

ASSOCIATION RULES MINING USING APRIORI ALGORITHM, SUPPORT FOR SALES ACTIVITIES IN SUPERMARKET

Tran Thi Xuan^{1*}, Nguyen Van Nui²

¹TNU - University of Economics and Business Administration

²TNU - University of Information and Communication Technology

ARTICLE INFO	ABSTRACT
Received: 07/10/2021	Currently, data mining is gaining popularity in the retail sector and is an effective analytical method for detecting useful and unknown information in retail data. The organization of goods and related business activities towards enhancing the customer satisfaction is one of the very important jobs. This study will focus on analyzing, mining and finding association rules based on past data, thereby proposing some recommendations to support the business operation of the supermarket to be more optimized. For example, if a supermarket wants to arrange its stores in the most reasonable way, they can look at the purchase history and arrange the sets of products that are often bought together into one store. Or a news website that wants to introduce users to the most related articles, the same rule can be applied. In this paper, we calculate and analyze the relationship between products to help a supermarket arrange reasonable items for customers to buy goods by using association rule mining algorithm Apriori.
Revised: 15/11/2021	
Published: 15/11/2021	
KEYWORDS	
Data mining	
Association rule mining	
Association rule	
Apriori	
Sale activity	

KHAI PHÁ LUẬT KẾT HỢP SỬ DỤNG THUẬT TOÁN APRIORI, HỖ TRỢ CHO HOẠT ĐỘNG BÁN HÀNG TẠI SIÊU THỊ

Trần Thị Xuân^{1*}, Nguyễn Văn Núi²

¹Trường Đại học Kinh tế và Quản trị kinh doanh – ĐH Thái Nguyên

²Trường Đại học Công nghệ Thông tin và Truyền thông – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 07/10/2021	Hiện nay, khai phá dữ liệu trở nên phổ biến trong lĩnh vực bán lẻ và là phương pháp phân tích hiệu quả cho phát hiện thông tin hữu ích và chưa biết trong dữ liệu bán lẻ. Việc sắp xếp tổ chức hàng hoá và các hoạt động kinh doanh có liên quan nhằm nâng cao sự hài lòng của khách hàng là một trong những công việc rất quan trọng. Nghiên cứu này sẽ tập trung phân tích, khai phá và tìm ra luật kết hợp dựa trên dữ liệu của quá khứ, từ đó đề xuất một số kiến nghị để hỗ trợ cho hoạt động kinh doanh của siêu thị được tối ưu hơn. Ví dụ một siêu thị muốn sắp xếp các gian hàng một cách hợp lý nhất, họ có thể nhìn vào lịch sử mua hàng và sắp xếp các tập sản phẩm thường được mua cùng nhau vào một gian hàng. Hoặc một trang web tin tức muốn giới thiệu cho người dùng các bài viết liên quan đến nhau nhất, cũng có thể áp dụng quy luật tương tự. Trong bài báo này, chúng tôi tính toán phân tích tìm mối liên hệ giữa các sản phẩm giúp một siêu thị có thể sắp xếp mặt hàng hợp lý để khách hàng thuận tiện khi mua hàng bằng phương pháp khai phá luật kết hợp của thuật toán Apriori.
Ngày hoàn thiện: 15/11/2021	
Ngày đăng: 15/11/2021	
TỪ KHÓA	
Khai phá dữ liệu	
Khai phá luật kết hợp	
Luật kết hợp	
Apriori	
Hoạt động bán hàng	

DOI: <https://doi.org/10.34238/tnu-jst.5122>

* Corresponding author. Email: tranxuantbhd@tueba.edu.vn

1. Giới thiệu chung

Khai phá dữ liệu là một trong những lĩnh vực nghiên cứu quan trọng và ngày càng phát triển với mục đích trích xuất thông tin từ số lượng lớn các tập dữ liệu tích lũy.

