

Bảng 2 liệt kê các độ đo liên kết trọng số và độ đo liên kết mở rộng sử dụng để thực nghiệm khuyến nghị cộng tác mới.

Bảng 2. *Tập các đặc trưng trong thực nghiệm khuyến nghị cộng tác mới*

STT	Tên tổ hợp độ đo liên kết	Các độ đo liên kết được sử dụng làm đặc trưng phân cụm
1	Weight1	$S_{CN}^{np}, S_{AA}^{np}, S_{JC}^{np}$
2	Weight2	$S_{CN}^{na}, S_{AA}^{na}, S_{JC}^{na}$
3	Weight3	$S_{CN}^{pt}, S_{AA}^{pt}, S_{JC}^{pt}$
4	Weight1_P_LDAcosin	$S_{CN}^{np}, S_{AA}^{np}, S_{JC}^{np}, S_{PLC}$
5	Weight2_P_LDAcosin	$S_{CN}^{na}, S_{AA}^{na}, S_{JC}^{na}, S_{PLC}$
6	Weight3_P_LDAcosin	$S_{CN}^{pt}, S_{AA}^{pt}, S_{JC}^{pt}, S_{PLC}$

Đối với ba tổ hợp đặc trưng Weight1_P_LDAcosin, Weight2_P_LDAcosin và Weight1_P_LDAcosin để xác định được số lượng chủ đề tối ưu, nghiên cứu đã tiến hành chạy thực nghiệm lần lượt với số lượng chủ đề trong tập {5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}.

a. Kết quả thực nghiệm hệ thống khuyến nghị cộng tác mới dựa trên phân lớp

Bảng 3. *Số chủ đề tối ưu ứng với các tổ hợp đặc trưng trên các bộ dữ liệu*

Tập dữ liệu	Weight1_ P_LDAcosin	Weight2_ P_LDAcosin	Weight3_ P_LDAcosin
D1	80	80	5
D2	10	40	40
D3	50	100	40
D4	50	50	40

Bảng 3 cho biết số chủ đề tối ưu ứng với từng tổ hợp đặc trưng trong mỗi tập dữ liệu.

Bảng 4. *Giá trị chỉ số Recall trung bình ứng với các tổ hợp đặc trưng trên các bộ dữ liệu*

Tập dữ liệu	Weight1	Weight1_ P_LDAcosin	Weight2	Weight2_ P_LDAcosin	Weight3	Weight3_ P_LDAcosin
D1	0.3571	**0.5714**	0.3571	**0.4464**	0.4464	**0.4464**
D2	0.6337	**0.6337**	0.4205	**0.5310**	0.3140	**0.3721**
D3	0.5926	**0.6481**	0.3611	**0.4722**	0.6667	**0.6667**
D4	0.3194	**0.4444**	0.6944	0.6806	0.4444	**0.4444**
TBC	0.4757	**0.5744**	0.4583	**0.5326**	0.4679	**0.4824**

Bảng 5. *Giá trị chỉ số F1-measure trung bình ứng với các tổ hợp đặc trưng trên các bộ dữ liệu*

Tập dữ liệu	Weight1	Weight1_ P_LDAcosin	Weight2	Weight2_ P_LDAcosin	Weight3	Weight3_ P_LDAcosin
D1	0.2418	**0.3954**	0.2435	**0.3054**	0.3109	**0.3466**
D2	0.2883	**0.3050**	0.2434	**0.3179**	0.1997	**0.2297**
D3	0.3172	**0.3529**	0.1920	**0.2549**	0.3211	0.2836
D4	0.2056	**0.2972**	0.4278	0.3948	0.2671	**0.3087**
TBC	0.2632	**0.3376**	0.2767	**0.3183**	0.2747	**0.2922**

Đối với khuyến nghị cộng tác mới dựa trên phân lớp, quan sát Bảng 4 và 5, dễ nhận thấy hầu hết giá trị trung bình các chỉ số đánh giá Recall và F1-measure tương ứng với các tổ hợp đặc trưng Weight#_P_LDAcosin so với Weight# đều cải thiện đáng kể trong tất cả các tập dữ liệu D1 – D4 ngoại trừ tổ hợp đặc trưng Weight2_P_LDAcosin và Weight3_P_LDAcosin không cải thiện hơn so với Weight2 và Weight3 lần lượt trong tập dữ liệu D4 và D3. Tuy nhiên, nếu xét trung bình chung (TBC) trên bốn tập dữ liệu (D1-D4) thì các giá trị của cả ba chỉ số đánh giá ứng với tổ hợp đặc trưng Weight#_P_LDAcosin đều cao hơn so với Weight#.

b. Kết quả thực nghiệm hệ thống khuyến nghị cộng tác mới dựa trên phân cụm bán giám sát mờ (SSSFC)

Bảng 6 cho biết số chủ đề tối ưu ứng với từng tổ hợp đặc trưng (Weight#_P_LDAcosin) trong mỗi tập dữ liệu.

