Đề xuất mô hình khuyến nghị cộng tác mới cho mạng đồng tác giả dựa trên chỉ số cộng tác và tương quan

Phạm Minh Chuẩn[1,2*], Lê Hoàng Sơn[3], Trần Đình Khang[2], Lê Thanh Hương[2]

[1]Trường Đại học Sư phạm Kỹ thuật Hưng Yên
[2]Trường Đại học Bách khoa Hà Nội
[3]Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội

Ngày nhận bài 11/9/2017; ngày chuyển phản biện 14/9/2017; ngày nhận phản biện 16/10/2017; ngày chấp nhận đăng 18/10/2017

Tóm tắt:

Trong bài báo này, các tác giả đề xuất một mô hình khuyến nghị cộng tác mới trên mạng đồng tác giả nhằm hỗ trợ các nhà nghiên cứu trong việc xác định các mối cộng tác đã có và tăng cường quan hệ hợp tác trong tương lai. Mô hình đề xuất dựa trên ý tưởng về cải tiến hệ tư vấn trong mạng đồng tác giả với hai chỉ số cộng tác và tương quan nhằm cải tiến hiệu năng khuyến nghị. Chỉ số cộng tác được xây dựng dựa trên liên kết giữa các tác giả và số bài báo đã viết trong quá khứ. Chỉ số tương quan được xác định từ việc phân tích chủ đề nội dung các bài báo thông qua phương pháp phân tích chủ đề LDA. Hệ sẽ khuyến nghị khả năng liên kết dựa trên ngưỡng đối với từng chỉ số tương quan và cộng tác. Hệ thống đề xuất được thử nghiệm và đánh giá trên mạng đồng tác giả được xây dựng từ tập các bài báo được đăng trên tạp chí "Biophysical Journal" từ năm 2006 đến 2017.

Từ khóa: *Chỉ số cộng tác, chỉ số tương quan, hệ thống khuyến nghị, mạng cộng tác, phân tích chủ đề.*

Chỉ số phân loại: *1.2*

Mở đầu

Ngày nay, với sự phát triển của mạng internet đã giúp mọi người trên toàn thế giới dễ dàng kết nối thông qua các mạng xã hội như Facebook, Twitter..., đồng thời cũng làm bùng nổ thông tin được lưu trữ trên mạng, dẫn đến người dùng rất khó khăn trong việc tìm kiếm, lựa chọn thông tin phù hợp [1]. Hệ khuyến nghị hay hệ tư vấn (Recommender Systems) [2] là một giải pháp trợ giúp người dùng ra quyết định lựa chọn và tìm kiếm thông tin phù hợp trong thời gian ngắn. Hệ tư vấn có ý nghĩa đặc biệt quan trọng trong bối cảnh cách mạng công nghiệp 4.0 khi nhiều nhà (khoa học, doanh nghiệp, chính phủ, người dân) có thể kết nối với nhau thông qua một cổng thông tin. Khi đó hệ tư vấn đóng vai trò cầu nối, giúp gợi ý cho doanh nghiệp về những công nghệ lõi phù hợp với đặc thù phát triển kinh tế do các nhà khoa học thiết kế, người dân cũng có thể tìm thấy các sáng chế, ý tưởng dựa trên việc đánh giá các sản phẩm phù hợp với nhu cầu thông qua cơ chế khuyến nghị trong hệ tư vấn. Trên mạng xã hội (chẳng hạn trên mạng Facebook), hệ tư vấn được thể hiện rõ ràng thông qua việc khuyến nghị người dùng trong việc xác định những người bạn cũ hoặc kết nối với những người bạn mới một cách nhanh chóng và hiệu quả. Trong tư vấn bán hàng trực tuyến (như trên Amazone), hệ tư vấn giúp xác định các mặt hàng phù hợp đối với sở thích người dùng. Còn rất nhiều ví dụ nữa minh họa tiềm năng ứng dụng của hệ tư vấn.

Trong bài báo này, chúng tôi quan tâm đến một ứng dụng cụ thể của hệ tư vấn trong việc tìm ra được nhóm hoặc những nhà khoa học phù hợp với mỗi người nghiên cứu (hay còn gọi là bài toán mạng đồng tác giả). Mạng đồng tác giả giúp ích rất nhiều trong công việc, hợp tác cũng như công bố kết quả trên những tạp chí hoặc hội thảo uy tín của các nhà khoa học. Đây là bài toán được quan tâm nhiều trong nước và trên thế giới trong cộng đồng khoa học nói riêng và giúp hỗ trợ chuyển giao các tri thức cho cộng đồng nói chung. Ý nghĩa của hệ thống khuyến nghị được đề xuất nhằm tạo nền tảng cho các hệ thống hỗ trợ ra quyết định, giúp các nhà khoa học có thể dễ dàng tìm kiếm những người cộng tác trong khoa học có nhiều điểm tương đồng về các hướng nghiên cứu và tạo ra các mối cộng tác mới về khoa học.

