Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

HỆ THỐNG GỢI Ý HỖ TRỢ TRA CỨU TÀI LIỆU

Trần Nguyễn Minh Thư và Huỳnh Quang Nghi
Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận: 16/12/2015
Ngày chấp nhận: 24/05/2016

Title:
Recommender system for assisting document search

Từ khóa:
Hệ thống gợi ý, lọc cộng tác, hệ thống gợi ý thư viện

Keywords:
Recommender system, collabortive filetring, recommender system for library

ABSTRACT

Document searching for research is a frequent and necessary task for all students as well as lecturers. It is an absolutely necessary process in any library or learning resource center. However, the search function is almost based on keywords so the search results are not rich and not really effective to meet the needs of readers. In order to assist the readers better in searching documents, this study proposed applying the collaborative filtering method of recommender system and the keyword search index functionality (Elastic Search) to the search function. The result of this study is a recommender system for library – RecoLRC – which ensures the success of the document searching by the keywords, and creating a list of recommendation based on the document names, keywords and book borrowing histories. RecoLRC enhances the efficiency of the document searching in The Learning Resource Center of Can Tho University.

TÓM TẮT

Tra cứu tài liệu phục vụ cho nghiên cứu là thường xuyên và cần thiết đối với tất cả sinh viên cũng như giảng viên, đây là một quá trình không thể thiếu trong bất kỳ thư viện hay trung tâm học liệu. Tuy nhiên, các chức năng tìm kiếm hiện nay chủ yếu dựa theo từ khoá nên kết quả tìm kiếm không được phong phú cũng như chưa thực sự hiệu quả đáp ứng nhu cầu của độc giả. Để hỗ trợ tốt nhất cho độc giả tại các thư viện, trung tâm học liệu, nghiên cứu này đề xuất ứng dụng phương pháp lọc cộng tác của hệ thống gợi ý kết hợp với công cụ chỉ mục từ khoá tìm kiếm Elastic Search vào chức năng tìm kiếm tài liệu. Kết quả là một hệ thống gợi ý tài liệu – RecoLRC đảm bảo tìm kiếm tốt theo từ khoá đồng thời tạo ra danh sách các gợi ý dựa trên tên tài liệu, từ khoá và lịch sử mượn sách. Hệ thống RecoRLC giúp nâng cao hiệu quả của việc tra cứu tại Trung tâm Học liệu-Trường Đại học Cần Thơ.

Trích dẫn: Trần Nguyễn Minh Thư và Huỳnh Quang Nghi, 2016. Hệ thống gợi ý hỗ trợ tra cứu tài liệu. Tạp chí Khoa học Trường Đại học Cần Thơ. 43a: 126-134.

1 GIỚI THIỆU

Ngày nay, với sự thuận tiện của Internet, các trang web giải trí, bán hàng trực tuyến, tìm kiếm thông tin trên máy tính và trên điện thoại ngày càng phát triển và phổ biến. Tuy nhiên, các phương tiện này còn bị nhiều hạn chế trong việc hiển thị thông tin so với không gian thực của nhà sách, thư viện, trung tâm thương mại,… Hệ thống gợi ý ra đời như là một công cụ hỗ trợ quyết định nhằm cung cấp cho người dùng những lựa chọn hữu ích và cá nhân hoá nhất. Hệ thống gợi ý được ứng dụng trong nhiều lĩnh vực như thương mại điện tử,

giải trí, tin tức, khoa học, giáo dục…*(G. Adomavicius and A. Tuzhilin, 2005).* Trong lĩnh vực thương mại, người dùng sẽ được hệ thống gợi ý các sản phẩm phù hợp với nhu cầu của từng cá nhân. Ví dụ như hệ thống gợi ý bán hàng của Amazon, Ebay,… Trong lĩnh vực giải trí, người dùng có thể được gợi ý các bộ phim, bài hát phù hợp mà người sử dụng không phải mất nhiều công tìm kiếm như hệ thống gợi ý phim MovieLens[1], last.fm, Film-Conseil. Trong lĩnh vực tin tức, người sử dụng được hệ thống hỗ trợ gợi ý các bài báo phù hợp với từng người riêng biệt ví dụ như netnews, yahoo news… Trong lĩnh vực khoa học và giáo dục, hệ thống gợi ý hỗ trợ người dùng tìm kiếm các bài báo khoa học như hệ thống tìm kiếm Citeseer[2] hay sinh viên tìm kiếm các tài liệu học tập phù hợp với cá nhân như hệ thống School e-Guide của tác giả M. Almulla *(G. Adomavicius and A. Tuzhilin, 2005; Trần Nguyễn Minh Thư, 2011; Huỳnh Xuân Hiệp và ctv., 2014).*

Trong hội thảo quốc tế lần thứ V chuyên về lĩnh vực hệ thống gợi ý tổ chức hàng năm bởi ACM tại Chicago năm 2011, bài viết của tác giả S. Gottwald *và ctv* đã tóm lượt một số hệ thống gợi ý cho việc quản lý thư viện *(S Gottwald, T Koch, 2011).* Tác giả đã giới thiệu và phân tích đặc điểm của 6 hệ thống gợi ý áp dụng trong các thư viện như BibTip, ExLibris bX, Foxtrot, TechLens, Fab, LIBRA. Dựa trên những phân tích đó, tác giả đưa ra hướng phát triển tiếp theo của hệ thống gợi ý thư viện ZIB. ZIB là Viện Nghiên cứu Toán học và Khoa học Máy tính của Đức. Họ đã xây dựng được mô hình đồ thị các siêu dữ liệu (meta-data) biểu diễn mối tương quan giữa các bài báo dựa trên các thông tin: tác giả, đồng tác giả, trích dẫn, tiêu đề, từ khoá, tổ chức, tờ báo, thể loại…

Đây là cơ sở khoa học chắc chắn khẳng định một hướng mới cho khả năng nâng cao hiệu quả tìm kiếm của các thư viện điện tử. Tuy nhiên, việc nghiên cứu và ứng dụng hệ thống gợi ý tại Việt Nam đang trong giai đoạn khởi đầu và các nghiên cứu này đa phần tập trung ứng dụng trong lĩnh vực giải trí, thương mại, nơi có thể mang lại lợi nhuận. Việc đầu tư vào trong giáo dục thì chưa được quan tâm nhiều vì vấn đề chính là vấn đề kinh phí thực hiện mặc dù việc tra cứu tài liệu phục vụ cho nghiên cứu là thường xuyên và cần thiết đối với tất cả sinh viên cũng như giảng viên.

