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

CÁC PHƯƠNG PHÁP ĐÁNH GIÁ HỆ THỐNG GỢI Ý

Trần Nguyễn Minh Thư và Phạm Xuân Hiền
Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận: 01/08/2015
Ngày chấp nhận: 25/02/2016

Title:
Evaluation methods for recommender systems

Từ khóa:
Hệ thống gợi ý, phương thức, đánh giá, đánh giá offline, đánh giá on-line

Keywords:
Recommender system, protocol, measure, off-line evalution, on-line evalution

ABSTRACT

Recommender system is a decisive support tool to provide users the most useful choice in the era of "information explosion". When a recommender system is built, the effectiveness of the system is usually more concerned. However, evaluating the effectiveness of the recommender system depends a lot on the purpose of building the systems, kind of data, and conditions to evaluate the system. These conditions can be online or based on available data (offline). In this article, we will focus on analyzing and introducing the evaluation methods based on a system of qualitative criteria (diversity, novelty, covers) as well as quantitative criteria (precision, recall, F1, MSE, RMSE). The process to evaluate a recommender system for each kind of database is also mentioned in this article.

TÓM TẮT

Hệ thống gợi ý là một công cụ hỗ trợ quyết định nhằm cung cấp cho người dùng những lựa chọn hữu ích nhất trong thời đại bùng nổ thông tin. Khi xây dựng một hệ thống gợi ý, người ta thường quan tâm đến tính hiệu quả của nó. Tuy nhiên, việc đánh giá tính hiệu quả của một hệ thống gợi ý còn tuỳ thuộc rất nhiều vào mục đích xây dựng hệ thống, loại dữ liệu và điều kiện để đánh giá hệ thống. Điều kiện đánh giá hệ thống có thể là trực tuyến (online) hay dựa vào dữ liệu có sẵn (offline). Trong bài báo này, chúng tôi sẽ tập trung phân tích và giới thiệu các phương pháp đánh giá một hệ thống gợi ý theo tiêu chí định tính (tính đa dạng, tính mới, tính bao phủ) cũng như định lượng (precision, recall, F1, MSE, RMSE). Đồng thời, các nghi thức đánh giá phù hợp đối với từng loại cơ sở dữ liệu cũng được đề cập trong bài báo này.

Trích dẫn: Trần Nguyễn Minh Thư và Phạm Xuân Hiền, 2016. Các phương pháp đánh giá hệ thống gợi ý. Tạp chí Khoa học Trường Đại học Cần Thơ. 42a: 18-27.

1 GIỚI THIỆU

Ngày nay, hệ thống gợi ý được nhiều người biết đến như một công cụ hỗ trợ hữu ích để giúp người dùng tìm được nhiều thông tin liên quan và phù hợp trong một cơ sở dữ liệu lớn một cách nhanh chóng. Các hệ thống gợi ý được ứng dụng trong nhiều lĩnh vực như thương mại điện tử, giải trí, khoa học, tin tức… 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 sức tìm kiếm như hệ thống gợi ý phim *MovieLens*[1], *last.fm*[2], *Film-Conseil*. Trong lĩnh vực khoa học, hệ thống gợi ý hỗ trợ người dùng tìm kiếm các bài báo khoa học như hệ thống tìm kiếm *Citeseer*[3] 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. Trong lĩnh vực tin tức, người đọc được hệ thống hỗ

[1] http ://www.movielens.org
[2] http://www.last.fm
[3] http://citeseerx.ist.psu.edu

trợ gợi ý các bài báo phù hợp với từng người đọc riêng biệt ví dụ như *netnews, yahoo news*…

Một hệ thống gợi ý không thể được triển khai nếu chưa qua đánh giá. Việc đánh giá một hệ thống gợi ý là một giai đoạn cần thiết vì hiệu quả của một hệ thống gợi ý không chỉ phụ thuộc vào đặc điểm dữ liệu mà còn phụ thuộc vào mục đích gợi ý (Herlocker J.L *et al,* 2004). Nghĩa là một hệ thống gợi ý với cùng một giải thuật thì kết quả gợi ý có thể có hiệu quả khác nhau trên những tập dữ liệu khác nhau. Liên quan đến mục đích của hệ thống, một vài hệ thống gợi ý chú trọng đến tính đa dạng của các mục dữ liệu trong danh sách gợi ý nhưng một số khác lại chú trọng đến tính mới của các mục dữ liệu. Tùy thuộc vào đặc trưng dữ liệu và mục đích của hệ thống gợi ý, các phương pháp đánh giá khác nhau có thể được sử dụng. Ngoài ra, nó còn phụ thuộc vào điều kiện để đánh giá hệ thống như dựa vào dữ liệu có sẵn để đánh giá (offline) hay triển khai hệ thống và đánh giá trực tuyến (online).

Trong phần hai, chúng tôi sẽ trình bày khái quát về một hệ thống gợi ý. Tiếp theo các nghi thức kiểm tra hệ thống, các phương pháp đánh giá một hệ thống gợi ý được trình bày chi tiết trong phần ba và bốn. Cuối cùng là phần kết luận.

2 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 (Adomavicius, G. 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 (Adomavicius, G. and A. Tuzhilin, 2005): 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 (Schafer J.B., *et al,* 2007).

