Khai thác xu hướng sở thích và quan hệ lòng tin để phát triển phương pháp khuyến nghị bài báo khoa học

Exploiting Trust Relationship and Research Trend of Researchers to Develop New Method for Scientific Paper Recommendation

Huỳnh Ngọc Tín, Hoàng Kiếm

Abstract: *In this paper, we propose a hybrid method for recommending potential scientific publications for researcher based on combination of trust relationships and research trend of researchers. The research trend let us know which research topic recently is interested in by a researcher while trust relationship let us know experts whom a researcher trust. Experiments are conducted on a big dataset crawled from Microsoft Academic Search[1]. The experimental results show that our proposed methods are more effective than the existing methods in recommending potential publications those are met with research interest of researchers.*

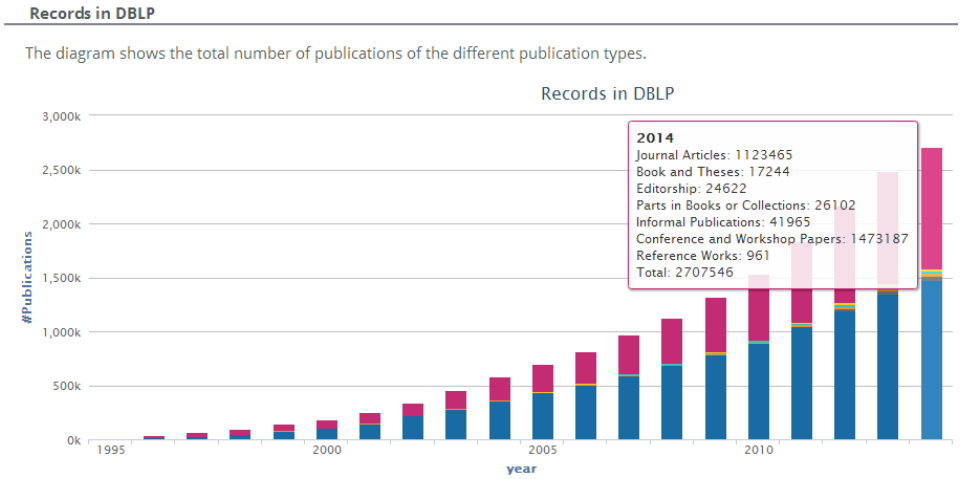
Từ khóa*: Hệ khuyến nghị (Recommender System), Khuyến nghị Bài báo (Paper Recommendation), Quan hệ Lòng tin (Trust Relationship), Xu hướng Nghiên cứu (Research Trend)*

I. GIỚI THIỆU

Tìm kiếm bài báo khoa học liên quan đến nghiên cứu để đọc, tham khảo, trích dẫn là việc làm thường xuyên của những người làm nghiên cứu khoa học, cụ thể là các nhà nghiên cứu. Hiện nay, các hệ thống tìm kiếm, thư viện số phổ biến trong lĩnh vực học thuật như ACM DL Portal, IEEE Xplore, Google Scholar, Microsoft Academic Search, DBLP,... đã đáp ứng hầu hết nhu cầu tìm kiếm tài liệu khoa học của các nhà nghiên cứu. Tuy nhiên, khối lượng khổng lồ các bài báo khoa học tăng lên hàng năm (Hình 1), làm cho các nhà nghiên cứu phải đương đầu với tình trạng quá tải thông tin, và mất nhiều thời gian hơn để tìm được những tài liệu liên quan. Bên cạnh đó, có thể có nhiều thông tin bài báo liên quan đến quan tâm nghiên cứu mà họ đã bỏ qua, hoặc không tìm thấy. Vấn đề đặt ra là "Làm thế nào để hầu hết các bài báo liên quan đến quan tâm nghiên cứu của các nhà nghiên cứu sẽ chủ động tìm đến họ, thay vì họ phải vất vả tự đi tìm thông tin liên quan?". Hệ khuyến nghị bài báo khoa học là giải pháp được các nghiên cứu gần đây quan tâm.

Các nghiên cứu dựa trên tiếp cận nội dung, gọi tắt tiếp cận nội dung, đã chứng tỏ được những thành công đối với bài toán này, điển hình là các nghiên cứu của Sugiyama và cộng sự năm 2010, 2011, 2013 [4-6]. Với tiếp cận nội dung, hệ thống sẽ mô hình hoá sở thích nghiên cứu của các nhà nghiên cứu dựa trên nội dung các bài báo mà họ công bố trong quá khứ. Sau đó, sở thích của họ sẽ được so khớp với nội dung của các bài báo quan sát được và một danh sách xếp hạng các bài báo liên quan sẽ được đề xuất. Tuy nhiên, đôi khi sở thích của nhà nghiên cứu thay đổi theo thời gian. Nếu chỉ dựa trên nội dung của tất cả các bài báo đã công bố trong quá khứ có thể không xác định đúng xu hướng quan tâm nghiên cứu của nhà nghiên cứu. Bên cạnh đó, thật sự không phù hợp nếu chọn một bài báo có nội dung liên quan, nhưng quá cũ, hoặc không đáng tin cậy để ưu tiên khuyến nghị. Do đó, cần xem xét những bài báo có chất lượng tốt, có độ tin cậy cao, của những chuyên gia có uy tín để ưu tiên khuyến nghị.

[1] http://academic.research.microsoft.com/



Hình 1. Sự gia tăng dữ liệu khoa học dựa trên Cơ sở dữ liệu khoa học DBLP
(Nguồn: http://www.informatik.uni-trier.de/~ley/statistics/recordsindblp.html, truy cập lần cuối 30/07/2014)

Câu hỏi đặt ra là như thế nào là những bài báo đáng tin cậy và như thế nào là những chuyên gia có uy tín? Trên thực tế, những chuyên gia uy tín thường là những người sẽ sản sinh ra nhiều công trình tốt, đáng tin cậy được cộng đồng trích dẫn và đặt lòng tin. Làm thế nào để lượng hóa được mức độ tin cậy hay lòng tin của người này đối với người khác? Và lòng tin ảnh hưởng như thế nào đến quyết định chọn bài báo để đọc, trích dẫn? Trong bài báo này, chúng tôi đề xuất phương pháp lượng hóa quan hệ lòng tin giữa các nhà nghiên cứu kết hợp với yếu tố xu hướng quan tâm nghiên cứu để phát triển các phương pháp cho khuyến nghị bài báo khoa học tiềm năng. Các đóng góp chính của bài báo có thể tóm tắt như sau:

- Khảo sát, đánh giá thực nghiệm các phương pháp khuyến nghị bài báo khoa học phổ biến hiện nay trên một tập dữ liệu lớn.
- Đề xuất và mô hình hóa quan hệ lòng tin trong lĩnh vực học thuật dựa trên quan hệ cộng tác và hành vi trích dẫn.
- Kết hợp xu hướng sở thích nghiên cứu và quan hệ lòng tin trong lĩnh vực học thuật để phát triển các phương pháp mới cho bài toán khuyến nghị bài báo khoa học liên quan.

Phần còn lại của bài báo được bố cục như sau: Phần II tóm tắt các nghiên cứu liên quan; Phần III trình bày các phương pháp phổ biến hiện nay cho khuyến nghị bài báo khoa học. Phần IV sẽ là các phương pháp đề xuất; Phần V tiến hành phân tích, đánh giá dựa trên kết quả thực nghiệm. Kết luận và hướng phát triển sẽ được trình bày trong mục VI.