Trung tâm Học liệu-Trường Đại học Cần Thơ[3] là thư viện điện tử đầu tiên tại Đồng bằng sông Cửu Long, cung cấp các bộ sưu tập phong phú về tài liệu (sách, tạp chí khoa học, tài liệu số, tài liệu nghe nhìn, cơ sở dữ liệu,...). Trung tâm Học liệu hiện có khoảng 150.000 nhan đề sách (titles), và các cơ sở dữ liệu (CSDL) trực tuyến như Proquyest central, Sage, Hinari… Độc giả có thể dùng hệ thống thư mục trực tuyến (OPAC) hiện có khoảng 150.000 biểu ghi thư mục (bibliographic records) để truy cập thông tin về sách in và CSDL luận văn của Trường Đại học Cần Thơ. Các CSDL khác được Trung tâm Học liệu đăng ký quyền truy cập thì vẫn còn truy cập rời rạc theo từng CSDL. Trung tâm Học liệu đóng vai trò rất quan trọng trong việc học tập, nghiên cứu của sinh viên và giảng viên vì nó vừa là người thầy, người bạn đồng hành đáng tin cậy ngoài giảng đường và phòng thí nghiệm.

Trung tâm Học liệu hiện tại của Trường Đại học Cần Thơ được xây dựng từ thư viện trung tâm của Trường Đại học Cần Thơ, mà thư viện trung tâm được tài trợ từ các tổ chức quốc tế như OMS, IRRI, MCC, SAREC, ALA (Mỹ) từ 1985 cũng như chương trình MHO. Do quá trình hình thành và phát triển trải qua một thời gian dài và được hỗ trợ từ nhiều tổ chức khác nhau nên việc lưu trữ cơ sở dữ liệu của trung tâm còn chưa được thống nhất. Hệ thống quản lý tài liệu của trung tâm học liệu trường Đại học Cần Thơ gồm 2 phần tách biệt mà mỗi phần được quản lý bởi một hệ quản trị cơ sở dữ liệu khác nhau với một số thông tin tương đối giống nhau. Cụ thể là các cơ sở dữ liệu của các loại sách in, sách tham khảo, giáo trình môn học… được lưu trên hệ quản trị cơ sở dữ liệu Oracle, còn các luận văn đại học, luận văn cao học… thì hiện được lưu trữ trên hệ quản trị cơ sở dữ liệu MySQL (Hình 1).

Việc lưu trữ cơ sở dữ liệu còn rời rạc như vậy dẫn đến việc tra cứu tài liệu của độc giả sẽ mất thời gian hơn. Thay vì độc giải chỉ tìm kiếm 1 lần thì với 2 cơ sở dữ liệu như hiện tại, độc giả cần phải thực hiện việc tìm kiếm hai lần để có đầy đủ các tài liệu cho nội dung cần nghiên cứu. Ngoài việc gây khó khăn cho độc giả thì với 2 cơ sở dữ liệu như vậy cũng làm cho việc quản lý, bảo trì phức tạp và mất nhiều thời gian hơn.

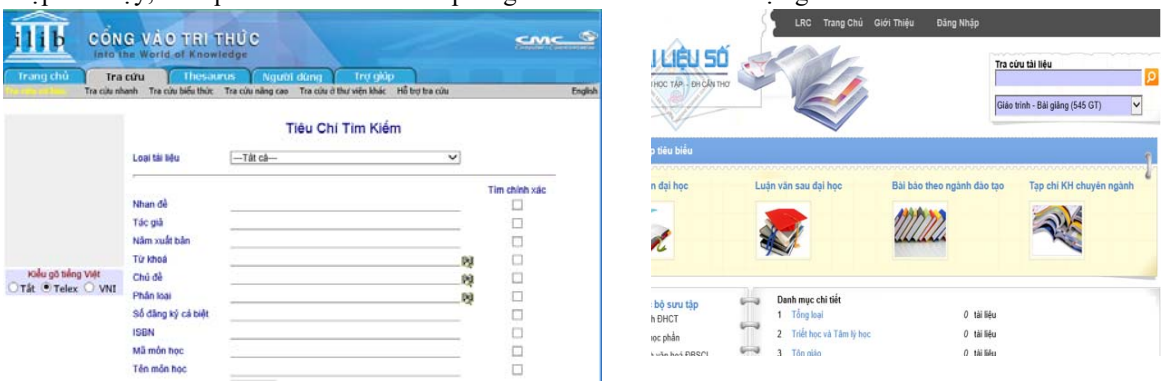
Bên cạnh cách tổ chức thì chức năng tra cứu tài liệu hiện tại của Trung tâm Học liệu được xây dựng hiện nay là tìm kiếm chính xác từ khoá được

[1] https://movielens.org/

[2] http://citeseerx.ist.psu.edu

[3] http://www.lrc.ctu.edu.vn/

nhập. Vì vậy, kết quả tìm kiếm chưa đáp ứng tốt nhu cầu của độc giả.



Hình 1: Giao diện tra cứu tài liệu trên cơ sở dữ liệu Oracle (bên trái) và MySql (bên phải)

Dựa vào việc phân tích thực trạng của Trung tâm Học liệu-Trường Đại học Cần Thơ, cũng như phân tích các hệ thống quản lý thư viện khác, chức năng tìm kiếm của Trung tâm Học liệu-Trường Đại học Cần Thơ cần phải cải tiến để phục vụ hiệu quả hơn cho độc giả. Để khắc phục vấn đề này, chúng tôi đề xuất sử dụng một kho dữ liệu (data warehouse) chung để xây dựng một hệ thống gợi ý hỗ trợ tra cứu tài liệu - RecoLRC. Trên hệ thống RecoRLC, độc giả không phải vào 2 địa chỉ url khác nhau để tìm kiếm tài liệu cùng một chủ đề nhưng khác thể loại như hình 1. Bên cạnh đó, chúng tôi cũng áp dụng hệ thống gợi ý vào chức năng tìm kiếm nhằm làm phong phú thêm kết quả tìm kiếm để hỗ trợ tốt nhất cho các độc giả trong quá trình tra cứu sách.