Sự phát triển của công nghệ thông tin như hiện nay đang dần thể hiện rõ hơn vai trò định hướng cho ngành bán lẻ, kinh doanh sản phẩm của các doanh nghiệp. Xu thế thị trường cạnh tranh ngày càng gay gắt đòi hỏi các doanh nghiệp cần phải có những chiến lược, giải pháp của riêng mình để đáp ứng tốt hơn mong muốn của khách hàng. Các doanh nghiệp cần tìm hiểu thông tin có giá trị và chi tiết các hàng hóa để bán tốt hơn và nâng cao hiệu quả của hoạt động thị trường. Hiện nay, doanh nghiệp bán lẻ có thể thu thập các quy trình thông qua phân tích mẫu là tìm kiếm dữ liệu với các liên kết nhằm cung cấp dịch vụ tốt nhất cho người tiêu dùng. Dữ liệu lớn được mô hình hóa, chọn lọc và khai phá để thu thập thông tin có thể hiểu là hữu ích cho con người. Khai phá dữ liệu là một triển vọng và là lĩnh vực cập nhật một phần của khoa học máy tính. Sự tồn tại của dữ liệu lớn rất quan trọng để sử dụng đúng cách trong việc trích xuất kiến thức ẩn trong kho dữ liệu data mart, hoặc kho lưu trữ. Thuật toán Apriori là một trong những thuật toán học máy không giám sát đối với các quy tắc tìm ra luật kết hợp. Thuật toán apriori có thể được áp dụng cho tập hợp các giao dịch của các nhóm khách hàng tìm mối liên hệ giữa các sản phẩm.

Trong những năm gần đây, kỹ thuật khai phá dữ liệu và phân lớp đã được áp dụng thành công trong việc đề xuất mô hình hỗ trợ nhau để nâng cao chất lượng dịch vụ bán lẻ [1]-[7].

Tác giả Eni Heni Hermaliani [1] đã sử dụng thuật toán Apriori để hỗ trợ tìm ra quy luật mua bán sản phẩm trái cây. Tác giả J.Silva [2] bằng cách sử dụng thuật toán Apriori để khai phá quy tắc liên kết để phân khúc khách hàng trong khu vực doanh nghiệp vừa và nhỏ. Nhóm tác giả M. Kavitha và Subbaiah [3] đã sử dụng thuật toán Apriori để trích xuất sản phẩm trong cửa hàng tạp hóa.

Mục đích nghiên cứu nhằm xác định mức độ mà thuật toán Apriori có thể giúp sự phát triển chiến lược tiếp thị, có được mô hình liên kết và xác định các sản phẩm bán chạy nhất.

Do vai trò rất quan trọng trong việc phát triển chiến lược tiếp thị, chủ đề nghiên cứu để tìm hiểu sâu rộng về các mô hình để xác định quy luật, xác định được sản phẩm bán chạy... đã tăng nhanh trong những năm qua. Gần đây, có một vài mô hình phân lớp được nghiên cứu, đề xuất để hỗ trợ các nhà nghiên cứu trong việc xây dựng mô hình xác định quy luật, sản phẩm bán chạy [1]-[15]. Tuy nhiên, ở thời điểm hiện tại, vẫn còn thiếu các mô hình tính toán phù hợp và công cụ dự đoán với độ chính xác cao có thể hỗ trợ hiệu quả cho việc tìm kiếm luật chính xác. Bên cạnh đó, do sự tiến bộ của khoa học kỹ thuật và ảnh hưởng của cách mạng công nghiệp 4.0, dữ liệu khách hàng đã kiểm chứng thực nghiệm đang ngày càng được bổ sung nhiều hơn. Chính vì vậy, việc thiếu hụt mô hình dự đoán là một vấn đề cấp thiết cần được quan tâm giải quyết.

Tiếp tục phát triển các ý tưởng nghiên cứu trước đây, trong bài viết này nhóm tác giả tập trung vào vấn đề phân tích tìm quy luật liên kết giữa các mặt hàng trong siêu thị dựa trên dữ liệu quá khứ mua hàng của khách bằng thuật toán Apriori, sử dụng bộ công cụ Weka [16].

