Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: sj.ctu.edu.vn

DOI:10.22144/ctu.jsi.2017.023

LỰA CHỌN MÔ HÌNH VÀ THAM SỐ CHO BÀI TOÁN TƯ VẤN LỌC CỘNG TÁC DỰA TRÊN ĐỒ THỊ ĐÁNH GIÁ

Phan Quốc Nghĩa[1], Đặng Hoài Phương[2] và Huỳnh Xuân Hiệp[3]

[1]*Phòng Khảo thí, Trường Đại học Trà Vinh*
[2]*Khoa Công nghệ Thông tin, Trường Đại học Bách khoa Đà Nẵng*
[3]*Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận bài: 15/09/2017
Ngày nhận bài sửa: 10/10/2017
Ngày duyệt đăng: 20/10/2017

Title:
Select models and parameters for collaborative filtering recommender problems based on evaluation charts

Từ khóa:
Giải thuật máy học, hệ tư vấn, kỹ thuật thống kê, mô hình tư vấn lọc cộng tác

Keywords:
Collaborative filtering, machine learning algorithm, recommender systems, recommender models, statistical techniques

ABSTRACT

Recommender system is considered one of the most effective solutions that can cope with information explosion due to the rapid development of Internet services and is widely applied in many fields. However, to design a recommender system can meet the needs of users, the selection of suitable models for the recommender system and choosing the appropriate value of parameters for the model are always big challenges of designers. This study proposes solutions to choose models and value of parameters suitable for specific collaborative filtering recommender systems. To evaluate the proposed solutions, experiments on three standard datasets of MovieLens, MSWeb, and Jester5k are conducted. Experimental results show that the proposed solutions can assist designers and researchers to quickly identify model and the value parameters model for their specific collaborative filtering recommender systems.

TÓM TẮT

Hệ tư vấn được xem là một giải pháp hiệu quả có thể ứng phó với vấn đề bùng nổ thông tin do sự phát triển quá nhanh của các dịch vụ Internet và được ứng dụng rộng rãi trong nhiều lĩnh vực. Tuy nhiên, để thiết kế một hệ tư vấn có thể đáp ứng được nhu cầu của người dùng thì việc lựa chọn mô hình phù hợp cho hệ thống tư vấn và lựa chọn các giá trị tham số thích hợp cho mô hình luôn là một thách thức lớn của người thiết kế. Trong nghiên cứu này, chúng tôi đề xuất giải pháp lựa chọn mô hình và các giá trị tham số phù hợp cho bài toán tư vấn lọc cộng tác cụ thể. Để đánh giá các giải pháp đề xuất, chúng tôi tiến hành thực nghiệm trên ba tập dữ liệu chuẩn gồm: MovieLens, MSWeb và Jester5k. Kết quả thực nghiệm cho thấy các giải pháp của chúng tôi đề xuất có thể hỗ trợ nhà thiết kế, nhà nghiên cứu xác định được mô hình cũng như các giá trị tham số của mô hình cho bài toán tư vấn cụ thể của họ một cách nhanh chóng.

Trích dẫn: Phan Quốc Nghĩa, Đặng Hoài Phương và Huỳnh Xuân Hiệp, 2017. Lựa chọn mô hình và tham số cho bài toán tư vấn lọc cộng tác dựa trên đồ thị đánh giá. Tạp chí Khoa học Trường Đại học Cần Thơ. Số chuyên đề: Công nghệ thông tin: 171-178.

1 GIỚI THIỆU

Hệ tư vấn lọc cộng tác (collaborative filtering) (Aggarwal, 2016; Schafer *et al.*, 2007; Nghe, 2016) đã được ứng dụng thành công trong lĩnh vực thương mại điện tử như Amazon (Greg *et al.*, 2003), Netflix (Carlos and Neil, 2015) và Pandora (Michael Howe, 2007). Nó là một trong những giải pháp hiệu quả để giải quyết vấn đề bùng nổ thông tin cho các hệ thống trực tuyến nơi mà số lượng người dùng tăng lên rất nhanh. Hệ tư vấn lọc cộng tác giúp người dùng có thể tìm nhanh hơn các sản phẩm mà họ cần mua dựa trên dữ liệu xếp hạng của người dùng cho các sản phẩm trong quá khứ. Để dự đoán được các sản phẩm mà người thích dựa trên ma trận xếp hạng, nhiều mô hình lọc cộng tác được đề xuất như mô hình lọc cộng tác dựa trên người dùng (User-based collaborative filtering) (Martin *et al.*, 2014; Michael *et al.*, 2010; Nghia *et al.*, 2016), mô hình lọc cộng tác dựa trên sản phẩm (Item-based collaborative filtering) (Martin *et al.*, 2014; Michael *et al.*, 2010), mô hình lọc cộng tác dựa trên luật kết hợp (Collaborative filtering based on association rules) (Ahmed, 2015; Nghia *et al.*, 2015) và nhiều mô hình khác. Bên cạnh đó, dựa trên loại dữ liệu xếp hạng của người dùng của từng bài toán tư vấn, các mô hình lọc cộng tác tiếp tục được phát triển sâu hơn để xử lý cho từng loại dữ liệu như mô hình lọc cộng tác dựa trên người dùng cho ma trận xếp hạng dạng số thực (realRatingMatrix), mô hình lọc cộng tác dựa trên người dùng cho ma trận xếp hạng dạng nhị phân (binaryRatingMatrix) (Michael Hahsler, 2015). Chính vì sự đa dạng của các mô hình lọc cộng tác và sự tương thích dữ liệu đầu vào của từng mô hình đã gây sự khó khăn cho việc phát triển các hệ tư vấn. Thứ nhất là làm thế nào để chọn được mô hình tư vấn lọc cộng tác phù hợp cho bài toán tư vấn cụ thể. Thứ hai là làm thế nào để chọn được các tham số phù hợp cho mô hình tư vấn lọc cộng tác đã chọn. Từ hai vấn đề trên, trong nghiên cứu này, chúng tôi đề xuất phương pháp có thể hỗ trợ nhà thiết kế, nhà nghiên cứu lựa chọn nhanh mô hình cũng như các tham số của mô hình cho bài toán tư vấn cụ thể.