II. NGHIÊN CỨU LIÊN QUAN

Liên quan đến khuyến nghị bài báo khoa học. Có một số bài toán con khác nhau mà các nghiên cứu hiện nay đang quan tâm. Bài toán khuyến nghị bài báo trích dẫn cho các nhà nghiên cứu khi viết bài. Một số nghiên cứu điển hình có thể kể đến như nghiên cứu Qi He và cộng sự, 2010, 2011 [2,3], Wenyi Huang và cộng sự, 2012 [16]. Các nghiên cứu này nhằm phát triển mô hình cho phép ánh xạ giữa các câu trong bài báo với tài liệu trích dẫn. Lawrence và cộng sự, 1999 [10], Huynh và cộng sự, 2012 [17], đã thực hiện các nghiên cứu nhằm phát triển các thuật toán khuyến nghị các bài báo tương tự khi người dùng duyệt qua một bài báo trong thư viện số.

Trong ngữ cảnh ứng dụng khác, Sugiyama và cộng sự, 2010, đã đề các xuất các phương pháp tiếp cận nội dung mới cho khuyến nghị bài báo khoa học phù hợp với quan tâm nghiên cứu của các nhà nghiên cứu [4]. Đóng góp chính của họ là khai thác quan tâm tiềm ẩn trong hồ sơ sở thích của các nhà nghiên cứu từ bài báo trong quá khứ kết hợp với các bài báo tham khảo và bài báo trích dẫn của các nhà nghiên cứu từ mạng

trích dẫn. Họ đã thu thập 597 bài báo từ hội nghị ACL (Association of Computational Linguistics) và lấy ý kiến 28 nhà nghiên cứu. 28 nhà nghiên cứu này sẽ xem danh sách 597 bài báo và cho biết bài báo nào liên quan hay không liên quan đến quan tâm nghiên cứu của họ. Tác giả đã dùng tập dữ liệu gán nhãn này để xây dựng tập đánh giá (Ground Truth). Bản chất của mạng trích dẫn này là một mạng rất thưa. Do đó, Sugiyama và cộng sự, 2013 đã tìm cách giảm bớt dữ liệu thưa bằng lọc cộng tác để khám phá bài báo trích dẫn tiềm năng và dùng các bài trích dẫn tiềm năng để tinh chỉnh việc dùng bài báo trích dẫn để mô hình hóa bài báo ứng viên. Kết quả thực nghiệm cho thấy việc khai thác bài báo trích dẫn tiềm năng đã cải tiến độ chính xác khuyến nghị [6].

Trong một nghiên cứu khác, Jianshan Sun và cộng sự, 2013 đã đề xuất các phương pháp mới cho khuyến nghị bài báo khoa học liên quan đến quan tâm nghiên cứu của nhà nghiên cứu bằng cách kết hợp thông tin nội dung của các bài báo quan tâm và các mối quan hệ xã hội của nhà nghiên cứu [7]. Họ đã rút trích danh sách các bài báo liên quan và các mối quan hệ xã hội của những nhà nghiên cứu từ trang mạng trực tuyến CiteULike[2] để xây dựng tập dữ liệu thực nghiệm bao gồm tập đánh giá (ground truth), tập huấn luyện (training set), cũng như tập kiểm tra (testing set). Kết quả thực nghiệm cho thấy phương pháp kết hợp thông tin nội dung và quan hệ xã hội rút trích từ các mạng trực tuyến CiteULike đã cải tiến chất lượng khuyến nghị so với phương pháp tiếp cận nội dung.

Joeran Beel và cộng sự, 2013 đã thực hiện một khảo sát hơn 170 bài báo, bằng sang chế, trang web được công bố trong lĩnh vực này và đã chỉ ra rằng: cho đến bây giờ vẫn chưa có sự đồng thuận, thống nhất về các tập dữ liệu cũng như phương pháp đánh giá khi thực hiện so sánh các phương pháp khuyến nghị bài báo khoa học khác nhau [1]. Điều đó dẫn đến một tình trạng, khó khăn chung, đó là chưa thể biết được những điểm mạnh và yếu thật sự của những phương pháp đề xuất hiện có.

Hiện nay, các công trình nghiên cứu của Sugiyama và cộng sự, 2010-2013 [4-6], Jianshan Sun và cộng sự, 2013 [7], là các nghiên cứu tương tự nhất với vấn đề mà chúng tôi đang nghiên cứu và trình bày trong bài báo này. Tuy nhiên, hầu hết các nghiên cứu này chưa thật sự quan tâm đến các mối quan hệ xã hội tiềm ẩn, cụ thể là quan hệ lòng tin khi thực hiện khuyến nghị bài báo khoa học cho nhà nghiên cứu.

Lòng tin (trust) có thể xem là thuộc tính của quan hệ xã hội. Theo Touhid Bhuiyan, 2013 [22], có nhiều định nghĩa khác nhau cho khái niệm lòng tin, nhưng định nghĩa được đa số cộng đồng trích dẫn và sử dụng là định nghĩa của nhà xã hội học Dasgupta. Lòng tin là sự mong đợi của một người về những hành động của người khác mà có ảnh hưởng đến quyết định, lựa chọn của họ [19]. Theo Piotr Sztompka, 1999 [25], lòng tin gồm hai thành phần chính là tin tưởng (belief) và cam kết (commitment). Tức một người sẽ tin tưởng rằng một người khác sẽ hành động theo một cách nhất định và đặt lòng tin vào họ, nhưng sự tin tưởng không thôi thì chưa đủ để có lòng tin. Lòng tin được đặt vào một ai đó khi sự tin tưởng đạt tới mức độ làm nền tảng cho một cam kết thực hiện một hành động cụ thể. Gần đây, lòng tin đã trở thành một chủ đề nghiên cứu quan trọng trong nhiều lĩnh vực như: xã hội học, tâm lý học, và cả tin học.

Stephen Marsh là một trong những người đi tiên phong trong việc khai thác lòng tin trong tính toán khoa học [18]. Gần đây, lòng tin đã thu hút nhiều quan tâm nghiên cứu của cộng đồng trong việc phát triển các hệ thống khuyến nghị trực tuyến. Người dùng thường sẽ tin tưởng và dễ dàng chấp nhận các khuyến nghị từ bạn bè, người thân hơn là những người lạ khác, ngay cả khi hệ khuyến nghị có những đề xuất hữu ích và chất lượng. Bên cạnh đó, lòng tin được sử dụng để cải tiến các phương pháp khuyến nghị truyền thống. Việc sử dụng quan hệ lòng tin giúp các hệ khuyến nghị có thể đương đầu với những khó khăn, thách thức như: ma trận đánh giá thưa, khởi động lạnh (cold-start).

[2] http://www.citeulike.org/

Paolo Massa và Paolo Avesani đã đề xuất thay thế bước tính toán tương tự người dùng trên ma trận đánh giá bằng độ đo lòng tin giữa những người. Họ đề xuất thuật toán lan truyền lòng tin trên mạng và tính mức độ lòng tin giữa những người dùng. Kết quả thực nghiệm trên tập dữ liệu Epinions cho thấy việc khai thác lòng tin cải tiến độ chính xác khuyến nghị [20]. Hao Ma và cộng sự đã nghiên cứu đề xuất phương pháp tối ưu dựa trên kết hợp cả các mối quan hệ lòng tin và không tin (distrust) nhằm cung cấp các khuyến nghị chính xác và thực tế cho người dùng. Nhóm tác giả cũng đã thực nghiệm trên tập dữ liệu Epinions và cho thấy hương pháp của họ tốt hơn hẳn các phương pháp hiện có trên tập dữ liệu này [21]. Lahiru S. Gallege và cộng sự đã nghiên cứu khai thác lòng tin để hướng đến phát triển hệ khuyến nghị cho các dịch phần mềm trực tuyến [23].

