*Tạp chí Tin học và Điều khiển học, T.29, S.4 (2013), 338–350*

# ĐỀ XUẤT CÁC PHƯƠNG PHÁP TÍNH ĐỘ TƯƠNG TỰ ĐỈNH DỰA TRÊN XU HƯỚNG ỨNG DỤNG CHO BÀI TOÁN KHUYẾN NGHỊ CỘNG TÁC

HUỲNH NGỌC TÍN, HOÀNG VĂN KIẾM

*Trường Đại Học Công Nghệ Thông Tin, ĐHQG TP HCM*

**Tóm tắt.** Khuyến nghị cộng tác trong nghiên cứu khoa học là bài toán tự động liệt kê những người, nhóm nghiên cứu cộng tác tiềm năng ứng với đầu vào là một người, nhóm nghiên cứu nào đó. Đây là bài toán được quan tâm bởi những chuyên gia trong lĩnh vực này trong thời gian gần đây. Tiếp cận phổ biến đã và đang được các nghiên cứu này áp dụng là dựa vào phân tích mạng xã hội, cụ thể là phân tích các mối quan hệ trong mạng đồng tác giả, cũng như ảnh hưởng của nó đối với việc tìm ra những người, nhóm nghiên cứu cộng tác tiềm năng. Tuy nhiên, các phương pháp hiện nay đều chưa quan tâm đến thông tin về xu hướng cộng tác, một yếu tố quan trọng trong việc hình thành các mối quan hệ cộng tác mới. Bài báo này đề xuất ba phương pháp mới để tính tương tự đỉnh trong mạng đồng tác giả: (1) Maximum Path based Relation Strength ($MPRS$); (2) Maximum Path based Relation Strength Plus ($MPRS+$); và (3) Relation Strength Similarity Plus ($RSS+$). Hai phương pháp $MPRS+$ và $RSS+$ có sử dụng thông tin về xu hướng cộng tác để cải tiến việc tính toán mức độ quan hệ của những người nghiên cứu trong mạng đồng tác giả. Các phương pháp đề xuất ứng dụng vào bài toán khuyến nghị cộng tác ở mức cá nhân người nghiên cứu. Thực nghiệm được tiến hành trên hai tập dữ liệu khoa học: i) cơ sở dữ liệu khoa học mở 'Digital Bibliography & Library Project' (DBLP); ii) tập dữ liệu do chúng tôi rút trích từ trang web của hệ thống Microsoft Academic Search (gọi tắt là MAS). Kết quả thực nghiệm cho thấy các phương pháp đề xuất cho kết quả tốt hơn các phương pháp tương tự đỉnh truyền thống và phổ biến hiện nay.

**Từ khóa.** Hệ khuyến nghị (recommender system), cộng tác nghiên cứu (research collaboration), khuyến nghị cộng tác (collaboration recommendation), phân tích mạng đồng tác giả (co-author network analysis), xu hướng cộng tác (collaborative trend).

**Abstract.** Collaboration recommendation is a problem that automatically selects and provides a list of potential researchers or research groups with respect to the input which is a researcher or a research group. Recently, this problem has attracted a lot of attention of many researchers in this area. A popular approach for collaboration recommendation problem is based on social network analysis, specifically co-author network analysis. However, the current methods do not consider collaborative trend in analyzing co-author network and the collaborative trend is one of the key factors for forming new co-authorships. In this paper, we propose three new methods: (1) Maximum Path based Relation Strength (MPRS); (2) Maximum Path based Relation Strength Plus (MPRS+); and (3) Relation Strength Similarity Plus (RSS+), for modeling and calculating vertex similarity in the co-author network. In our trend-based methods (MPRS+, RSS+), information of collaborative trend is used to improve the calculation of relation strength for researchers in the co-author network. The proposed methods are applied for researcher collaboration recommendation. Experiments are conducted on two dataset: i) Digital Bibliography & Library Project (DBLP), one popular and public science database;

ii) the dataset extracted from the Microsoft Academic Search website. The experiment results show that our proposed methods are more effective than the existing vertex similarity methods in predicting co-author collaboration.

**Key words.** Recommender system, research collaboration, collaboration recommendation, co-author network analysis, collaborative trend.

