ĐÁNH GIÁ CÁC THUẬT TOÁN DỰA TRÊN NGƯỜI DÙNG SỬ DỤNG TRONG HỆ THỐNG KHUYẾN NGHỊ

Lê Văn Thịnh*
Trường Cao đẳng Công thương miền Trung

Tóm tắt

Hiện nay, nhu cầu mua hàng trực tuyến của người tiêu dùng ngày càng tăng mạnh, để đáp ứng sự hài lòng của người sử dụng, các nhà cung cấp dịch vụ đã đưa ra nhiều giải pháp để hỗ trợ người sử dụng tìm kiếm các mặt hàng tốt nhất mà họ đang cần mua. Trong bài báo này, chúng tôi nghiên cứu đánh giá một số thuật toán dựa trên người dùng trong lọc cộng tác để đưa ra khuyến nghị cho người sử dụng. Kết quả khuyến nghị này được dựa trên những hành vi của những người sử dụng trước đó.Thí nghiệm đã được thực hiện trên hai bộ dữ liệu MovieLens và EachMovie. Kết quả cho thấy thuật toán Euclidean cho ra kết quả tốt nhất. Thuật toán này có thể ứng dụng trong các hệ thống bán hàng trực tuyến để nâng cao hiệu quả tìm kiếm sản phẩm.

Từ khóa: *Thuật toán dựa trên người dùng, chất lượng dịch vụ, hệ thống khuyến nghị*

Abstract

Evaluation of user-based algorithms used in the recommendation system

Nowadays, the consumers' demand for online shopping is rapidly increasing. To satisfy the users' satisfaction, service providers have come up with many solutions to support the users in searching for the best items. In this paper, we examine a number of user-based algorithms in collaborative filtering for user recommendations, which is based on the previous users' behaviors. The experiment was performed on the two data sets called "MovieLens" and "EachMovie". The results showed that the Euclidean algorithm produces the best results. This algorithm might be used in online trading systems to improve the searching efficiency.

Keywords: *User-based algorithms, quality of service, recommendation system.*

1. Giới thiệu

Hiện nay nhu cầu mua hàng trực tuyến đang tăng, có rất nhiều trang web bán hàng trực tuyến tại Việt Nam. Do đó, người tiêu dùng có rất nhiều lựa chọn. Tuy nhiên, với số lượng nhà cung cấp dịch vụ ngày càng tăng và nhiều nhà cung cấp dịch vụ có những sản phẩm tương tự nên đã gây khó khăn cho người tiêu dùng trong việc lựa chọn sản phẩm tốt nhất.

Trong nghiên cứu này chúng tôi thực hiện thí nghiệm đánh giá một số thuật toán dựa trên người dùng (User-based) trong lọc cộng tác, tìm ra thuật toán tối ưu nhất để ứng dụng trong quá trình hỗ trợ người tiêu dùng lựa chọn sản phẩm tốt nhất trong bối cảnh có hàng nghìn nhà cung cấp dịch vụ khác nhau. Đóng góp này giúp cho các nhà cung cấp dịch vụ sử dụng thuật toán có độ tin cậy cao trong quá trình đưa ra khuyến cáo.

2. Các thuật toán dựa trên người dùng

Thuật toán dựa trên người dùng [5,1] hay còn gọi là phương pháp tiếp cận vùng lân

* Email: *thinhcdcn@gmail.com*

cận dựa trên người dùng, đây là phương pháp được sử dụng thông dụng nhất trong lọc cộng tác và được tuân thủ theo hai bước như sau:

(1) Tính toán tương tự giữa người dùng đang hoạt động và các người dùng lân cận.

(2) Chọn một tập hợp con của người sử dụng hàng xóm (vùng lân cận) tương tự với người dùng đang hoạt động, sau đó dự đoán bằng cách sử dụng dựa trên xếp hạng của người dùng hàng xóm.

Một số thuật toán được sử dụng đánh giá có liên quan đến người sử dụng.