Bảng 6. *Số chủ đề tối ưu ứng với các tổ hợp đặc trưng trên các bộ dữ liệu*

Tập dữ liệu	Weight1_ P_LDAcosin	Weight2_ P_LDAcosin	Weight3_ P_LDAcosin
D1	5	50	30
D2	90	80	10
D3	80	20	90
D4	100	70	30

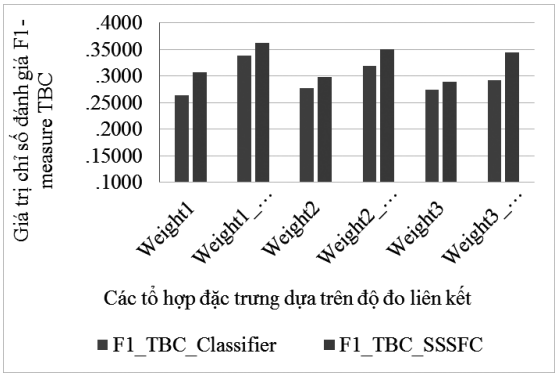
Bảng 7. *Giá trị chỉ số Recall trung bình ứng với các tổ hợp đặc trưng trên các bộ dữ liệu*

Tập dữ liệu	Weight1	Weight1_ P_LDAcosin	Weight2	Weight2_ P_LDAcosin	Weight3	Weight3_ P_LDAcosin
D1	0.5714	**0.6607**	0.5714	**0.6071**	0.5000	**0.5714**
D2	0.5659	**0.5833**	0.4709	**0.5291**	0.4845	**0.5911**
D3	0.5648	**0.6574**	0.6019	**0.6574**	0.4630	**0.6574**
D4	0.6528	**0.8194**	0.5278	**0.6944**	0.6944	**0.8056**
TBC	0.5887	**0.6802**	0.5430	**0.6220**	0.5355	**0.6564**

Bảng 8. *Giá trị chỉ số F1-measure trung bình ứng với các tổ hợp đặc trưng trên các bộ dữ liệu*

Tập dữ liệu	Weight1	Weight1_ P_LDAcosin	Weight2	Weight2_ P_LDAcosin	Weight3	Weight3_ P_LDAcosin
D1	0.5714	**0.6607**	0.5714	**0.6071**	0.5000	**0.5714**
D2	0.5659	**0.5833**	0.4709	**0.5291**	0.4845	**0.5911**
D3	0.5648	**0.6574**	0.6019	**0.6574**	0.4630	**0.6574**
D4	0.6528	**0.8194**	0.5278	**0.6944**	0.6944	**0.8056**
TBC	0.5887	**0.6802**	0.5430	**0.6220**	0.5355	**0.6564**

Quan sát các Bảng 7 và 8, dễ nhận thấy giá trị trung bình các chỉ số đánh giá Recall và F1-measure tương ứng với các tổ hợp đặc trưng Weight#_P_LDAcosin so với Weight# đều cải thiện đáng kể trong tất cả các tập dữ liệu D1 – D4.



Hình 3. *So sánh giá trị chỉ số đánh giá F1-measure TBC giữa khuyến nghị cộng tác mới dựa trên phân lớp và SSSFC*

Hình 3 cho biết giá trị F1-measure theo trung bình chung trên bốn tập dữ liệu (D1-D4) ứng với hai hệ thống khuyến nghị cộng tác mới dựa trên phân lớp và phân cụm bán giám sát mờ. Dễ nhận thấy, giá trị của chỉ số đánh giá trong tất cả các tổ hợp đặc trưng, ứng với khuyến nghị cộng tác mới dựa trên phân cụm bán giám sát mờ đề cao hơn so với dựa trên phân lớp. Điều này chứng tỏ việc áp dụng phương pháp phân cụm bán giám sát mờ vào bài toán khuyến nghị cộng tác mới hiệu quả hơn khi sử dụng với phương pháp phân lớp.

5. Kết luận

Trong bài báo này, chúng tôi đã tiến hành thực nghiệm hệ thống khuyến nghị cộng tác mới dựa trên phân cụm bám giám sát mờ và có so sánh với hệ thống khuyến nghị cộng tác dựa trên phân lớp. Thông qua kết quả thực nghiệm, nhận thấy độ đo liên kết mở rộng dựa trên nội dung tóm tắt bài báo (SPLC) [6] khi kết hợp với các độ đo liên kết trọng số đều cho kết quả cải thiện đáng kể so với các tổ hợp chỉ bao gồm các độ đo liên kết trọng số trong 4 tập dữ liệu D1 - D4 đối với cải hai hệ thống khuyến nghị cộng tác mới.