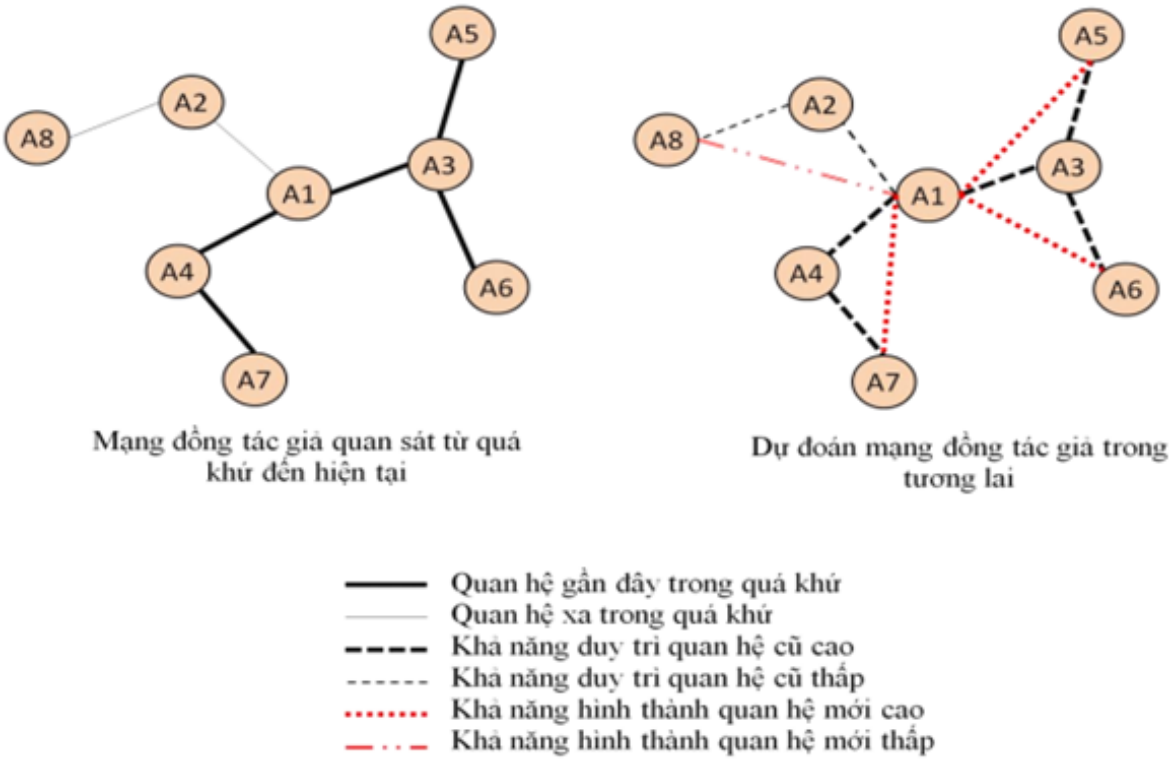
## 1. GIỚI THIỆU

Mục đích của nghiên cứu này là phát triển các phương pháp khuyến nghị cộng tác trong nghiên cứu khoa học. Có nhiều định nghĩa khác nhau cho cộng tác, nhưng nhìn chung 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 [1]. Trong nghiên cứu khoa học, có thể quan niệm cộng tác nghiên cứu là quá trình làm việc cùng nhau của những người nghiên cứu để đạt được một mục đích chung trong việc tìm ra các tri thức khoa học mới [2].

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 với một mục đích công việc cụ thể? Đối với các sinh viên, những người nghiên cứu trẻ hay một người nghiên cứu kinh nghiệm đi vào một lĩnh vực mới thì thông thường rất khó có thể biết được ai sẽ là người cộng tác phù hợp, vì họ thiếu thông tin và tri thức về lĩnh vực quan tâm. Bên cạnh đó, ngay cả những người nghiên cứu có kinh nghiệm chắc chắn cũng không thể nắm hết tất cả thông tin về những người thuộc lĩnh vực của mình trong cộng đồng nghiên cứu trên toàn thế giới. Vì vậy, việc nghiên cứu phát triển các hệ thống khuyến nghị cộng tác trong nghiên cứu khoa học là thật sự cần thiết.

Về các hệ khuyến nghị trong nghiên cứu khoa học có thể kể đến các nghiên cứu phát triển trên các hệ thống như CiteSeer [8], Microsoft Academic Search, ArtnetMinner [7]. Hầu hết các nghiên cứu gần đây của các nhóm này đều quan tâm đến hướng phân tích mạng xã hội (cụ thể là mạng đồng tác giả) [3, 4]. Trong các nghiên cứu gần đây, thì tiếp cận phân tích mạng xã hội đã cho thấy đây là một hướng tiếp cận tiềm năng và bước đầu khá thành công trong việc phát triển các phương pháp khuyến nghị trong nghiên cứu khoa học [3, 4, 12, 13, 14, 17, 21]. Tuy nhiên, các nghiên cứu liên quan kể trên đều chưa quan tâm đến yếu tố xu hướng cộng tác khi phân tích các mối quan hệ trong mạng. Trong khi xu hướng cộng tác có ảnh hưởng rất lớn đến việc hình thành các mối quan hệ cộng tác mới. Vì vậy, mục đích chính của nghiên cứu trong bài báo này là đưa vào xu hướng cộng tác để cải tiến các phương pháp phân tích mạng đồng tác giả phổ biến hiện nay, ứng dụng vào bài toán khuyến nghị cộng tác ở mức cá nhân người nghiên cứu.

Tiếp cận trong bài báo này dựa trên giả thuyết là: "Xu hướng cộng tác của một người trong quá khứ gần sẽ là yếu tố tác động chính đến các mối quan hệ cộng tác trong tương lai". Với giả thuyết đưa ra thì thông thường trong tương lai, xu hướng một người tiếp tục duy trì các mối quan hệ cộng tác với những người đã có quan hệ cộng tác trong quá khứ gần lớn hơn là các mối quan hệ cộng tác quá lâu trong quá khứ (ví dụ trong hình vẽ 1 thì khả năng A1 duy trì quan hệ công tác với A3, A4 lớn hơn so với A2). Đồng thời, xu hướng mà một người hình thành các quan hệ cộng tác mới dựa trên 'việc bắc cầu' các mối quan hệ cộng tác trong quá khứ gần lớn hơn "việc bắc cầu" các mối quan hệ cộng tác trong quá khứ xa. Chẳng hạn, hình vẽ 1 cho thấy khả năng A1 hình thành các mối quan hệ mới với A5, A6, A7 lớn hơn so với A8.




*Hình 1.* Ảnh hưởng của xu hướng cộng tác trong quá khứ đến các cộng tác trong tương lai

Làm thế nào để mô hình hóa, phân tích xu hướng cộng tác, cũng như đánh giá mức độ ảnh hưởng của nó đối với các quan hệ trong tương lai? Trong nghiên cứu của mình, nhóm tác giả đã tìm cách mô hình hóa và phân tích yếu tố xu hướng và áp dụng cho bài toán khuyến nghị cộng tác trong nghiên cứu khoa học.

Các đóng góp chính của bài báo:

+ Tiếp cận dựa trên xu hướng để phân tích mạng đồng tác giả.

+ Đề xuất một phương pháp để mô hình hóa, phân tích xu hướng cộng tác.

+ Khảo sát, thực nghiệm, đánh giá các phương pháp đề xuất dựa trên xu hướng với các phương pháp phân tích mạng đồng tác giả phổ biến khác.

+ Tiến hành thực nghiệm trên tập dữ liệu khoa học DBLP, và một tập do nhóm tác giả thu thập và rút trích từ website của hệ thống Microsoft Academic Search (MAS).

Phần còn lại của bài báo được bố cục như sau: Mục 2 sẽ là tóm tắt khảo sát các nghiên cứu liên quan; Mục 3 trình bày về các phương pháp tương tự đỉnh phổ biến dùng trong phân tích mạng đồng tác giả; Mục 4 trình bày tiếp cận dựa trên thông tin xu hướng và một phương pháp đề xuất để mô hình hóa, tính toán xu hướng cộng tác trong mạng đồng tác giả, cũng như đề xuất các phương pháp tính toán tương tự tỉnh trong mạng đồng tác giả và ứng dụng cho bài toán khuyến nghị cộng tác; Mục 5 trình bày kết quả thực nghiệm trên các tập dữ liệu phổ biến và một số thảo luận, cũng như nhận định về kết quả thực nghiệm. Và cuối cùng là phần kết luận và hướng phát triển.

## 2. CÁC NGHIÊN CỨU LIÊN QUAN

Adomavicius và Tuzhilin đã khảo sát tổng quan về hệ khuyến nghị và đã phân chia các phương pháp khuyến nghị truyền thống thành ba nhóm chính: (1) Lọc dựa trên nội dung (Content-Based Filtering); (2) Lọc cộng tác (Collaborative Filtering - CF); (3) Các phương pháp kết hợp (Hybrid) [6].