2 MÔ HÌNH TƯ VẤN LỌC CỘNG TÁC

Mô hình tư vấn lọc cộng tác sử dụng dữ liệu xếp hạng của người dùng cho các sản phẩm để dự đoán các giá trị xếp hạng cho các sản phẩm mà người dùng chưa xếp hạng hoặc tạo ra danh sách các sản phẩm cần tư vấn cho người dùng (Martin *et al.*, 2014; Michael Hahsler, 2015; Nghe, 2016). Mô hình được mô tả như sau: $U = \{u_1, u_2, \dots, u_m\}$ là tập các người dùng; $I = \{i_1, i_2, \dots, i_n\}$ là tập các sản phẩm; $R = \{r_{jk}\}$ là ma trận xếp hạng của m người dùng cho n sản phẩm. Trong đó, mỗi dòng của ma trận chứa các giá trị của một người dùng, mỗi cột của ma trận chứa các giá trị xếp hạng cho một sản phẩm, r_{jk} là giá trị xếp hạng của người dùng u_j cho sản phẩm i_k (Michael Hahsler, 2011). Do tính đặc trưng của dữ liệu xếp hạng, ma trận xếp hạng chỉ chứa một số rất nhỏ các giá trị xếp hạng của người dùng và rất nhiều ô của ma trận có giá trị rỗng. Mục tiêu của mô hình lọc cộng tác là tìm cách dự đoán các giá trị còn rỗng của ma trận từ các giá trị xếp hạng đã có; từ đó xác định danh sách các sản phẩm cần tư vấn cho người dùng cụ thể.

2.1 Mô hình tư vấn lọc cộng tác dựa trên người dùng

Mô hình lọc cộng tác dựa trên người dùng (UBCF) là mô hình lọc cộng tác thuộc nhóm giải pháp tư vấn dựa trên bộ nhớ (memory-based approach) (Martin *et al.*, 2014; Michael Hahsler, 2011; Nghia *et al.*, 2016). Mô hình này tìm ra kết quả tư vấn theo quy tắc truyền miệng (word of mouth) dựa trên ma trận xếp hạng của người dùng. Quy tắc này cho rằng những người dùng có cùng sở thích sẽ xếp hạng các sản phẩm ở mức tương tự nhau. Vì vậy, các giá trị chưa xếp hạng của người dùng u_a có thể được dự đoán bằng cách kết hợp các giá trị xếp hạng của những người dùng có sở thích tương đồng với người dùng u_a. Khi đó, hai người dùng có sở thích tương đồng được xem như là láng giềng của nhau và số lượng láng giềng được xác định bởi một tham số cho trước (k nearest neighbors) hoặc dựa trên ngưỡng giá trị tương đồng cho trước. Thông thường, giá trị tương đồng giữa hai người dùng được đo bằng hai độ đo phổ biến: hệ số tương quan Pearson (Pearson correlation coefficient) và độ đo tương đồng Cosine (Cosine similarity). Ví dụ, giá trị tương đồng giữa hai người dùng u và v được xác định như sau (Martin *et al.*, 2014):

$$S_{\text{Pearson}}(u, v) = \frac{\sum_{i \in I}(r_{v,i} - \bar{r}_v)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i \in I}(r_{v,i} - \bar{r}_v)^2}\sqrt{\sum_{i \in I}(r_{u,i} - \bar{r}_u)^2}} \quad (1)$$

Với $S(u, v)$ là giá trị tương đồng giữa người dùng u và người dùng v; I là tập các sản phẩm được xếp hạng bởi cả hai người dùng; $r_{v,i}$ là giá trị xếp hạng của người dùng v cho sản phẩm i; $\bar{r}_v$ là giá trị xếp hạng trung bình của người dùng v; $r_{u,i}$ là giá trị xếp hạng của người dùng u cho sản phẩm i; $\bar{r}_u$ là giá trị xếp hạng trung bình của người dùng u;

$$S_{Cosine}(u, v) = \frac{\vec{r}_u \cdot \vec{r}_v}{\|\vec{r}_u\|^2 \, X \, \|\vec{r}_v\|^2} = \frac{\sum_{i=1}^{m} r_{u,i} r_{v,i}}{\sqrt{\sum_{i=1}^{m} r_{u,i}^2}\sqrt{\sum_{i=1}^{m} r_{v,i}^2}} \quad (2)$$

Với $S(u, v)$ là giá trị tương đồng giữa người dùng u và người dùng v; m là số chiều của vector

(số sản phẩm); $r_{u,i}$ là giá trị xếp hạng của người dùng u cho sản phẩm i; $r_{v,i}$ là giá trị xếp hạng của người dùng v cho sản phẩm i;

Sau khi xác định được danh sách láng giềng của người dùng cần tư vấn (u_a), các giá trị xếp hạng của họ được tích hợp lại để tính ra giá trị xếp hạng dự đoán người dùng u_a đối với sản phẩm i. Thông thường, kết quả xếp hạng dự đoán được tính dựa trên trọng số trung bình của giá trị xếp hạng của k láng giềng người dùng và được biểu diễn bằng công thức sau:

$$P(u,i) = \bar{r}_u + \frac{\sum_{u' \in N} S(u,u')(r_{u',i} - \bar{r}_{u'})}{\sum_{u' \in N} |S(u,u')|} \quad (3)$$