2. Xây dựng, huấn luyện mô hình

2.1. Thu thập, tiền xử lý dữ liệu

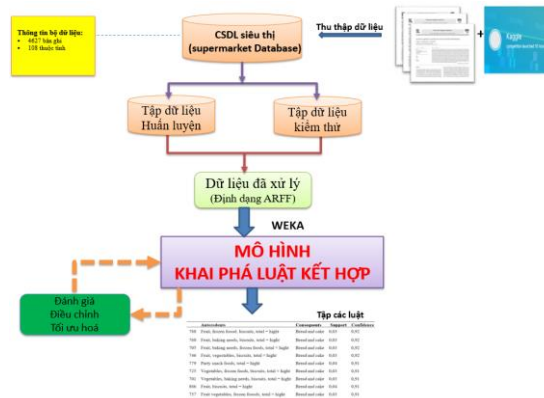
Bài báo sử dụng bộ dữ liệu Kaggle [17] đánh giá hiệu quả các kỹ thuật học máy. Kaggle có nhiều bộ dữ liệu khác nhau cho các lĩnh vực nhằm hỗ trợ cho nghiên cứu về học máy và khoa học dữ liệu. Kaggle đã được các nhà nghiên cứu trên thế giới sử dụng rộng rãi. Bộ dữ liệu này sau bước tiền xử lý, dữ liệu bao gồm 4627 thông tin về giao dịch mua hàng với 108 thuộc tính là các mặt hàng và tổng giá trị giao dịch.

2.2. Xây dựng và huấn luyện mô hình

Mô hình tổng thể của nghiên cứu này được thể hiện chi tiết ở Hình 1 bên dưới.

Trong nghiên cứu này, để tìm ra các luật kết hợp hỗ trợ hiệu quả cho hoạt động kinh doanh tại siêu thị, điều kiện trước tiên đó là thỏa mãn 2 giá trị cho trước là độ hỗ trợ cực tiểu (minimum

support) và độ tin cậy cực tiểu (minimum confidence) từ một cơ sở dữ liệu có sẵn, công việc thực hiện được chia làm hai bước:



Hình 1. Mô hình tổng thể của hệ thống

(1) Tìm tất cả các tập chỉ mục phổ biến: một tập chỉ mục là phổ biến được xác định qua việc tính độ hỗ trợ và thỏa mãn độ hỗ trợ cực tiểu.

(2) Sinh ra các luật kết hợp mạnh từ các tập chỉ mục phổ biến: Các luật phải thỏa mãn độ hỗ trợ cực tiểu và độ tin cậy cực tiểu.

Giả sử có tập chỉ mục phổ biến là L_k , $L_k = \{I_1, I_2, I_3, \dots, I_k\}$, các luật kết hợp của tập chỉ mục này được sinh như sau: khởi tạo luật đầu tiên $\{I_1, \{I_1, I_2, I_3, \dots, I_{k-1}\}, \rightarrow \{I_k\}$, sau đó tiến hành kiểm tra độ tin cậy (confidence) để xác định luật trên có thỏa mãn hay không.

Thực hiện cắt bỏ phần tử cuối cùng của vế trái, chuyển sang vế phải để tạo thành luật mới, rồi lại kiểm tra độ tin cậy. Quá trình trên được thực hiện cho tới khi vế trái trở thành tập rỗng. Do bước thứ 2 khá đơn giản nên hầu hết các nghiên cứu về khai phá luật kết hợp đều tập trung vào bước một.

Đối với bước thứ nhất trong khai phá luật kết hợp, ta lại có thể chia ra làm 2 bước con: sinh tập chỉ mục ứng viên (candidate frequent itemsets) và sinh tập chỉ mục phổ biến (frequent itemsets).

