TIẾN TỚI CÁC DỊCH VỤ THƯ VIỆN SỐ THÔNG MINH VÀ TÙY BIẾN

Md Maruf Hasan
Trường công nghệ, Đại học Shinawatra, Thái Lan

Ekawit Nantajeewarawat
Viện công nghệ Sirindhorn, Đại học Thammasat, Thái Lan

1. Mở đầu

Thư viện số là một sưu tập tài liệu dưới dạng điện tử có tổ chức và có thể truy cập được qua các giao diện tìm kiếm và lướt tìm [1]. Dựa vào một thư viện cụ thể, người dùng có thể truy cập các bài tạp chí, sách, báo cáo, hình ảnh, tệp âm thanh và băng đĩa hình. Các hệ thống thư viện số điển hình thường cung cấp các giao diện tìm kiếm và xem lướt. Tìm kiếm ngụ ý rằng, người dùng đã biết chính xác muốn tìm cái gì, còn lướt tìm cần trợ giúp người dùng lướt qua các thuật ngữ tìm kiếm tương quan để tìm ra cái mới hoặc đáng quan tâm. Giao diện tìm kiếm có thể đi từ phép tìm theo từ khóa cơ bản tới phép tìm nâng cao đặc thù cho một lĩnh vực, v.v…. Giao diện xem lướt bao gồm các hình thức tìm kiếm xác thực dựa trên các bảng phân loại và siêu dữ liệu nhất định như xem lướt theo tác giả, môn loại, v.v…. [2], [3].

Tìm kiếm thông tin trong khung cảnh thư viện số vốn khác với những gì trong phạm vi tìm kiếm trên Web hoặc tìm tin nói chung. Một đặc trưng phổ biến của những trường hợp sau là chúng không cung cấp bất kỳ sự trợ giúp cá biệt cho người dùng tin riêng lẻ hoặc chỉ hỗ trợ rất ít [3], [4]. Quả thực, điều làm cho thư viện số trở nên độc đáo chính là tính sẵn có của nội dung dưới dạng điện tử (mà có thể được xử lý một cách tự động và suy diễn được) và tính sẵn có của diện người dùng và các mẫu sử dụng. Trái với WWW, các phép tìm theo từ khóa kiểu Google hoặc theo thư mục (directory) kiểu Yahoo chắc chắn không thỏa đáng để khai thác thông tin trong thư viện số một cách hiệu quả.

Thách thức của việc tích hợp nội dung thư viện số với diện người dùng, mẫu sử dụng, v.v… có thể được giải quyết hiệu quả bằng cách sử dụng các thuật toán thông minh. Trong công trình nghiên cứu này, chúng tôi đưa ra *Kiến trúc thư viện số 3 lớp*, tạo điều kiện thuận lợi cho các dịch vụ thư viện số thông minh và tùy biến bằng cách tích hợp nội dung thư viện số với bản thể học lĩnh vực (domain-ontology), diện người dùng và mẫu sử dụng nhờ tận dụng các thuật toán và kỹ thuật thông minh một cách độc đáo.