Trong những năm gần đây, ngày càng có nhiều nhóm tiến hành nghiên cứu về việc tăng cường chất lượng dự báo trong hệ khuyến nghị cho mạng đồng tác giả, có thể kể đến những nghiên cứu tiêu biểu như của Yu và cs (2014), Makarov và cs (2016)... [1, 3-7]. Những nghiên cứu này đặt nền móng cho việc phát triển các hệ khuyến nghị cộng tác trên mạng đồng tác giả với mục đích chính là giúp các nhà nghiên cứu tăng cường cộng tác đã có và thiết lập những mối cộng tác với những nhà nghiên cứu mà chưa từng có mối cộng tác. Lopes và cs (2010) [2] đã đề xuất một mô hình khuyến nghị cộng tác với hai chỉ số cộng tác và tương quan để đưa ra khuyến nghị cho người dùng. Lee và cs (2011) [8] nghiên cứu mối tương đồng giữa các tác giả thông qua thông tin của các bài báo được công bố bởi họ như từ khóa đại diện và vị trí của tác giả trong bài báo. Phương pháp khuyến nghị dựa trên nội dung và lọc cộng tác dựa trên mối quan hệ trong mạng đồng tác giả đã được đề xuất kèm theo chiến lược lai ghép trong hệ khuyến nghị. Xia và cs (2014) [7] đề xuất phương pháp MVCWalker trong việc khuyến nghị những

*Tác giả liên hệ: Tel: 0983081120; Email: chuanpm@gmail.com

A approach for a new collaboration recommendation in co-authorship networks based on Global Cooperation and Global Correlation

Minh Chuan Pham[1,2*], Hoang Son Le[3], Dinh Khang Tran[2], Thanh Huong Le[2]

[1]Hung Yen Univertity of Technology and Education
[2]Hanoi University of Science and Technology
[3]VNU University of Science

Received 11 Septemrber 2017; accepted 18 October 2017

Abstract:

In this paper, we propose a new collaboration recommendation in co-authorship networks to assist researchers in specifying existing research collaborations and strengthening them in the future. It is based on Global Cooperation and Global Correlation to further improve the recommendation performance. Global Cooperation relies on the connection between authors and their common research works. Global Correlation is determined through a topic modeling method, namely Latent Dirichlet Allocation (LDA). The proposed system determines the outcome based on specified thresholds for the Global Cooperation and Global Correlation. It is experimentally validated on a dataset of co-authorship networks published in the "Biophysical Journal" from 2006 to 2017.

Keywords: ***Collaborative networks, global cooperation, global correlation, recommendation system, topic modeling.***

Classification number: ***1.2***

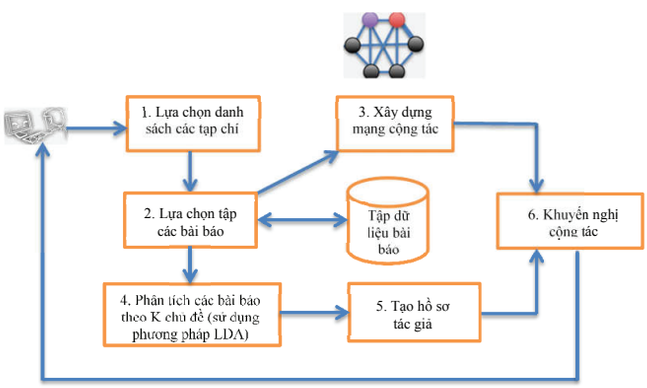
người cộng tác hữu hiệu nhất thông qua trọng số liên kết giữa các tác giả theo 3 yếu tố: Vị trí của tác giả trong bài báo, thời gian cộng tác gần nhất và số lần cộng tác. Ngoài các nghiên cứu tiêu biểu trên, còn nhiều nghiên cứu khác, tuy nhiên phần lớn ý tưởng chung là đề xuất mô hình hoặc phương pháp dự báo nhằm làm tăng cường các mối cộng tác đã có hoặc tạo ra các mối cộng tác mới phù hợp nhất.

Trong bài báo này, chúng tôi đề xuất một mô hình khuyến nghị cộng tác mới trên mạng đồng tác giả dựa trên chỉ số cộng tác và tương quan. Mô hình này là cải tiến của mô hình trong nghiên cứu của Lopes và cs (2010) [2], cụ thể: 1) Đề xuất cách tính chỉ số cộng tác mới không những dựa trên số bài báo được viết chung bởi hai tác giả mà còn xem xét đến số lượng tác giả trong mỗi bài báo. Điều này xuất phát từ đề xuất của Newman (2001) [5] trong việc tính trọng số liên kết giữa hai tác giả trong một bài báo; 2) Đưa ra phương pháp để xác định hồ sơ của mỗi tác giả dựa trên các bài báo mà họ đã viết thông qua phương pháp phân tích chủ đề, ví dụ Latent Dirichlet Allocation (LDA) [9] được sử dụng rất nhiều trong các lĩnh vực khai phá dữ liệu, phân lớp văn bản và trích rút thông tin…

Trong mô hình khuyến nghị mới, chỉ số cộng tác được xây dựng dựa trên liên kết giữa các tác giả và số bài báo đã viết trong quá khứ. Chỉ số tương quan được xác định từ việc phân tích chủ đề nội dung các bài báo thông qua phương pháp LDA. Hệ sẽ khuyến nghị khả năng liên kết dựa trên ngưỡng đối từng chỉ số tương quan và cộng tác. Hệ thống đề xuất được thử nghiệm và đánh giá trên mạng đồng tác giả được xây dựng từ tập các bài báo được đăng trên tạp chí "Biophysical Journal" từ năm 2006 đến 2017.