Cụ thể một hệ thống gợi ý có thể được miêu tả như trong Hình 1 (Trần Nguyễn Minh Thư, 2011). Tập hợp người dùng có thể là một người phụ nữ, một người đàn ông hay là một đứa trẻ. Người dùng này có thể mua, xem, chọn lựa, đọc hay đánh giá mục dữ liệu. "Người dùng" được xem như là tác nhân của hệ thống tác động lên các "mục dữ liệu". "Mục dữ liệu" có thể là quần áo, phim ảnh, sách vở, bài báo, bài hát, cd, trang web, rượu… Mối quan hệ giữa người dùng và mục dữ liệu có thể là quan hệ yêu thích, mong muốn, mua, đọc… Sau đó, hệ thống sẽ cung cấp một danh sách các mục dữ liệu đề nghị cho người dùng. Những mục dữ liệu đề nghị phải dựa trên tiêu chí phù hợp với sở thích, thói quen của người dùng. Mục tiêu cuối cùng của một hệ thống gợi ý là đưa ra một danh sách các mục dữ liệu tiềm năng phù hợp với nhu cầu, mong muốn của người dùng.

Hình 1: Sơ đồ tổng quát của một hệ thống gợi ý

Hệ thống gợi ý thông thường được xây dựng dựa trên 3 bước tuần tự (Sarwar B., *et al*, 2001; Sarwar, B and G. Karypis, 2000; Breese, J.S. and D. Heckerman, 1998). Bước thứ nhất chính là bước trình bày /tổ chức/ chọn lọc lại dữ liệu sẽ sử dụng để xây dựng hệ thống dựa trên dữ liệu có sẵn trong hệ thống (representation). Những dữ liệu có sẵn này (các mục dữ liệu, danh sách các người dùng, mối liên hệ giữa các người dùng) được thể hiện lại dưới nhiều dạng khác nhau như véc tơ các từ khóa hay các mối quan hệ, ... Tiếp đến, mối tương quan hoặc sự tương tự giữa những người dùng, giữa những mục dữ liệu sẽ được tính toán. Các chỉ số Pearson, véc tơ tương tự, các phương pháp gom nhóm,... được sử dụng ở bước này. Cuối cùng, hệ thống sẽ đưa ra một danh sách các mục dữ liệu đề nghị hoặc giá trị đánh giá dự đoán của mục dữ liệu (ví dụ như giá trị đánh giá dự đoán của một bộ phim hay một quyển sách).

Một hệ thống gợi ý được đánh giá bằng cách phân tích trên tập dữ liệu đã tồn tại (off-line evalution), hoặc lấy thông tin trực tiếp từ người sử dụng hệ thống (on-line evalution), hoặc kết hợp cả hai cách trên (Herlocker J.L *et al,* 2004; Mortensen M., 2007).

Trong cách đánh giá off-line, tập dữ liệu có sẵn được chia thành 2 phần, một phần dùng để huấn luyện, một phần dùng để kiểm tra (Fouss, F. and M. Saerens, 2008). Trong cách đánh giá này, người ta thường sử dụng nghi thức k-fold, hold-out. Đánh giá off-line được thực hiện nhanh chóng, ít tốn kém và có thể thực hiện trên tập dữ liệu lớn thậm chí có khả năng lặp lại các sự tương tác của hệ thống gợi ý. Tuy nhiên, nó chỉ dự đoán với tỉ lệ người dùng nhất định và không đo được sự hài lòng của người dùng thực và độ chính xác sẽ không cao như đánh giá online.

Trong cách đánh giá on-line, người dùng tương tác với hệ thống và nhận được những gợi ý thực sự. Hệ thống hỏi và thu nhận các câu trả lời từ đó đưa ra những gợi ý phù hợp thực tế đối với người sử dụng, đo được sự hài lòng của người dùng thực. Tuy nhiên, cách này mất nhiều thời gian và khó có điều kiện triển khai vì cần phải có sự hợp tác của người sử dụng.

3 NGHI THỨC KIỂM TRA

Nghi thức kiểm tra khá phổ biến được đề cập đến là hold-out và k-fold. Trong cả 2 nghi thức kiểm tra, tập dữ liệu đều được phân thành một tập học và một tập kiểm tra. Tuy nhiên, nghi thức hold-out chia tách tập dữ liệu thành hai phần, một phần dùng để học và một phần dùng để kiểm tra. Thông thường, lấy ngẫu nhiên 2/3 tập dữ liệu để học và 1/3 tập dữ liệu còn lại dùng để kiểm tra, có thể lặp lại quá trình này k lần rồi tính giá trị trung bình (Adomavicius, G. And Y. Kwon, 2008). Nghi thức k-fold chia tập dữ liệu ban đầu thành k phần (fold) bằng nhau, quá trình học và kiểm tra được thực hiện k lần, mỗi lần sử dụng k-1 folds để học và 1 fold để kiểm tra, sau đó tính trung bình của k lần kiểm tra (Adomavicius, G. And Y. Kwon, 2008; Herlocker J.L *et al,* 2004). Các nghi thức kiểm tra thông dụng áp dụng trong hệ thống các nghiên cứu với các tập dữ liệu được trình bày trong Bảng 1.