2.1. Pearson Correlation degree

Pearson correlation là độ đo tính sự tương đồng giữa hai người dùng dựa trên tương quan thống kê [4]. Độ tương đồng của hai người dùng *u* và *v* được xác định bằng công thức:

$$S(u,v) = \frac{\sum_{i \in I}(r_{u,i} - \vec{r}_u)(r_{v,i} - \vec{r}_v)}{\sqrt{\sum_{i \in I}(r_{u,i} - \vec{r}_u)^2}\sqrt{\sum_{i \in I}(r_{v,i} - \vec{r}_v)^2}}$$

Với $S(u,v)$ là giá trị tương đồng giữa người dùng *u* và người dùng *v*; I là tập các sản phẩm hay các mục dữ liệu được xếp hạng bởi cả hai người dùng; $r_{u,i}$ là giá trị xếp hạng của người dùng *u* cho sản phẩm *i*; $r_{v,i}$ là giá trị xếp hạng của người *v* dùng cho sản phẩm i;

2.2. CosineSimilarity

Cosine Similarity là độ đo tính sự tương đồng giữa hai người dùng dựa trên không gian vector đại số tuyến tính [3]. Các giá trị xếp hạng của từng người dùng trên *m* sản phẩm hay mục dữ liệu được biểu diễn bằng một vector *m* chiều. Độ tương đồng của hai người dùng u và v được xác định bằng khoảng cách Cosine giữa hai vector $\vec{r}_u$ và vector $\vec{r}_v$ theo công thức sau:

$$S(u,v) = CoSine(\vec{r}_u, \vec{r}_v) = \frac{\sum_{i=1}^{m} r_{u,i} r_{v,i}}{\sqrt{\sum_{i=1}^{m} r^2{}_{u,i}}\sqrt{\sum_{i=1}^{m} r^2{}_{v,i}}}$$

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 *v* dùng cho sản phẩm *i*.

2.3. Euclidean Distance Similarity

Euclidean Distance Similarity là độ đo tính sự tương đồng giữa hai người dùng dựa trên khoảng cách giữa hai điểm trong không gian Euclide. Công thức của khoảng cách Euclide [4] như sau:

$$S(u,v) = \sqrt{\sum_{i=1}^{m}(r_{u,i} - r_{v,i})^2}$$

Với $S(u,v)$ là giá trị tương đồng giữa người dùng *u* và người dùng *v*; $r_{u,i}$ là giá trị xếp hạng của người dùng *u* cho sản phẩm *i*; $r_{v,i}$ là giá trị xếp hạng của người *v* dùng cho sản phẩm *i*;

2.4. Spearman rank correlation coefficient

Spearman rank correlation coefficient là độ đo tính sự tương đồng giữa hai người

dùng dựa hệ số tương quan người dùng được sử dụng công thức [6] như sau:

$$S(u,v) = \frac{\sum_i (r_{u,i} - \vec{r_u})(r_{v,i} - \vec{r_v})}{\sqrt{\sum_i (r_{u,i} - \vec{r_u})^2 (r_{v,i} - \vec{r_v})^2}}$$

Với $S(u,v)$ là giá trị tương đồng giữa người dùng *u* và người dùng *v*; $r_{u,i}$ là giá trị xếp hạng của người dùng *u* cho sản phẩm *i*; $r_{v,i}$ là giá trị xếp hạng của người *v* dùng cho sản phẩm *i*;

2.5. Log-likelihood similarity

Log-likelihood similarity [1] dùng để tính toán độ tương tự giữa các người sử dụng dựa trên thống kê sự suất hiện xung quanh đối với các người dùng.

$$d(u,v) = 2 * \left(\left(r_{u,i} * \ln \left({}^{r_{u,i}}/_{E1} \right) \right) + \left(r_{v,i} * \left(\ln \left({}^{r_{v,i}}/_{E2} \right) \right) \right) \right)$$

$$E1 = {}^{N1 * (r_{u,i} + r_{v,i})}/_{(N1 + N2)}$$

$$E2 = {}^{N2 * (r_{u,i} + r_{v,i})}/_{(N1 + N2)}$$

Trong đó *N1* và *N2* là đại diện nhóm người dùng tương ứng với người dùng *u* và *v*.

2.6. Mahattan distance

Mahattan distance là độ đo tính khoảng cách giữa hai người dùng trong không gian vector *n* chiều và được sử dụng công thức [5] như sau:

$$d(u,v) = \sum_{i=1}^{n} |r_{u,i} - r_{v,i}|$$

Với $d(u,v)$ là giá trị tương đồng khoảng cách giữa người dùng *u* và người dùng *v*; $r_{u,i}$ là giá trị xếp hạng của người dùng *u* cho sản phẩm *i*; $r_{v,i}$ là giá trị xếp hạng của người *v* dùng cho sản phẩm *i*;