Đề xuất hệ thống khuyến nghị cộng tác trên mạng đồng tác giả

Trong mô hình khuyến nghị đề xuất, chúng tôi đưa ra công thức tính chỉ số cộng tác tổng thể (Global Cooperation) dựa trên loại trọng số liên kết [5]. Ngoài ra, đối với chỉ số tương quan tổng thể (Global Correlation) chúng tôi cũng đề xuất một cách xác định khác lấy ý tưởng từ Chuan và cs (2017) [3] áp dụng trong việc xây dựng các độ đo tương đồng dựa trên phương pháp LDA [9]. Mô hình tổng thể của hệ thống khuyến nghị cộng tác trên mạng đồng tác giả được thể hiện trong hình 1.



Hình 1. Mô hình tổng thể của hệ thống khuyến nghị cộng tác đề xuất.

Trong mô hình khuyến nghị cộng tác (hình 1), quá trình thực hiện sẽ diễn ra bởi 6 bước chính, gồm: 1) Lựa chọn danh sách các tạp chí để xây dựng mạng cộng tác thực hiện trong quá trình khuyến nghị; 2) Lựa chọn ra tập các bài báo trên các tạp chí đã chọn từ nguồn dữ liệu số lưu trữ thông tin của các bài báo trên mạng internet; 3) Xây dựng mạng cộng tác thông qua các tác giả được lựa chọn từ tập các bài báo nhận được trong bước 2, gồm liên kết giữa các tác giả viết chung bài, số bài báo viết chung và nội dung các bài báo; 4) Sử dụng phương pháp phân tích chủ đề LDA [9] để biểu diễn mỗi bài báo dưới dạng một véc tơ K chiều; 5) Tạo hồ

sơ cho các tác giả dựa trên công thức (8) (ở phần sau) dựa trên kết quả trong bước 4 để biểu diễn hồ sơ của mỗi tác giả là một véc tơ K chiều, được xác định bởi trung bình chung của tập véc tơ ứng với các bài báo mà mỗi tác giả đã viết; 6) Tính toán các chỉ số cộng tác theo công thức (2) và chỉ số tương quan theo công thức (7). Lựa chọn các tác giả dựa trên hai chỉ số cộng tác tổng thể và tương quan tổng thể để khuyến nghị.

Trong mô hình của Lopes và cs [2], hồ sơ của mỗi tác giả trong mạng cộng tác được sử dụng bởi hệ thống khuyến nghị được xây dựng dựa trên những thông tin có sẵn về các tác giả trong cơ sở dữ liệu bài báo và sự phân loại các bài báo của tác giả [10]. Bài báo này dựa trên một cách thức khác để xây dựng hồ sơ của các tác giả dựa trên phương pháp phân tích chủ đề [9].

Chỉ số cộng tác tổng thể dựa trên trọng số liên kết

Trong nghiên cứu của Lopes và cs (2010) [2], nhóm tác giả đã mô hình hóa một mạng xã hội (Social Network - SN) đối với quan hệ cộng tác a là một cặp: $SN_a = (N, E)$, trong đó N và E tương ứng là tập các đỉnh và tập các cạnh có hướng. Mỗi một cạnh $e \in E$ có dạng $<v_i, t, \omega, v_j>$, trong đó cạnh có hướng từ v_i đến v_j; t là ký hiệu kiểu cộng tác giữa v_i và v_j; ω là trọng số tác động trên quan hệ cộng tác t nhận giá trị số trong khoảng (0, 1). Khi đó, chỉ số cộng tác tổng thể ($\omega_{t\text{–}Ca}$) được tính theo công thức sau:

$$\omega_{t\_Ca(v_i \to v_j)} = \frac{|v_j co\_authorship|}{|v_i author|} \quad (1)$$

Trong đó: $\omega_{t\text{–}Ca\,(v_i \to v_j)}$ tương ứng là chỉ số cộng tác (trọng số từ $v_i \to v_j$ là khác với trọng số từ $v_j \to v_i$); $|v_j co\_authorship|$ là số lần mà tác giả v_j đã cộng tác viết báo với tác giả v_i; $|v_j author|$ tương ứng với tổng số bài báo mà tác giả v_i đã công bố.

Ngoài ra, nếu giá trị $\omega_{t\text{–}Ca\,(v_i \to v_j)}$ càng cao thì có nghĩa rằng mức độ liên quan (phù hợp) giữa v_j với v_i càng nhiều.