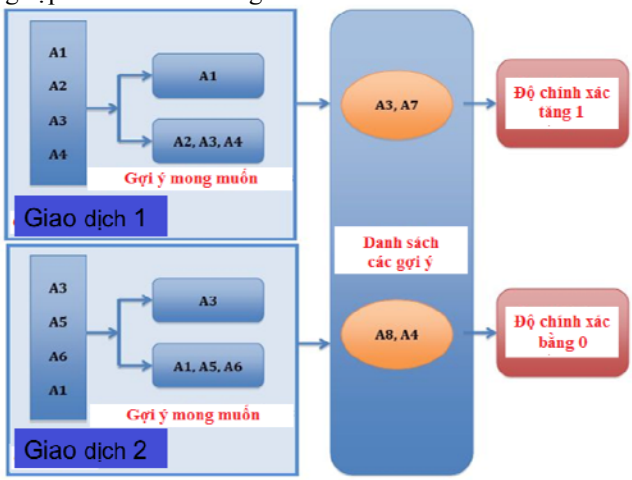
Bảng 1: Các nghi thức kiểm tra

STT	Hệ thống áp dụng	Tập dữ liệu	Nghi thức đánh giá
1.	*Sarwar B., et al, 2001*	MovieLens	Holdout 20% - 80%
2.	*Park, Y. and A. Tuzhilin, 2008*	MovieLens, BookCrossing	10 fold cross validation
3.	*Hsu, C. and H. Chung, 2004*	TaFeng, B&Q	Holdout P1.7% - 93.3%
4.	*Koren.Y, 2009; Takács G., et al, 2007; Kozma, L and T. Raiko, 2009*	Netflix	Holdout 2.9% - 97.1%
5.	*Dias et al, 2008 Ming Li, 2007*	Leshop	Leave One Out

Việc lựa chọn một nghi thức kiểm tra cho hệ thống gợi ý còn phụ thuộc sâu sắc vào đặc điểm của cơ sở dữ liệu. Một biến thể của phương pháp hold-out thường được sử dụng trong lĩnh vực thương mại điện tử của các hệ thống gợi ý chính là phương thức Given-N và AllButOne. Hai phương thức này được đề xuất bởi Hsu và các cộng sự khi đánh giá hệ thống gợi ý trong lĩnh vực thương mại điện tử (Hsu, C. and H. Chung, 2004).

Đối với phương thức Given-N, nguyên tắc của nó là xét tất cả các giao dịch có ít nhất N+1 mục dữ liệu. Danh sách các sản phẩm trong giao dịch được chia thành 2 tập, tập được gọi là Given (N sản phẩm) và một tập kiểm tra (phần còn lại của giao dịch). Sau khi hệ thống gợi ý đề nghị những sản

phẩm, ta so sánh chúng với các gợi ý thực tế (sản phẩm nằm trong phần kiểm tra), độ chính xác của hệ thống sẽ tăng lên 1 đơn vị khi sản phẩm gợi ý trùng với sản phẩm trong tập kiểm tra và bằng 0 khi sản phẩm gợi ý không trùng với sản phẩm trong tập kiểm tra. Hình 2 minh hoạ nghi thức đánh giá Given-N.



Hình 2: Nghi thức đánh giá Given-N

Đối với phương thức AllButOne, nó có thể triển khai cho các giao dịch có ít nhất 1 sản phẩm. Tương tự như Given-N, danh sách các sản phẩm trong giao dịch cũng được chia thành 2 tập, tập được gọi là Given và một tập kiểm tra. Trong đó, tập Given có số lượng bằng "tổng số giao dịch của các mặt hàng -1", tập kiểm tra thì luôn luôn bằng 1. Sau khi hệ thống gợi ý đề nghị những sản phẩm, chúng ta so sánh chúng với gợi ý thực tế (sản phẩm trong tập kiểm tra), độ chính xác của hệ thống sẽ tăng lên 1 đơn vị khi sản phẩm gợi ý trùng với sản phẩm trong tập kiểm tra và bằng 0 khi sản phẩm gợi ý không trùng với sản phẩm trong tập kiểm tra.

4 CÁC PHƯƠNG PHÁP ĐÁNH GIÁ

Có hai nhóm tiêu chí đánh giá: *các tiêu chí định lượng và tiêu chí định tính*. Các tiêu chí định lượng được dành riêng cho việc đánh giá số lượng các gợi ý liên quan, chúng tương ứng với độ chính xác (Trần Nguyễn Minh Thư, 2011). Với sự phát triển không ngừng, bên cạnh các tiêu chí định lượng thì người ta nghiên cứu thêm các tiêu chí đánh giá mới (tiêu chí định tính) nhằm có những đánh giá chính xác hơn để cải thiện hệ thống gợi ý. Các tiêu chí định tính được sử dụng để đánh giá chung về chất lượng của hệ thống gợi ý.

4.1 Tiêu chí định lượng

4.1.1 Đánh giá độ chính xác của các dự đoán

Việc đánh giá tính chính xác các dự đoán có thể sử dụng sai số bình phương trung bình (MSE - Mean Square Error), căn của sai số bình phương trung bình (RMSE - Root Mean Square Error), sai số tuyệt đối trung bình (MAE - Mean Absolute Error) (Herlocker J.L *et al,* 2004; Koren.Y, 2009; Trần Nguyễn Minh Thư, 2011). Tính chính xác của các dự đoán được đo trên n quan sát, trong đó p_i là giá trị dự đoán đánh giá của mục i và r_i là giá trị đánh giá thực tế của mục i.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)$$

Các chỉ số này thích hợp cho một cơ sở dữ liệu không phải nhị phân và cho một giá trị dự đoán là số. Nó giúp đo lường mức độ sai số của các dự đoán. Các giá trị đo lường này bằng 0 khi hệ thống đạt được hiệu quả tốt nhất. Giá trị này càng cao thì hiệu quả của hệ thống càng thấp.