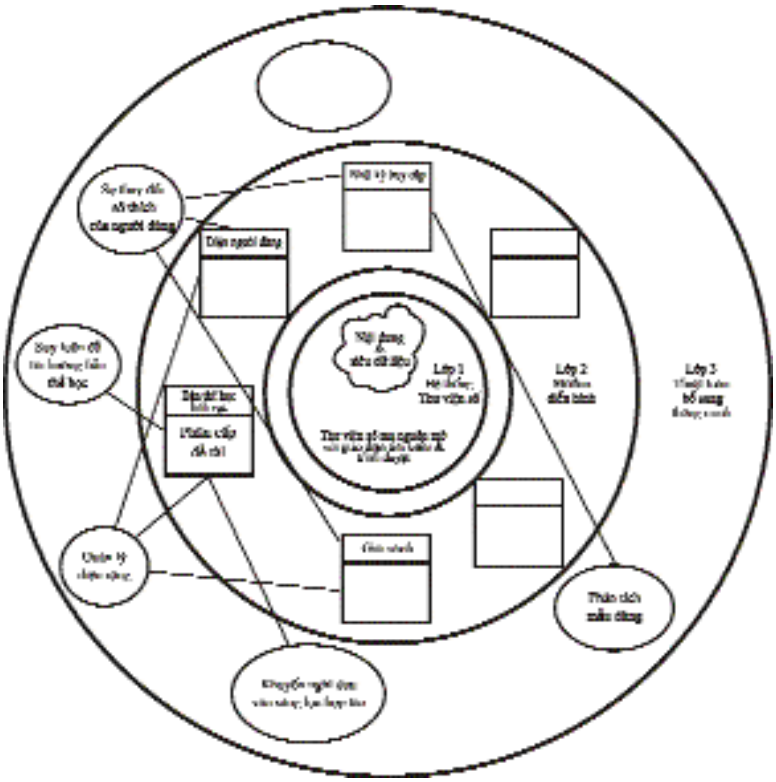
2. Kiến trúc thư viện số 3 lớp

Một số mô hình kiến trúc thư viện số và dịch vụ được mô tả trong tài liệu liên quan [5]. Trong những kiến trúc thư viện số 3 lớp được đề nghị, chúng tôi tập trung vào tính môđun và khả năng bảo trì. Cốt lõi của kiến trúc thư viện số 3 lớp là hệ thống thư viện số mã nguồn mở điển hình được bao quanh bởi một loạt môđun bổ sung để nắm bắt và trình bày thông tin tiếp theo về nội dung, người sử dụng và cách dùng. Lớp ngoài cùng gồm những dịch vụ được phát

triển để sử dụng một cách thông minh các thông tin do một hoặc nhiều môđun bổ sung ở lớp giữa trình bày và nắm bắt, và với sự trợ giúp của các thuật toán và kỹ thuật hữu hiệu (Hình 1).

Chúng tôi sử dụng bản thể học-lĩnh vực để dẫn giải và tổ chức các tài liệu thư viện số và để thể hiện diện người dùng. Chúng tôi cũng nắm nhật ký truy cập của người dùng và phân tích chúng để nhận được tần số đăng nhập, lịch sử tìm kiếm và lướt tìm ghi rõ thời gian, v.v… sao cho những thay đổi về thời gian trong các mẫu truy cập có thể được tính toán dễ dàng. Cần lưu ý rằng, các diện người dùng ban đầu phần lớn đều không đầy đủ và thiếu chính xác [3], [6]. Bằng cách ánh xạ và liên kết chúng với bản thể học lĩnh vực, chúng tôi tinh chỉnh lại diện người dùng ban đầu. Việc nâng cao hơn nữa diện người dùng được tiến hành liên tục bằng cách xem xét các dữ liệu sử dụng của người dùng. Qua việc phân tích các mẫu sử dụng và bằng cách nhận dạng các thay đổi theo thời gian trong sở thích của người dùng (nghĩa là mô hình hóa sở thích- dòng thời gian), chúng tôi tiếp tục nâng cao diện này, sử dụng phương pháp điều chỉnh tỉ trọng, hơi giống *Cơ chế kích hoạt lan tỏa* (Spreading Activation mechanism) [7]. Chúng tôi phác họa kiến trúc thư viện số 3 lớp trong Hình 1.

Hình 1 cho thấy sự thay đổi sở thích của người dùng theo thời gian được mô hình hóa bằng cách sử dụng nhật ký truy nhập, diện người dùng và nội dung giá sách; diện động được tính toán nhờ các dữ liệu diện, nội dung giá sách và suy luận đề tài định hướng bản thể luận,…



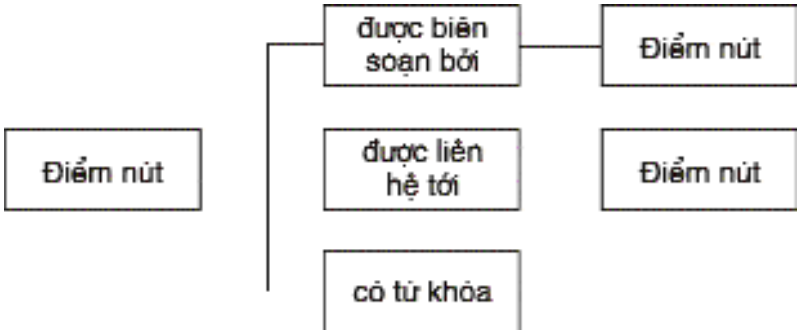
Hình 1. *Kiến trúc thư viện số 3 lớp cho các dịch vụ thư viện số thông minh và tùy biến*

3. Tài liệu và phương pháp

Trong công trình nghiên cứu, chúng tôi sử dụng hệ thống thư viện số Greenstone [8] trong Lớp 1. Một trong những môđun chính được bổ sung ở Lớp 2 là bản thể học lĩnh vực. Vì mục đích thí nghiệm thăm dò, chúng tôi tạo một bộ sưu tập thư viện số nhỏ với khoảng 300 tài liệu có liên quan đến tin học, bao quát một số đề tài nhỏ trong lĩnh vực tin học và được tổ chức/phân loại theo Hệ thống phân loại tính toán (ACM-CCS). Cả hai giao diện xem lướt và tìm kiếm đều sẵn sàng cho sưu tập nhỏ này. Chúng tôi đã chấp nhận và làm theo một phiên bản đã thay đổi về bản thể luận của ACM-CCS [9] trong việc tổ chức tài liệu thư viện số và thể hiện diện người dùng thư viện này.