2.2 Mô hình tư vấn lọc cộng tác dựa trên sản phẩm

Mô hình lọc cộng tác dựa trên sản phẩm (IBCF) là mô hình lọc cộng tác thuộc nhóm giải pháp tư vấn dựa trên mô hình (Model-based approach) (Martin *et al.*, 2014; Michael Hahsler, 2011; Nghia *et al.*, 2017). Mô hình này tìm ra các sản phẩm cần tư vấn dựa trên mối quan hệ giữa các sản phẩm được suy ra từ ma trận xếp hạng của người dùng. Giả thuyết của mô hình là người dùng sẽ thích các sản phẩm có sự tương đồng với các sản phẩm khác mà họ đã mua hoặc xếp hạng cao trong quá khứ. Bước đầu tiên của mô hình là xây dựng ma trận tương đồng S với kích cỡ n x n cho tất cả các cặp sản phẩm tương đồng dựa trên các độ đo tương đồng như Pearson, Jaccard, Cosine. Sau đó, mô hình dựa trên ma trận tương đồng này để tính tổng trọng số các giá trị xếp hạng của người dùng cho các sản phẩm liên quan. Từ đó, mô hình sẽ dự đoán ra sản phẩm nào mà người dùng xếp hạng cao nhất. Tuy nhiên, để tiết kiệm không gian lưu trữ và thời gian tính toán, mô hình lọc cộng tác dựa trên sản phẩm đưa ra giải pháp giảm kích thước của mô hình bằng cách xây dựng ma trận tương đồng kích cỡ n × k thay vì kích cỡ n x n. Trong đó, k là số sản phẩm tương đồng nhất với sản phẩm i đang xét và k rất nhỏ so với n. Khi đó giá trị xếp hạng trung bình của người dùng u cho sản phẩm i được xác định như sau (Michael Hahsler, 2011):

$$\hat{r}_{ui} = \frac{1}{\sum_{j \in S(i)} s_{ij}} \sum_{j \in S(i)} s_{ij} r_{uj} \quad (4)$$

Với S(i) là tập k sản phẩm tương đồng với sản phẩm i; s_{ij} là giá trị tương đồng của sản phẩm i với sản phẩm j; r_{uj} là giá trị xếp hạng của người dùng u cho sản phẩm j.

2.3 Mô hình lọc cộng tác dựa trên luật kết hợp

Mô hình lọc cộng tác dựa trên luật kết hợp (AR) là mô hình tư vấn sử dụng luật kết hợp để sinh ra kết quả tư vấn cho người dùng (JinHyun *et al.*, 2016; Nghia *et al.*, 2015). Mô hình chỉ áp dụng trên dữ liệu xếp hạng dạng nhị phân. Các sản phẩm được dự đoán phụ thuộc vào tập luật kết hợp được sinh ra dựa trên ma trận xếp hạng của người dùng. Trong đó, mỗi người dùng được xem như một giao dịch bao gồm các sản phẩm được họ xếp hạng bằng 1. Khi đó, một giao dịch k được định nghĩa $T_k = \{i_j \in I | r_{jk} = 1\}$ và ma trận xếp hạng trở thành cơ sở dữ liệu giao dịch $D = \{T_1, T_2, \dots, T_u\}$ với u là số người dùng. Để xây dựng mô hình, các luật kết hợp được sinh ra từ cơ sở dữ liệu giao dịch với định dạng $X \rightarrow Y$ với $X, Y \subseteq I$ và $X \cap Y = \emptyset$ (Michael Hahsler, 2011, 2015). Do số luật kết hợp được sinh ra rất lớn nên mô hình này chỉ chọn các luật kết hợp theo hai yêu cầu sau: các luật có vế phải chứa 1 phần tử (số phần tử của Y bằng 1) và đạt trên ngưỡng của các độ đo hấp dẫn (ví dụ: Support, Confidence và Implication Index). Dựa trên tập luật kết hợp đã chọn, mô hình tìm ra các sản phẩm cần tư vấn cho người dùng u_a gồm hai bước sau: Đầu tiên, tìm tất cả các luật kết hợp có vế trái chứa các sản phẩm mà người dùng u_a đã xếp hạng bằng 1 trong cơ sở dữ liệu giao dịch. Sau đó, chọn N sản phẩm từ vế phải của các luật có giá trị hấp dẫn cao nhất trong danh sách các luật đã chọn để tư vấn cho người dùng u_a.

2.4 Các mô hình lọc cộng tác khác

Ngoài ba mô hình phổ biến đã trình bày ở trên, trong thực tế, các hệ tư vấn lọc cộng tác còn sử dụng nhiều mô hình khác tùy thuộc vào bài toán tư vấn (dữ liệu đầu vào) như mô hình dựa vào phương pháp phân tích giúp giảm số chiều của dữ liệu PCA (Principal Component Analysis) (Michael Hahsler, 2015; Manolis and Konstantinos, 2008), mô hình dựa trên sản phẩm phổ biến (Recommender based on item popularity) (Harald Steek, 2011; Michael Hahsler, 2015), mô hình sinh kết quả tư vấn ngẫu nhiên (Produce random recommendations) (Michael Hahsler, 2015), mô hình dựa trên phương pháp triển khai phân tích ma trận SVD (Singular Value Decomposition) (Nghe and Hiep, 2012; Nghe and Phong, 2014; Xun Zhou *et al.*, 2015).