Tại cuộc thi nhằm cải thiện độ chính xác của hệ thống gợi ý do Netflix[4] tổ chức, các hệ thống gợi ý đã được đánh giá bởi chỉ số RMSE. Các chỉ số MAE, MSE và RMSE đã được sử dụng để đánh giá hệ thống gợi ý mà kết quả là giá trị dự đoán các đánh giá như hệ thống MovieLens, BookCrossing. Những chỉ số này rất dễ sử dụng để đánh giá, tuy

[4] http://www.netflixprize.com/

nhiên MAE là biện pháp sử dụng nhiều nhất vì khả năng giải thích trực tiếp của nó.

4.1.2 Đánh giá việc sử dụng các dự đoán

Ngoài việc đánh giá tính chính xác của các dự đoán, một số chỉ số khác như *precision, recall và F_score*, R_{score} được dùng để đánh giá việc sử dụng của các dự đoán trong trường hợp cơ sở dữ liệu nhị phân *(*Herlocker J.L et al, 2004; Sarwar, B and G. Karypis, 2000; Breese, J.S. and D. Heckerman, 1998*)*. Các chỉ số này đánh giá các gợi ý phù hợp cho mỗi người dùng thay vì đánh giá số điểm liên quan đến từng đề nghị. Đề nghị được coi là phù hợp khi người dùng chọn mục dữ liệu từ danh sách những đề nghị đã được gợi ý cho người dùng.

Precision là tỷ lệ giữa số lượng các gợi ý phù hợp và tổng số các gợi ý đã cung cấp (đã tạo ra). Precision bằng 100% có nghĩa là tất cả các kiến nghị đều phù hợp.

$$\text{Precision} = \frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng gợi ý tạo ra}}$$

Recall được định nghĩa bởi tỉ lệ giữa số lượng các gợi ý phù hợp và số lượng các mục dữ liệu mà người dùng đã chọn lựa (xem, nghe, mua, đọc). Recall được sử dụng để đo khả năng hệ thống tìm được những mục dữ liệu phù hợp so với những gì mà người dùng cần.

$$\text{Recall} = \frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng sản phẩm mua bởi người dùng}}$$

Precision và Recall được xem là hữu ích trong việc đánh giá một gợi ý. Tuy nhiên, trong một số trường hợp thì precision và recall có giá trị tỉ lệ nghịch với nhau. Ví dụ như số lượng gợi ý mà hệ thống tạo ra là 10, số lượng gợi ý phù hợp là 3, số lượng sản phẩm mua bởi người dùng là 3 thì độ chính xác thấp (30%), tuy nhiên giá trị recall lại cao (100%) nghĩa là độ chính xác thấp nhưng người dùng lại hài lòng bởi vì họ mua có 3 sản phẩm và hệ thống gợi ý đúng cả 3 sản phẩm đó. Trong tình huống đó, chỉ số ***F-score*** được sử dụng để đánh giá hiệu quả tổng thể của hệ thống bằng cách kết hợp hài hòa hai chỉ số Recall và Precision.

$$\text{F-score} = \frac{2 \times \text{Precison} \times \text{Recall}}{\text{Precison} + \text{Recall}}$$

R_{score} hay Breese score (Breese, J.S. and D. Heckerman, 1998) cũng là một trong những chỉ số đánh giá khả năng sử dụng dự đoán nhưng chỉ số này chính xác đến thứ tự của các gợi ý được xây dựng. R_{score} đánh giá vị trí của sản phẩm được chọn bởi người dùng trong danh sách sản phẩm gợi ý được tạo ra bởi hệ thống. Ví dụ, một hệ thống gợi ý cho người dùng 10 sản phẩm sắp xếp theo thứ tự ưu tiên từ cao đến thấp. Nếu người dùng chọn sản phẩm đầu tiên trong danh sách thì hệ thống gợi ý hiệu quả hơn khi người dùng chọn sản phẩm có thứ tự thứ 10. Chỉ số Rscore được tính dựa vào tỉ lệ giữa thứ tự của mục gợi ý đúng ($Rankscore_p$) và thứ tự của mục gợi ý đúng tốt nhất ($Rankscore_{max}$) như công thức sau:

$$Rankscore = \frac{Rankscore_p}{Rankscore_{max}}$$

$$Rankscore_p = \sum_{i \in h} 2^{-\frac{Rank(i)-1}{\alpha}}$$

$$Rankscore_{max} = \sum_{i=}^{|T|} 2^{-\frac{i-1}{\alpha}}$$

Trong đó:

– h là tập các sản phẩm gợi ý đúng.

– Rank trả về thứ tự sắp xếp của một sản phẩm trong danh sách gợi ý

– T là tập tất cả các sản phẩm người dùng quan tâm

– α là chu kỳ nữa phân kỳ (xác suất mà mục dữ liệu trong danh sách gợi ý được chọn là 50%).

Các chỉ số *Precision, Recall và F_score*, R_{score} thường được sử dụng đối với các hệ thống gợi ý trong lĩnh vực thương mại điện tử. Các chỉ số đánh giá, công thức tương ứng và một số hệ thống gợi ý/ nghiên cứu đã áp dụng các chỉ số tương ứng đó được tổng hợp trong Bảng 2.