(1) Các phương pháp dựa trên nội dung tìm cách so khớp và khuyến nghị các items gần và giống nhất với các items mà người dùng thích và quan tâm trong quá khứ. Các items được so khớp tương tự dựa trên đặc trưng về nội dung.

(2) Các phương pháp lọc cộng tác (CF) sẽ tìm cách lọc ra những người đồng sở thích dựa vào những items mà họ đã quan tâm. Và hệ thống sẽ khuyến nghị những items cho người dùng dựa vào các items của những người đồng sở thích.

(3) Các phương pháp Hybrid là các phương pháp kết hợp cả (1) và (2). Có thể nói, các phương pháp tiếp cận truyền thống chưa quan tâm đến những mối quan hệ xã hội của người dùng. Trong khi các mối quan hệ xã hội là yếu tố quan trọng ảnh hưởng đến sở thích và hành vi của các cá nhân. Gần đây, tiếp cận phân tích mạng xã hội đã chứng tỏ được những thành công trong việc phát triển các hệ khuyến nghị [12, 13, 14].

Liên quan đến khuyến nghị trong nghiên cứu khoa học nói chung, thì khuyến nghị cộng tác đóng vai trò quan trọng và gần đây đã bắt đầu thu hút nhiều quan tâm. Chen và đồng nghiệp đã phát triển hệ thống tìm kiếm chuyên gia cộng tác CollabSeer dựa trên cấu trúc mạng đồng tác giả [3]. Tang và đồng nghiệp đã nghiên cứu đề xuất các phương pháp so khớp chuyên gia dựa trên nhiều ràng buộc khác nhau và ứng dụng vào bài toán khuyến nghị chuyên gia phản biện bài báo khoa học, khuyến nghị giảng viên cho một môn học [11]. Trong một nghiên cứu khuyến nghị cộng tác khác, Tang và đồng nghiệp đề xuất các phương pháp khuyến nghị cộng tác cho các chuyên gia trong các nghiên cứu liên ngành [5]. Bên cạnh đó các nghiên cứu liên quan đến bài toán tìm kiếm, so khớp chuyên gia cũng cung cấp các phương pháp nền tảng cho khuyến nghị cộng tác nghiên cứu. Balog và đồng nghiệp, Gollapalli và đồng nghiệp đã trình bày, thực nghiệm, đánh giá các phương pháp phổ biến mà dùng để biểu diễn thông tin và tính toán tương tự giữa các chuyên gia [9, 10, 15].

Các nghiên cứu dựa trên tiếp cận phân tích mạng xã hội, cụ thể là mạng đồng tác giả cho các bài toán khuyến nghị trong nghiên cứu khoa học đã chứng tỏ được khả năng tiềm ẩn và ưu điểm của nó thông qua các thực nghiệm, đánh giá [3, 12, 13, 14, 16]. Tuy nhiên hầu hết các nghiên cứu này đều chưa quan tâm đến yếu tố xu hướng nghiên cứu, cũng như xu hướng cộng tác khi phát triển các phương pháp khuyến nghị. Trong khi yếu tố xu hướng quan hệ có ảnh hưởng lớn đến việc hình thành các mối quan hệ cộng tác mới. Trong bài báo này, sẽ đưa vào thông tin xu hướng cộng tác để cải tiến các phương pháp phân tích mạng đồng tác giả phổ biến hiện nay và ứng dụng cho bài toán khuyến nghị cộng tác ở mức cá nhân người nghiên cứu.

## 3. CÁC PHƯƠNG PHÁP TƯƠNG TỰ ĐỈNH PHỔ BIẾN

Đối với phân tích mạng đồng tác giả, thì việc tính toán tương tự của các đỉnh là bước khá quan trọng giúp tiên đoán, khám phá những liên kết tiềm năng. Các phương pháp tính toán tương tự đỉnh truyền thống có thể chia thành hai nhóm: các phương pháp dựa trên cấu trúc cục bộ và các phương pháp dựa trên cấu trúc toàn cục của mạng.

### 3.1. Tương tự đỉnh dựa trên cấu trúc cục bộ

Các phương pháp dựa trên cấu trúc cục bộ dùng thông tin lân cận cục bộ để tính độ tương tự của hai đỉnh bất kỳ trong mạng. Ý tưởng chung của các độ đo cục bộ là "Hai đỉnh càng tương tự nhau nếu chúng có chung nhiều lân cận". Tức chỉ có những đỉnh lân cận trực tiếp của hai đỉnh được xét đến khi tính toán sự tương tự, trong khi các đỉnh khác không được quan tâm xem xét. Một số phương pháp tương tự đỉnh cục bộ phổ biến có thể kể đến như Jaccard [3], Cosine [3], Adamic-Adar [17].

### 3.2. Tương tự đỉnh dựa trên cấu trúc toàn cục

Các phương pháp này dựa trên cấu trúc toàn cục của cả mạng thay vì chỉ xét cấu trúc cục bộ như các phương pháp kể trên. Một số phương pháp toàn cục phổ biến thường dùng như SimRank [18], P-Rank [19]. Các phương pháp này dựa trên ý tưởng "Hai đỉnh càng giống nhau nếu như những lân cận trực tiếp trong mạng là tương tự nhau". Vì thế việc tính độ tương tự đỉnh dùng SimRank, P-Rank là một quá trình đệ qui.

Ta hoàn toàn có thể áp dụng các phương pháp cục bộ lẫn toàn cục kể trên cho việc tính toán tương tự đỉnh trong mạng đồng tác giả để tìm ra các ứng viên tiềm năng cho khuyến nghị cộng tác. Tuy nhiên, các phương pháp kể trên đều chưa quan tâm đến yếu tố xu hướng khi lượng hóa các mối quan hệ. Vì vậy, nhóm nghiên cứu đã đưa vào thông tin xu hướng để cải tiến các phương pháp tương tự đỉnh khi phân tích mạng đồng tác giả.