3 ĐÁNH GIÁ MÔ HÌNH LỌC CỘNG TÁC

Đánh giá mô hình tư vấn lọc cộng tác được dựa trên giả thuyết nếu mô hình chạy tốt trên dữ liệu kiểm tra (các sản phẩm được người dùng xếp hạng) thì nó sẽ cho kết quả dự đoán tốt cho dữ liệu mới (các sản phẩm chưa được người dùng xếp hạng) (Isinkaye *et al.*, 2015; Feng Zhang *et al.*, 2016). Trong đó, ma trận xếp hạng của người dùng được chia làm hai phần dựa trên dòng (users): phần được dùng để mô hình học gọi là tập huấn luyện và phần được dùng để kiểm tra kết quả dự đoán của mô

hình gọi là tập kiểm tra. Một mô hình được đánh giá là tốt nếu nó đưa ra các giá trị xếp hạng gần giống với các giá trị xếp hạng mà người dùng đã xếp hạng cho các sản phẩm trong tập kiểm tra hoặc các sản phẩm được mô hình chọn làm kết quả tư vấn cho người dùng là các sản phẩm được người dùng đó mua hoặc xếp hạng cao trong tập kiểm tra (Herlocker *et al.*, 2004; Michael Hahsler, 2011). Để đánh giá độ chính xác của mô hình tư vấn lọc cộng tác, người ta sử dụng một trong hai phương pháp sau: đánh giá dựa trên giá trị xếp hạng dự đoán và đánh giá dựa trên kết quả dự đoán.

Trong bài viết này, chúng tôi sử dụng phương pháp đánh giá dựa trên kết quả dự đoán của mô hình. Phương pháp này đánh giá độ chính xác của mô hình bằng cách so sánh các sản phẩm của mô hình đưa ra với các sản phẩm được người dùng xếp hạng cao. Độ chính xác của mô hình được xác định thông qua các chỉ số: độ chính xác (Precision), độ bao phủ (Recall) và trung bình điều hòa giữa độ chính xác và độ bao phủ (F-measure) (Michael Hahsler, 2011; Suresh and Michele, 2015). Giá trị của các chỉ số này được tính dựa trên ma trận hỗn độn 2x2 (Confusion matrix). Mô hình được đánh giá là tốt khi các chỉ số trên có giá trị cao.

Bảng 1: Ma trận hỗn độn (Confusion matrix)

Xếp hạng của người dùng	**Kết quả của mô hình**	
	Giới thiệu	**Không giới thiệu**
Xếp hạng cao	**TP**	**FN**
Xếp hạng thấp	**FP**	**TN**

Trong đó:

TP: Những sản phẩm được mô hình giới thiệu đã được người dùng xếp hạng cao.

FP: Những sản phẩm được mô hình giới thiệu đã được người dùng xếp hạng thấp.

FN: Những sản phẩm không được mô hình giới thiệu đã được người dùng xếp hạng cao.

TN: Những sản phẩm không được mô hình khuyến nghị đã được người dùng xếp hạng thấp.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (7)$$

4 LỰA CHỌN MÔ HÌNH DỰA TRÊN ĐỒ THỊ ĐÁNH GIÁ

Để xác định được mô hình phù hợp cho bài toán tư vấn lọc cộng tác người thiết kế hệ thống thường phải mất nhiều thời gian cho việc lựa chọn mô hình. Do mỗi mô hình chỉ phù hợp với một số bài toán cụ thể. Để giúp người thiết kế hệ thống chọn được mô hình nhanh hơn, chúng tôi đề xuất giải thuật lựa chọn mô hình tư vấn lọc cộng tác dựa trên đồ thị đánh giá như sau:

Giải thuật: Chọn mô hình lọc cộng tác

Input: Ma trận dữ liệu xếp hạng; Danh sách các mô hình;

Output: Đồ thị đánh giá;

Begin

<Dữ liệu đánh giá> = Xulydulieu(<Ma trận xếp hạng>, <Phương pháp xử lý dữ liệu>);

<Danh sách mô hình>= <các mô hình lọc cộng tác cần kiểm tra>;

<Kết quả đánh giá> = Danhgiamohinh(<Dữ liệu đánh giá>, <Danh sách mô hình>);

<Đồ thị đánh giá> = Vedothi(<Kết quả đánh giá>, <Kiểu đồ thị>);

Return(<Đồ thị đánh giá>);

End;

Trong giải thuật trên, đầu tiên, người thiết kế hệ thống xác định cách thức xây dựng dữ liệu cho bài toán tư vấn như cắt tập dữ liệu thành hai phần theo tỷ lệ cho trước (Splitting), cắt tập dữ liệu ngẫu nhiên nhiều lần (Bootstrap sampling), cắt tập dữ liệu thành k phần tương tự nhau (K-fold cross-validation) (Michael Hahsler, 2011; Suresh and Michele, 2015); tiếp theo, người thiết kế hệ thống chọn danh sách các mô hình lọc cộng tác cần thử nghiệm trên dữ liệu của bài toán; tiếp đến, thực hiện đánh giá đồng thời các mô hình và vẽ đồ thị đánh giá để so sánh độ chính xác của các mô hình. Bước này giúp người thiết kế hệ thống rút ngắn thời gian chọn mô hình. Thay vì phải thử nghiệm trên từng mô hình rồi so sánh kết quả thì chỉ cần thử nghiệm một lần duy nhất sẽ nhận được kết quả so sánh; cuối cùng, đọc kết quả từ đồ thị đánh giá để xác định mô hình hiệu quả nhất cho bài toán tư vấn.