Bảng 2: Các phương pháp đánh giá

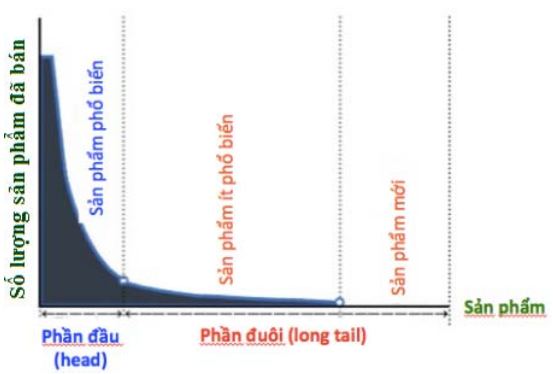
STT	Chỉ số	Công thức	Hệ thống đã áp dụng
1.	MAE	$\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)$	MovieLens
2.	MSE	$\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2$	Netflix
3.	RMSE	$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2}$	BookCrossing
4.	Precision	$\frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng gợi ý tạo ra}}$	EachMovie
5.	Recall	$\frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng sản phẩm mua bởi người dùng}}$	Yeong, *et al, 2005*
6.	F_{score}	$\text{F-score} = \frac{2 \times \text{Precison} \times \text{Recall}}{\text{Precison} + \text{Recall}}$	MovieLens
7.	R_{score} hay Breese score	$Rank\text{score} = \frac{Rank\text{score}_p}{Rank\text{score}_{\max}}$ $Rank\text{score}_p = \sum_{i \in h} 2^{-\frac{Rank(i)-1}{\alpha}}$ $Rank\text{score}_{\max} = \sum_{i=}^{\lvert T \rvert} 2^{-\frac{i-1}{\alpha}}$	TaFeng, B&Q (*Breese, J.S. and D. Heckerman, 1998, Hsu, C. and H. Chung, 2004)*

4.2 Tiêu chí định tính

Trong những giai đoạn đầu phát triển thì hệ thống gợi ý chỉ sử dụng các độ đo chính xác định lượng như đã đề cập. Tuy nhiên, người dùng ngày càng có yêu cầu cao hơn và nhiều hơn về chất lượng của các gợi ý. Nếu chỉ xét độ chính xác thì không đủ để đánh giá hiệu quả của một hệ thống gợi ý nên cần đưa thêm thuộc tính chất lượng các gợi ý thay vì chỉ sử dụng độ chính xác của các gợi ý. Các chỉ số đánh giá chất lượng có thể là tính đa dạng, tính mới lạ, khả năng cầu may (một gợi ý không mong đợi bởi người dùng nhưng cuối cùng lại phù hợp cho người dùng), phạm vi bao phủ của gợi ý (độ bao phủ của dự báo hay gợi ý). Một số chỉ số định tính sẽ được phân tích chi tiết trong nội dung tiếp theo (Herlocker J.L *et al,* 2004; Slaney.M, 2006; Takács G., *et al,* 2007; Yu, C and L. Lakshmanan, 2009).

4.2.1 Tính mới của các gợi ý

Việc đánh giá tính mới của gợi ý là hiển nhiên để đáp ứng nhu cầu của người sử dụng các sản phẩm mới được tạo ra liên tục. Khái niệm "sản phẩm mới" có thể có nhiều ý nghĩa khi đề cập đến hệ thống gợi ý (Herlocker J.L *et al*, 2004; Karypis.g, 2001). Tính mới của mục dữ liệu theo quan điểm thời gian (trong trường hợp xuất hiện một sản phẩm mới) hoặc liên quan đến lịch sử của người sử dụng (một sản phẩm mà chưa bao giờ được mua). Điều này xảy ra như một trường hợp đặc biệt mà mục dữ liệu trong hệ thống chưa có thông tin liên quan đến người sử dụng, như thể hiện "sản phẩm mới" trong Hình 3. Vấn đề này cũng được xác định như là một trong những khó khăn của hệ thống gợi ý – vấn đề "thiếu thông tin" (cold start problem).



Hình 3: Sự phổ biến của sản phẩm

Một số hệ thống cung cấp rất chính xác gợi ý nhưng không hữu dụng trong thực tế vì không quan

tâm đến các tiêu chí định tính trong quá trình xây dựng hệ thống. Ví dụ như hệ thống gợi ý "sữa tươi" cho khách hàng trong một siêu thị ở châu Âu. Đề nghị này là chính xác bởi vì hầu như tất cả các khách hàng đều mua sữa, nhưng nó không phải là hữu ích cho tất cả người dùng đã quen thuộc với sản phẩm này. Các nhà cung cấp nhận thức được điều này trong một thời gian dài và tổ chức các kệ đựng hàng hoá cho phù hợp. Do đó, việc giới thiệu cho người mua một thực phẩm mới có khả năng để làm hài lòng người mua mà họ không bao giờ nghĩ là cần thiết. Ví dụ thứ hai, liên quan đến sự cần thiết phải có thuộc tính mới trong các gợi ý. Giả sử một người dùng mua quà tặng trước Giáng sinh hai tuần. Trong trường hợp này, việc xây dựng danh sách các gợi ý căn cứ vào mặt hàng phổ biến của các năm trước thì không phù hợp bởi vì người dùng thường tìm cách tặng các sản phẩm mới cho người thân.