Trong bài báo này, ngoài xem xét số lượng bài báo mà hai tác giả đã từng cộng tác, chúng tôi còn dựa trên một loại trọng số được đề xuất trong nghiên cứu của Newman (2001) [5]. Ý nghĩa của loại trọng số này xuất phát từ một thực tế là mối liên kết giữa hai tác giả trong một bài báo phụ thuộc vào số lượng tác giả trong bài báo đó. Nghĩa là nếu số lượng tác giả trong một bài báo càng ít thì mức độ liên kết giữa các tác giả trong bài báo đó càng cao và ngược lại. Khi đó, chỉ số cộng tác tổng thể mà chúng tôi đề xuất được tính theo công thức sau:

$$\omega_{t\_Ca(v_i \to v_j)} = \frac{1}{2}\left(\frac{|v_j co\_authorship|}{|v_i author|} + \frac{\sum_{p \in P_{ij}} \frac{1}{n_p - 1}}{\sum_{f \in P_i} \frac{1}{n_f - 1}}\right) \quad (2)$$

Trong đó, P_{ij} là tập các bài báo được viết chung bởi cả hai tác giả v_i và v_j; P_i là tập các bài báo được viết bởi tác giả v_i.

Để minh họa cho công thức (2), chúng tôi xem xét một ví dụ giữa 3 tác giả u, v, z như sau: Giả sử tập bài báo của tác giả u là $P_u = \{p_1, p_2, p_3, p_4, p_5\}$ tương ứng với số lượng tác giả trong từng bài báo là {2, 2, 4, 3, 3} và tập các bài báo được viết chung bởi hai tác giả u, v là $P_{uv} = \{p_1, p_2, p_3\}$ và tập các bài báo được viết chung bởi hai tác giả u và z là $P_{uz} = \{p_2, p_3, p_4\}$. Khi đó, theo công thức (1) chúng ta sẽ tính được mức độ cộng tác tổng thể là $\omega_{t\text{–}Ca\,(u \to z)} = \omega_{t\text{–}Ca\,(v \to z)} = 3/5 = 0{,}6$, khi áp dụng công thức (2) thì mức độ cộng tác tổng thể do chúng tôi đề xuất giữa u và z; v và z sẽ được tính như biểu thức (3) và (4).

$$\omega_{t\_Ca(u \to z)} = 0.6/2 + \frac{1}{2} \times \frac{\frac{1}{2-1} + \frac{1}{2-1} + \frac{1}{4-1}}{\frac{1}{2-1} + \frac{1}{2-1} + \frac{1}{4-1} + \frac{1}{3-1} + \frac{1}{3-1}} = 0{,}3 + \frac{\frac{7}{6}}{\frac{10}{3}} = 0{,}3 + \frac{7}{20} = 0{,}65 \quad (3)$$

$$\omega_{t\_Ca(v \to z)} = 0{,}6/2 + \frac{1}{2} \times \frac{\frac{1}{2-1} + \frac{1}{4-1} + \frac{1}{3-1}}{\frac{1}{2-1} + \frac{1}{2-1} + \frac{1}{4-1} + \frac{1}{3-1} + \frac{1}{3-1}} = 0{,}3 + \frac{\frac{11}{12}}{\frac{10}{3}} = 0{,}3 + \frac{11}{40} = 0{,}575 \quad (4)$$

Kết quả tính được theo biểu thức (3) và (4) cho thấy, mức độ cộng tác giữa hai tác giả ngoài việc phụ thuộc vào số bài báo viết chung thì còn phụ thuộc vào số lượng tác giả trong mỗi bài báo mà hai tác giả đã viết chung. Nếu trong một bài báo, số lượng tác giả tham gia càng ít thì mức độ liên kết giữa các tác giả trong bài báo đó càng cao và ngược lại.

Chỉ số tương quan tổng thể dựa trên phân tích chủ đề LDA

Đối với khuyến nghị cộng tác, điều quan trọng nằm ở việc xác định được mối tương quan tổng thể giữa các tác giả. Mức độ tương quan tổng thể có thể được xác định thông qua mức độ khác biệt trên các lĩnh vực nghiên cứu. Trong nghiên cứu của Lopes và cs [2], nhóm tác giả đã đề xuất cách thức xác định mức độ tương quan tổng thể như công thức (5).

$$\text{global\_correlation}(v_i,v_j)=\frac{\sum_{k=1}^{n}\omega_{Ra}(v_i,x_k)\times\omega_{Ra}(v_j,x_k)}{\sqrt{\sum_{k=1}^{n}\left(\omega_{Ra}(v_i,x_k)\right)^2\times\sum_{k=1}^{n}\left(\omega_{Ra}(v_j,x_k)\right)^2}} \quad (5)$$

Trong đó, n là số lĩnh vực; $\omega_{Ra}(v_i,x_k)$ là trọng số ứng với lĩnh vực nghiên cứu x_k mà tác giả v_i đóng góp vào so với toàn bộ bài báo của tác giả v_i và được tính theo công thức (6).

$$\omega_{Ra}(v_i,x)=\frac{\left|v_i author_{research\_area\_x}\right|}{\left|v_i author\right|} \quad (6)$$

Với $\left|v_i author_{research\_area\_x}\right|$ là số bài báo mà tác giả v_i đăng trong lĩnh vực x.