Xuất phát từ nhu cầu thực tiễn trên, hệ thống gợi ý sách RecoRLC được đề xuất nhằm giải quyết những hạn chế của việc tìm kiếm chính xác theo từ khoá hiện nay. Nghiên cứu sử dụng MongoDB để tích hợp cơ cở dữ liệu MyQSL và Oracle đang áp dụng tại Trung tâm Học liệu. Dựa trên cơ sở dữ liệu chung này chúng tôi cài đặt công cụ tìm kiếm chỉ mục Elastic Search kết hợp với phương pháp lọc cộng tác dựa trên lịch sử mượn sách để góp phần làm phong phú kết quả tìm kiếm của độc giả.

Trong phần hai, chúng tôi sẽ trình bày tổng quan về việc ứng dụng hệ thống gợi ý trong các thư viện điện tử. Các thông tin về Trung tâm Học liệu-Trường Đại học Cần Thơ cũng như giải pháp hệ thống gợi ý áp dụng tại Trung tâm Học liệu được trình bày chi tiết trong phần ba. Bài báo kết thúc bởi phần kết luận và hướng phát triển trong tương lai.

2 NỘI DUNG NGHIÊN CỨU

2.1 Giới thiệu về hệ thống gợi ý

Hệ thống gợi ý là hệ thống hỗ trợ ra quyết định nhằm gợi ý các thông tin liên quan đến người dùng một cách dễ dàng và nhanh chóng, phù hợp với từng người dùng *(G. Adomavicius and A. Tuzhilin, 2005)*. Ví dụ với trang web Amazon, một trong những trang web thương mại điện tử nổi tiếng nhất, khi người dùng truy cập vào trang web này họ sẽ được gợi ý những sản phẩm tiềm năng nhất từ hàng triệu sản phẩm trong hệ thống. Hệ thống gợi ý như một công cụ cung cấp những thông tin hữu ích và riêng biệt theo từng cá nhân trên một hệ thống chứa đựng một lượng lớn thông tin. Các hệ thống gợi ý được thiết kế nhằm cung cấp cho người dùng những đề nghị liên quan, những đề nghị hiệu quả nhất có thể từ thông tin của các mục dữ liệu, từ hồ sơ người sử dụng và từ mối liên hệ giữa những đối tượng này.

Cấu trúc của một hệ thống gợi ý gồm có ba thành phần chính *(G. Adomavicius and A. Tuzhilin, 2005; Huỳnh Xuân Hiệp và ctv., 2014)*: tập hợp các người dùng $U = \{u_1,...,u_p\}$ bao gồm các thông tin của người dùng được lưu trên hệ thống; tập hợp các mục dữ liệu $I = \{i_1,...,i_p\}$ bao gồm định danh và các thuộc tính của mục dữ liệu; tập hợp các "mối quan hệ" $R = (U_i, I_j)$ giữa "người dùng" và "mục dữ liệu", đây là tập hợp các giao dịch liên kết giữa tập hợp người dùng U và tập hợp mục dữ liệu I và những mô tả của mối liên kết này. Dựa trên cách thức xây dựng hệ thống gợi ý, người ta chia hệ thống gợi ý thành 3 loại chính: hệ thống gợi ý dựa trên nội dung; hệ thống gợi ý dựa trên phương pháp lọc cộng tác; hệ thống gợi ý lai – kết hợp phương pháp lọc cộng tác và phương pháp dựa trên nội dung.

2.2 Hệ thống gợi ý áp dụng trong các hệ thống quản lý thư viện

Chức năng quan trọng nhất của một thư viên hay một trung tâm học liệu chính là chức năng tìm kiếm, giúp độc giả nhanh chóng tìm được tài liệu theo đúng nhu cầu nghiên cứu. Nội dung chính trong phần này là tham khảo tìm hiểu xem chức năng tìm kiếm của các hệ thống thư viện hay trung tâm học liệu hiện nay được thực hiện như thế nào.

Hệ thống gợi ý Fab nằm trong dự án thư viện điện tử của trường đại học Stanford phát triển từ những năm 1997 *(M. Balabanovic and Y. Shoham, 1997)*. Hệ thống kết hợp giữa phương pháp lọc cộng tác và phương pháp dựa trên nội dung để xây dựng chức năng tìm kiếm tài liệu cho độc giả. Gợi ý được thực hiện dựa trên đánh giá của người dùng hiện tại và đánh giá của những người dùng tương tự. Phương pháp lọc cộng tác của hệ thống gợi ý cũng ra đời từ nghiên cứu này, giúp hệ thống gợi ý trở thành một lĩnh vực nghiên cứu độc lập.

Hệ thống gợi ý LIBRA viết tắt của từ "Learning Intelligent Book Recommending Agent" ra đời năm 2000 *(E.Vozalis and K. G. Margaritis, 2003).* Hệ thống gợi ý sách thay vì bài báo khoa học như hệ thống Fab. Tuy nhiên, độc giả bắt buộc phải đánh giá 10 kết quả tìm kiếm đầu tiên, sau đó tập kết quả sẽ được sắp xếp lại theo "sở thích" người dùng. Hệ thống được xây dựng dựa trên phương pháp lọc cộng tác dựa trên người dùng. Giải thuật Bayes thơ ngây được sử dụng để tạo ra bảng xếp hạng cho những quyển sách của người sử dụng. Chỉ số tương quan Pearson được sử dụng để tìm kiếm những người dùng tương tự dựa trên bảng sếp hạng các quyển sách của mỗi người dùng.