Thuộc tính mới được nhấn mạnh như là một chỉ số cần thiết để đánh giá tính hiệu quả của hệ thống gợi ý. G. Shani *và ctv* cùng với những nghiên cứu của mình đã chỉ ra 3 điểm quan trọng liên quan đến hiệu quả của hệ thống gợi ý. Thứ nhất, tính chính xác và tính mới lạ phải được tính đến để xây dựng được các gợi ý hiệu quả. Thứ hai, yếu tố thời gian là điều cần thiết trong việc đánh giá tính mới của mục dữ liệu. Thứ ba, danh sách gợi ý phù hợp nhất phải kết hợp một tỉ lệ các mục dữ liệu mới và các gợi ý phù hợp khác.

4.2.2 Tính đa dạng (Diversity) của các gợi ý

Sự đa dạng của hệ thống gợi ý đo lường khả năng cung cấp một danh sách các mục dữ liệu được phân phối từ nhiều loại khác nhau. Có thể phân chia sự đa dạng của các gợi ý thành hai loại đa dạng: sự đa dạng cá nhân và đa dạng tổng thể. Loại đa dạng cá nhân quan tâm đến các khái niệm về đa dạng từ quan điểm của người sử dụng. Chỉ số này được tính toán dựa trên trung bình sự khác nhau giữa tất cả các cặp mục dữ liệu đã gợi ý. Ngược lại, sự đa dạng tổng thể là quan tâm đến các mục dữ liệu đã gợi ý hơn là quan tâm đến người dùng. Nếu sự đa dạng tổng thể của hệ thống giới thiệu là lớn, thì sự đa dạng của các gợi ý cá nhân cũng là rất lớn, nhưng điều này không đúng cho chiều ngược lại. Ví dụ, hệ thống cung cấp 3 gợi ý khác nhau cho tất cả người dùng, thì sự đa dạng cá nhân là tương đối cao nhưng sự đa dạng tổng thể là rất thấp (Adomavicius, G. And Y. Kwon, 2008; Adomavicius, G. and Y. Kwon, 2010; Bradley, 2001; Takács G., *et al,* 2007; Ziegler C., *et al,* 2005).

Trong các hệ thống gợi ý truyền thống, sự đa dạng của các gợi ý chưa được quan tâm đến mặc dù chỉ số này rất quan trọng. Trong một số trường hợp, sự đa dạng sẽ trở thành một điều cần thiết. Ví dụ như sự đang dạng của các điểm tham quan cho các kỳ nghỉ lễ trong hệ thống gợi ý các địa điểm du lịch. Với thực tế đó, đã có nhiều nghiên cứu cải thiện hiệu quả của hệ thống gợi ý hướng đến sự đa dạng và các nghiên cứu này cũng đã khẳng định "Nếu chỉ tính đến độ chính xác của các gợi ý để đánh giá chất lượng của một hệ thống là không đủ để đảm bảo sự phù hợp, hiệu quả của những gợi ý cho người dùng" G. Adomavicius.

Hai yếu tố có tác động trực tiếp đến sự đa dạng của gợi ý là các thuật toán sử dụng để xây dựng hệ thống và các đặc tính của cơ sở dữ liệu. G. Adomavicius và Y. Kwon đã nghiên cứu sự ảnh hưởng của hai yếu tố này đến sự đa dạng của các gợi ý (Adomavicius, G. and Y. Kwon, 2010). Mối quan hệ này được biểu diễn trong *Bảng 3*: Bảng này cho thấy kết quả của hai trường hợp riêng biệt: các gợi ý được tạo ra, hoặc từ các mặt hàng phổ biến nhất (i), hoặc từ đuôi dài (long tail) (ii). Theo thông tin trong bảng này, sự đa dạng lớn hơn nhiều khi các gợi ý được tạo ra từ các mặt hàng thuộc đuôi dài (695 sản phẩm khác nhau). Trong trường hợp này, người ta cũng có thể nhận thấy rằng độ chính xác giảm. Điều đó chứng tỏ rằng sự đa dạng của các gợi ý sẽ tập trung vào các mục dữ liệu nằm thuộc phần đuôi dài, tuy nhiên cần phải có một tỉ lệ hợp lý để không làm giảm quá nhiều độ chính xác của hệ thống.

Bảng 3: Sự tương quan giữa độ chính xác và tính đa dạng

	Độ chính xác (precision)	Sự đa dạng
Sản phẩm nằm ở phần đầu (head)	82%	49 sản phẩm khác nhau
Sản phẩm nằm ở phần đuôi (tail)	68%	695 sản phẩm khác nhau

Với những thông tin trên, tầm quan trọng về sự đa dạng của các gợi ý đã được khẳng định. Tuy nhiên, sự đa dạng không phải là chỉ tiêu quan trọng nhất, độ chính xác cũng phải được tính đến. Do đó, phải có một sự thỏa hiệp giữa độ chính xác và sự đa dạng (nghĩa là giữa định lượng và định tính). Tính đa dạng có thể được quan tâm trực tiếp khi xây dựng danh sách các gợi ý hoặc danh sách gợi ý này sẽ được tính lại thứ tự sắp xếp. Một vài chỉ số để tính lại thứ tự của các mục dữ liệu như sự phổ biến của các mục dữ liệu, trung bình các đánh giá

cho mỗi mục dữ liệu, phần trăm người dùng đã có cùng một đánh giá cho một mục dữ liệu. Các giải thuật này chứng tỏ được tính hiệu quả ở sự đa dạng nhưng không làm giảm đáng kể độ chính xác của hệ thống. Ví dụ, giải thuật đề nghị của Adomavicius và Y. Kwon*(Adomavicius, G. and Y. Kwon, 2010)* đánh giá trên cơ sở dữ liệu MovieLens, thì hệ thống tăng thêm 20% tính đa dạng nhưng chỉ làm mất đi 1% độ chính xác. Ngoài ra, một số kỹ thuật khác cũng tính đến khái niệm "đa dạng", ví dụ như sự đa dạng của các đối tượng hoặc đa dạng về âm nhạc.