Sơ đồ bản thể luận của ACM-CCS được trình bày trong Hình 2. Cần lưu ý rằng, thuộc từ hasKeyword (có từ khóa) rất có ích trong suy luận đề tài từ lịch sử sử dụng (nghĩa là ghi lại các từ khóa đã được dùng trong tìm tin) và thông qua việc phân tích toàn văn (nghĩa là tự động tách các cụm từ khóa bằng phương pháp xử lý ngôn ngữ tự nhiên). Chúng tôi sử dụng công cụ tách cụm từ khóa tự động (KEA) để tự động tách các cụm từ khóa cho sưu tập của chúng tôi [10]. Làm như vậy, chúng tôi có thể *gián tiếp* kết hợp môt đề tài này với một đề tài khác bằng cách tính mức độ tương tự của các từ khóa.

Các môđun then chốt khác trong Lớp 2 bao gồm môđun nhật ký truy cập của người dùng, là môđun ghi lại tần số đăng nhập và các hoạt động khác của người dùng như số lần tương tác giữa tìm kiếm và xem lướt trong từng phiên tìm (mà được sử dụng như là một tham số cho mô hình hóa sở thích-dòng thời gian sẽ được giải thích sau); Môđun giá sách-ghi lại thông tin về nội dung giá sách của mỗi người dùng, v.v...



Hình 2. *Bản thể học lĩnh vực : Sơ đồ cho Bản thể học ACM-CCS dựa vào phân loại CCS*

Cuối cùng, trong Lớp 3, chúng tôi xác định những dịch vụ thư viện số thông minh tận dụng thông tin nắm bắt được trong các lớp thấp hơn như được phác họa dưới đây.

3.1. Diện người dùng và quản lý diện năng động

Trong phạm vi của thư viện số, điển hình là việc người dùng được nhắc để

đặc tả các sở thích của họ khi họ đăng nhập lần đầu tiên. Người dùng trong nghiên cứu thử nghiệm của chúng tôi là những sinh viên và nhân viên tin học của khoa và người dùng thường xuyên thư viện số ACM, vì vậy, họ đã quen với lĩnh vực tin học và phân loại theo ACM-CCS. Chúng tôi giới thiệu hệ phân cấp các đề tài của ACM-CCS cho người dùng vào lúc đăng nhập và yêu cầu họ lựa chọn diện của mình bằng cách kiểm tra các đề tài thích hợp trong hệ phân cấp xem có thích hợp hay không. Tuy nhiên, diện người dùng *thô* như vậy thường không đầy đủ hoặc không chính xác và thường thay đổi theo thời gian [3], [6].