Trong lĩnh vực học thuật, theo hiểu biết của chúng tôi thì khái niệm lòng tin chưa được được đề cập và khai thác để phát triển các phương pháp khuyến nghị nhằm hỗ trợ các nhà nghiên cứu tìm kiếm thông tin. Vì vậy, bài báo này đề xuất khái niệm lòng tin trong lĩnh vực học thuật và khai thác quan hệ lòng tin của các nhà nghiên cứu để phát triển các phương pháp cho khuyến nghị bài báo khoa học. Phần tiếp theo trình bày chi tiết các phương pháp phổ biến, cũng như phương pháp đề xuất.

III. CÁC PHƯƠNG PHÁP KHUYẾN NGHỊ BÀI BÁO

III.1. Tiếp cận nội dung (CB)

Tiếp cận nội dung được đánh giá là tiếp cận phù hợp nhất cho các đối tượng khuyến nghị dạng văn bản [8]. Với tiếp cận nội dung, vector biểu diễn hồ sơ nghiên cứu của các nhà nghiên cứu và vector biểu diễn nội dung bài báo sẽ được xây dựng và so khớp.

Phương pháp 1 (CB): Phương pháp mô hình hóa sở thích của nhà nghiên cứu dựa trên nội dung các bài báo đã công bố được dủng như phương pháp cơ sở (base line) để so sánh với các phương pháp đề xuất.

Phương pháp 1: CB

Đầu vào:

R = {r} tập các nhà nghiên cứu quan sát được

P = {p} tập bài báo của các nhà nghiên cứu.

Đầu ra: $\forall r \in R$, trả về Top-N những $p \in P$.

- Bước 1: Tiền xử lý các bài báo $p \in P$
  - Rút trích phần tiêu đề và tóm tắt.
  - Loại bỏ stopwords, và stemming.
- Bước 2: Vector hóa nội dung các bài báo dùng TFIDF
  - $\forall p \in P$: xây dựng vector biểu diễn nội dung bài báo p là $\overrightarrow{f_p}$ dùng phương pháp gán trọng số TFIDF.
- Bước 3: Vector hóa sở thích nhà nghiên cứu
  - $\forall r \in R$: xây dựng vector profile $\overrightarrow{P_r}$ cho mỗi nhà nghiên cứu *r* dựa vào các bài báo mà *r* đã công bố.

$$\overrightarrow{P_r} = \sum_{i=1}^{n} \overrightarrow{f_{p_i}} \quad (1)$$

Trong đó, *n: Tổng số bài báo mà r đã công bố.*

- Bước 4: So khớp nội dung bài báo với sở thích của nhà nghiên cứu

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

$$Sim_{CB}(r,p) = Cosine(\overrightarrow{w_p}, \overrightarrow{P_r}) \quad (2)$$

- Xếp hạng và chọn TopN những bài báo có độ tương tự cao nhất với *r*, mà *r* chưa biết đến trước đây để thực hiện khuyến nghị cho *r*.

Cuối lặp.

Phương pháp 2 (CB+R+C): Mô hình hóa sở thích của các nhà nghiên cứu **dựa trên nội dung các bài báo công bố, tham khảo, và trích dẫn.**

Phương pháp này được đề xuất bởi Sugiyama và cộng sự, 2010 [4]. Họ quan niệm, quan tâm nghiên cứu của nhà nghiên cứu không chỉ thể hiện thông qua nội dung của các bài báo mà họ công bố, mà còn được thể hiện thông qua nội dung của các bài báo mà họ tham khảo (ký hiệu R), được trích dẫn (ký hiệu C). Do đó, Sugiyama và cộng sự đã tổng hợp vector đặc trưng của tất cả các bài báo công bố kết hợp với vector đặc của bài tham khảo, trích dẫn để mô hình hoá quan tâm nghiên cứu của các nhà nghiên cứu. Phương pháp CB+R+C có thể tóm tắt như sau:

Phương pháp 2: CB+R+C
Đầu vào:
R = {r} tập các nhà nghiên cứu quan sát được
P = {p} tập bài báo của các nhà nghiên cứu.
Đầu ra: $\forall r \in$ R, trả về Top-N những $p \in$ P.

- Bước 1: Tương tự phương pháp 1.
- Bước 2: Mô hình hóa nội dung bài báo.

$$\overrightarrow{F_p} = \overrightarrow{f_p} + \sum_{i=1}^{m} Sim(p, Ref(p)_i) * \overrightarrow{f_{Ref(p)_i}} + \sum_{i=1}^{n} Sim(p, Cite(p)_i) * \overrightarrow{f_{Cite(p)_i}} \quad (3)$$

Trong đó,
m: Tổng số bài mà p đã tham khảo,
n: Tổng số bài đã trích dẫn bài p,
$Ref(p)_i$:bài báo tham khảo thứ i của p,
$Cite(p)_i$:bài báo thứ i đã trích dẫn bài p.

- Bước 3: Vector hóa sở thích nhà nghiên cứu
  - $\forall r \in$ R: xây dựng vector profile $\overrightarrow{P_r}$

$$\overrightarrow{P_r} = \sum_{i=1}^{n} \overrightarrow{F_{p_i}} \quad (4)$$

n: Tổng số bài báo mà r đã công bố.

- Bước 4: Tương tự phương pháp 1.

Để lọc bớt những bài báo không liên quan khi xem xét các bài báo tham khảo và trích dẫn, Sugiyama và cộng sự, 2010 đã đề xuất sử dụng một tham số ngưỡng tương tự ($Th_j \in [0,1]$) để quyết định chọn ra những bài tham khảo, trích dẫn dùng để kết hợp với các bài báo khác khi xây dựng mô hình sở thích của nhà nghiên cứu [4]. Tức $Sim(p, Ref(p)_i) > Th_j$, $Sim(p, Cite(p)_i) > Th_j$, thì khi đó vector đặc trưng của $Ref(p)_i$ và $Cite(p)_i$ sẽ được kết hợp với vector đặc trưng của p.

Phương pháp 3 (CB-Recent): Khuyến nghị dựa trên sở thích gần đây của nhà nghiên cứu**.**

Các phương pháp mô hình hóa sở thích của các nhà nghiên cứu thông thường chỉ tập trung vào việc mã hóa nội dung các bài báo mà họ công bố, tham khảo hoặc được trích dẫn. Trên thực tế, sở thích của người dùng sẽ dần thay đổi theo thời gian. Sugiyama và cộng sự, 2010 cũng đã phát triển các phương pháp mô hình sở thích nghiên cứu gần đây của nhà nghiên cứu cho khuyến nghị bài báo khoa học [4]. Các bước thực hiện có thể tóm tắt như sau:

Phương pháp 3: CB-Recent
Đầu vào:
R = {r} tập các nhà nghiên cứu quan sát được
P = {p} tập bài báo của các nhà nghiên cứu.
Đầu ra: $\forall r \in$ R, trả về Top-N những $p \in$ P.

Các bước thực hiện:

- Bước 1: Tương tự phương pháp 2.
- Bước 2: Tương tự phương pháp 2.
- Bước 3: Vector hóa sở thích nhà nghiên cứu dựa trên xu hướng
  - $\forall r \in$ R: xây dựng vector profile $\overrightarrow{P_r}$ cho mỗi nhà nghiên cứu r.

$$\overrightarrow{P_r} = \sum_{i=1}^{n} e^{\alpha * (t_{cur} - t(p_i))} * \overrightarrow{F_{p_i}} \quad (5)$$

Trong đó,
α: hệ số ảnh hưởng của yếu tố xu hướng. ($\alpha \in [0,1]$. Trường hợp đơn giản $\alpha = 1$)
t_{cur}: năm hiện tại thực hiện khuyến nghị.
$t(p_i)$: năm công bố của bài báo p_i.
n: Tổng số bài báo mà r công bố trong quá khứ.