4.2.3 Độ bao phủ (coverage) của các gợi ý

Độ bao phủ của hệ thống gợi ý là thước đo số lượng lĩnh vực mà danh sách các sản phẩm gợi ý được tạo ra thuộc về chúng, số lĩnh vực này có bao trùm được hệ thống hay không (Herlocker J.L *et al,* 2004, Takács G., *et al,* 2007). Độ bao phủ của các gợi ý thấp thì thường ít được đánh giá cao bởi người dùng bị giới hạn thông tin về các lĩnh vực của hệ thống và họ cần được tư vấn đa lĩnh vực. Độ bao phủ đã được sử dụng trong đánh giá hệ thống gợi ý bởi một số nhà nghiên cứu như Good *et al*. 1999, Herlocker *et al*. 1999, Sarwar *et al*. 1998.

Hầu hết độ bao phủ được đo bằng số các mặt hàng mà dự đoán có thể được hình thành như là một tỷ lệ phần trăm của tổng số các mặt hàng. Cách dễ nhất để đo loại này là chọn một cách ngẫu nhiên cặp user/item, yêu cầu một dự đoán cho mỗi cặp, và đo tỷ lệ phần trăm mà dự đoán được cung cấp. Giống như chỉ số precision và recall phải được xem xét đồng thời, độ bao phủ (Coverage) thường được kết hợp với chỉ số "accuracy", vì không thể tăng giá độ bao phủ mà không quan tâm đến việc tạo ra những gợi ý không thuộc hệ thống. Một cách khác để tính độ bao phủ là chỉ xem xét độ bao phủ trên những mặt hàng mà người dùng quan tâm. Độ bao phủ tính theo cách này không được đo trên toàn bộ các sản phẩm mà chỉ quan tâm đến những sản phẩm mà khách hàng đã biết hay đã từng xem qua. Ưu điểm của cách tính này là nó đáp ứng tốt nhu cầu của người dùng.

Độ bao phủ được đo bằng sự phong phú của hồ sơ người dùng để đưa ra gợi ý. Ví dụ trong lọc cộng tác, người dùng phải đánh giá các item trước khi nhận các gợi ý. Việc đo lường này là một loại hình đặc trưng trong nghi thức đánh giá off-line.

4.2.4 Sự hài lòng của người sử dụng

Sự hài lòng của người sử dụng là một khía cạnh hơi mơ hồ và phụ thuộc vào từng cá nhân khác nhau và do đó rất khó để đo lường. Theo định nghĩa của *(Herlocker J.L et al, 2004)* thì sự hài lòng của người dùng được định nghĩa là mức độ mà một người dùng được hỗ trợ trong việc đối phó với các vấn đề quá tải thông tin. Herlocker *et al* đã phân loại một số phương pháp đánh giá sự hài lòng của người sử dụng.

Phương pháp đánh giá "rõ ràng" (Explicit) và "ngầm hiểu" (Implicit): phương pháp đánh giá một cách rõ ràng nghĩa là hệ thống đo độ hài lòng của người sử dụng bằng cách yêu cầu trực tiếp; phương pháp đánh giá ngầm hiểu thì cần phải đặt ra những giả định và dịch những quan sát được thành những giả định, ví dụ như sự gia tăng doanh số của một cửa hàng chứng tỏ sự hài lòng của khách hàng tăng lên.

– Kết quả so với quá trình: việc đánh giá có thể chỉ tập trung vào kết quả, nhưng nó cũng có thể tập trung vào quá trình áp dụng hệ thống gợi ý.

Đánh giá trong khoảng thời gian ngắn (Short term) và khoảng thời gian dài (Long term): đánh giá người sử dụng trong một khoảng thời gian ngắn có thể sẽ thiếu sót thông tin mà nó sẽ trở nên chính xác hơn sau một khoảng thời gian nhất định. Sở thích của người dùng cần phải xem xét đánh giá của người dùng qua một khoảng thời gian dài. Các nghiên cứu điều tra sự hài lòng của người dùng đối với hệ thống gợi ý là rất hiếm và nghiên cứu tập trung trên sự hài lòng của các gợi ý thì càng hiếm hơn. Tiêu chí đánh giá này được sử dụng trong nghiên cứu của Cosley (Cosley D., *et al.*, 2003) và Herlocker (Herlocker J. L., *et al.*, 2000).