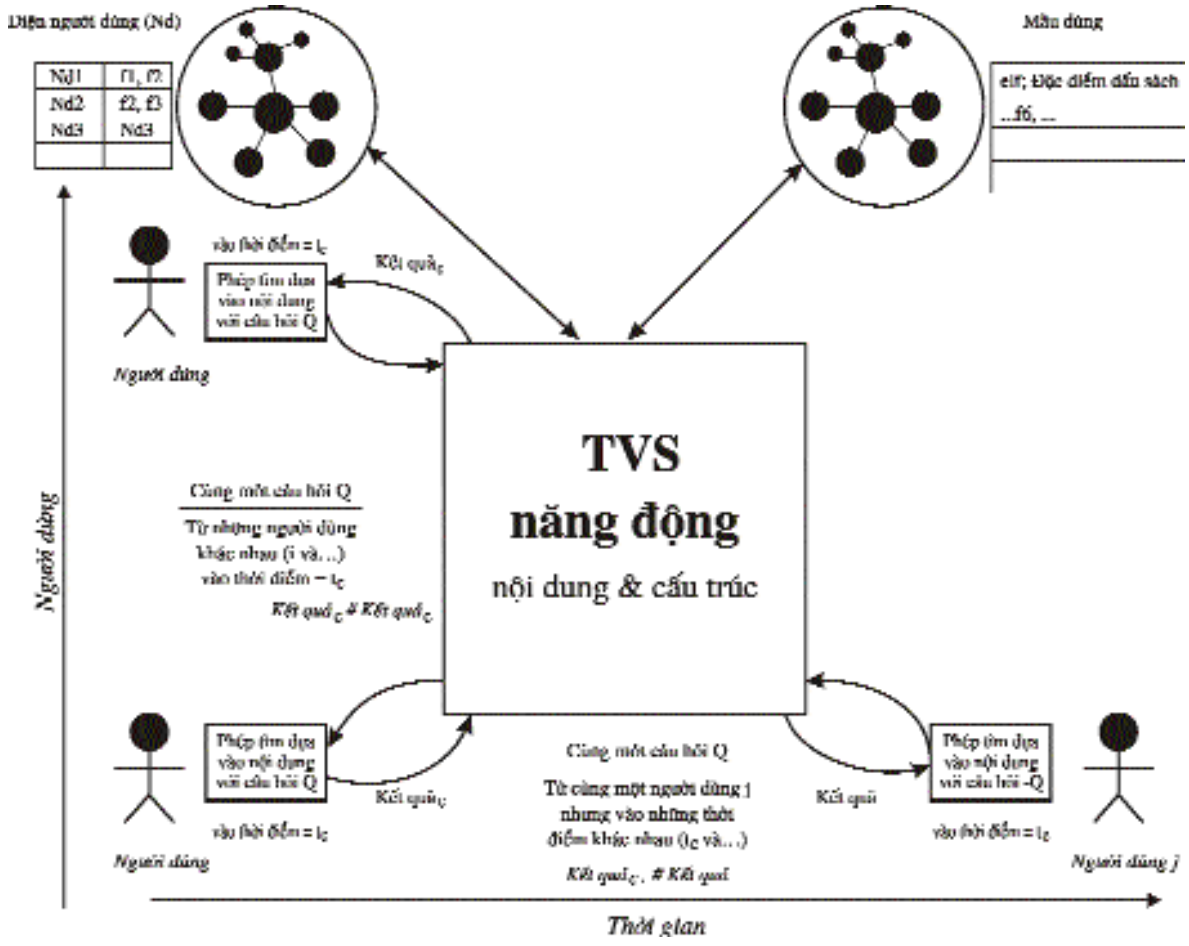
Chúng tôi làm tăng ngay lập tức diện người dùng *ban đầu* bằng cách chuẩn hóa và truyền tỉ trọng nhờ sử dụng một cơ chế giống như kích hoạt lan tỏa [7], xem xét mối tương quan giữa các đề tài. Chẳng hạn, khi chọn một đề tài không phát triển (không trổ lá - non leaf), chúng tôi truyền tỉ trọng chuẩn đồng nhất cho tất cả các đề tài nhánh (chồi con-child note) và tham chiếu qua lại (nếu có). Sự gia tăng như vậy khá thô thiển, và chúng tôi gọi nó là *diện người dùng thô*. Tuy nhiên, bằng cách xác định và lần theo những mối quan hệ nội dung và lịch sử sử dụng, chúng tôi tiếp tục tinh chỉnh diện người dùng bằng cách sử dụng cơ chế kích hoạt lan tỏa. Chẳng hạn, ghi lại những cuộc tìm kiếm theo từ khóa của người dùng và đối chiếu với danh sách từ khóa hướng đề tài để xác định lĩnh vực đề tài thích hợp nhất (tính tương thích được tính toán dựa trên tỉ số tương tự của từ khóa). Chúng tôi gọi những diện người dùng như vậy là *diện người dùng động*. Trong thí nghiệm của mình, chúng tôi tính lại diện người dùng một cách đoạn tuyến sau khi kết thúc một phiên đăng nhập. Một diện động như vậy phản ánh một cách chính xác có luận cứ sở thích và phạm vi của người dùng. Bằng cách giữ thêm những thông tin như nhật ký truy nhập của người dùng bao gồm cả tập hợp từ khóa của họ và những tài liệu có trên giá sách kèm theo dấu ấn thời gian, chúng tôi cũng thử mô hình hóa sự thay đổi sở thích của người dùng qua thời gian. Phương pháp hình thức hóa toán học sự thay đổi sở thích được giải thích dưới đây.