Trong nghiên cứu của Lopes và cs (2010) [2], nhóm tác giả xác định lĩnh vực của một bài báo dựa trên một ontology được xây dựng sẵn. Điều này sẽ gặp khó khăn khi số lượng bài báo lớn, phân bố ở nhiều lĩnh vực khác nhau và việc xây dựng tập mẫu để huấn luyện cũng không dễ dàng. Trên thực tế có thể cùng một mảng nghiên cứu được phân vào các lĩnh vực khác nhau và một lĩnh vực nghiên cứu có thể được diễn đạt với các tên khác nhau. Để giải quyết vấn đề này, chúng tôi áp dụng phương pháp LDA [9]. LDA đã được áp dụng nhiều trong các lĩnh vực khai phá dữ liệu, phân lớp văn bản và trích rút thông tin... Chúng tôi sử dụng LDA để phân tích mỗi bài báo vào K chủ đề khác nhau, thông tin của mỗi bài báo được sử dụng để phân tích chủ đề bao gồm tên, các từ khóa và nội dung tóm tắt của bài báo với mong muốn xác định được lĩnh vực nghiên cứu của mỗi tác giả thông qua nội dung của các bài báo một cách chính xác nhất và có tính tương đồng cao về ngữ nghĩa thông qua phương pháp LDA.

Giả sử hai tác giả u, v có hai tập bài báo là $P_u = \{p_{u1},...,p_{um}\}$ và $P_v = \{p_{v1},...,p_{vn}\}$ (m, n nguyên dương), sau khi phân tích theo K chủ đề, chúng ta nhận được các véc tơ biểu diễn cho từng bài báo trong không gian K chiều như sau: $X_u = \{x_{u1},...,x_{um}\}$ và $X_v = \{x_{v1},...,x_{vn}\}$. Khi đó, mức độ tương quan tổng thể mà chúng tôi đề xuất được tính theo công thức (7).

$$\text{global\_correlation}(u,v)=\frac{\sum_{i=1}^{K}x_u(i)\times x_v(i)}{\sqrt{\sum_{i=1}^{K}\left(x_u(i)\right)^2\times\sum_{i=1}^{K}\left(x_v(i)\right)^2}} \quad (7)$$

Trong đó, x_u và x_v là hai véc tơ trung bình chung được tính thông qua tập hai véc tơ X_u và X_v như công thức (8).

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

Khuyến nghị cộng tác

Hệ thống khuyến nghị cộng tác nhằm đưa ra những gợi ý giúp những cặp người dùng (tác giả) có thể đưa ra quyết định xem có nên hay không nên tăng cường mối cộng tác nghiên cứu dựa trên hai chỉ số cộng tác tổng thể và tương quan tổng thể [2].

Trong mô hình khuyến nghị cộng tác đề xuất, chúng tôi tập trung tăng cường cho những cặp tác giả đã từng có liên kết có chỉ số cộng tác thấp (nhỏ hơn một giá trị ngưỡng) nhưng lại có chỉ số tương quan tổng thể cao (lớn hơn một ngưỡng nào đó). Giả sử chúng ta có tập các tác giả đã cộng tác với tác giả u là P_u, khi đó tập các tác giả cần khuyến nghị cộng tác tăng cường với tác giả u được xác định như biểu thức (9) dưới đây.

$$\text{RS(u)}=\left\{v\in P_u : \omega_{t\_Ca(u\to v)}\le\alpha \text{ and } \text{global\_correlation(u,v)}>\beta\right\} \quad (9)$$

Trong đó, các hằng số α, β được xác định thông qua thực nghiệm.

Minh họa hệ thống khuyến nghị cộng tác

Để minh họa cho mô hình khuyến nghị cộng tác đề xuất, chúng tôi tiến hành thử nghiệm một mạng đồng tác giả được xây dựng từ tập các bài báo được đăng trên tạp chí "Biophysical Journal" [11] từ năm 2006 đến 2017. Sở dĩ chúng tôi lựa chọn tập các bài báo đã đăng trên tạp chí này là do số lượng bài báo được công bố trong các năm từ 2006 đến 2017 đủ lớn và mỗi tác giả được mã hóa sẽ tránh việc nhầm lẫn về tên các tác giả vì có thể hai tác giả cùng tên viết tắt thì chưa chắc thuộc về cùng một tác giả. Tổng số bài báo thu được là 7.845, tổng số tác giả là 22.106 và tổng số liên kết là 72.186. Tuy nhiên, để đánh giá được mô hình khuyến nghị cộng tác đã đề xuất, chúng tôi xây dựng kịch bản thực nghiệm như sau:

(1) Xây dựng một đơn đồ thị vô hướng G, bao gồm 22.106 đỉnh (mỗi đỉnh là một tác giả), hai tác giả viết chung ít nhất một bài báo trong khoảng 2006-2017 thì sẽ có một cạnh nối giữa hai tác giả (đỉnh) đó.