2.7. TanimotoCoefficient

TanimotoCoefficient là độ đo tính sự trùng lặp của các sản phẩm ưu thích giữa hai người *u* và *v*. Công thức [6] tính như sau:

$$T(u,v) = \frac{\sum r_{u,i} r_{v,i}}{(\sum r_{u,i}^2 + \sum r_{v,i}^2 - \sum r_{u,i} r_{v,i})}$$

Với $T(u,v)$ là giá trị tương đồng khoảng cách giữa người dùng *u* và người dùng *v*; $r_{u,i}$ là giá trị xếp hạng của người dùng *u* cho sản phẩm *i*; $r_{v,i}$ là giá trị xếp hạng của người *v* dùng cho sản phẩm *i*;

3. Thí nghiệm

3.1. Phương pháp đánh giá

Việc đánh giá tính chính xác các dự đoán trong các hệ thống gợi ý thường được sử dụng phương pháp căn của sai số bình phương trung bình (RMSE - Root mean squared error) và sai số tuyệt đối trung bình (MAE - Mean Absolute Error) [2] như sau:

- **S**ai số tuyệt đối trung bình

$$MAE = \frac{\sum_{i=1}^{n}|p_{i,j} - r_{i,j}|}{n}$$

- Căn của sai số bình phương trung bình

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_{i,j} - r_{i,j})^2}{n}}$$

Trong đó: Tính chính xác của các dự đoán được đo trên *n* quan xác; $P_{i,j}$ là giá trị dự đoán của của người dùng *i* trên sản phẩm *j*; $R_{i,j}$ là giá trị xếp hạng thực tế ;

3.2. Dữ liệu thực nghiệm

Điểm chung các thuật toán dựa trên người sử dụng trong lọc cộng tác là dựa vào kết quả đánh giá xếp hạng của các người sử dụng trên các sản phẩm được thu thập trong quá khứ để tìm kiếm đưa ra mặt hàng tương tự tốt nhất cho người sử dụng hiện tại. Do vậy, trong nghiên cứu này, chúng tôi sử dụng hai bộ dữ liệu chuẩn sau để thử nghiệm đánh giá các thuật toán trên: MovieLens 100k là bộ dữ liệu chuẩn dùng để đánh giá giải thuật lọc cộng tác, dữ liệu này được tập hợp từ các đánh giá của người dùng tại website http://movielens.umn.edu. Bộ dữ liệu chứa các đánh giá (rating) của 943 người dùng cho 1682 bộ phim, mỗi người dùng đánh giá ít nhất 20 bộ phim và bộ dữ liệu EachMovie bao gồm 72916 người dùng được đánh giá tổng số 2811983 cho 1628 bộ phim khác nhau (https://grouplens.org/datasets/eachmovie).

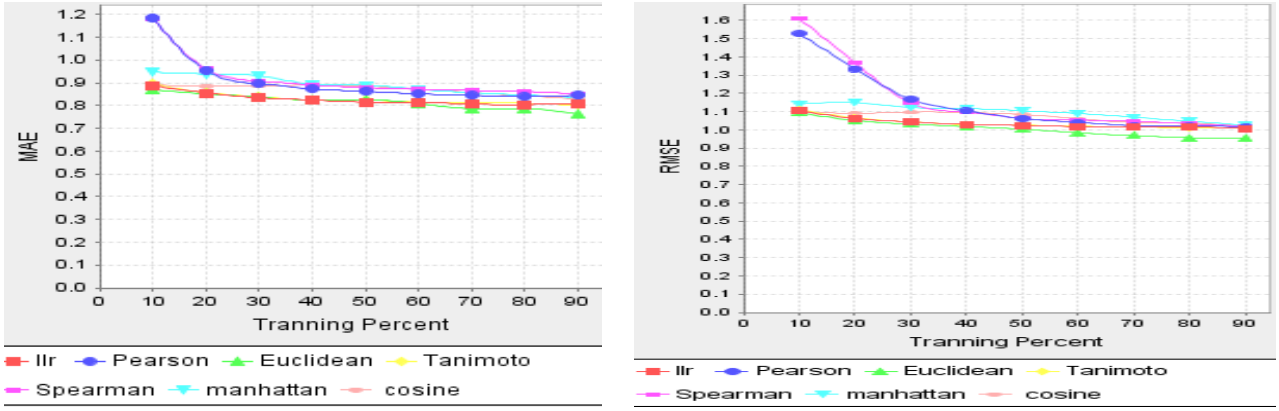
Để huấn luyện, mỗi tập dữ liệu chúng tôi chia ra làm 2 phần, phần thứ nhất chúng tôi chọn ngẫu nhiên 80% tập dữ liệu làm dữ liệu đầu vào và phần còn lại là 20% chúng tôi dùng để test. Để chứng minh độ tin cậy của thuật toán cho ra kết quả tốt nhất, chúng tôi tiến hành thí nghiệm trên bộ dữ liệu MovieLens bằng cách thay đổi tập dữ liệu huấn luyện là 70% và 30% còn lại dùng để test. Và chúng tôi ký hiệu các thuật toán như sau để hiển thị trên kết quả thí nghiệm:

Euclidean Distance Similarity**(***Euclidean*)
Pearson Correlation degree(*Pearson*)
CosineSimilarity(*Cosine*)
Spearman rank Correlation coefficient(*Spearman*)
Mahattan distance(*Manhattan*)
LogLikelihoodSimilarity(*llr*)
TanimotoCoefficient(*Tanimoto*)

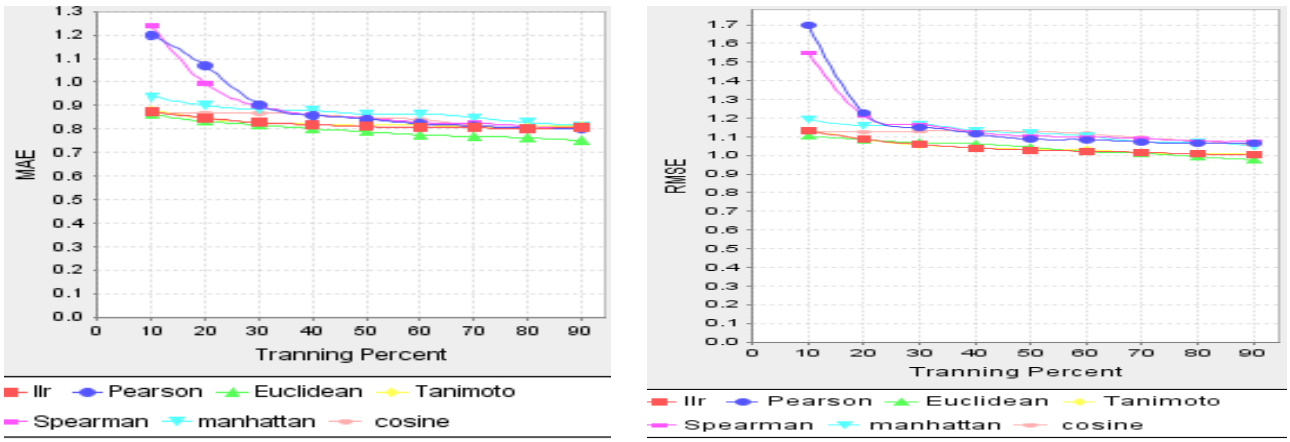
3.3. Kết quả thí nghiệm

Trong thí nghiệm chúng tôi chọn người sử dụng gần nhất với N =200. Nhìn vào kết quả MAE và RMSE (đối với bộ dữ liệu MovieLens) tại hình 1 cho thấy thuật toán Euclidean cho ra kết quả tốt nhất.

Tương tự đối với bộ dữ liệu EachMovie kết quả MAE và RMSE tại hình 2 thuật toán Euclidean cũng cho ra kết quả tốt nhất. Để quan sát kết quả rõ hơn chúng ta hãy nhìn vào kết quả giá trị chi tiết cho từng thuật toán tại bảng 1.