3.2. Mô hình hóa sự thay đổi sở thích của người dùng

Khi người sử dụng thư viện số thực hiện những nhiệm vụ mới hoặc chuyển sang những dự án mới, một thư viện số thông minh cần có khả năng thích nghi với những sự thay đổi như vậy về khung cảnh và sự ưu tiên [11], [12]. Nói cách khác, một người dùng tìm tin với *cùng một từ khóa nhưng vào một thời điểm khác* không nhất thiết phải nhận được cùng một kết quả tìm kiếm (ít nhất việc phân loại cấp độ tài liệu phải thay đổi). Tương tự như vậy, hai người

dùng khác nhau tìm kiếm thư viện số, sử dụng cùng một từ khóa ở cùng một thời điểm có thể không nhận được cùng một kết quả tìm tin. Ít nhất thì việc phân loại cấp độ kết quả tìm cũng phải chú ý tới diện của họ. Chúng tôi minh họa kịch bản thư viện số tùy biến như thế trong Hình 3.

Để đạt được tính tùy biến như vậy, chúng tôi sử dụng kỹ thuật *dung hòa theo số mũ đơn* [13] để mô hình hóa sự thay đổi sở thích trong thư viện số của chúng tôi.



Hình 3. *Những đặc điểm của dịch vụ thông minh và tùy biến trong thư viện số phản ánh sự thay đổi sở thích của người dùng (a) Trục hoành minh họa: kết quả tìm đối với một người dùng cụ thể sử dụng cùng một từ khóa ở thời điểm khác nhau sẽ cho những kết quả tìm khác nhau. (b) Trục tung minh họa: kết quả tìm đối với những người dùng khác nhau dựa trên cùng một từ khóa có thể cho những kết quả tìm khác nhau.*

Kỹ thuật dung hòa theo số mũ đơn đã được sử dụng rộng rãi trong dự báo tài chính và những ứng dụng kỹ thuật, nơi mà sự dự đoán dựa vào dung hòa là cơ bản. Trong một thư viện số, sở thích của người dùng, nhu cầu thông tin và hành vi tìm tin thay đổi theo thời gian, và do đó, việc dung hòa là điều không tránh khỏi để phản ánh những thay đổi đó (thay đổi sở thích). Chúng tôi thực hiện chính sách ấn định tỉ trọng cao hơn cho những sử dụng gần đây, vì những đề tài liên quan đến sử dụng mới đây phản ánh khung cảnh hiện tại của người dùng. Chúng tôi cũng ấn định những tỉ trọng tính theo số mũ cho những sử dụng trước đây để phản ánh khung cảnh tổng thể của người dùng khi phải tính khung cảnh tổng thể của họ.

Việc dung hòa được thực hiện theo công thức *lặp* sau đây:

$$\mathbf{S}_t = \alpha\, \mathbf{Y}_{t-1} + (1-\alpha)\, \mathbf{S}_{t-1} \quad 0 < \alpha \leq 1$$

ở đây: α là tham số dung hòa,

$\mathbf{S}_t$ = vectơ được dung hòa (vectơ đề tài)

$\mathbf{Y}_{t-1}$ = những quan sát phiên tìm tin đang thực hiện (cũng là vectơ đề tài)

Trong phạm vi thư viện số, tần số đăng nhập của người dùng và hoạt động của người dùng có liên quan trực tiếp đến việc lựa chọn hằng số dung hòa α. Do vậy, những giá trị của α khác nhau đối với từng người dùng ở mỗi phiên tìm tin và được ước tính nhờ một *thước đo hoạt động* đã chuẩn hóa f_i.

$$f_i = N_i \;/\; Max_N$$

ở đây: - N_i = # của các giao tác đối với người dùng

- Max_N = # của các giao tác đối với người dùng tích cực nhất trong khoảng thời gian này

Khi hằng số dung hòa α gần bằng 1, tỉ trọng giảm nhanh và khi hằng số dung hòa α gần bằng 0, tỉ trọng giảm chậm. Do đó, dựa trên tham số hoạt động f_i , chúng tôi lựa chọn giá trị thích hợp của hằng số dung hòa α đối với người dùng.

3.3. Các dịch vụ khuyến nghị thông qua phân tích mẫu sử dụng

Kỹ thuật sàng lọc hợp tác (SLHT) đã được dùng nhiều trong thương mại điện tử. Ý tưởng này có thể mở rộng dễ dàng trong khung cảnh của thư viện số.