5 LỰA CHỌN THAM SỐ DỰA TRÊN ĐỒ THỊ ĐÁNH GIÁ

Làm thế nào để xác định được giá trị phù hợp cho các tham số khi thực thi các mô hình tư vấn lọc cộng tác là một khâu quan trọng trong quá trình thiết kế hệ tư vấn. Ví dụ, để áp dụng mô hình lọc cộng tác dựa trên người dùng trên tập dữ liệu MovieLens ta nên chọn độ đo tương đồng nào thì mô hình sẽ cho kết quả chính xác cao hoặc mô hình sẽ cho kết quả cao nhất với bao nhiêu người dùng tương đồng. Từ mô hình tư vấn lọc cộng tác đã chọn, chúng tôi đề xuất giải thuật xác định giá trị phù hợp cho các tham số của mô hình lọc cộng tác dựa trên đồ thị đánh giá như sau:

Giải thuật: Chọn giá trị tham số cho mô hình lọc cộng tác

Input: Ma trận dữ liệu xếp hạng, mô hình tư vấn và danh sách giá trị của tham số;

Output: Đồ thị đánh giá;

Begin

<Dữ liệu đánh giá> = Xulydulieu(<Ma trận xếp hạng>, <Phương pháp xử lý dữ liệu>);

<Danh sách giá trị tham số> = <Các giá trị tham số cần kiểm tra>;

<Kết quả đánh giá> = Danhgiamohinh(<Dữ liệu đánh giá>, <Mô hình tư vấn>, <Danh sách giá trị tham số>);

<Đồ thị đánh giá> = Vedothi(<Kết quả đánh giá>, <Kiểu đồ thị>);

Return(<Đồ thị đánh giá>);

End;

Trong giải thuật này, đầu tiên, dữ liệu đánh giá được xử lý tương tự như trong giải thuật chọn mô hình tư vấn; tiếp theo, người thiết kế hệ thống chọn danh sách giá trị của tham số cần kiểm tra trên dữ liệu của bài toán; tiếp đến, thực hiện đánh giá mô hình trên tất cả giá trị của tham số và vẽ đồ thị đánh giá để so sánh độ chính xác của mô hình trên từng giá trị của tham số. Bước này giúp người thiết kế hệ thống rút ngắn thời gian chọn giá trị tham số tốt nhất cho mô hình. Thay vì phải thử nghiệm trên từng giá trị tham số rồi so sánh kết quả thì chỉ cần đánh giá một lần duy nhất sẽ nhận được kết quả so sánh trên tất cả giá trị tham số; cuối cùng, đọc kết quả từ đồ thị đánh giá để xác định giá trị tham số tốt nhất cho mô hình.

6 THỰC NGHIỆM

6.1 Dữ liệu thực nghiệm

Trong phần thực nghiệm, chúng tôi sử dụng các tập dữ liệu MovieLens (Maxwell and Joseph, 2015), MSWeb (Jack *et al.*, 1998) và Jester5k (Ken Goldberg *et al.*, 2001). Đầu tiên, MovieLens là tập dữ liệu được thu thập từ kết quả xếp hạng của 943 người dùng cho 1.664 bộ phim thông qua trang web MovieLens (movielens.umn.edu) trong thời gian 7 tháng (từ 19/9/1997 đến 22/4/1998). Tập dữ liệu này được tổ chức theo định dạng ma trận gồm 943 hàng, 1.664 cột và 1.569.152 ô chứa giá trị xếp hạng. Trong đó, có hơn 93 phần trăm giá trị xếp hạng có giá trị bằng 0 và hơn 6 phần trăm còn lại có giá trị xếp hạng có giá trị từ 1 đến 5. Tiếp theo, MSWeb là tập dữ liệu về người dùng Microsoft truy cập các trang web trong thời gian một tuần trong tháng 2 năm 1998 được lấy mẫu và xử lý từ file log của địa chỉ www.microsoft.com. Tập dữ liệu này bao gồm 38.000 người dùng không định danh truy cập trên 285 địa chỉ web gốc và được xử lý và tổ chức thành ma trận nhị phân với 32.710 hàng, 285 cột và 98.653 giá trị xếp hạng. Cuối cùng, Jester5k là tập dữ liệu được thu thập từ kết quả xếp hạng của 5.000 người dùng thông qua hệ tư vấn về truyện cười (Jester Online Joke Recommender System) trong khoảng thời gian từ tháng 4 năm 1999 đến tháng 5 năm 2003. Tập dữ liệu này được tổ chức theo định dạng ma trận gồm 5.000 hàng, 100 cột và 362.106 giá trị xếp hạng. Trong đó, mỗi người dùng xếp hạng ít nhất cho 36 truyện cười. Các giá trị xếp hạng của người dùng cho truyện cười nằm trong khoảng từ -10 đến 10.

6.2 Xử lý dữ liệu thực nghiệm

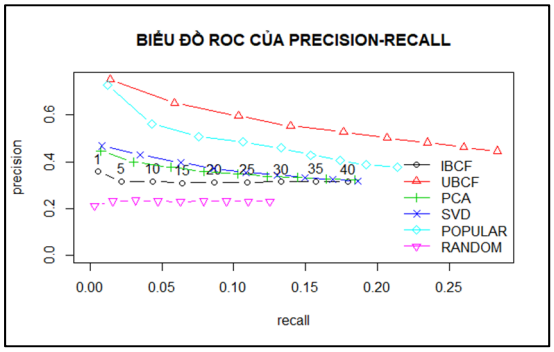
Trong thực nghiệm này, chúng tôi sử dụng kỹ thuật k-fold cross-validation (với k=5) để xử lý các tập dữ liệu thực nghiệm. Kỹ thuật này đảm bảo mỗi người dùng ít nhất một lần xuất hiện trong tập kiểm tra tương ứng với k lần đánh giá mô hình. Trong mỗi lần đánh giá, mô hình sử dụng một tập con làm tập kiểm tra và k-1 tập con còn lại dùng làm tập huấn luyện để các mô hình học. Kết quả đánh giá các mô hình là kết quả trung bình của k lần đánh giá.