- Bước 4: Tương tự phương pháp 2.

III.2. Tiếp cận lọc cộng tác (CF)

Khác với tiếp cận nội dung, tiếp cận lọc cộng tác (tiếp cận CF) không bị hạn chế về mặt phân tích nội dung văn bản. Những phương pháp CF dùng thông tin từ ma trận đánh giá quan sát được từ người dùng và đối tượng khuyến nghị. Tiếp cận CF có thể áp dụng cho nhiều dạng đối tượng, nhiều kiểu nội dung khác nhau, ngay cả với những đối tượng khuyến nghị không tương tự với những đối tượng quan sát trong quá khứ. Theo Su & Khoshgoftaar, 2009, các phương pháp CF được đánh giá là các phương pháp thành công nhất trong việc xây dựng các hệ thống khuyến nghị [11].

Với bài toán khuyến nghị bài báo khoa học liên quan cho các nhà nghiên cứu, giả sử các bài báo được các nhà nghiên cứu tham khảo, trích dẫn là các bài có liên quan đến quan tâm nghiên cứu của họ. Khi đó, chúng ta có thể xây dựng ma trận đánh giá M dựa trên quan hệ trích dẫn, nhằm thể hiện sự quan tâm của các nhà nghiên cứu đối với các bài báo trong kho dữ liệu.

M có dòng là các nhà nghiên cứu và cột là các bài báo. Giá trị $v(i,j)$ ở dòng i, cột j trong ma trận M thể hiện sự quan tâm của researche r_i với bài báo p_j

$$v(i,j) = \frac{CitationCount(r_i,p_j)}{TotalCitation(r_i)} \quad (6)$$

$CitationCount(r_i,p_j)$: số lần mà nhà nghiên cứu r_i đã trích dẫn bài báo p_j trong quá khứ.

$TotalCitation(r_i)$: tổng số trích dẫn của r_i

Dựa trên quan điểm này, chúng ta có thể xây dựng phương pháp lọc cộng tác cho bài toán khuyến nghị bài báo khoa học liên quan.

Phương pháp 4 (CF-kNN): tiên đoán mức độ liên quan của các bài báo khoa học với các nhà nghiên cứu dựa trên tiếp cận CF, có thể tóm tắt như sau:

Phương pháp 4: CF-kNN
Đầu vào:
R = {r} tập các nhà nghiên cứu quan sát được
P = {p} tập bài báo của các nhà nghiên cứu.
Đầu ra: $\forall r \in$ R, trả về Top-N những $p \in$ P.

Các bước thực hiện:

- Bước 1: Xây dựng ma trận M có giá trị tại dòng i, cột j thể hiện mức độ liên quan của các $p_j \in P$ với $r_i \in R$, $v(r_i,p_j)$.
- Bước 2: Xác định những người đồng sở thích, và tiên đoán các giá trị $v(i,j)$ còn lại chưa xác định trong M

Lặp: $\forall r_i \in R$

- Dùng thuật toán kNN để xác định k người có sở thích tương tự r_i. Độ tương tự của $r \in R$ với r_i có thể tính theo hệ số tương quan Pearson dựa trên ma trận M như sau:

$$Sim_{Pearson}(r_i,r) = \frac{\sum_{p_j \in P_{r,r_i}} (v(r_i,p_j) - \overline{v_{r_i}}) * (v(r,p_j) - \overline{v_r})}{\sqrt{\sum_{p \in P_{r,r_i}} (v(r_i,p_j) - \overline{v_{r_i}})^2} * \sqrt{\sum_{p \in P_{r,r_i}} (v(r,p_j) - \overline{v_r})^2}} \quad (7)$$

Trong đó,

P_{r,r_i}: Tập các bài báo mà r, r_i đồng trích dẫn trong quá khứ.

$\overline{v_{r_i}}$: giá trị trung bình trích dẫn của nhà nghiên cứu r_i trên các bài báo p_j.

- Tổng hợp giá trị từ k người đồng sở thích, để tiên đoán những giá trị $v(i,j)$ chưa xác định trong M.
- Lặp: $\forall v(r_i,p_j) = 0$

$$v(r_i,p_j) = k * \sum_{r' \in kNN(r_i)} Sim(r',r_i) * v(r',p_j) \quad (8)$$

Trong đó,

$kNN(r_i)$: Tập k lân cận gần nhất của r_i
k: hệ số chuẩn hóa,
$k = 1/\sum_{r' \in kNN(r_i)} |Sim(r',r_i)|$
Cuối lặp.

- Chọn ra TopN những $v(r_i,p_j)$ chưa xác định để khuyến nghị cho r_i.
(Không khuyến nghị lại các bài báo p_j mà r_i đã biết)

Cuối lặp.

Mặc dù được đánh giá là tiếp cận thành công trong việc phát triển các phương pháp, hệ thống khuyến nghị, nhưng các phương pháp CF cũng có những hạn chế của nó. Adomavicius & Tuzhilin, 2005 [8], Bobadilla và cộng sự, 2013 [9], đã chỉ ra những hạn chế của các phương pháp CF như sau:

- Ma trận đánh giá thưa: ảnh hưởng nhiều đến việc phân tích ma trận để tiên đoán những giá trị đánh giá chưa xác định trong ma trận.
- Đối tượng khuyến nghị mới: không thể thực hiện khuyến nghị cho người dùng những đối tượng khuyến nghị mới. Tức đối tượng khuyến nghị chưa được ai quan tâm đánh giá, mặc dù có thể đối tượng mới đó rất gần với sở thích của người dùng.
- Người dùng mới: không thể khuyến nghị cho những người dùng mới chưa có thông tin quan sát trong ma trận đánh giá.

Việc áp dụng tiếp cận CF cho bài toán khuyến nghị bài báo khoa học liên quan đã gặp phải những hạn chế đã đề cập, đặc biệt ma trận đánh giá thể hiện sự quan tâm của các nhà nghiên cứu với các đối tượng khuyến nghị bài báo khoa học là một ma trận rất thưa. Như vậy, mặc dù rất tiềm năng nhưng tiếp cận CF không phải là tiếp cận phù hợp cho bài toán khuyến nghị bài báo khoa học liên quan cho các nhà nghiên cứu.

III.3. Kết hợp tuyến tính CB-Recent và CF-kNN

Hình thức kết hợp đơn giản nhất là kết hợp tuyến tính kết quả của CB-Recent và CF-kNN.

Phương pháp 5: (CB-Recent+CF) kết hợp tuyến tính CB và CF

$$Sim_{Hybrid}(r_i, p_j) = \alpha * Sim_{CB}(r_i, p_j) + (1-\alpha) * v(r_i, p_j)) \quad (9)$$
$$(\forall r_i \in R,\ \forall p_j \in P)$$

IV. ĐỀ XUẤT CÁC PHƯƠNG PHÁP KHAI THÁC QUAN HỆ LÒNG TIN CỦA CÁC NHÀ NGHIÊN CỨU.

Lòng tin đã thu hút nhiều quan tâm nghiên cứu của cộng đồng trong việc phát triển các hệ thống khuyến nghị trực tuyến, như các hệ thống khuyến nghị phim FilmTrust [24], hệ khuyến nghị sản phẩm Epinions[3]. Tuy nhiên Trong lĩnh vực học thuật, theo hiểu biết của chúng tôi thì khái niệm lòng tin chưa được được đề cập và khai thác để phát triển các phương pháp khuyến nghị nhằm hỗ trợ các nhà nghiên cứu tìm kiếm thông tin.