Trong phạm vi của thư viện số, chúng tôi có thể tận dụng ưu điểm của các thuật toán sàng lọc hợp tác vừa dựa vào mẫu sử dụng, vừa dựa vào nội dung [14], [15], [16], [17]. Điều thú vị cần ghi nhận là, trong phạm vi của thư viện số, SLHT có thể *đa diện* - nghĩa là nó không chỉ có thể khuyến nghị các tài liệu mới trong thư viện số cho người dùng bằng cách phân tích mức độ tương tự về diện và nhìn vào giá sách của chúng; mà còn có thể phân tích mức độ tương tự của tài liệu trên giá sách (sử dụng phép phân tích từ khóa) và tăng thêm diện người dùng.

Chúng tôi sử dụng thuật toán SLHT nhậy cảm với thời gian (tương tự với [17] cho mục đích này) để tích hợp các hiệu ứng tùy biến vào các khuyến nghị đã đưa ra.

Phần tiếp theo của bài dịch sẽ được đăng trong số tiếp theo của Tạp chí (BBT).

Tài liệu tham khảo

1. Chowdhury, G.G., Chowdhury, S.: Introduction to Digital Libraries. Facet Publishing, London (2003)

2. Feng, L., Jeusfeld, M.A., Hoppenbrouwers, J.: Beyond information searching and Browsing: Acquiring Knowledge from Digital Libraries, (Retrieved March 25) (2007)

3. Marchionini, G. : Information seeking in electronic environments. Cambridge Series on Human-Computer Interaction. Cambridge Univ. Pr, Cambridge (1997)

4. Straccia, U.: Collaborative Working in the Digital Environment. Cyclades (Retrieved March 12) (2007), http: // dlibcenter.iei.pi.cnr.it/

5. Hurley, B.J., Price-Wilkin, J. , Proffitt, M., Besser, H.: The Making of America II Testbed Project: A Digital Library Service Model. The Digital Library Federation Washington DC (1999)

6. Brusilovsky, P.: Adaptive Hypermedia. User Modeling and User-Adapted interaction 11(1-2), 87-110 (2001)

7. Crestani, F.: Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review 11(6), 453-482 (1997)

8. Greenstone Digital Library Software. Project. Retrieved 2/2/2007, from http:// www.greenstone.org/

9. ACM-CCS Add-on Ontology. University of Minho Web Site (Accessed March12,2006) http://dspace-dev.dsi.uminho.pt:8080/en/research_about.jsp

10. Witten L.H., Paynter, G.W., Frank, E.Gutwin, C., Nevill-Manning, C.G.: KEA: Practical Automatic Keyphrase Extraction. In: 4th ACM Conference on Digital Libraries DữLIệU 1999, pp. 254-255. ACM, New York (1999)

11. Pitkow, J., Schutze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., Breuel, T.: Personalized Search. Communication of the ACM 45(9), 50-55 (2002)

12. Dumais, S, Cutrell, E., Chen, H.: Optimizing Search by Showing Results in Context. In: ACM Conference on Human Factors in Computing Systems (CH 2001), Seatle, W.A, pp. 277-284. ACM Pr, N.Y. (2001)

13. Forecasting with Single Exponential Smoothing. NIT/SEMATECH e-Handbook of Statistical Methods. Retrieved 10/02/2007, from http : // www.itl.nist.gov./div898/handbook

14. Liao, I.E., Liao S.C., Kao, K.F., Harn, I.F. : A Personal Ontology Model for Library Recommendation System. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006, LNCS, vol. 4312, pp. 173-182, Springer, Heidelberg (2006)

15. Middleton, S.E., De Roure, D.C., Shadbolt, N.R.: Capturing Knowledge of User Preference: Ontologies on Recommender Systems. In: 1st International Conference on Knowledge Capture (K-CAP2001), pp. 100-107 (2001)

16. Sarwar, B., Karypis, G.,, Konstan, J., Riedl, J.: Item-based Collaborative Filtering Recommendation Algorithms. In: 10th International World Wide Web Conference (WWW 2-10) Hongkong, pp. 285-295 (2001)

17. Ding, Y, Li X.: Time Weight Collaborative Filtering. In: 14th ACM International Conference on Information and Knowledge Management, pp. 485-492 (2005)

18. Olston, C., Chi, E.H.: ScentTrails: Integrating Browsing and Searching on the Web. ACM Transaction on Computer-Human Interaction 10(3), 177-197 (2003)

19. Perugini, S., Ramakrishnan, N., Personalizing Web Sites with Mixed-Initiative Interaction. IEEE IT Professional 5(2), 9-15 (2003)

VVS dịch

Tài liệu gốc: *"Digital libraries: Universal and Ubiquitous Access to Information", pp.105-109*