6.3 Công cụ thực nghiệm

Để triển khai thực nghiệm, chúng tôi sử dụng công cụ ARQAT được triển khai trên ngôn ngữ R. Đây là gói công cụ được nhóm nghiên cứu phát triển từ nền tảng của công cụ ARQAT phát triển trên Java (Hiep *et al*., 2005). Trong đó, chúng tôi tích hợp các mô hình tư vấn lọc cộng tác từ gói công cụ recommenderlab (Michael Hahsler, 2011, 2015) và cài đặt thêm các chức năng: xử lý dữ liệu; tích hợp các mô hình lọc cộng tác cần đánh giá; tích hợp các giá trị tham số cần đánh giá vào mô hình lọc cộng tác; đánh giá các mô hình tích hợp và xây dựng đồ thị đánh giá.

6.4 Lựa chọn mô hình và tham số trên tập dữ liệu MovieLenese

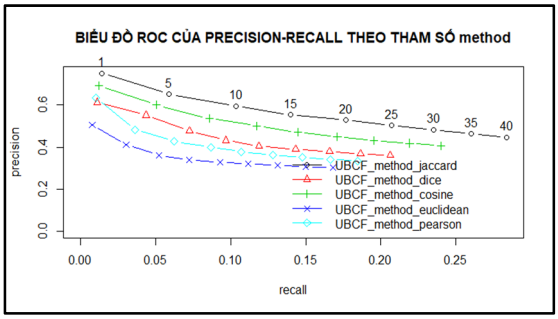
Để chọn được mô hình tư vấn lọc cộng tác nào cho kết quả tốt nhất trên tập dữ liệu MovieLens, chúng tôi xây dựng danh sách các mô hình cần đánh giá gồm: mô hình lọc cộng tác dựa trên sản phẩm (IBCF), mô hình lọc cộng tác dựa trên người dùng (UBCF), mô hình dựa phương pháp phân tích giúp giảm số chiều của dữ liệu (PCA), mô hình dựa trên phương pháp triển khai phân tích ma trận (SVD), mô hình dựa trên sản phẩm phổ biến (POPULAR), mô hình sinh kết quả tư vấn ngẫu nhiên (RANDOM). Sau đó, tiến hành đánh giá các mô hình (với mỗi người dùng được giới thiệu từ 1 đến 40) và xây dựng đồ thị đánh giá để so sánh độ chính xác của các mô hình cần khảo sát. Trong phần xây dựng đồ thị so sánh, chúng tôi vẽ biểu đồ ROC của cặp chỉ số Precision và Recall. Hình 1 cho thấy mô hình UBCF cho kết quả cao nhất trên tập MovieLens so với 5 mô hình còn lại.



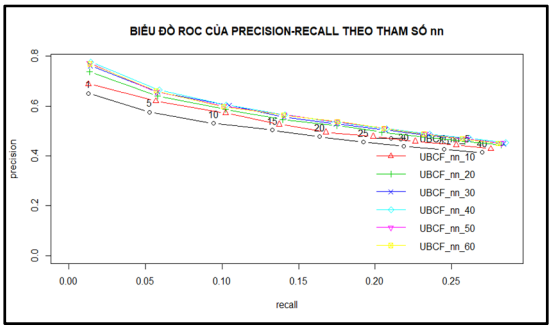
Hình 1: Biểu đồ so sánh độ chính xác các mô hình trên tập dữ liệu MovieLens

Dựa trên kết quả chọn mô hình, chúng tôi chọn mô hình UBCF để tiến hành thực nghiệm phần lựa chọn các giá trị tham số cho mô hình. Trong phần này, thực nghiệm được tiến hành trên hai tham số để lựa chọn độ đo tương đồng giữa các người dùng (method) và số lượng người dùng tương đồng cho mỗi người dùng cần tư vấn (nn).

Đối với tham số method, mô hình được đánh giá trên các độ đo tương đồng sau: vector_method = c("jaccard", "dice", "cosine", "euclidean", "pearson"). Từ kết quả so sánh trình bày trong Hình 2 cho thấy mô hình tư UBCF có độ chính xác cao nhất khi sử dụng độ đo tương đồng "jaccard" trên tập dữ liệu MovieLens.



Hình 2: Biểu đồ so sánh độ chính xác của mô hình UBCF theo các độ đo tương đồng

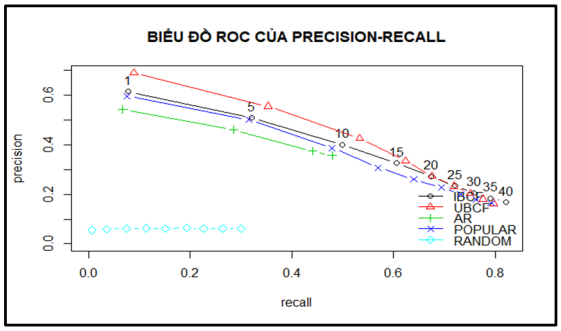


Hình 3: Biểu đồ so sánh độ chính xác của mô hình UBCF theo số người dùng tương đồng

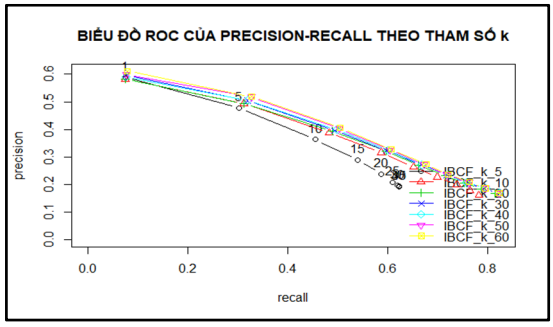
Đối với tham số nn, mô hình được đánh giá trên các giá trị sau: vector_nn = c(5, 10, 20, 30, 40, 50, 60). Hình 3 cho thấy mô hình UBCF có độ chính xác cao nhất khi tham số nn = 40 (số người dùng tương đồng cho mỗi người cần tư vấn) trên tập dữ liệu MovieLens.