## 4. CÁC PHƯƠNG PHÁP ĐỀ XUẤT

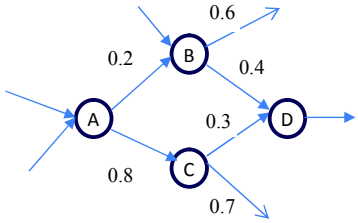
Phần này trình bày các phương pháp đề xuất để tính toán tương tự đỉnh trong mạng đồng tác giả. Các phương pháp đề xuất dựa trên sự kết hợp giữa lý thuyết đồ thị và lý thuyết xác suất. RSS+ là phương pháp cải tiến từ RSS (Relation Strength Similarity) [3]. RSS tính mức độ quan hệ của hai đỉnh $X, Y$ bất kỳ trong mạng bằng tổng trọng số các đường đi có thể từ $X$ đến $Y$. MPRS (Maximum Path based Relation Strength) dựa trên đường đi có trọng số cực đại. Hai phương pháp RSS+ và MPRS+ đưa vào xu hướng cộng tác cho tính toán tương tự đỉnh.

Ví dụ, trong hình 2 bên dưới thì A có viết tổng cộng 10 bài báo. Trong đó A viết chung với B là 2 bài, và với C là 8 bài. Khi đó trọng số các cung AB và AC trong mạng đồng tác giả có thể gán lần lượt là: 0.2, 0.8. Hình vẽ bên cạnh minh họa việc tính tương tự dựa trên MPRS và RSS cho hai đỉnh A và D.

- $Sim_{MPRS}(A, D) = MAX((0.2 \times 0.4), (0.8 \times 0.3)) = 0.24$
- $SimRSS(A, D) = SUM((0.2 \times 0.4), (0.8 \times 0.3)) = 0.32$
- Đối với các phương pháp RSS+ và MPRS+ thì trọng số của một cung nối trong mạng đồng tác giả không chỉ phụ thuộc vào số bài viết chung, mà còn phụ thuộc vào thời gian viết chung bài đó là lúc nào.

### 4.1. Tương tự đỉnh dựa trên đường đi cực đại (MPRS)

Phương pháp tương tự đỉnh dựa trên "đường đi cực đại" giữa hai đỉnh $X, Y$ bất kỳ trong mạng đồng tác giả, gọi tắt là MPRS (Maximum Path based Relation Strength). Khi đó độ




*Hình 2.* Minh họa cách tính của các phương pháp tương tự đỉnh

tương tự giữa $X, Y$ gọi là $Sim_{MPRS}(X, Y)$ có thể được tính theo các bước sau:

Gọi $Direct\_Sim_{MPRS}(X, Y)$ là trọng số cạnh nối giữa hai đỉnh $X, Y$ bất kỳ. Khi đó

$$Direct\_Sim_{MPRS}(X, Y) = \begin{cases} \dfrac{f_{XY}}{\sum\limits_{\forall Z \in N_X} f_{XZ}}, & \text{nếu } X \text{ và } Y \text{ có liên kết trực tiếp} \\ 0, & \text{ngược lại} \end{cases} \quad (1)$$

trong đó, $f_{XY}$ là một hàm tính số lần đồng tác giả của $X$ và $Y$, $N_X$ là tập các đỉnh lận cận của $X$.

Trong trường hợp $X$ và $Y$ không có liên kết trực tiếp. Nếu trong mạng có một đường đi đơn $p$ từ $X$ đến $Y$ qua $k$ đỉnh là $Z_1, Z_2, Z_3, ..., Z_k$ (với $Z_1$ là $X$, $Z_k$ là $Y$), thì trọng số đường đi có thể tính như sau

$$WeightOf\_DirectPath_p(X, Y) = \prod_{i=1}^{k} Direct\_Sim_{MPRS}(Z_i, Z_{i+1}). \quad (2)$$

Trong trường hợp mạng đang xét có $m$ đường đi đơn từ $X$ đến $Y$ là $p_1, p_2, ..., p_m$ thì khi đó mức độ quan hệ của $X$ và $Y$, tức $Indirect\_Sim_{MPBVS}$ có thể được tính như sau

$$Indirect\_Sim_{MPRS}(X, Y) = MAX_{i=1..m} WeightOf\_DirectPath_{p_i}(X, Y)$$

Trong những mạng kích thước lớn thì việc tính $Indirect\_Sim_{MPRS}(X, Y)$ có thể "quá tải". Đồng thời, độ tương tự đỉnh của hai đỉnh bất kỳ có thể ít ý nghĩa và giá trị của $WeightOf\_DirectPath_p(X, Y)$ tiệm cận 0 nếu như đường đi đơn từ $X$ đến $Y$ qua quá nhiều đỉnh trung gian. Vì thế trong quá trình thực nghiệm, đã sử dụng một giá trị ngưỡng là $r$ như một thông số heuristic để kiểm soát quá trình xác định và tính trọng số các đường đi đơn từ $X$ đến $Y$ trong những mạng kích thước lớn. Tức là chỉ xem xét các đường đi đơn từ $X$ đến $Y$ có "bán kính" (số đỉnh trên đường đi) nhỏ hơn hay bằng $r$. Như vậy $WeightOf\_DirectPath_p(X, Y)$ được tính như sau

$$WeightOf\_DirectPath_p(X, Y) = \begin{cases} \prod\limits_{i=1}^{k-1} Direct\_Sim_{MPRS}(Z_i, Z_{i+1}), & \text{nếu } \ k \leq r, \\ 0, \ \text{ngược lại.} \end{cases} \quad (3)$$

Tóm lại, tương tự của hai đỉnh $X, Y$ trong mạng theo phương pháp $MPRS$ có thể tính như sau

$$Sim_{MPRS}(X, Y) = Indirect\_Sim_{MPRS}(X, Y).$$

### 4.2. Tương tự đỉnh dựa trên đường đi cực đại và yếu tố xu hướng ($MPRS+$)

Tương tự $MPRS$, nhưng với $MPRS+$ sẽ xem xét yếu tố xu hướng cộng tác để lượng hóa mức độ quan hệ của hai đỉnh trong mạng đồng tác giả. Độ tương tự giữa hai đỉnh bất kỳ theo $MPRS+$ được tính như sau:

Gọi $Direct\_Sim_{MPRS+}(X,Y)$ là trọng số cạnh nối giữa hai đỉnh $X,Y$ bất kỳ. Khi đó

$$Direct\_Sim_{MPRS+}(X,Y) = \begin{cases} \dfrac{f_{(Trend)XY}}{\sum\limits_{\forall Z \in N_X} f_{(Trend)XZ}}, \text{ nếu } X \text{ và } Y \text{ có liên kết trực tiếp} \\ 0, \text{ ngược lại} \end{cases} \quad (5)$$

trong đó, $f_{(Trend)XY}$ là một hàm phụ thuộc yếu tố xu hướng cộng tác, $N_X$ là tập hợp các đỉnh lận cận của $X$.