Tiếp theo không thể không kể đến một công trình dài hơi của nhóm tác giả Joseph A. Konstan *và ctv* trong nghiên cứu "TechLens: Exploring the Use of Recommenders to Support Users of Digital Libraries" *(J. Konstan et al., 2005; Roberto Torres et al., 2004).* Các nhà khoa học đã xây dựng hệ thống gợi ý dành riêng cho thư viện điện tử dựa trên các giải thuật mà họ đã thành công trong lĩnh vực thương mại nhằm phục vụ cho nhu cầu tìm kiếm các tài liệu khoa học của các sinh viên cũng như các học giả. Nghiên cứu bắt đầu thực nghiệm trên cơ sở dữ liệu trích dẫn của CiteSeer. Hệ thống gợi ý được xây dựng dựa trên phương pháp lọc cộng tác sử dụng các thông tin trích dẫn, từ khóa và tóm tắt của mỗi tài liệu… và đồ thị của các trích dẫn trong các tài liệu khoa học. Bằng cách sử dụng phương pháp này, họ có thể nắm bắt được các tài liệu khoa học nào có liên quan đến nhau để đưa ra gợi ý về các tài liệu có liên quan đó, trong khi nếu chỉ dựa trên việc tìm kiếm nội dung đơn thuần thì có thể bị bỏ lỡ. Họ cũng tìm thấy các kết quả gợi ý với hiệu quả khá cao từ việc kết hợp cả lọc cộng tác và lọc dựa trên nội dung với nhau. Hệ thống gợi ý thư viện này được triển khai 2 giai đoạn: giai đoạn 2001-2002: áp dụng trên cơ sở dữ liệu trích dẫn của CiteSeer và sử dụng thư viện số ACM là cốt lõi. Năm 2004-2005, họ tiến hành tích hợp hệ thống gợi ý vào thư viện số, họ thử nghiệm trên một chiếc máy tính dược xây dựng như một thư viện số. Ngoài việc lặp lại các tính năng đã có, demo thử nghiệm lần này của họ còn có những tính năng cụ thể hơn như các công cụ để thu thập dữ liệu thư mục cá nhân và danh sách người dùng ...

Nhằm nâng cao hiệu quả của hệ thống TechLens, những nghiên cứu sâu hơn đã được thực hiện thông qua bài báo "Enhancing Digital Libraries with TechLens+" *(Roberto Torres et al., 2004).* Các thuật toán áp dụng cho TechLens+ không chỉ xem xét mặt nội dung của tài liệu mà còn xem xét cả về mặt ngữ cảnh của tài liệu, thông qua các trích dẫn của nó đến các tài liệu khác. Các tác giả tin rằng các thuật toán mà họ phát triển sẽ là một trợ thủ đắc lực trong cả lĩnh vực thư viện kỹ thuật số hiện tại và tương lai.

Trong công trình nghiên cứu này, họ đã phát triển 10 thuật toán gợi ý. Mỗi thuật toán đều có đầu vào là tập các văn bản đại diện của tài liệu và đầu ra là một danh sách có thứ tự các tài liệu được khuyến nghị. Các thiết lập đầu vào được tạo ra từ danh sách tài liệu và tất cả thuật toán của họ đều sử dụng lọc cộng tác (CF) và lọc nội dung (CBF) hoặc kết hợp cả hai phương pháp. Sự kết hợp của hai phương pháp CF và CBF tuân theo hai nguyên lý sau:

Tuần tự: Đầu ra của thành phần thứ nhất sẽ là đầu vào của thành phần thứ hai. Các khuyến nghị đầu ra của các thành phần thứ nhất sẽ được sử dụng như là đầu vào của thành phần thứ thứ hai và kết quả sau cùng là danh sách được đề nghị chính thức.

Song song: Hai thành phần chạy song song và kết hợp các khuyến nghị với nhau thành một danh sách được đề nghị chính thức.

Các thuật toán thuần (chỉ sử dụng độc lập CF hoặc CBF) được làm cơ sở so sánh cho các thuật toán lai mà họ phát triển. Các thuật toán được đánh giá online và offline trên tập dữ liệu được rút trích từ CiteSeer với hơn 500.000 bài báo và 2 triệu trích dẫn với trung bình 14 nối kết cho mỗi bài báo. Với phương pháp đánh giá "leave one out", kết quả

thực nghiệm cho thấy phương pháp kết hợp CBF và CF dạng fusion là đạt kết quả tốt nhất, tiếp theo phải kể đến phương pháp thuần lọc cộng tác. Kết quả của phương pháp Denser-CF gây ngạc nhiên vì không tốt như mong đợi, có thể việc thêm các trích dẫn cho 1 tài liệu có thể đã làm tăng nhiễu thay vì tăng thông tin hữu ích.

Từ năm 2007, Stefan Pohl *và ctv (Stefan Pohl et al., 2007)* đã bắt đầu áp dụng hệ thống gợi ý cho lĩnh vực thư viện điện tử, tuy nhiên nó chỉ dừng lại ở mức gợi ý các bài báo liên quan thay vì toàn bộ dữ liệu của thư viện như bài báo khoa học, tạp chí, sách, tài liệu online,... Nghiên cứu của các tác giả được thực nghiệm trên tập dữ liệu arXiv thu thập từ hơn 650 triệu truy cập của hơn 350.000 tài liệu khoa học trong khoảng từ năm 1994 đến tháng 6 năm 2006. Thông tin thu thập được bao gồm: địa chỉ IP, thời gian truy cập và tài liệu đã truy cập. Tác giả đã chứng minh được không chỉ có các trích dẫn đóng vai trò quan trọng mà thông tin lịch sử truy cập tài liệu cũng đóng vai trò quan trọng (http-server logs) trong việc gợi ý cho độc giả tài liệu phù hợp. Co-access tốt hơn co-citiation ở đặc tính "bao phủ".