(2) Xác định một thành phần liên thông lớn nhất của đồ thị G (tức một đồ thị con G' liên thông lớn nhất của G).

(3) Lựa chọn tập các tác giả chính là tập các đỉnh xuất hiện trong G'. Và chỉ lựa chọn tập các tác giả có số lượng bài báo từ năm 2006 đến 2017 lớn hơn 4, kết quả đã lựa chọn được 615 tác giả thỏa mãn điều kiện có số bài báo lớn hơn 4.

(4) Xây dựng tập dữ liệu để kiểm chứng mô hình khuyến nghị cộng tác. Gọi T1 = [2006-2011] tập các năm từ 2006 đến 2011; và T2 = [2012-2017]; chúng tôi sử dụng tập các

bài báo xuất hiện trong những năm T1 để xây dựng mô hình khuyến nghị cộng tác; tập bài báo xuất hiện trong những năm T2 để kiểm chứng mô hình khuyến nghị cộng tác. Để đánh giá mức độ chính xác cho mô hình khuyến nghị cộng tác, chúng tôi lựa chọn ra tập các tác giả thỏa mãn điều kiện trong bước 3 và có cộng tác với ít nhất 14 tác giả trong những năm T1, đồng thời trong những năm T2 lại tiếp tục có mối cộng tác với ít nhất 4 tác giả đã từng cộng tác trong những năm T1. Ví dụ, giả sử tác giả A trong những năm T1 có mối cộng tác với 14 tác giả {A1, A2, A3, A4, A5,…, A14}, và trong những năm T2 lại tiếp tục cộng tác với 4 tác giả {A1, A2, A4, A5} thì tác giả A sẽ được lựa chọn để đánh giá mức độ chính xác của mô hình khuyến nghị cộng tác. Với cách lựa chọn như vậy, chúng tôi đã lựa chọn được 65/615 tác giả dùng để đánh giá mô hình.

(5) Tổng số bài báo trong những năm T1 là 4.856, những bài báo này sẽ được sử dụng để phân tích theo K (trong bài báo này chúng tôi chọn $K = 50$) chủ đề (LDA [9]), làm cơ sở cho việc xây dựng hồ sơ tác giả và tính toán chỉ số tương quan tổng thể.

Chúng tôi sử dụng tiêu chí đánh giá độ bao phủ (Recall), độ chính xác (Precision) và F1-measure được xác định bởi các công thức từ (10) đến (12) để đánh giá mô hình khuyến nghị cộng tác.

$$\text{Recall} = \frac{|TP|}{|TP|+|FN|} \quad (10)$$

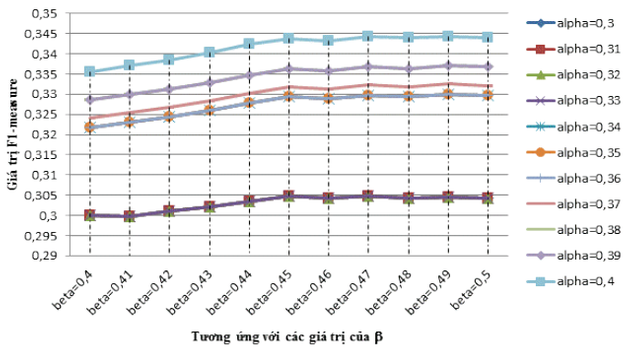
$$\text{Precision} = \frac{|TP|}{|TP|+|FP|} \quad (11)$$

$$\text{F1-measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}} \quad (12)$$

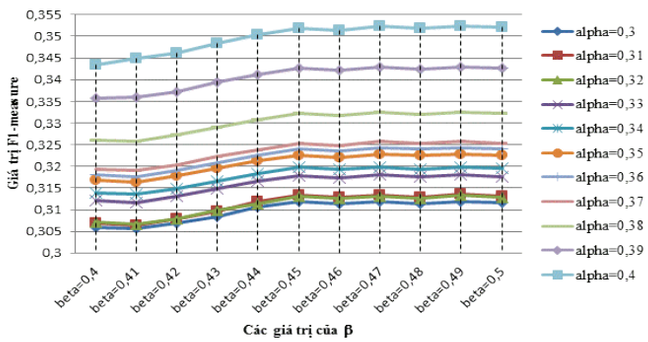
Trong đó, TP là tập tác giả được khuyến nghị cộng tác tăng cường là đúng; FN là tập các tác giả cộng tác tăng cường nhưng không được khuyến nghị cộng tác; FP là tập các tác giả được khuyến nghị cộng tác tăng cường nhưng không đúng.

Chúng tôi tiến hành thực nghiệm với hai trường hợp, gồm: 1) Sử dụng chỉ số cộng tác tổng thể do nhóm tác giả Lopes và cs [2] đã đề xuất trong biểu thức (1); 2) Sử dụng chỉ số cộng tác tổng thể do chúng tôi đề xuất trong biểu thức (2).