Với hàm $f_{(Trend)XY}$, có thể dùng hệ số $k$ và thông số $t$ để đánh giá mức độ ảnh hưởng của các quan hệ đồng tác giả dựa trên yếu tố xu hướng. Hàm $f_{(Trend)XY}$ có thể định nghĩa như sau

$$f_{(Trend)XY} = f(t)_{XY} = k * n_1(t) + (1-k) * n_2(t), \quad (6)$$

trong đó, $k$ là hệ số ảnh hưởng đến xu hướng cộng tác, $n_1(t)$ cho biết số lần mà $X$ đồng tác giả với $Y$ trong $t$ năm gần đây, $n_2(t)$ cho biết số lần mà $X$ đồng tác giả với $Y$ trước đây hơn $t$ năm.

Trong thực nghiệm, ta thay đổi các tham số $k$ và $t$ để đánh giá hiệu năng của phương pháp đề xuất.

### 4.3. Tương tự đỉnh dùng phương pháp $RSS+$ (cải tiến từ $RSS$)

Chen và đồng nghiệp đã đề xuất một phương pháp tương tự đỉnh dựa trên mức độ quan hệ giữa hai đỉnh bất kỳ trong mạng đồng tác giả, gọi là $RSS$ (Relation Strength Similarity) [3]. $RSS$ là một độ đo bất đối xứng, áp dụng cho mạng có trọng số. Chen và đồng nghiệp cũng đã dùng độ đo này để khám phá các liên kết tiềm năng trong mạng đồng tác giả [4, 20]. Tuy nhiên, yếu tố xu hướng thì chưa được họ quan tâm trong $RSS$. Ở đây, ta đã đưa xu hướng cộng tác vào $RSS$ và cải tiến $RSS$ thành $RSS+$. Với $RSS+$ thì độ tương tự đỉnh có thể được tính như sau:

Gọi $Direct\_Sim_{MPRS}(X,Y)$ trọng số cạnh nối giữa hai đỉnh $X,Y$ bất kỳ. Khi đó,

$$Direct\_Sim_{RSS+}(X,Y) = \begin{cases} \dfrac{f_{(Trend)XY}}{\sum\limits_{\forall Z \in N_X} f_{(Trend)XZ}}, \text{ nếu } X \text{ và } Y \text{ có liên kết trực tiếp} \\ 0, \text{ ngược lại} \end{cases} \quad (7)$$

trong đó, $f_{(Trend)XY}$ là một hàm phụ thuộc yếu tố xu hướng cộng tác, $f_{(Trend)XY}$ được tính tương tự như phương pháp $MPRS+$, $N_X$ là tập hợp các đỉnh lận cận của $X$.

Trong trường hợp $X$ và $Y$ không có liên kết trực tiếp. Nếu trong mạng có một đường đi đơn $p$ từ $X$ đến $Y$ qua $k$ đỉnh là $Z_1, Z_2, Z_3, ..., Z_k$ (với $Z_1$ là $X$, $Z_k$ là $Y$), thì trọng số đường đi có thể tính như sau

$$WeightOf\_DirectPath_p(X,Y) = \prod_{i=1}^{k} Direct\_Sim_{RSS+}(Z_i, Z_{i+1}). \quad (8)$$

Trong trường hợp mạng đang xét có $m$ đường đi đơn từ $X$ đến $Y$ là $p_1, p_2, ..., p_m$ thì khác với phương pháp $MPRS$ là chọn đường đi có trọng số cực đại (đường đi cộng tác có tích xác suất cộng tác qua các đỉnh trung gian là lớn nhất). Với phương pháp $RSS$, cũng như $RSS+$ mức độ quan hệ của $X$ và $Y$ trong trường hợp này là tổng của các phân bố xác suất cộng tác qua các con đường cộng tác trung gian có thể, tức $Indirect\_SimRSS+$ có thể được tính như sau

$$Indirect\_Sim_{RSS+}(X,Y) = \sum_{i=1}^{m} WeightOf\_DirectPath_{p_i}(X,Y). \tag{9}$$

Tương tự như $MPRS+$, với những mạng có kích thước lớn ta chỉ xem xét các đường đi đơn từ $X$ đến $Y$ có "bán kính" (số đỉnh trên đường đi) nhỏ hơn hay bằng $r$. Như vậy $WeightOf\_DirectPath_{RSS+}(X,Y)$ được tính như sau

$$WeightOf\_DirectPath_p(X,Y) = \begin{cases} \prod_{i=1}^{k-1} Direct\_Sim_{RSS+}(Z_i, Z_{i+1}), \text{ nếu } \ k \leq r, \\ 0, \text{ ngược lại.} \end{cases} \tag{10}$$

Tóm lại, độ đo $RSS+$ của hai đỉnh $X, Y$ bất kỳ trong mạng có thể tính như sau

$$Sim_{RSS+}(X,Y) = Direct\_Sim_{RSS+}(X,Y) + Indirect\_Sim_{RSS+}(X,Y). \tag{11}$$

## 5. THỰC NGHIỆM VÀ 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. Tang và cộng sự thực nghiệm trên tập dữ liệu của hệ thống ArnetMiner cho bài toán khuyến nghị cộng tác liên ngành [5]. Chen và cộng sự triển khai các thực nghiệm của họ trên tập dữ liệu của CiteSeer cho bài toán khuyến nghị cộng tác, cũng như khám phá các liên kết tiềm năng trên mạng đồng tác giả [2, 3, 20]. Các nhóm nghiên cứu kể trên chỉ đề cập đến số liệu thực nghiệm rút ra như thế nào, chứ họ chưa công bố tập dữ liệu. Với tính phổ biến của DBLP và hệ thống tìm kiếm Microsoft Academic Search, trong nghiên cứu của mình, nhóm tác giả đã tiến hành thực nghiệm trên tập DBLP và tập dữ liệu rút trích từ website của Microsoft Academic Search. Dữ liệu, cũng như mã nguồn sử dụng trong thực nghiệm của bài báo có thể tham khảo và download tại trang web <https://sites.google.com/site/tinhuynhuit/dataset> .