Việc chọn một bài báo để tham khảo, bên cạnh yếu tố nội dung bài báo có liên quan, các nhà nghiên cứu còn quan tâm đến uy tín của những tác giả của bài báo đó. Hay nói cách khác nhà nghiên cứu đang đặt lòng tin vào một số nhà nghiên cứu, chuyên gia uy tín khác trong lĩnh vực. Đây là những khiếm khuyết của các phương pháp phổ biến hiện nay. Ở đây, chúng tôi đề xuất kết hợp khai thác nội dung bài báo với các quan hệ lòng tin của nhà nghiên cứu để phát triển các phương pháp mới cho khuyến nghị bài báo khoa học tiềm năng cho nhà nghiên cứu.

IV.1. Phương pháp 6: Lòng tin dựa trên quan hệ đồng tác giả và quan hệ trích dẫn (CB-RecentTrust1)

Giả sử rằng, lòng tin của một nhà nghiên cứu đối với một bài báo phụ thuộc vào mức độ lòng tin của chính nhà nghiên cứu đó kết hợp với lòng tin của những đồng tác giả của họ đối với việc trích dẫn các tác giả của bài báo đang xem xét. Chi tiết phương pháp có thể tóm tắt qua các bước sau:

Phương pháp 6: CB-RecentTrust1
Đầu vào:
R = {r} tập các nhà nghiên cứu quan sát được
P = {p} tập bài báo của các nhà nghiên cứu.
Đầu ra: $\forall r \in$ R, trả về Top-N những $p \in$ P.

Bước 1: Xây dựng mạng trích dẫn CiNet_Author gồm 2 thành phần chính là A, R.
CiNet_Author (A, R).
- A: Tập các đỉnh, mỗi đỉnh là một nhà nghiên cứu
- R: Tập các cạnh (cặp đỉnh) có hướng thể hiện quan hệ trích dẫn, hướng từ x→ y thể hiện quan hệ x đã trích dẫn y, hay x đặt lòng tin lên y, khi trích dẫn y. Trọng số của cạnh có thể lượng hóa như sau:

$$w_{cite}(a_i, a_j, t_0) = \frac{\sum_{t_i=t_0}^{t_{cur}} NumCitation(a_i, a_j, t_i)}{e^{\gamma*(t_{cur}-t_i)} * TotalCitation(a_i, t_0)} \quad (10)$$

Trong đó,
- $NumCitation(a_i, a_j, t_i)$: Số lần mà a_i đã trích dẫn a_j trong năm t_i.
- $TotalCitation(a_i, t_0)$: Tổng số trích dẫn của a_i tính từ thời điểm t_0 đến thời điểm hiện tại
- t_{cur}: năm hiện tại
- t_0: thời điểm bắt đầu xem xét yếu tố xu hướng.
- γ: hệ số xu hướng. (trường hợp đơn giản γ=1)

Bước 2: Xây dựng mạng đồng tác giả CoNet (A, R).
- A: Tập các đỉnh, mỗi đỉnh là một nhà nghiên cứu
- R: tập các cặp đỉnh có hướng thể hiện quan hệ đồng tác giả, hướng từ x→ y thể hiện quan hệ x đồng tác giả với y.

Bước 3: Kết hợp quan hệ trích dẫn của tác giả a_i với quan hệ trích dẫn của các đồng tác giả của a_i để lượng hóa quan hệ lòng tin giữa 2 nhà nghiên cứu là a_i và a_j tính từ thời điểm t_0, $w_{trust}(a_i, a_j, t_0)$

$$w_{trust}(a_i, a_j, t_0) = w_{cite}(a_i, a_j, t_0) + \frac{\sum_{a_u \in CoAuthor(a_i)} w_{coauthor}(a_i, a_u, t_0) * w_{cite}(a_u, a_j, t_0)}{|CoAuthor(a_i)|} \quad (11)$$

Bước 4: Lượng hóa mức độ tin tưởng của một nhà nghiên cứu a_i với bài báo p_j:

$$w_{trust}(a_i, p_j, t_0) = MAX\left(w_{trust}(a_i, a_j, t_0)\right) \quad (12)$$

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

Bước 5: Kết hợp trọng số lòng tin với độ tương tự sở thích nghiên cứu gần đây của nhà nghiên cứu.
Lặp $\forall a_i \in R,\ \forall p_j \in P$

$$RatingValue(a_i, p_j) = \alpha * w_{trust}(a_i, p_j, t_0) + (1-\alpha) * Sim_{CB}(a_i, p_j, t_0) \quad (13)$$

Bước 6: Với mỗi a_i $a_i \in R$, lấy Top-N bài báo tiềm năng có $RatingValue(a_i, p_j)$ cao nhất để khuyến nghị cho a_i.

[3] www.epinions.com

IV.2. Phương pháp 7: Lòng tin dựa trên quan hệ trích dẫn tiềm ẩn. (CB-RecentTrust2)

Trên thực tế, một nhà nghiên cứu thường sẽ lần theo các bài báo trong mục tham khảo của các bài báo mà họ quan tâm để tìm kiếm các bài báo tiềm năng liên quan. Hành động đó thể hiện một quan hệ trích dẫn tiềm ẩn của các nhà nghiên cứu đối với các bài báo liên quan dựa trên việc bắt cầu quan hệ trích dẫn. Nếu xét ở góc độ lòng tin, có thể nói, nhà nghiên cứu có thể đặt lòng tin vào những nhà nghiên cứu khác dựa trên việc bắt cầu quan hệ lòng tin. Chi tiết của phương pháp khai thác quan hệ lòng tin dựa trên quan hệ trích dẫn tiềm ẩn có thể tóm tắt như sau:

Phương pháp 7: CB-RecentTrust2
Đầu vào:
R = {r} tập các nhà nghiên cứu quan sát được
P = {p} tập bài báo của các nhà nghiên cứu.
Đầu ra: $\forall r \in$ R, trả về Top-N những $p \in$ P.

Bước 1: Tương tự phương pháp 6.
Bước 2: Tổng hợp quan hệ trích dẫn của tác giả a_i với quan hệ trích dẫn của các tác giả mà a_i đã trích dẫn để lượng hóa quan hệ lòng tin giữa 2 nhà nghiên cứu là a_i và a_j tính từ thời điểm t_0, $w_{trust}(a_i, a_j, t_0)$

$$w_{trust}(a_i, a_j, t_0) = w_{cite}(a_i, a_j, t_0) + * \frac{\sum_{a_u \in CitedAuthor(a_i)} w_{Cite}(a_i, a_u, t_0) * w_{cite}(a_u, a_j, t_0)}{|CitedAuthor(a_i)|} \quad (14)$$

Bước 3: Áp dụng tiếp bước 4, 5, 6 phương pháp 6.

V. THỰC NGHIỆM ĐÁNH GIÁ VÀ THẢO LUẬN

Phần này trình bày kết quả đánh giá, so sánh các phương pháp khác nhau cho khuyến nghị bài báo khoa học liên quan cho nhà nghiên cứu trên tập dữ liệu lớn thu thập từ trang web Microsoft Academic Search.