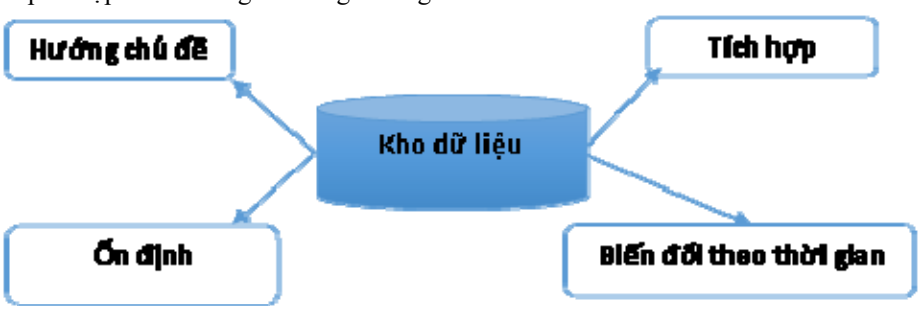
Trong Tạp chí Quốc tế về Khoa học Xã hội và Nhân văn, Vol. 5, số 11, tháng 11 năm 2015: vai trò ngữ cảnh đối với các hệ thống gợi ý thư viện kỹ thuật số đã được các tác giả Zohreh Dehghani Champiri *và ctv* thảo luận rất chi tiết (Zohreh Dehghani Champiri *et al.,* 2015). Việc xác định và áp dụng các thông tin theo ngữ cảnh trong từng lĩnh vực cũng như người sử dụng khác nhau là những thách thức cần được giải quyết và những thách thức này đã xuất hiện trong cách tiếp cận trong những năm gần đây. Nghiên cứu của Zohreh D. C. nhấn mạnh tầm quan trọng của ngữ cảnh và sự cần thiết phải khai thác thông tin ngữ cảnh trong các hệ thống gợi ý để đưa ra được các khuyến nghị chính xác và phù hợp hơn cho người dùng. Trong thư viện truyền thống, thủ thư chủ yếu là cố gắng tìm kiếm các tài liệu phù hợp với yêu cầu/nhu cầu của độc giả dựa trên kiến thức và sự am hiểu về thư viện mà họ đang quản lý. Đôi khi, họ phải có được nhiều thông tin về độc giả để việc tìm kiếm được chính xác hơn như hoạt động của họ, công việc, kiến thức nền tảng, kinh nghiệm, chủ đề và mục đích cũng như các điều kiện như thời gian, địa điểm, tình hình,… Trong tương tác giữa người dùng với thư viện kỹ thuật số, hệ thống gợi ý có thể đóng vai trò của thủ thư để giới thiệu tài nguyên đến độc giả vì thế hệ thống không chỉ cần phải nhận thức được ngữ cảnh của độc giả mà còn phải nhận thức các vấn đề mà người dùng đang tìm kiếm thông tin để giải quyết chúng.

Qua những thông tin lược sử về quá trình phát triển của các hệ thống gợi ý áp dụng vào hệ thống quản lý thư viện, ta thấy được ứng dụng các giải thuật của hệ thống gợi ý vào trong chức năng tìm kiếm của hệ thống quản lý thư viện là vô cùng cần thiết và đem lại hiệu quả cao so với chức năng tìm kiếm thông thường.

2.3 Kho dữ liệu

Vào những năm 1980, có nhiều ứng dụng trực tuyến được xây dựng phục vụ cho nhu cầu của xã hội. Tuy nhiên, các ứng dụng này rời rạc nhau, dẫn đến việc thông tin cung cấp không được đầy đủ. Kho dữ liệu xuất hiện nhằm tích hợp các ứng dụng riêng lẻ, để có thông tin đầy đủ hơn, phục vụ nhu cầu ngày càng cao của con người (Moh'd Alsqour *et al*., 2012).

Theo William H. Inmon (Inmon, 2005), một kho dữ liệu là một tập hợp cơ sở dữ liệu hướng chủ đề (subject-oriented), tích hợp (integrated), dữ liệu gắn với thời gian (time-variant), ổn định (nonvolatile), được thiết kế để hỗ trợ ra quyết định của các nhà quản lý (Hình 2).



Hình 2: Đặc điểm của kho dữ liệu

Tính hướng chủ đề (subject-oriented) thể hiện trong kho dữ liệu chính là cơ sở dữ liệu tổ chức để đáp ứng một chủ đề nhất định, loại bỏ những thông tin dư thừa không cần thiết cho bài toán đặt ra; tính tích hợp (integrated) của kho dữ liệu thể hiện sự tập hợp dữ liệu từ nhiều nguồn khác nhau ví dụ như dữ liệu sách, giáo trình từ cơ sở dữ liệu Oracle và luận văn từ cơ sở dữ liệu MySQL của Trung tâm Học liệu-Trường Đại học Cần Thơ; tính biến đổi theo thời gian (time-variant): dữ liệu trong kho dữ liệu gắn với thời gian, có tính lịch sử; tính ổn định (nonVolatile) kho dữ liệu tách tời vật lý với môi trường tác nghiệp nên dữ liệu trong kho là dữ liệu chỉ đọc không chỉnh sửa hoặc thêm mới được.

Có hai phương pháp tiếp cận chính để xây dựng một kho dữ liệu: phương pháp Immon (từ trên xuống: Top-Down) và phương pháp tiếp cận Kimball (từ dưới lên: Bottom-up). Với phương pháp tiếp cận Immon, các kho dữ liệu được thiết kế trước, sao đó chuyển yêu cầu xuống đến các cơ sở dữ liệu nhỏ bên dưới, để các cơ sở dữ liệu nhỏ phải đảm bảo cấu trúc theo đúng yêu cầu của kho dữ liệu. Với phương pháp tiếp cận Kimball, các cơ sở dữ liệu nhỏ bên dưới được xây dựng trước, dựa vào các cơ sở dữ liệu nhỏ này kho dữ liệu sẽ được xây dựng sau dựa trên việc tích hợp các cơ sở dữ liệu nhỏ có sẵn (Moh'd Alsqour *et al*., 2012).