Về phương pháp đánh giá cho hệ khuyến nghị, đây là một vấn đề vẫn đang được nghiên cứu. Đáng tin cậy nhất là khảo sát người dùng, phân tích phản hồi của người dùng thông qua hệ thống, hoặc lấy ý kiến chuyên gia. Để làm được điều đó thì ta cần phải có hệ thống triển khai sử dụng trên thực tế. Bên cạnh đó, một số nghiên cứu liên quan hiện nay dùng kết quả tiên đoán liên kết đồng tác giả để đánh giá hiệu năng của các phương pháp khuyến nghị cộng tác [2, 3, 5, 20]. Ở đây cũng dùng kết quả tiên đoán liên kết đồng tác giả để đánh giá, so sánh hiệu năng các phương pháp đề xuất với các phương pháp khác.

### 5.1. Thiết lập dữ liệu thực nghiệm cho DBLP và MAS

Trong nghiên cứu này, ta sử dụng dữ liệu từ các bài báo công bố năm 2001 đến năm 2011 để tiến hành thực nghiệm. Dữ liệu 5 năm đầu (2001-2005) được dùng để xây dựng mạng huấn

luyện (training network). Để khách quan, ta chia mạng huấn luyện thành ba nhóm bậc khác nhau: cao, trung bình và thấp. Những tác giả bậc cao là những tác giả có số bậc thuộc nhóm 1/3 bậc cao nhất của tất cả các bậc, những tác giả bậc thấp là những tác giả có số bậc thuộc nhóm 1/3 bậc thấp nhất của tất cả các bậc, còn lại là những tác giả thuộc nhóm tác giả bậc trung bình. Với mỗi loại nhóm bậc tác giả, ta chọn ngẫu nhiên 100 tác giả để tiến hành thực nghiệm. Với mỗi tác giả, mức độ tương tự với tất cả những người còn lại trong mạng được tính và $Top-n$ những người tương tự nhất được trả về theo các phương pháp khác nhau. Độ chính xác Precision cho tiên đoán liên kết đồng tác giả được tính dựa vào mạng đồng tác giả tương lai gần (2006-2008) và mạng đồng tác giả trong tương lai xa (2009-2011).

## 5.2. Kết quả thực nghiệm

### *5.2.1. Khảo sát tham số xu hướng k và t*

Trong thực nghiệm, ta ước lượng tham số $k$ và $t$ dựa trên kinh nghiệm và khảo sát hiệu năng của phương pháp $RSS+$ khi thực nghiệm với nhiều $k, t$ khác nhau. Lần lượt $t$ nhận các khoảng thời gian là 1, 2, 3, 4 năm, và $k$ lần lượt được thực nghiệm với 0.6, 0.7, 0.8 và 0.9. Kết quả tốt nhất thu được với $k = 0.9$ và $t = 1$ trong hầu hết các trường hợp (Bảng 1). Với việc

*Bảng 1.* Khảo sát hệ số $k$ và $n_1(t)$ khi lượng hóa yếu tố xu hướng

| **RSS+ khi lượng hóa xu hướng và tiên đoán tương lai gần ([2006-2008])** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **$n_1(t)$ = 1 ([2005])** | | | | | | $n_1(t)$ = 2 ([2004-2005]) | | | | |
| K | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
| 0.6 | 0.7200 | 0.6745 | 0.6350 | 0.5998 | 0.5673 | 0.7200 | 0.6761 | 0.6350 | 0.6023 | 0.5747 |
| 0.7 | 0.7533 | 0.6878 | 0.6451 | 0.6158 | 0.5922 | 0.7333 | 0.6861 | 0.6496 | 0.6133 | 0.5922 |
| 0.8 | 0.7400 | 0.6928 | 0.6596 | 0.6284 | 0.6070 | 0.7433 | 0.6945 | 0.6496 | 0.6208 | 0.5915 |
| **0.9** | **0.7567** | **0.7045** | **0.6540** | **0.6225** | **0.5956** | 0.7400 | 0.6978 | 0.6440 | 0.6149 | 0.5983 |
| $n_1(t)$ = 3 ([2003-2005]) | | | | | | $n_1(t)$ = 4 ([2002-2005]) | | | | |
| K | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
| 0.6 | 0.7167 | 0.6661 | 0.6283 | 0.5931 | 0.5633 | 0.7033 | 0.6494 | 0.6127 | 0.5814 | 0.5545 |
| 0.7 | 0.7167 | 0.6811 | 0.6350 | 0.5973 | 0.5740 | 0.7033 | 0.6511 | 0.6138 | 0.5847 | 0.5579 |
| 0.8 | 0.7200 | 0.6895 | 0.6440 | 0.6023 | 0.5734 | 0.7000 | 0.6561 | 0.6205 | 0.5898 | 0.5612 |
| 0.9 | 0.7167 | 0.6878 | 0.6350 | 0.6040 | 0.5774 | 0.6933 | 0.6561 | 0.6239 | 0.5872 | 0.5606 |

ước lượng các tham số $k, t$ bằng thực nghiệm cho hàm xu hướng $f_{(Trend)}$, ta được kết quả tốt nhất với $k = 0.9$ và $t = 1$. Như vậy trong các thực nghiệm kế tiếp, thì hàm xu hướng được chọn là

$$f_{(Trend)XY} = f(t)_{XY} = 0.9 * n_1(1) + (1 - 0.9) * n_2(1)$$

trong đó, $n_1(1)$ cho biết số lần mà $X$ đồng tác giả với $Y$ trong 1 năm gần đây, $n_2(1)$ cho biết số lần mà $X$ đồng tác giả với $Y$ trước đây hơn 1 năm.

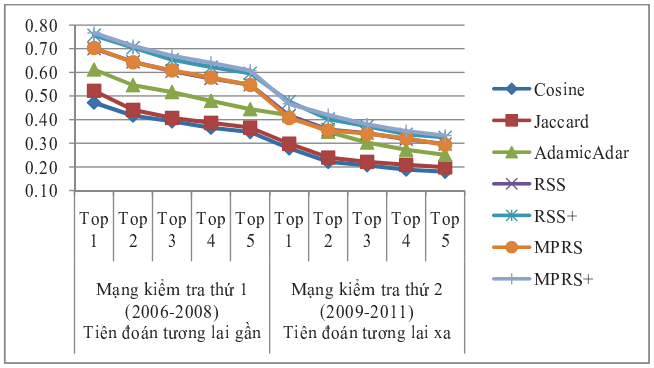
### *5.2.2. Thực nghiệm trên tập DBLP*

Với việc thiết lập dữ liệu thực nghiệm được mô tả như trong mục 5.1 và hàm $f(Trend)$ đã chọn trong mục 5.2.1, cho phép đánh giá độ chính xác tiên đoán liên kết đồng tác giả (Precision) với Top-1, Top-2, Top-3, Top-4, Top-5 những đỉnh tương tự nhất được trả về. Với Top-5 những đỉnh tương tự nhất được trả về thì $MPRS+, RSS+$ có độ chính xác (Precision) lần lượt là 0.61, 0.60 cho tiên đoán đồng tác giả trong tương lai gần (2006-2008) và lần lượt là 0.33, 0.32 cho tiên đoán đồng tác giả trong tương lai xa (2009-2011). Trong khi các phương pháp tương tự đỉnh phổ biến hiện nay cao nhất là $RSS$ chỉ đạt 0.55 với tiên đoán tương lai gần và 0.30 với tiên đoán tương lai xa (Bảng 2, Hình 3). Như vậy, kết quả thực nghiệm trên

tập DBLP cho thấy các phương pháp đề xuất $RSS+, MPRS, MPRS+$ cho kết quả tốt hơn so với các phương pháp tương tự đỉnh phổ biến hiện nay.