Ngoài ra, việc áp dụng hệ thống phân cụm bán giám sát mờ vào khuyến nghị cộng tác mới cho hiệu quả khuyến nghị tốt hơn so với dựa trên phân lớp. Điều này cho thấy, phương pháp phân cụm bán giám mờ có tiềm năng áp dụng cho hệ thống khuyến nghị cộng tác.

Tài liệu tham khảo

[1]. Lopes G. R., Moro M. M., Wives L. K. and De Oliveira J. P. M., Collaboration recommendation on academic social networks. *International Conference on Conceptual Modeling*, 2010.

[2]. Hasan M. Al, Chaoji V., Salem S. and Zaki M., Link prediction using supervised learning. *SDM06: workshop on link analysis, counter-terrorism and security*, 2006.

[3]. Chen B., Li F., Chen S., Hu R.and Chen L., Link prediction based on non-negative matrix factorization. *PloS one*, p. e0182968, 2017, **vol. 12, no. 8**.

[4]. Y. Guisheng, Y. Wansi and D. Yuxin, "A new link prediction algorithm: node link strength algorithm," in *Computer Applications and Communications (SCAC)*, *2014 IEEE Symposium*, 2014, pp. 5-9.

[5]. Gupta S., Pandey S.and. Shukla K. K, Comparison analysis of link prediction algorithms in social network. *International Journal of Computer Applications*, 2015, **vol. 111, no. 16.**

[6]. Chuan P. M., Ali M., Khang T. D., Huong L. T. and Dey N. Link prediction in co-authorship networks based on hybrid content similarity metric, Applied Intelligence, 2018, **48(8)**, 2470-2486.

[7]. J. S. Breese, D. Heckerman and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence,* 1998.
[8]. C. Basu, H. Hirsh and W. Cohen, "Recommendation as classification: Using social and content-based information in recommendation," in *Aaai/iaai*, 1998, pp. 714-720.
[9]. T. Bogers and A. Van den Bosch, "Recommending scientific articles using citeulike," in *In Proceedings of the 2008 ACM conference on Recommender systems,* 2008.
[10]. R. Burke, "Hybrid recommender systems: Survey and experiments," *User modeling and user-adapted interaction*, 2002, **vol. 12, no. 4**, pp. 331-370.
[11]. R. D. Burke (2007) "Hybrid web recommender systems," in *P. Brusilovsky, A. Kobsa, & W. Nejdl, editors, The Adaptive Web, Methods and Strategies of Web Personalization, volume 4321 of Lecture Notes in Computer Science*, *Springer*, 2007, pp. 377-408.
[12]. C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining,* 2011, August, pp. 448-456, ACM.
[13]. J.C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms," Plenum, New York, 1981.
[14]. E. Yasunori, H. Yukihiro, Y. Makito and M. Sadaaki, "On semi-supervised fuzzy c-means clustering," in *Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on,* IEEE, 2009, pp. 1119-1124.
[15]. T. Murata and S. Moriyasu, "Link prediction of social networks based on weighted proximity measures," in *the IEEE/WIC/ACM international conference on In Web Intelligence*, 2007.
[16]. H. R. De Sá and R. B. Prudêncio, "Supervised link prediction in weighted networks," in Neural Networks (IJCNN), *The 2011 International Joint Conference on*, IEEE, 2011, pp. 2281-2288.
[17]. I. Günes, S. Gündüz-Öüdücü and Z. Çataltepe, "Link prediction using time series of neighborhood-based node similarity scores," Data Mining and Knowledge Discovery, 2016, **vol. 30, no. 1**, pp. 147-180.
[18]. F. Xia, Z. Chen, W. Wang, J. Li and L. T. Yang, "Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors," *IEEE Transactions on Emerging Topics in Computing*, 2014, **vol. 2, no. 3**, pp. 364-375.

COLLABORATIVE RECOMMENDATION SYSTEMS BASED ON SEMI-SUPERVISED FUZZY CLUSTERING METHOD AND APPLING IN CO-AUTHOR NETWORKS

Abstract:

The collaborative recommendation problem among researchers is currently being emphasized. Most of the existing reseaches deal with collaborative recommendation problems based on collaborative and non-collaborative binary classification. However, due to the sparseness of the co-authors network, the data set used for training is often subject to imbalance leading to low classification efficiency. This paper proposes a collaboration recommendation system based on a fuzzy semi-supervised clustering to overcome the disadvantages of binary clustering for sparse and unbalanced data. Experimental results for the proposed collaborative recommendation system were empirically tested on a practical data set, suggesting that in most cases a more effective fuzzy semi-observer clustering collaboration recommendations system would be more effective compared with the binary classification system.

Keywords: *collaborative recommendation, classification, fuzzy semi-supervised clustering.*