Hình 1. *Các thuật toán User – Based đối với bộ dữ liệu MovieLens*

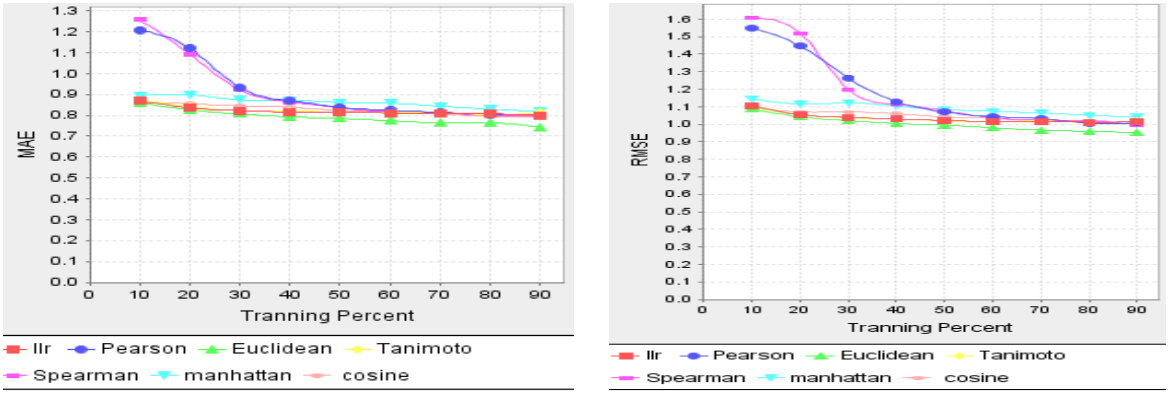


Hình 2. *Các thuật toán User – Based đối với bộ dữ liệu EachMovie*

Bảng 1: *Kết quả các giá trị tốt nhất cho từng thuật toán*

Các thuật toán	Bộ dữ liệu			
	MovieLens		EachMovie	
	MAE	**RMSE**	**MAE**	**RMSE**
LLr	0.8063	1.0069	0.8091	1.0021
Pearson	0.8455	1.0163	0.8004	1.0647
Tanimoto	0.8002	1.0148	0.8105	1.0057
Spearman	0.8504	1.0196	0.8036	1.0747
Manhattan	0.8355	1.0253	0.8145	1.0503
Cosine	0.8275	1.0164	0.8058	1.0561
Edulidean	**0.7639**	**0.9561**	**0.7504**	**0.9771**

Hình 3 là kết quả mô phỏng của tập dữ liệu MovieLens sau khi thay đổi tập dữ liệu huấn luyện và test. Kết quả cho thấy thuật toán Euclidean cũng cho ra kết quả tốt nhất với giá trị MAE là 0.7421và RMSE là 0.9519.



Hình 3. *Các thuật toán User – Based đối với bộ dữ liệu MovieLens*

4. Kết luận

Sự gia tăng các dịch vụ mua hàng trực tuyến trong thời gian gần đây đã giúp cho người sử dụng có nhiều thuận lợi về thời gian, cũng như có nhiều lựa chọn sản phẩm từ nhiều nhà cung cấp dịch vụ khác nhau, tuy nhiên có quá nhiều nhà cung cấp dịch vụ có chức năng tương tự nhau, điều này đã gây ra khó khăn cho người sử dụng trong việc lựa chọn. Do đó, bài báo đã thực hiện đánh giá mô phỏng các thuật toán dựa trên người dùng trong lọc cộng tác để chọn lựa thuật toán tối ưu nhất giúp cho các nhà cung cấp dịch vụ ứng dụng trong hệ thống khuyến nghị để hỗ trợ người sử dụng chọn lựa các sản phẩm tốt nhất mà người dùng đang cần. Kết quả thí nghiệm cho thấy thuật toán Euclidean đã cho ra kết quả tốt nhất. Do đó, việc ứng dụng thuật toán này trong các hệ thống khuyến nghị sẽ cho ra độ tin cậy cao hơn❒

TÀI LIỆU THAM KHẢO

[1] Feng Ge. 2011, A User-Based Collaborative Filtering Recommendation Algorithm Based on Folksonomy Smoothing, *International Conference, CSE 2011*.

[2] Herlocker JL, Konstan JA, Terveen LG and Riedl JT. 2004, Evaluating collaborative filtering recommender systems, *ACM Transactions on Information Systems*, 22(1), ISSN 1046-8188, page:.5–53.

[3] 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.

[4] Michael D. Ekstrand, John T. Riedl and Joseph A. Konstan, 2010, Collaborative Filtering Recommender Systems, *Foundations and Trends in Human–Computer Interaction*, Vol. 4, No. 2, page:.81–173.

[5] Zhi-Dan Zhao, Ming-sheng Shang, 2010, User-Based Collaborative-Filtering Recommendation Algorithms on Hadoop. *In proceeding of: Third International Conference on Knowledge Discovery and Data Mining, WKDD 2010, Phuket, Thailand*.

[6] Jonathan l. Herlocker, joseph a. Konstan, loren g. Terveen, and john t. Riedl. 2004, Evaluating Collaborative Filtering Recommender Systems, *ACM Transactions on Information Systems*, Vol. 22, No. 1, page: 5-53.

(*Ngày nhận bài: 20/12/2018; ngày phản biện: 28/12/2018; ngày nhận đăng: 03/06/2019)*