*Bảng 2.* Kết quả tiên đoán liên kết đồng tác giả trên tập thực nghiệm DBLP

| Phương pháp | $f(trend)_{XY} = f(1)_{XY} = 0.9.n_1(1) + (1-0.9).n_2(1)$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mạng kiểm tra thứ 1 (2006-2008) Tiên đoán tương lai gần | | | | | Mạng kiểm tra thứ 2 (2009-2011) Tiên đoán tương lai xa | | | | |
| | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
| Cosine | 0.47 | 0.42 | 0.39 | 0.37 | 0.35 | 0.28 | 0.22 | 0.21 | 0.19 | 0.18 |
| Jaccard | 0.52 | 0.44 | 0.41 | 0.39 | 0.37 | 0.30 | 0.24 | 0.22 | 0.21 | 0.20 |
| AdamicAdar | 0.61 | 0.55 | 0.52 | 0.48 | 0.44 | 0.42 | 0.35 | 0.30 | 0.27 | 0.25 |
| RSS | 0.70 | 0.64 | 0.60 | 0.57 | 0.55 | 0.42 | 0.36 | 0.34 | 0.32 | 0.30 |
| **RSS+** | **0.76** | **0.70** | **0.65** | **0.62** | **0.60** | **0.48** | **0.40** | **0.37** | **0.34** | **0.32** |
| **MPRS** | **0.70** | **0.64** | **0.61** | **0.58** | **0.55** | **0.41** | **0.35** | **0.34** | **0.32** | **0.29** |
| **MPRS+** | **0.77** | **0.71** | **0.67** | **0.64** | **0.61** | **0.47** | **0.42** | **0.38** | **0.35** | **0.33** |




*Hình 3.* Kết quả tiên đoán liên kết đồng tác giả trên tập thực nghiệm DBLP

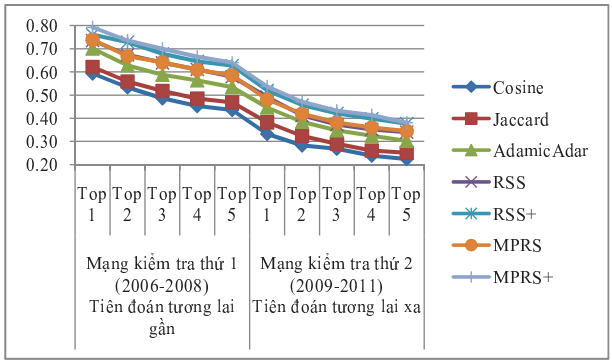
### *5.2.3. Thực nghiệm trên tập MAS*

Tương tự với việc tiến hành thực nghiệm trên tập DBLP như mô tả ở trên (mục 5.2.2). Phần này trình bày kết quả thực nghiệm trên một tập dữ liệu khác được rút trích từ trang web của hệ thống Microsoft Academic Search (MAS). Với Top-5 những đỉnh tương tự nhất được trả về thì $MPRS+, RSS+$ có độ chính xác (Precision) lần lượt là 0.64, 0.63 cho tiên đoán đồng tác giả trong tương lai gần (2006-2008) và lần lượt là 0.38, 0.38 cho tiên đoán đồng tác giả trong tương lai xa (2009-2011). Trong khi các phương pháp tương tự đỉnh phổ biến hiện nay cao nhất là $RSS$ chỉ đạt 0.58 với tiên đoán tương lai gần và 0.34 với tiên đoán tương lai xa (bảng 3, hình 4).

Như vậy, kết quả thực nghiệm trên cả hai tập dữ liệu thực nghiệm là $DBLP$ và $MAS$ (mục 5.2.2 và mục 5.2.3) đều cho thấy các phương pháp đề xuất $RSS+, MPRS, MPRS+$ cho kết quả tốt hơn so với các phương pháp tương tự đỉnh phổ biến hiện nay. Đặc biệt yếu tố xu hướng cộng tác trong hai phương pháp $MPRS+$ và $RSS+$ đã giúp cải tiến đáng kể kết quả dựa trên đánh giá tiên đoán liên kết đồng tác giả.

*Bảng 3.* Kết quả tiên đoán liên kết đồng tác giả trên tập thực nghiệm MAS

| Phương pháp | $f(trend)_{XY} = f(1)_{XY} = 0.9.n_1(1) + (1-0.9).n_2(1)$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mạng kiểm tra thứ 1 (2006-2008) Tiên đoán tương lai gần | | | | | Mạng kiểm tra thứ 2 (2009-2011) Tiên đoán tương lai xa | | | | |
| | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 | Top 1 | Top 2 | Top 3 | Top 4 | Top 5 |
| Cosine | 0.59 | 0.53 | 0.49 | 0.45 | 0.44 | 0.33 | 0.28 | 0.27 | 0.24 | 0.22 |
| Jaccard | 0.62 | 0.56 | 0.52 | 0.49 | 0.47 | 0.38 | 0.32 | 0.29 | 0.26 | 0.25 |
| AdamicAdar | 0.70 | 0.63 | 0.59 | 0.56 | 0.53 | 0.45 | 0.38 | 0.35 | 0.32 | 0.30 |
| RSS | 0.73 | 0.67 | 0.64 | 0.61 | 0.58 | 0.49 | 0.41 | 0.37 | 0.35 | 0.34 |
| **RSS+** | 0.76 | 0.73 | 0.68 | 0.65 | 0.63 | 0.52 | 0.46 | 0.42 | 0.40 | 0.38 |
| **MPRS** | 0.74 | 0.67 | 0.64 | 0.61 | 0.59 | 0.48 | 0.42 | 0.38 | 0.36 | 0.34 |
| **MPRS+** | 0.79 | 0.74 | 0.70 | 0.67 | 0.64 | 0.54 | 0.47 | 0.43 | 0.41 | 0.38 |