V.1. Tập dữ liệu và thiết lập thực nghiệm

Joeran Beel và cộng sự, 2013, đã chỉ ra rằng: đến bây giờ vẫn chưa có sự thống nhất về các tập dữ liệu cũng như phương pháp đánh giá khi thực hiện so sánh các phương pháp khác nhau cho khuyến nghị bài báo khoa học [1]. Trong nghiên cứu này, chúng tôi đã thu thập thông tin các bài báo khoa học từ trang Microsoft Academic Search để xây dựng tập dữ liệu thực nghiệm. Để cùng góp phần với cộng đồng trong việc đa dạng, và dần chuẩn hóa các tập dữ liệu thực nghiệm cho bài toán này, chúng tôi đã phổ biến tập dữ liệu tại sites.google.com/site/tinhuynhuit/dataset.

Trong thực nghiệm, chọn ngẫu nhiên 1000 nhà nghiên cứu có bài báo công bố trước 2006 và sau 2006 như dữ liệu đầu vào. Các bài báo của họ công bố trước năm 2006 (xem như dữ liệu quá khứ) được chọn làm dữ liệu huấn luyện. Các bài báo được 1000 nhà nghiên cứu trích dẫn từ 2006 đến 2008 xem như dữ liệu trong tương lai làm Ground-Truth để kiểm chứng chất lượng các phương pháp khuyến nghị. Tức là, nếu phương pháp khuyến nghị một bài báo tiềm năng cho nhà nghiên cứu, mà trong tương lai nhà nghiên cứu có trích dẫn bài báo này thì xem như đó là một khuyến nghị đúng, ngược lại là sai. Ground-Truth bao gồm 52.254 bài được 1000 nhà nghiên cứu này trích dẫn trong năm từ 2006 đến 2008. Cách chia trục thời gian thành dữ liệu quá khứ và dữ liệu tương lai, sau đó dùng dữ liệu tương lai làm Ground-Truth để đánh giá chất lượng phương pháp khuyến nghị được áp dụng phổ biến trong những nghiên cứu hiện nay như J. Tang và cộng sự, 2012 [13], K. Sugiyama và cộng sự, 2010, 2013 [4,6], J. Sun và cộng sự, 2013 [7].

V.2. Độ đo đánh giá độ chính xác khuyến nghị

Thông thường, Top-N những đối tượng tiềm năng trả về từ hệ thống sẽ được dùng để đánh giá độ chính xác của phương pháp khuyến nghị. Hầu hết các độ đo đánh giá được dùng phổ biến trong các nghiên cứu hiện nay đều có nguồn gốc từ lĩnh vực truy vấn thông tin (IR). Tương tự các nghiên cứu của Sugiyama và cộng sự [4-6], ở đây chúng tôi tập trung phân tích kết quả thực nghiệm với độ đo NDCG [14] và MRR [15].

V.2.1. Độ đo NDCG (Normalized Discounted Cumulative Gain)

DCG là một độ đo liên quan đến chất lượng xếp hạng. DCG đo lường tính hữu ích của đối tượng dựa trên vị trí của nó trong danh sách xếp hạng trả về. Tính hữu ích sẽ được tích lũy từ đầu cho đến cuối

danh sách xếp hạng trả về. Và giá trị trung bình của DCG (tức NDCG) qua tất cả các người dùng sẽ được dùng để thể hiện độ chính xác khuyến nghị.

Ở đây chúng ta chỉ quan tâm TopN những kết quả trả về là có liên quan hay không liên quan. Vì vậy, NDCG@TopN được dùng để đánh giá. Với TopN là số lượng các bài báo trong danh sách xếp hạng được khuyến nghị cho các nhà nghiên cứu.

$$DCG(i)\begin{cases} G(1), nếu\ i = 1 \\ DCG(i-1) + \dfrac{G(i)}{\log(i)}, với\ i \neq 1 \end{cases} \quad (15)$$

Trong đó, *i* là vị trí xếp hạng thứ *i*. Ở đây *G(i)=1*, nếu kết quả khuyến nghị là liên quan (đúng), ngược lại *G(i)=0*.

V.2.2. Độ đo MRR (Mean Reciprocal Rank)

Reciprocal Rank (RR) là một độ đo xem xét vị trí xếp hạng của đối tượng liên quan đầu tiên được trả về. MRR là trung bình của RR thông qua nhiều truy vấn khác nhau. Hay trong bài toán của chúng ta MRR là trung bình kết quả khuyến nghị xét qua nhiều nhà nghiên cứu.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{Q}\frac{1}{Rank_i} \quad (16)$$

|Q|: Tổng số nhà nghiên cứu được thực hiện khuyến nghị

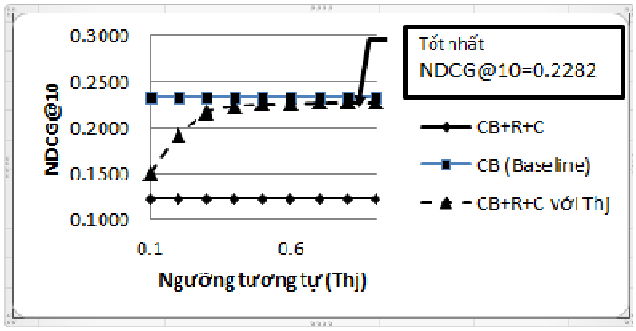
$Rank_i$: vị trí xuất hiện đầu tiên của bài báo khuyến nghị liên quan trong danh sách xếp hạng trả về.

V.3. Kết quả thực nghiệm

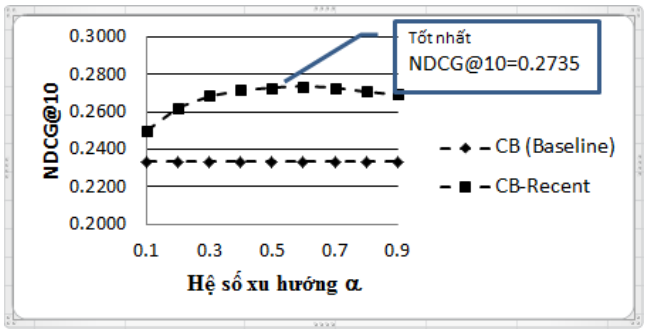
V.3.1. Phân tích các phương pháp phổ biến

Phần này trình bày kết quả thực nghiệm so sánh, phân tích các phương pháp phổ biến như phương pháp dựa trên nội dung như CB, CB+R+C, CB-Recent, phương pháp lọc công tác CF, phương pháp lai tuyến tính CB+CF.

Với phương pháp CB+R+C, để quyết định chọn bài báo tham khảo (R), trích dẫn (C) kết hợp với bài báo công bố dựa trên ngưỡng tương tự (Th_j), chúng tôi cũng đã tiến hành thay đổi Th_j, Th_j nhận các giá trị rời rạc 0.1, 0.2, …, 0.9. Kết quả tốt nhất đạt được tại Th_j = 0.8, với NDCG@10 = 0.2282, vẫn thấp hơn so với phương pháp cơ sở CB có NDCG@10=0.2334 (Hình 2).