2.4 Công cụ tìm kiếm theo chỉ mục Elastic Search

Trong thời đại bùng nổ thông tin, việc tìm kiếm "thông tin" trở nên rất cần thiết và yêu cầu cao hơn. Elasticsearch là một hệ thống hỗ trợ tạo chỉ mục và tìm kiếm riêng biệt, mạnh mẽ, thời gian đáp ứng và kết quả tìm kiếm phù hợp với yêu cầu. Elasticsearch[4] được phát hành phiên bản đầu tiên vào tháng 2/2010 bởi Shay Banon. Khi phát triển Compass lên phiên bản thứ ba vào năm 2004 Shay Banon đã gần như viết lại Compass để tạo ra một sản phẩm đổi mới hoàn toàn, đó chính là Elasticsearch. Elasticsearch là một máy chủ tìm kiếm dựa trên Lucene, có khả năng tìm kiếm toàn văn với một giao diện web HTTP và JSON schema-free. Elasticsearch được phát triển bằng Java và được phát hành dưới dạng mã nguồn mở theo các điều khoản của Giấy phép Apache. Phiên bản mới nhất của Elasticsearch 2.1.1 vừa được phát hành trong tháng 12 năm 2015. Để có thể xếp thứ 2 trong danh sách các bộ máy tìm kiếm, ElasticSearch có một số ưu điểm như sau:

Có thể tìm kiếm tất cả các loại dữ liệu, đáp ứng gần như thời gian thực, dễ sử dụng cũng như dễ cài đặt; hỗ trợ thêm, sửa, xoá hay thay đổi các độ đo, thông số thông qua HTTP và JSON.

Hỗ trợ lập chỉ mục với nhiều ngôn ngữ khác nhau và có hỗ trợ lập chỉ mục cho cả tiếng Việt.

Chia dữ liệu hệ thống (sharding) phục vụ cho 2 ngôn ngữ khác nhau là Tiếng Anh và Tiếng Pháp. Chúng ta có thể chia thông tin trên 2 máy chủ khác nhau, sau đó người dùng tìm kiếm tiếng Anh sẽ lấy kết quả từ nút (node) Tiếng Anh và tìm kiếm bằng tiếng Pháp sẽ lấy từ nút (node) Tiếng Pháp.

Khả năng tăng sức chịu đựng các truy xuất cùng lúc và giảm các rủi ro khi các nút khác bị sự cố thông qua khả năng nhân bản (replication).

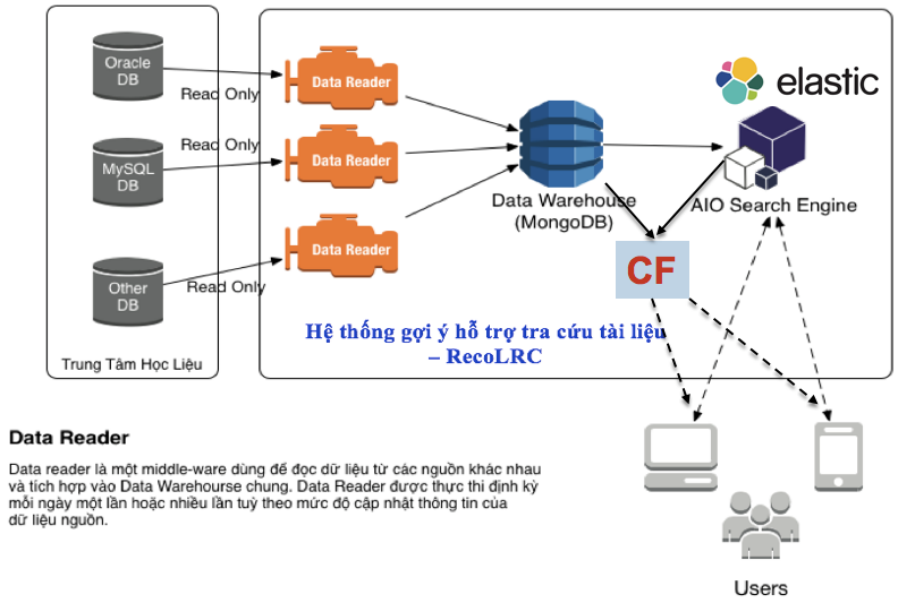
Thời gian lập chỉ mục nhanh nên rất hợp với các loại dữ liệu lớn và có tần suất cập nhật cao.

3 ỨNG DỤNG HỆ THỐNG GỢI Ý TẠI TRUNG TÂM HỌC LIỆU-TRƯỜNG ĐẠI HỌC CẦN THƠ

Dựa vào việc phân tích thực trạng của Trung tâm Học liệu-Đại học Cần Thơ, chúng tôi xây dựng "Hệ thống gợi ý hỗ trợ tra cứu tài liệu tại Trung tâm Học liệu Đại học Cần Thơ", nhằm hỗ trợ tốt nhất cho các độc giả trong quá trình tra cứu sách. Với phương pháp tiếp cận Kimball, chúng tôi tích hợp cơ sở dữ liệu Oracle và MySQL hiện có của Trung tâm Học liệu thành một kho dữ liệu (data warehouse). Để hỗ trợ tích hợp các cơ sở dữ liệu khác nhau cho hệ thống RecoLRC, chúng tôi xây dựng các "DataReader" để đọc dữ liệu tương ứng từ các cơ sở dữ liệu khác nhau đưa vào cơ sở dữ liệu chung (Data Warehouse) được xây dựng bằng MongoDB. Lịch trình đọc dữ liệu bởi các "DataReader" sẽ được thực hiện định kỳ theo lịch biểu đã đặt sẵn hoặc thực hiện khi cần thiết.

Với cơ sở dữ liệu chung, chúng tôi sử dụng các tiêu chí sau để xây dựng hệ thống gợi ý: tên tài liệu, từ khoá và lịch sử mượn sách của thư viện. Hệ thống RecoLRC được xây dựng kết hợp phương pháp lọc cộng tác (dựa vào lịch sử mượn sách của độc giả) và Elastic Search (dựa vào tên tài liệu, từ khoá, tóm tắt). Thông tin về năm xuất bản và nhà xuất bản được sử dụng để cập nhật lại bảng xếp hạng của các tài liệu gợi ý.