Đối với chỉ số tương quan tổng thể sử dụng theo công thức (7) do chúng tôi đề xuất. Do không có đủ dữ liệu mẫu để xây dựng một ontology về các lĩnh vực như nhóm tác giả Lopes và cs [2] đã thực hiện, nên chúng tôi không thể thực nghiệm theo chỉ số tương quan tổng thể trong biểu thức (5).

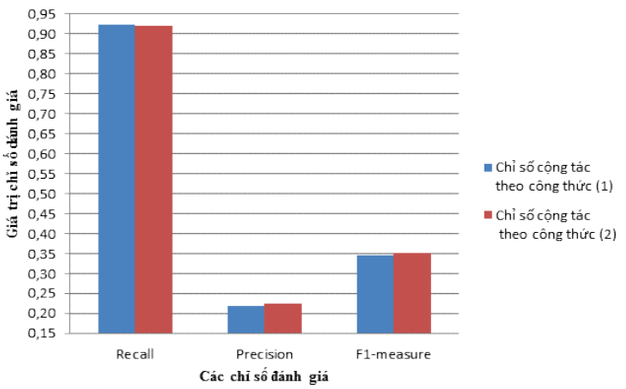


Hình 2. Kết quả trung bình chung của F1-measure đối với các giá trị ngưỡng α và β thực nghiệm trong trường hợp chỉ số cộng tác tính theo công thức (1).



Hình 3. Kết quả trung bình chung của F1-measure đối với các giá trị ngưỡng α và β thực nghiệm trong trường hợp chỉ số cộng tác tính theo công thức (2).

Để xác định được giá trị của α và β, chúng tôi đã tiến hành thực nghiệm với các giá trị khác nhau của α = {0,3, 0,31,…, 0,4} và β = {0,4, 0,41,…, 0,5}, kết quả F1-measure trung bình chung của 65 tác giả được biểu diễn trong hình 2 và hình 3 tương ứng với hai trường hợp thực nghiệm. Quan sát hình 2 và hình 3 chúng ta nhận thấy, giá trị trung bình chung F1-measure trong cả hai trường hợp đều đạt giá trị cao khi α = 0,4 và β ≥ 0,45, và F1-measure trung bình đạt lớn nhất khi α = 0,4 và β = 0,47, các giá trị α, β tối ưu nhận được ứng với giá trị F1-measure lớn nhất thông qua chạy thực nghiệm lần lượt với các giá trị của α, β trong khoảng (0, 1). Hình 4 cho biết giá trị trung bình của Recall, Precision và F1-measure trong cả hai trường hợp ứng với α, β tối ưu. Các giá trị trung bình của Recall, Precision và F1-measure trong trường hợp thứ 2 (sử dụng công thức (2) chỉ số cộng tác toàn phần do chúng tôi đề xuất) đều nhỉnh hơn so với trường hợp 1, cụ thể Precision đạt 0,225309 so với 0,218866 và F1-measure 0,352285 so với 0,344331, ngoài ra đối với giá trị Recall trong cả hai trường hợp đều khá cao và xấp xỉ nhau 0,922564 và 0,921026. Tuy kết quả cải thiện chưa nhiều nhưng có thể thấy việc áp dụng tính chỉ số cộng tác tổng thể theo công thức (2) cũng đã làm cho giá trị của chỉ số này mịn hơn (phân tách hơn), giúp việc lựa chọn các ứng cử viên khuyến nghị cộng tác thêm chính xác hơn.



Hình 4. Kết quả trung bình chung của Recall, Precision và F1-measure trong cả hai trường hợp.

Để so sánh một cách chi tiết hơn giá trị của chỉ số F1-measure đối với từng tác giả được thực hiện khuyến nghị cộng tác khi áp dụng chỉ số cộng tác theo công thức (1) và (2), chúng tôi đã liệt kê giá trị F1-measure của những tác giả có sự khác biệt khi áp dụng chỉ số cộng tác tổng thể theo công thức (1) và (2) trong bảng 1. Cụ thể, có 18/65 tác giả kết quả F1-measure nhận được có sự khác biệt, trong đó đối với chỉ số cộng tác theo công thức (2) có 15 tác giả nhận được giá trị F1-measure cao hơn so với công thức (1) và có 3 tác giả nhận được giá trị F1-measure thấp hơn so với công thức (1). Về tỷ lệ phần trăm cải thiện, đối với công thức (2) tỷ lệ cải thiện thấp nhất là 3,57% và cao nhất là 42,86%. Tuy nhiên, ba tác giả có giá trị F1-measure ứng với công thức (2) thấp hơn công thức (1) lần lượt chiếm tỷ lệ thấp hơn là 5, 25,93 và 18,18%.

Bảng 1. So sánh giá trị chỉ số F1-measure giữa chỉ số cộng tác theo công thức (1) và (2) ứng với từng tác giả được khuyến nghị.