Hình 2. Kết quả thực nghiệm phương pháp CB+R+C với tham số ngưỡng tương tự Th_j (Phương pháp 2)

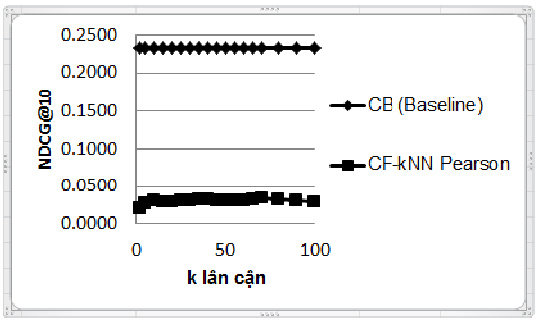


Hình 3. Kết quả thực nghiệm phương pháp CB-Recent với các hệ số xu hướng α khác nhau (Phương pháp 3)

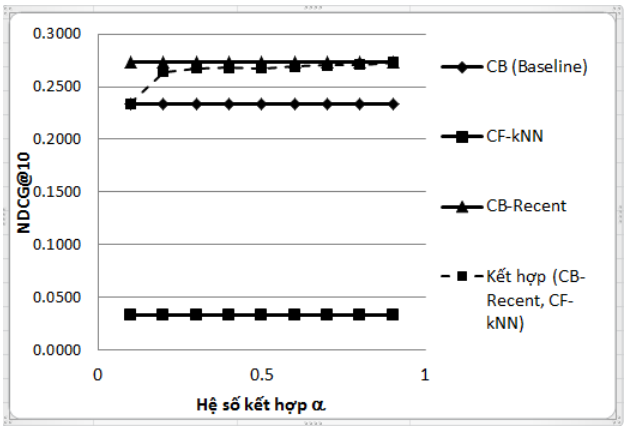
Phương pháp mô hình hóa sở thích của nhà nghiên cứu CB-Recent (phương pháp 3) đã cho thấy ưu điểm vượt trội, cải tiến đáng kể độ chính xác khuyến nghị trong thực nghiệm so sánh. Trong thực nghiệm với phương pháp 3, chúng tôi thay đổi hệ số xu hướng α, nhận các giá trị 0.1, 0.2, …, 0.9. Kết quả tốt nhất đạt được tại α = 0.6, với NDCG@10 = 0.2735, cao hơn hẳn so với phương pháp cơ sở CB với NDCG@10=0.2334 (Hình 3).

Đối với phương pháp lọc cộng tác CF, ở bước gom cụm các nhà nghiên cứu đồng sở thích với kNN, để chọn được giá trị k tốt nhất, chúng tôi đã tiến hành thay đổi k với các giá trị khác nhau từ 3 đến 100. Với mỗi giá trị k, chúng tôi xem xét độ chính xác khuyến nghị với độ đo NDCG@5, NDCG@10, MRR. Hình 4 là kết quả áp dụng phương pháp CF với các hệ số k khác nhau so với phương pháp nội dung CB

(Baseline). Kết quả thực nghiệm cho thấy tiếp cận lọc cộng tác CF không phải là tiếp cận phù hợp cho bài toán này. Ma trận trích dẫn quá thưa đã ảnh hưởng lớn đến độ chính xác của phương pháp lọc cộng tác. Việc kết hợp tuyến tính CB-Recent (tốt nhất trong các phương pháp CB) và CF-kNN cho kết quả trong Hình 5.



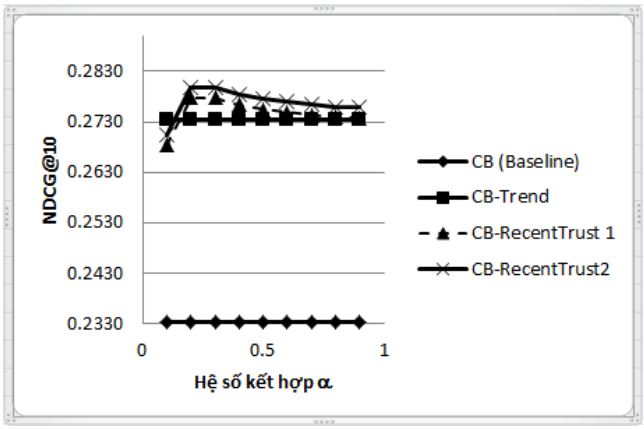
Hình 4. Kết quả thực nghiệm phương pháp lọc cộng tác CF-kNN với các giá trị k khác nhau (Phương pháp 4)



Hình 5. Kết quả thực nghiệm phương pháp kết hợp tuyến tính CB và CF (Phương pháp 5)

V.3.2. Phân tích các phương pháp đề xuất

Sau khi lượng hóa quan hệ lòng tin thông qua việc tổng hợp lòng tin của những quan hệ đồng tác giả và những quan hệ trích dẫn, chúng tôi kết hợp tuyến tính với xu hướng sở thích (CB-Recent). Trong thực nghiệm chúng tôi cho hệ số α nhận các giá trị rời rạc lần lượt là 0.1, 0.2, …, 0.9 để tìm giá trị tốt nhất cho kết hợp. Hình 6 trực quan kết quả kết hợp. Phương pháp CB-RecentTrust2 cho kết quả trội hơn so với CB-RecentTrust1 và CB-Recent.



Hình 6. Phương pháp kết hợp xu hướng sở thích và quan hệ lòng tin (CB-RecentTrust1 - Phương pháp 6)

Tổng hợp so sánh đánh giá các phương pháp đề xuất và phương pháp phổ biến hiện nay trình bày trong Bảng 1.

Bảng 1. Tóm tắt so sánh, đánh giá các phương pháp đề xuất và các phương pháp phổ biến hiện nay

Phương pháp khuyến nghị	**Độ đo đánh giá**		
	NDCG @5	**NDCG @10**	**MRR**
Phương pháp 1 (CB: Baseline)	0.2945	0.2334	0.5128
Phương pháp 2 (CB+R+C)	0.1464	0.1230	0.3061
Phương pháp 2 (CB+R+C, Th_j = 0.8)	0.2877	0.2282	0.4985
Phương pháp 3 (CB-Recent, α=0.6)	**0.3577**	**0.2735**	**0.6142**
Phương pháp 4 (CF-kNN với k=40)	0.0357	0.0330	0.0934
Phương pháp 5 (CB+CF, α=0.9)	**0.3570**	**0.2728**	**0.6140**
Phương pháp 6 (CB-RecentTrust1)	**0.3610**	**0.2778**	**0.6164**
Phương pháp 7 (CB-RecentTrust2)	**0.3617**	**0.2799**	**0.6169**

V.4. Nhận định và thảo luận

Thông qua kết quả thực nghiệm trên tập dữ liệu tương đối lớn, chúng ta có thể đưa ra các nhận định đối với các phương pháp khuyến nghị bài báo liên quan như sau:

- Tiếp cận lọc công tác CF cho thấy không phải là tiếp cận phù hợp cho bài toán này, trong khi tiếp cận nội dung là tiếp cận phù hợp nhất mà các nghiên cứu hiện nay đang thực hiện.

- Yếu tố xu hướng (sở thích gần đây của nhà nghiên cứu) đã cải tiến đáng kể độ chính xác khuyến nghị.
- Tiếp cận kết hợp nội dung với các mối quan hệ xã hội tiềm ẩn, cụ thể ở đây quan hệ lòng tin đã góp phần cải tiến độ chính xác khuyến nghị bài báo liên quan cho nhà nghiên cứu.

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Mục đích của nghiên cứu này là khảo sát, đánh giá các phương pháp khuyến nghị phổ biến, đồng thời đề xuất phương pháp mới cho bài toán khuyến nghị bài báo khoa học phù hợp với quan tâm nghiên cứu của nhà nghiên cứu. Các phương pháp đề xuất dựa trên việc kết hợp xu hướng sở thích nghiên cứu và quan hệ lòng tin của nhà nghiên cứu. Thực nghiệm tiến hành trên tập dữ liệu khoa học rút trích từ hệ thống Microsoft Academic Search. Độ chính xác khuyến nghị được đánh giá dựa trên dữ liệu gán nhãn (ground truth) là các bài báo mà nhà nghiên cứu tham khảo, trích dẫn trong tương lai, nhưng chưa trích dẫn trong quá khứ. Kết quả thực nghiệm cho thấy phương pháp đề xuất đã cải tiến độ chính xác khuyến nghị so với phương pháp cơ sở và các phương pháp phổ biến nhất hiện nay (Bảng 1).