[4]https://www.elastic.co/products



Hình 3: Mô hình hệ thống hỗ trợ tìm kiếm tài liệu RecoLRC

Để sử dụng phương pháp lọc cộng tác, chúng tôi dựa vào bảng lịch sử mượn sách của độc giả tại Trung tâm Học liệu. "Tài liệu" ở đây được xác định bởi mã "nhan đề tài liệu". "Nhan đề tài liệu" là tập hợp nhiều quyển tài liệu giống nhau, do Trung tâm Học liệu trang bị nhiều quyển sách giống nhau để phục vụ đồng thời được nhiều độc giả. Dựa vào các thông tin "mã số độc giả", "mã nhan đề tài liệu" và "thời gian mượn" tài liệu, chúng tôi xác định được các giao dịch mượn tài liệu. Với thông tin này chúng tôi xây dựng phương pháp lọc cộng tác dựa trên item / tài liệu để tạo ra danh sách gợi ý. Ví dụ để tìm được các tài liệu gợi ý cho độc giả đang chọn tài liệu "Kỹ thuật và thủ thuật lập trình PHP" bằng phương pháp lọc cộng tác dựa trên tài liệu, ta thực hiện các bước sau:

Bước 1: Tìm kiếm những tài liệu tương đồng với tài liệu "Kỹ thuật và thủ thuật lập trình PHP" dựa vào chỉ số tương tự Pearson.

Bước 2: Tính giá trị dự đoán đánh giá cho các tài liệu tương đồng tìm được từ bước 1 dựa vào công thức.

$$pred(u,i) = \frac{\sum_{j=1}^{nbItem} sim(i,j) \times r_{uj}}{\sum_{j=1}^{nbItem} sim(i,j)}$$

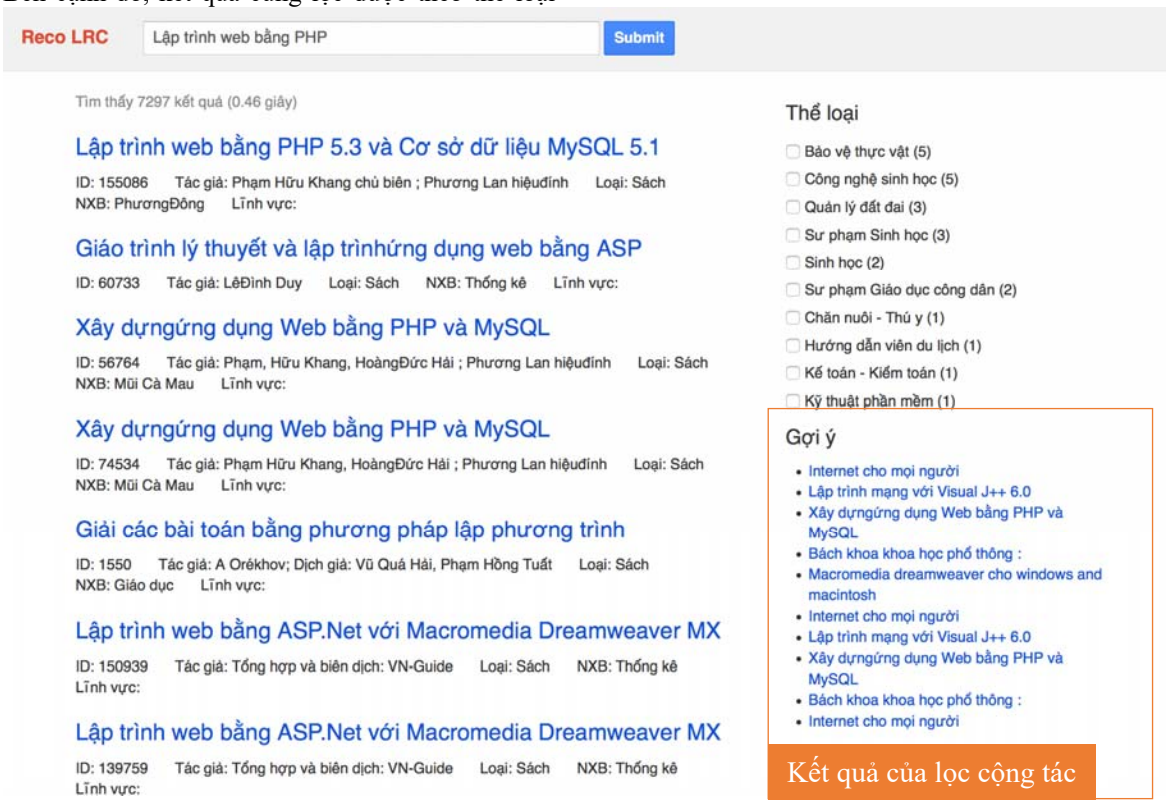
với $nbItem$ là số lượng các tài liệu được xem bởi người dùng u, r_{uj} là số lần xem của người dùng u trên tài liệu j, $sim(i,j)$ được định nghĩa như là độ tương tự giữa các tài liệu (items) i và j.

Với danh sách gợi ý được tạo ra từ phương pháp lọc cộng tác, chúng tôi kết hợp thêm thông tin năm xuất bản và kết quả tìm kiếm được từ Elastic Search để tạo ra một danh sách các gợi ý phong phú và hiệu quả cung cấp cho độc giả.

Hệ thống gợi hỗ trợ tra cứu tài liệu RecoLRC sẽ có liên kết với trang web hiện tại của Trung tâm Học liệu tại nút "Tìm kiếm". Khi độc giả nhấp vào nút "tìm kiếm" trên trang web hiện tại, thì độc giả sẽ được chuyển đến trang để gõ nhập từ khóa tìm kiếm như Hình 4.

Hình 4: Khung nhập từ khoá tìm kiếm tài liệu

Ví dụ người dùng gõ vào từ khoá tìm kiếm là "Lập trình web bằng Php", kết quả tìm kiếm hiển thị hai phần: tìm kiếm theo nội dung dựa vào công cụ tìm kiếm Elastic Search, kết quả tìm kiếm dựa trên phương pháp lọc cộng tác. (1) Với nội dung tìm được bởi Elastic Search, bộ phân tích ngữ nghĩa từ vựng tiếng Việt được tích hợp vào nhằm hỗ trợ việc tìm kiếm theo từ khoá chính xác hơn. Bên cạnh đó, kết quả cũng lọc được theo thể loại tài liệu và có tổng kết số lượng tài liệu tương ứng với từng thể loại. Độc giả có thể chọn hoặc bỏ chọn các kết quả không đúng lĩnh vực mình đang tìm kiếm để có được danh sách tài liệu phù hợp nhất với yêu cầu. (2) Với nội dung lọc cộng tác, danh sách các tài liệu có được từ phương pháp lọc cộng tác với 3 tài liệu đầu tiên thu được từ công cụ tìm kiếm ElasticSearch được giới thiệu tới độc giả. Kết quả được hiển thị giống như Hình 5.