STT	Id tác giả	Chỉ số cộng tác theo công thức (1)	Chỉ số cộng tác theo công thức (2)	Mức độ cải thiện của công thức (2) so với (1) (%)
1	2	0,421053	0,47619	13,10%
2	7	0,344828	0,357143	3,57%
3	16	0,333333	0,363636	9,09%
4	23	0,4	0,47619	19,05%
5	25	0,1	0,142857	42,86%
6	27	0,157895	0,15	-5,00%
7	31	0,705882	0,736842	4,39%
8	34	0,266667	0,352941	32,35%
9	36	0,285714	0,333333	16,67%
10	41	0,416667	0,434783	4,35%
11	43	0,296296	0,344828	16,38%
12	49	0,3	0,222222	-25,93%
13	50	0,105263	0,117647	11,76%
14	54	0,416667	0,434783	4,35%
15	55	0,428571	0,5	16,67%
16	56	0,428571	0,5	16,67%
17	63	0,333333	0,375	12,50%
18	64	0,333333	0,272727	-18,18%

Kết luận

Trong bài báo này, chúng tôi đã đề xuất một mô hình khuyến nghị cộng tác mới cho mạng đồng tác giả, nhằm trợ giúp các nhà nghiên cứu có cơ sở để quyết định xem mối cộng tác nào cần tăng cường hơn nữa. Mô hình mới dựa trên chỉ số cộng tác và chỉ số tương quan toàn phần nhằm tăng cường hiệu quả cho hệ thống khuyến nghị cộng tác. Kết quả thực nghiệm trên mạng đồng tác giả được xây dựng từ tập các bài báo được đăng trên tạp chí "Biophysical Journal" từ năm 2006 đến 2017 cho thấy, F1-measure đối với phương pháp đề xuất đạt giá trị cao khi $\alpha = 0,4$ và $\beta \geq 0,45$; F1-measure trung bình đạt lớn nhất khi $\alpha = 0,4$ và $\beta = 0,49$. Và giá trị trung bình chung F1-measure khi áp dụng chỉ số cộng tác với biểu thức (2) mà chúng tôi đề xuất là 0,35229 so với 0,34433 khi chỉ số cộng tác tính theo biểu thức (1).

Tuy nhiên, mô hình đề xuất còn nhiều tiềm năng để phát triển, chẳng hạn việc tính toán chỉ số tương quan tổng thể có thể xem xét thêm các yếu tố khác như: Lĩnh vực nghiên cứu đang quan tâm, địa chỉ... Do vậy, trong thời gian tới, chúng tôi sẽ tiếp tục nghiên cứu để đề xuất được mô hình khuyến nghị cộng tác hợp lý và hiệu quả hơn nữa để có thể áp dụng vào thực tế.

TÀI LIỆU THAM KHẢO

[1] Q. Yu, C. Long, Y. Lv, H. Shao, P. He, Z. Duan (2014), "Predicting co-author relationship in medical co-authorship networks", *PloS one*, **9(7)**, e101214.

[2] G.R. Lopes, M.M. Moro, L.K. Wives, J.P.M. De Oliveira (2010), "Collaboration recommendation on academic social networks", *International Conference on Conceptual Modeling*, pp.190-199.

[3] P.M. Chuan, C.N. Giap, L.H. Son, B. Chintan, T.D. Khang (2017), "Enhance Link Prediction in Online Social Networks Using Similarity Metrics, Sampling and Classification", *Proceedings of the 4th International Conference on Information System Design and Intelligent Applications (INDIA)* (Accepted).

[4] I. Makarov, O. Bulanov, L.E. Zhukov (2016), "Co-author Recommender System", *International Conference on Network Analysis*, pp.251-257.

[5] M.E. Newman (2001), "Scientific collaboration networks. I. Network construction and fundamental results", *Physical review E.*, **64(1)**, pp.16-31.

[6] M. Pavlov, R. Ichise (2007), "Finding experts by link prediction in co-authorship networks", *Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics*, pp.42-55.

[7] F. Xia, Z. Chen, W. Wang, J. Li, L.T. Yang (2014), "Mvcwalker: Random walk-based most valuable collaborators recommendation exploiting academic factors", *IEEE Transactions on Emerging Topics in Computing*, **2(3)**, pp.364-375.

[8] D.H. Lee, P. Brusilovsky, T. Schleyer (2011), "Recommending collaborators using social features and mesh terms", *Proceedings of the Association for Information Science and Technology*, pp.1-10.

[9] D.M. Blei (2012), "Probabilistic topic models", *Communications of the ACM*, **55(4)**, pp.77-84.

[10] S. Loh, D. Lichtnow, T. Borges, G. Piltcher, M. Freitas (2006), "Constructing domain ontologies for indexing texts and creating users' profiles", *In Work. on Ontologies and Metamodeling in Software and Data Engineering, Brazilian Symp. on Databases, UFSC, Florianópolis*, pp.72-82.

[11] https://www.journals.elsevier.com/biophysical-journal/, Accessed on 10/7/2017.