6.5 Lựa chọn mô hình và tham số trên tập dữ liệu MSWeb

Tương tự phần thực nghiệm trên tập dữ liệu MovieLens, chúng tôi chọn các mô hình để đánh giá trên tập dữ liệu MSWeb gồm: IBCF, UBCF, AR, POPULAR, RANDOM. Sau đó, tiến hành đánh giá các mô hình và xây dựng đồ thị đánh giá tương tự như phần thực nghiệm trên tập dữ liệu MovieLens. Từ kết quả so sánh trong Hình 4 cho thấy mô hình UBCF có độ chính xác cao nhất trong các mô hình được đánh giá khi số lượng trang web giới thiệu cho người dùng nhỏ hơn hoặc bằng 20. Tuy nhiên, khi tăng số lượng trang web giới thiệu cho người dùng lớn hơn 20 thì mô hình IBCF có độ chính xác cao nhất trong các mô hình được đánh giá. Do đó, việc lựa chọn mô hình nào để ứng dụng cho hệ thống còn phụ thuộc vào số lượng trang web cần giới thiệu cho khách hàng.



Hình 4: Biểu đồ so sánh độ chính xác các mô hình trên tập dữ liệu MSWeb



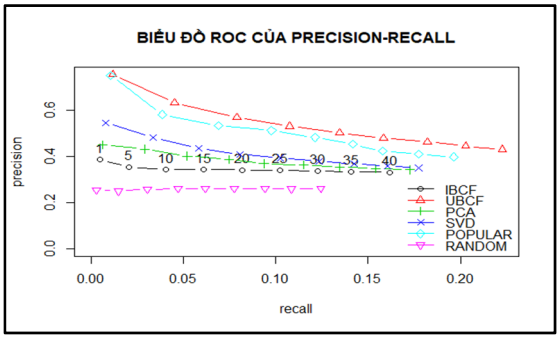
Hình 5: Biểu đồ so sánh độ chính xác của mô hình IBCF theo số sản phẩm tương đồng

Trong phần thực nghiệm chọn giá trị tham số trên tập dữ liệu MSWeb, chúng tôi chọn mô hình IBCF để tiến hành thực nghiệm phần lựa chọn giá trị cho tham số dùng để xác định số lượng sản phẩm tương đồng cho mỗi người dùng cần tư vấn

(k). Mô hình được đánh giá với tham số k gồm các giá trị sau: vector_k = c(5, 10, 20, 30, 40, 50, 60). Hình 5 cho thấy mô hình IBCF có độ chính xác cao nhất khi tham số k = 60 (số sản phẩm tương đồng cho mỗi người cần tư vấn) trên tập dữ liệu MSWeb.

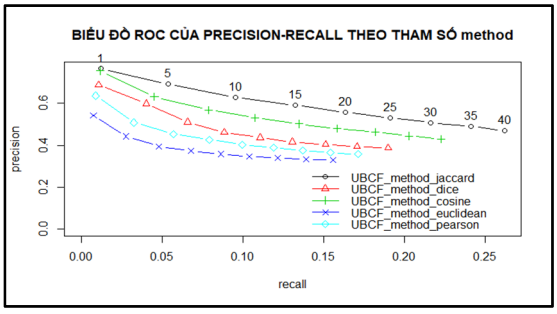
6.6 Lựa chọn mô hình và tham số trên tập dữ liệu Jester5k

Tương tự như hai tập dữ liệu trên, để chọn được mô hình nào cho kết quả tốt nhất trên tập dữ liệu Jester5k, chúng tôi xây dựng danh sách các mô hình cần đánh giá gồm: IBCF, UBCF, PCA, SVD, POPULAR, RANDOM. Kết quả đánh giá các mô hình được trình bày trong Hình 6. Kết quả này cho thấy mô hình UBCF cho kết quả cao nhất trên tập dữ liệu Jester5k so với 5 mô hình còn lại.



Hình 6: Biểu đồ so sánh độ chính xác các mô hình trên tập dữ liệu Jester5k

Từ kết quả chọn mô hình, chúng tôi chọn mô hình UBCF để tiến hành thực nghiệm phần lựa chọn các giá trị tham số cho mô hình. Trong phần này, thực nghiệm được tiến hành trên tham số dùng để xác định độ đo tương đồng giữa các người dùng gồm: vector_method = c("jaccard", "dice", "cosine", "euclidean", "pearson"). Từ kết quả so sánh trình bày trong Hình 7 cho thấy mô hình UBCF có độ chính xác cao nhất khi sử dụng độ đo tương đồng "jaccard" trên tập dữ liệu Jester5k.



Hình 7: Biểu đồ so sánh độ chính xác của mô hình UBCF theo các độ đo tương đồng

7 KẾT LUẬN

Trong nghiên cứu này, chúng tôi đề xuất các giải pháp lựa chọn mô hình và các giá trị tham số phù hợp cho bài toán tư vấn lọc cộng tác cụ thể. Qua kết quả thực nghiệm trên ba tập dữ liệu MovieLens, MSWeb và Jester5k, giải pháp mà chúng tôi đề xuất có thể hỗ trợ nhà thiết kế, nhà nghiên cứu xác định được chính xác mô hình cũng như các giá trị tham số của mô hình cho bài toán tư vấn cụ thể của họ một cách nhanh chóng.

TÀI LIỆU THAM KHẢO

Ahmed Mohammed K. Alsalama, 2015. A Hybrid Recommendation System Based On Association Rules. Engineering and Technology International Journal of Computer. Electrical, Automation, Control and Information Engineering. 9(1):55-62.