Hình 5: Trang web hỗ trợ tìm kiếm tài liệu RecoLRC

Khi độc giả muốn xem thông tin tài liệu của quyển sách/ giáo trình nào đó thì hệ thống RecoLRC sẽ liên kết lại với trang web hiện tại của trung tâm học liệu dựa vào thông tin mã của tài liệu.

Tóm lại, với hệ thống hỗ trợ tra cứu tài liệu RecoLRC, độc giả tại Trung tâm Học liệu-Trường Đại học Cần Thơ có thể rút ngắn thời gian tìm kiếm dữ liệu. Từ hai cơ sở dữ liệu khác nhau Oracle và MySQL, hệ thống RecoLRC đã tích hợp lại thành một kho dữ liệu chung (Data warehouse) sử dụng MongoDB. Đầu vào của trang web chính là "khung tìm kiếm", độc giả gõ vào khung tìm kiếm từ khoá cần tìm không cần phải chọn tiêu đề tài liệu, nhà xuất bản hay tác giả; kết quả tìm kiếm sẽ tìm tất cả thông tin lưu trữ của tài liệu tại Trung tâm Học liệu. Đầu ra của hệ thống RecoLRC sẽ được tạo ra từ hai nguồn dữ liệu: danh sách các tài liệu tìm được từ kết quả của công cụ tìm kiếm Elasticsearch có tích hợp phân tích ngữ nghĩa tiếng Việt và danh sách các tài liệu tìm được từ phương pháp lọc cộng tác với 3 tài liệu đầu tiên trong danh sách tài liệu tìm được từ công cụ tìm kiếm Elasticsearch.

4 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày một hướng tiếp cận trong việc tìm kiếm tài liệu tại Trung tâm Học liệu-Trường Đại học Cần Thơ, sử dụng công cụ tìm kiếm chỉ mục Elastic Search kết hợp với phương pháp lọc cộng tác. Kho dữ liệu chung MongoDB được sử dụng để tích hợp cơ sở dữ liệu MySQL và Oracle vào cùng một cơ sở dữ liệu chung. Dựa trên

cơ sở dữ liệu chung này, chúng tôi tích hợp vào chức năng tìm kiếm công cụ hỗ trợ tạo chỉ mục từ khoá tìm kiếm Elastic Search cũng như phương pháp lọc cộng tác dựa trên mục dữ liệu (tài liệu). Kết quả.

Trong tương lai, chúng tôi tiếp tục nghiên cứu để hệ thống RecoLRC phát triển được tốt hơn. Thông tin người dùng sẽ được thu thập để tạo ra những gợi ý theo hướng cá nhân hoá hơn, làm hài lòng hơn nữa cho độc giả sử dụng hệ thống. Bên cạnh thông tin người dùng, chúng tôi cũng sẽ tiếp cận thông tin đồ thị các trích dẫn của các tài liệu như giải pháp đã đề nghị trong hệ thống TechLens nhằm nâng cao hiệu quả của danh sách tài liệu gợi ý cho độc giả.

LỜI CÁM TẠ

Nhóm tác giả xin chân thành cảm ơn lãnh đạo và các thành viên tổ Công nghệ thông tin của Trung tâm Học liệu-Trường Đại học Cần Thơ đã hỗ trợ trong việc cung cấp cơ sở dữ liệu cũng như hỗ trợ trong công tác triển khai hệ thống RecoLRC.

TÀI LIỆU THAM KHẢO

G. Adomavicius and A. Tuzhilin, 2005. 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, 734-749.

Trần Nguyễn Minh Thư, 2011. Abstraction et règles d'association pour l'amélioration des systèmes de recommendation à partir de données de préférences binaires. Phd thesis.

Huỳnh Xuân Hiệp, Nguyễn Thái Nghe và Trần Nguyễn Minh Thư, 2014. Giáo trình Mô hình hóa quyết định. Nhà xuất bản Đại học Cần Thơ. Cần Thơ.

S Gottwald, T Koch, 2011. Recommender systems for libraries, ACM Recommender Systems 2011 Chicago.

Balabanovic, M. and Y. Shoham, 1997 Fab: Content-based, collaborative recommendation. Communications of the ACM, 40(3), pp.66-72.

E.Vozalis and K. G. Margaritis, 2003. Analysis of Recommender Systems' Algorithms, HERCMA, Athens, Greece.

J. Konstan, N. Kapoor, S. McNee, and J. Butler, 2005. Techlens: Exploring the use of recommenders to support users of digital libraries. Communications of the ACM.

Roberto Torres, Sean M. McNee, Mara Abel, Joseph A. Konstan, and John Riedl, 2004. Enhancing digital libraries with TechLens+. In Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries (JCDL '04). ACM, New York, NY, USA, 228-236.

Stefan Pohl, Filip Radlinski, and Thorsten Joachims, 2007. Recommending related papers based on digital library access records. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL '07). ACM, New York, NY, USA, 417-418.

Zohreh Dehghani Champiri, Siti Salwah Binti Salim, and Seyed Reza Shahamiri, 2015. The Role of Context for Recommendations in Digital Libraries, International Journal of Social Science and Humanity vol. 5, no. 11, pp. 948-954.

Inmon, William H., 2005, Building the Data Warehouse, 4th Edition, Wiley Publishing, Indianapoli.

Moh'd Alsqour, Kamal Matouk and Mieczyslaw L. Owoc, 2012, A survey of data warehouse architectures: preliminary results, Proceedings of the Federated Conference on Computer Science and Information Systems, pp 1121-1126, FedCSIS 2012, Wroclaw, Poland.