5 KẾT LUẬN

Bài báo đưa ra cái nhìn tổng quan về hệ thống gợi ý cũng như phân tích chi tiết các vấn đề liên quan đến việc đánh giá một hệ thống gợi ý. Các nghi thức đánh giá, các chỉ số đánh giá cho từng mục đích hay tập dữ liệu khác nhau cũng được trình bày chi tiết. Các phương pháp đánh giá cũng được hệ thống theo định lượng dựa trên các công thức MSE, RMSE, MAE, Precision, Recall, F-score và theo định tính dựa trên tính mới, tính đa dạng của các gợi ý. Độ bao phủ của các gợi ý và sự hài lòng của người sử dụng cũng được phân tích rõ ràng. Từ đó giúp việc chọn lựa các phương pháp đánh giá để triển khai hệ thống được phù hợp và hiệu quả đối với từng dữ liệu cụ thể. Tóm lại, để xây dựng hệ thống gợi ý chính xác và hữu dụng, chúng ta cần quan tâm đến cách thức đánh giá hệ thống cũng như các chỉ số đánh giá phù hợp.

Tuy nhiên, bài báo này chỉ tổng hợp kết quả nghiên cứu các phương pháp đánh giá hệ thống gợi

ý trong khoảng thời gian từ năm 2000- 2010, chưa tổng hợp các phương pháp đánh giá mới nhất. Bên cạnh đó, chúng tôi cũng chỉ quan tâm đến các chỉ số đánh giá tính hiệu quả của hệ thống mang lại mà chưa quan tâm đến tiêu chí về thời gian xây dựng hay thời gian đáp ứng của hệ thống. Chúng tôi sẽ tiếp tục phát triển thêm ở các nghiên cứu khác.

TÀI LIỆU THAM KHẢO

Adomavicius, G. 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.

Adomavicius, G. And Y. Kwon, 2008. Overcoming accuracy-diversity tradeoff in recommender systems: a variance-based approach. In Proceedings of the 18th Workshop on Information Technology and Systems, WITS 2008, Paris, France.

Adomavicius, G. and Y. Kwon, 2010. Improving recommendation diversity using ranking-based techniques. IEEE Transactions on Knowledge and Data Engineering.

Bradley, 2001. Improving recommendation diversity. In Proceedings of the 12th National Conference in Artificial Intelligence and Cognitive Science. D. O'donoghue, Ed., Maynooth, Ireland, pp. 75–84.

Breese, J.S. and D. Heckerman, 1998. Empirical analysis of predictive algorithms for collaborative filtering. Morgan Kaufmann, pp. 43–52.

Cosley D., et al, 2003. Is seeing believing?: how recommender system interfaces affect users' opinions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03). ACM, New York, NY, USA, 585-592. DOI=10.1145/642611.642713 http://doi.acm.org/10.1145/642611.642713

Dias et al, 2008. The value of personalised recommender systems to e-business : a case study. Recsys '08. New York, NY, USA : ACM, pp. 291–294.

Fouss, F. and M. Saerens, 2008. Evaluating performance of recommender systems: an experimental comparison. Web intelligence. IEEE, pp. 735–738.

Herlocker J. L., et al, 2000. Explaining collaborative filtering recommendations. In Proceedings of the 2000 Conference on Computer Supported Cooperative Work, 241–250.

Herlocker J.L et al, 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst., vol. 22, no. 1, pp. 5–53.

Hsu, C. and H. Chung, 2004. Mining skewed and sparse transaction data for personalized shopping recommendation. Mach. Learn., vol. 57, no. 1-2, pp. 35–59.

Karypis.g, 2001. Evaluation of item-based top-n recommendation algorithms. Cikm '01: proceedings of the tenth international conference on information and knowledge management. New york, ny, usa : acm, pp. 247–254.

Koren.Y, 2009. The Bellkor solution to the netflix grand prize. Http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1. 162.2118.

Kozma, L and T. Raiko, 2009. Binary principal component analysis in the netflix collaborative filtering task. In Proceedings of the IEEE Workshop on Machine Learning for Signal Processing.

Mortensen M., 2007. Design and evaluation of a recommender system. Http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.103.2726

Park, Y. and A. Tuzhilin, 2008. The long tail of recommender systems and how to leverage it. Recsys. ACM, pp. 11–18.

Sarwar B., et al, 2001. Item-based collaborative filtering recommendation algorithms. Proc. 10th international conference on the world wide web, pp. 285–295.

Sarwar, B and G. Karypis, 2000. Analysis of recommendation algorithms for ecommerce. EC '00. USA : ACM, pp. 158–167.

Schafer J.B., et al, 2007. Collaborative filtering recommender systems. The Adaptive Web, Ser. Lecture Notes in Computer Science, P. Brusilovsky, A. Kobsa, and W. Nejdl, eds. Springer Berlin, heidelberg, vol. 4321, pp. 291–324.

Slaney.M, 2006. Measuring playlist diversity for recommendation systems. In Proceedings of the 1st ACM Workshop on Audio and Music Computing Multimedia, Ser. AMCMM '06. New York, Ny, USA: ACM, pp. 77–82.

Takács G., et al, 2007. On the gravity recommendation system. Proc. Of the KDD CUP and Workshop 2007 (KDD 2007), pp. 22–30.

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.

Yeong, et al, 2005. Mining changes in customer buying behavior for collaborative recommendations. Expert Syst. Appl. 28, 2 (February 2005), 359-369. DOI=10.1016/j.eswa.2004.10.015 http://dx.doi.org/10.1016/j.eswa.2004.10.015.

Yu, C and L. Lakshmanan, 2009. It takes variety to make a world: diversification in recommender systems. In EDBT '09 Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology.

Ziegler C., et al, 2005. Improving recommendation lists through topic diversification. Proceedings of the 14th International Conference on World wide web, ser. WWW '05. New York, NY, USA : ACM, pp. 22–32.