*Hình 4.* Kết quả tiên đoán liên kết đồng tác giả trên tập thực nghiệm MAS

## 6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Mục đích của nghiên cứu này là tập trung phát triển phương pháp mới dựa trên phân tích mạng đồng tác giả. Các phương pháp hướng đến tiên đoán các mối quan hệ cộng tác tiềm năng trong tương lai để khuyến nghị cho người dùng. Các phương pháp đề xuất của bài báo dựa trên thông tin xu hướng kết hợp lý thuyết đồ thị và xác suất khi phân tích mạng đồng tác giả. Thực nghiệm được tiến hành trên các tập dữ liệu khoa học như: DBLP, tập download từ Microsoft Academic Search. Kết quả cho thấy các phương pháp dựa trên yếu tố xu hướng đề xuất cho kết quả tốt hơn hẳn các phương pháp khác trong cả hai tập thực nghiệm. Với tập DBLP, độ chính xác tiên đoán đồng tác giả trong tương lai gần với các phương pháp dựa trên xu hướng cho Top-5 là 0.60 cho phương pháp $RSS+$, 0.61 cho phương pháp $MPRS+$. Trong khi các phương pháp hiện tại chỉ đạt 0.55 (cao nhất) cho Top-5 với phương pháp $RSS$ (Bảng 2, Hình 3).

Nhằm khai thác các mối quan hệ cộng tác tiềm năng, công việc tiếp theo là cải tiến các phương pháp hiện có dựa trên tiếp cận phân tích mạng xã hội, xem xét khai thác các mối quan hệ trung gian khác như: quan hệ của các cơ quan, trường, viện, quốc gia. Đó là những yếu tố ảnh hưởng đến các mối quan hệ cộng tác tiềm năng trong nghiên cứu khoa học. Hiện nay, yếu tố nội dung vẫn chưa được xem xét, thực hiện trong bài báo này. Vì vậy, cần những nghiên cứu đề xuất, thực nghiệm và so sánh với các phương pháp dựa trên nội dung, cũng như các phương pháp kết hợp cả nội dung và lọc cộng tác dựa trên mạng xã hội.

**Lời cảm ơn.** Nghiên cứu này được tài trợ bởi Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh với đề tài mã số C2012-07.

## TÀI LIỆU THAM KHẢO


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

[2] J. S. Katz and B. R. Martin, What is research collaboration, *Research Policy* **26** (1) (1997) 1–18.

[3] H. Chen, L. Gou, X. Zhang, and C. L. Giles, CollabSeer: a search engine for collaboration discovery, *Proceedings of the 11th annual International ACM/IEEE Conference on Digital libraries (JCDL),* New York, USA, 2011 (231–240).

[4] H. Chen, L. Gou, X. Zhang, and C. L. Giles, Discovering missing links in networks using vertex similarity measures, *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12),* New York, USA, 2012 (138–143).

[5] J. Tang, S. Wu, J. Sun, and H. Su, Cross-domain collaboration recommendation, *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*, New York, USA, 2012 (1285–1293).

[6] G. Adomavicius and A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, *IEEE Transactions on Knowledge and Data Engineering* **17** (6) (2005) 734–749.

[7] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, ArnetMiner: extraction and mining of academic social networks, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08). ACM,* New York, NY, USA, 2008 (990–998).

[8] C. L. Giles, K. D. Bollacker, and S. Lawrence, CiteSeer: an automatic citation indexing system, *Proceedings of the third ACM conference on Digital libraries (DL '98), ACM*, New York, USA, 1998 (89–98).

[9] K. Hofmann, K. Balog, T. Bogers, and M. de Rijke, Contextual factors for finding similar experts, *Journal of American Society for Information Science and Technology* **61** (5) (2010) 994–1014.

[10] K. Balog and M. de Rijke, Finding similar experts, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07), ACM,* New York, USA, 2007 (821–822).

[11] W. Tang, J. Tang, T. Lei, C. Tan, B. Gao, and T. Li, On optimization of expertise matching with various constraints, *Journal Neurocomputing* **76** (1) (2012) 71–83.

[12] F. E. Walter, S. Battiston, and F. Schweitzer, A model of a trust-based recommendation system on a social network, *Journal Autonomous Agents and Multi-Agent Systems* **16** (1) (2008) 57–74.

[13] I. Konstas, V. Stathopoulos, and J. M. Jose, On social networks and collaborative recommendation, *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, (SIGIR '09), ACM,* New York, NY, USA, 2009 (195–202).

[14] E. Davoodi, M. Afsharchi, and K. Kianmehr, A social network-based approach to expert recommendation system, *Proceedings of the 7th International Conference on Hybrid Artificial Intelligent Systems, Volume Part I*, Salamanca, Spain, 2012 (91–102).

[15] S. D. Gollapalli, P. Mitra, and C. L. Giles, Similar researcher search in academic environments, *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '12), ACM*, New York, USA, 2012 (167–170).

[16] H. Luong, T. Huynh, S. Gauch, K. Hoang, Exploiting social networks for publication venue recommendations, *Proceedings of the 4th International Conference on Knowledge Discovery and Information Retrieval*, Barcelona, Spain, 2012 (239–245).

[17] L. A. Adamic, E. Adar, Friends and neighbors on the web, *Journal Social Networks* **25** (3) (2003) 211–230.

[18] G. Jeh and J. Widom, Simrank: A measure of structural-context similarity, *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02), ACM*, New York, USA, 2002 (538–543).

[19] P. Zhao, J. Han, and Y. Sun, P-rank: A comprehensive structural similarity measure over information networks, *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09), ACM*, New York, USA, 2009 (553–562).

[20] H. Chen, L. Gou, X. Zhang, and C. L. Giles, Capturing missing edges in social networks using vertex similarity, *Proceedings of the sixth International Conference on Knowledge Capture (K-CAP '11), ACM,* New York, USA, 2011 (195–196).

[21] G. R. Lopes, M. M. Moro, L. K. Wives, and J. P. M. D. Oliveira, Collaboration recommendation on academic social networks, *Proceedings of the International Conference on Advances in Conceptual Modeling: Applications and Challenges (ER'10),* Springer-Verlag, Berlin, Heidelberg, 2010 (190–199).

*Ngày nhận bài 05 - 7 - 2013*
*Nhận lại sau sửa ngày 25 - 11 - 2013*