Việc kết hợp quan hệ lòng tin với xu hướng sở thích của nhà nghiên cứu đã cải tiến độ chính xác khuyến nghị, nhưng kết quả đạt được chưa thật sự đáng kể. Bên cạnh đó, đôi khi nhà nghiên cứu còn quan tâm đến tính mới, cũng như chất lượng bài báo. Những vấn đề này đang được chúng tôi tiếp tục nghiên cứu giải quyết nhằm cung cấp các phương pháp khuyến nghị bài báo khoa học hiệu quả và phù hợp với nhu cầu thông tin của các nhà nghiên cứu.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (VNU-HCM) trong đề tài mã số C2014-26-03.

TÀI LIỆU THAM KHẢO

[1] J. BEEL, S. LANGER, M. GENZMEHR, B. GIPP, C. BREITINGER, A. NURNBERGER, "*Research paper recommender system evaluation: A quantitative literature survey*". In Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, RepSys '13, pages 15-22, New York, NY, USA, 2013.

[2] Q. HE, D. KIFER, J. PEI, P. MITRA, C. L. GILES, "*Citation recommendation without author supervision*", In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11, pages 755-764, New York, NY, USA, 2011.

[3] Q. HE, J. PEI, D. KIFER, P. MITRA, C. L. GILES, "*Context-aware citation recommendation*", In Proceedings of the 19th International Conference on World Wide Web, WWW '10, pages 421-430, New York, NY, USA, 2010.

[4] K. SUGIYAMA, M.-Y. KAN, "*Scholarly paper recommendation via user's recent research interests*", In Proceedings of the 10th Annual Joint Conference on Digital Libraries, JCDL '10, pages 29-38, New York, NY, USA, 2010.

[5] K. SUGIYAMA, M.-Y. KAN, "*Serendipitous recommendation for scholarly papers considering relations among researchers*". In Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, JCDL '11, pages 307-310, New York, NY, USA, 2011.

[6] K. SUGIYAMA, M.-Y. KAN, "*Exploiting potential citation papers in scholarly paper recommendation*", In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, pages 153-162, New York, NY, USA, 2013.

[7] J. SUN, J. MA, Z. LIU, Y. MIAO, "*Leveraging content and connections for scientific article recommendation in social computing contexts*", The Computer Journal, Oxford University Press, pages bxt086, 2013.

[8] G. ADOMAVICIUS, A. TUZHILIN, "*Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions*", IEEE Trans. on Knowl. and Data Eng., vol.17, No.6, pages 734-749, 2005.

[9] J. BOBADILLA, F. ORTEGA, A. HERNANDO, A. GUTIÉRREZ, "*Recommender systems survey*", *Know.-Based Syst*, vol.46 (July 2013), pages 109-132, 2013.

[10] T. HUYNH, K. HOANG, L. DO, H. TRAN, H. LUONG, S. GAUCH, "*Scientific Publication Recommendations Based on Collaborative Citation Networks*", In Proceedings of the 3rd International Workshop on Adaptive Collaboration (AC 2012) as part of The 2012 International Conference on Collaboration Technologies and Systems (CTS 2012). Colorado, USA, pages 316-321, 2012.

[11] X. SU, T.M. KHOSHGOFTAAR, "*A survey of collaborative filtering techniques*". *Adv. in Artif. Intell.* vol. 2009, no. 4, pages 2-2, 2009.

[12] K. STEFANIDIS, I. NTOUTSI, K. NØRVÅG, H.-P. KRIEGEL, "*A framework for time-aware recommendations*", In Proceedings of 23rd International Conference, DEXA 2012, Vienna, Austria, September 3-6, 2012, vol. 7447 of Lecture Notes in Computer Science, pages 329-344, 2012.

[13] J. TANG, S. WU, J. SUN, H. SU, "*Cross-domain collaboration recommendation*", In Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, KDD '12. New York, NY, USA: ACM, pages. 1285–1293, 2012.

[14] K. JÄRVELIN, J. KEKÄLÄINEN, "*IR Evaluation Methods for Retrieving Highly Relevant Documents"*, In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), pages 41–48, 2000.

[15] E. M. VOORHEES, "*The TREC-8 Question Answering Track Report*", In Proceedings of the 8th Text REtrieval Conference (TREC-8), pages 77–82, 1999.

[16] W. HUANG, S. KATARIA, C. CARAGEA, P. MITRA, C. L. GILES, L. ROKACH, "*Recommending citations: translating papers into references*", In Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12). ACM, New York, NY, USA, pages 1910-1914, 2012.

[17] S. LAWRENCE, C. L. GILES, K. BOLLACKER, "*Digital libraries and autonomous citation indexing*" Computer, vol. 32, pages 67–71, June 1999.

[18] S. P. MARSH, "*Formalising trust as a computational concept*", Ph.D. dissertation, University of Stirling, April 1994.

[19] P. DASGUPTA, "*Trust as a commodity*", In Trust: Making and Breaking Cooperative Relations, D. Gambetta, Ed. Blackwell, pages 49–72, 1988.

[20] P. MASSA, P. AVESANI, "*Trust-aware recommender systems*", In Proceedings of the 2007 ACM Conference on Recommender Systems, RecSys '07. New York, NY, USA, pages 17–24, 2007.

[21] H. MA, M. R. LYU, I. KING, "*Learning to recommend with trust and distrust relationships*", In Proceedings of the Third ACM Conference on Recommender Systems, RecSys '09, pages 189-196, New York, NY, USA, 2009.

[22] T. BHUIYAN, "*Trust for Intelligent Recommendation*", Springer Publishing Company, Incorporated, 2013.

[23] L. S. GALLEGE, D. U. GAMAGE, J. H. HILL, R. R. RAJE, "*Towards trust-based recommender systems for online software services*", In Proceedings of the 9th Annual Cyber and Information Security Research Conference, CISR'14, New York, NY, USA, pages 61–64, 2014.

[24] J. GOLBECK, J. HENDLER, "*Filmtrust: Movie recommendations using trust in web-based social networks*", In Proceedings of the IEEE Consumer communications and networking conference, CCNC 2006, vol. 96, pages 282-286, 2006.

[25] P. SZTOMPKA, "*Trust: A sociological theory"*, Cambridge University Press, 1999.

Nhận bài ngày: 23/8/2014.

SƠ LƯỢC VỀ TÁC GIẢ

HOÀNG KIẾM

Sinh ngày: 10/08/1950

Nhận học vị Tiến sỹ Khoa học năm 1992, được phong chức danh Giáo sư năm 1996.

Hiện đang công tác tại Trường Đại học CNTT, Đại học Quốc gia TP. HCM.

Lĩnh vực nghiên cứu: Công nghệ Tri thức và Ứng dụng, Khai thác Dữ liệu, Sinh Tin học, Dịch máy, Xử lý ảnh.

Email: kiemhv@uit.edu.vn

HUỲNH NGỌC TÍN

Sinh ngày: 31/03/1975

Tốt nghiệp Cử nhân Tin học năm 1998, Thạc sỹ năm 2003 tại Trường ĐH Khoa học Tự nhiên, ĐH Quốc gia TP. HCM

Hiện là giảng viên khoa Công nghệ Phần mềm, Trường ĐH CNTT, ĐH Quốc gia TP.HCM.

Lĩnh vực nghiên cứu: Khai thác dữ liệu văn bản, Tìm kiếm thông tin thông minh, Hệ khuyến nghị, Phân tích mạng xã hội.

Điện thoại: 0989241516; Email : tinhn@uit.edu.vn