C. Aggarwal, 2016. Recommender Systems: The Textbook. Springer International Publishing Switzerland 2016. DOI 10.1007/978-3-319-29659-31.

Carlos A. Gomez-uribe and Neil Hunt, 2015. The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Transactions on Management Information Systems. 6(4) :1-19.

F. Maxwell Harper and Joseph A. Konstan, 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS). 5(4):1-19.

F.O. Isinkaye, Y.O. Folajimi, and B.A. Ojokoh, 2015. Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal. 16(3):261-273.

Feng Zhang, TiGong, Victor E. Lee and Gansen Zhao, Chunming Rong and Guangzhi Qu, 2016. Fast algorithms to evaluate collaborative filtering recommender systems. Knowledge-Based Systems. 96:96-103.

Greg Linden, Brent Smith, and Jeremy York, 2003. Amazon.com Recommendations Item-to-Item Collaborative Filtering. IEEE Internet Computing. 7(1):76-80.

Harald Steck, 2011. Item popularity and recommendation accuracy. In Proceedings of the fifth ACM conference on Recommender systems. 11:125-132.

Herlocker JL, Konstan JA, Terveen LG, Riedl JT, 2004. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems. 22(1):5-53.

Hiep Xuan Huynh, Fabrice Guillet, Henri Briand, 2005. ARQAT: An Exploratory Analysis Tool for Interestingness Measures. In International Symposium on Applied Stochastic Models and Data Analysis. 10:334-344.

J.B. Schafer, D. Frankowski, J. Herlocker, S. Sen, 2007. Collaborative filtering recommender systems. In: P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.) The Adaptive Web. Springer Berlin Heidelberg2007:291-324.

Jack S. Breese, David Heckerman and Carl M. Kadie, 1998. Anonymous web data from www.microsoft.com. Microsoft Research, Redmond WA, 98052-6399, USA. https://kdd.ics.uci.edu/databases/msweb/msweb.html.

JinHyun Jooa, SangWon Bangb, and GeunDuk Parka, 2016. Implementation of a Recommendation System Using Association Rules and Collaborative Filtering. Procedia Computer Science. 91:944-952.

Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins, 2001. Eigentaste: A Constant Time Collaborative Filtering Algorithm. Information Retrieval. 4(2):133-151.

Manolis G. Vozalis and Konstantinos G. Margaritis, 2008. A Recommender System using Principal Component Analysis. 11th Panhellenic Conference in Informatics. 11:271-283.

Martin P. Robillard, Walid Maalej, Robert J. Walker and Thomas Zimmermann, 2014. Recommendation Systems in Software Engineering. Springer Heidelberg New York Dordrecht London. ISBN 978-3-642-45135-5 (eBook).

Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan, 2010. Collaborative Filtering Recommender Systems. Foundations and Trends in Human–Computer Interaction. 4(2):81-173.

Michael Hahsler, 2011. recommenderlab: A Framework for Developing and Testing Recommendation Algorithms. The Intelligent Data Analysis Lab at SMU. http://lyle.smu.edu/IDA/recommenderlab/.

Michael Hahsler, 2015. Lab for Developing and Testing Recommender Algorithms. Copyright (C) Michael Hahsler (PCA and SVD implementation (C) Saurabh Bathnagar). http://R-Forge.R-project.org/projects/recommenderlab/.

Michael Howe, 2007. Pandora's Music Recommender. ttps://www.semanticscholar.org/.

Nguyễn Thái Nghe, 2016. Hệ thống gợi ý: Kỹ thuật và ứng dụng. https://www.researchgate .net/publication/310059523.

Nguyễn Thái Nghe, Huỳnh Xuân Hiệp, 2012. Ứng dụng kỹ thuật phân rã ma trận đa quan hệ trong xây dựng hệ trợ giảng thông minh. Kỷ yếu Hội thảo quốc gia lần thứ XV: Một số vấn đề chọn lọc của CNTT&TT (@2012), Nhà xuất bản Khoa học và Kỹ thuật. ISBN: 893-5048-931578. pp. 470-477.

Nguyễn Thái Nghe, Nguyễn Tấn Phong, 2014. Xây dựng hệ thống gợi ý bài hát dựa trên phản hồi tiềm ẩn. Tạp chí Khoa học Trường Đại học Cần Thơ. 34 (2014): 81-91.

Phan Quốc Nghĩa, Nguyễn Minh Kỳ, Đặng Hoài Phương, Huỳnh Xuân Hiệp, 2016. Hệ tư vấn lọc cộng tác theo người dùng dựa trên độ đo hàm ý thống kê. Kỷ yếu Hội nghị Quốc gia lần thứ IX về Nghiên cứu cơ bản và ứng dụng Công Nghệ thông tin. FAIR'9:231-239.

Phan Quốc Nghĩa, Nguyễn Minh Kỳ, Nguyễn Tấn Hoàng, Huỳnh Xuân Hiệp, 2015. Hệ tư vấn dựa trên tiếp cận hàm ý thống kê. Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin. FAIR'8:297-308.

Phan Quốc Nghĩa, Đặng Hoài Phương, Huỳnh Xuân Hiệp, 2017. Mô hình tư vấn lọc cộng tác tích hợp dựa trên tương đồng sản phẩm, Tạp chí Khoa học và Công nghệ - Đại học Đà Nẵng. 1(110):55-58.

Suresh K. Gorakala and Michele Usuelli, 2015. Building a Recommendation System with R. Published by Packt Publishing Ltd. ISBN 978-1-78355-449-2. www.packtpub.com.

Xun Zhou, Jing He, Guangyan Huang, and Yanchun Zhang, 2015. SVD-based incremental approaches for recommender systems. Journal of Computer and System Sciences. 81(4):717-733.