Trong đa số trường hợp, số lượng tập chỉ mục phổ biến sinh ra là rất lớn, kéo theo số lượng luật kết hợp tạo ra thường là hàng nghìn, thậm chí hàng triệu luật. Người dùng cuối gần như không thể hiểu hoặc đánh giá hết được một lượng lớn luật phức tạp như trên, do đó hạn chế phần nào giá trị của kết quả thu được. Hiện nay đã có rất nhiều thuật toán hiệu quả được đưa ra để giải quyết vấn đề này, bằng cách chỉ sinh luật phù hợp với nhu cầu của người dùng (interest rules), sinh luật “không dư thừa” (“non-redundant” rules), hoặc chỉ sinh luật thỏa mãn một tiêu chuẩn cụ thể nào đó như coverage, leverage, lift hoặc strength.

Cho tập hợp $I = \{I_1, I_2, I_3, \dots, I_n\}$ gồm n phần tử khác nhau, I được gọi là tập chỉ mục (itemset), T là một giao tác (transaction) chứa một tập các phần tử thuộc I ($T \subseteq I$), D là một cơ sở dữ liệu chứa m giao tác T khác nhau.

Một luật kết hợp là một phát biểu có dạng $X \rightarrow Y$, trong đó $X \subseteq I$, $Y \subseteq I$ và $X \cap Y = \emptyset$. Vế phải X được gọi là tiền đề, còn vế trái Y gọi là kết luận của luật. Có hai độ đo cơ bản cho luật kết hợp, đó là độ hỗ trợ (support) và độ tin cậy (confidence).

Độ hỗ trợ một tập chỉ mục X trong D , kí hiệu $\text{supp}(X)$, được tính bằng phần trăm số giao tác T trong D có chứa X (hay còn gọi là hỗ trợ X).

Giả sử độ hỗ trợ của một phần tử là 0,1%, điều đó có nghĩa là chỉ có 0,1% số giao tác có chứa phần tử đó.

Độ hỗ trợ của một luật kết hợp $r = X \rightarrow Y$, kí hiệu $\text{supp}(r)$, biểu thị tần số luật có trong các giao tác. Độ hỗ trợ thể hiện trong bao nhiêu phần trăm dữ liệu thì những điều ở vế trái và vế phải cùng xảy ra. Như vậy, độ hỗ trợ chính là xác suất $P(X \cup Y)$:

Độ tin cậy của một luật kết hợp $r = X \rightarrow Y$, kí hiệu $\text{conf}(r)$, là số phần trăm các giao tác trong D chứa cả X và Y trên số giao tác trong D chứa X. Độ tin cậy chính là xác suất có điều kiện $P(Y|X)$, nó thể hiện nếu về trái xảy ra thì có bao nhiêu khả năng về phải cũng xảy ra:

Độ tin cậy biểu thị độ mạnh của một luật kết hợp, giả sử độ tin cậy của luật r bằng 90%, có nghĩa là 90% số giao tác có chứa X thì cũng chứa Y.

Do cơ sở dữ liệu có kích thước lớn và người dùng thường chỉ quan tâm tới một tập các phần tử nhất định, do vậy người ta đưa ra các ngưỡng giá trị cho độ hỗ trợ và độ tin cậy nhằm loại bỏ các luật không phù hợp với yêu cầu của người dùng hoặc các luật vô dụng. Hai ngưỡng này được gọi là độ hỗ trợ cực tiểu (minimum support) và độ tin cậy cực tiểu (minimum confidence).

Tập chỉ mục X có $\text{supp}(X) \geq \text{minsupp}$, với minsupp là độ hỗ trợ cực tiểu, được gọi là tập chỉ mục phổ biến (frequent itemset hay large itemset). Một số tính chất điển hình của tập chỉ mục phổ biến:

Nếu $A \subseteq B$ với A, B là các tập chỉ mục thì $\text{supp}(A) \geq \text{supp}(B)$.

Một tập chứa một tập không phổ biến thì cũng là tập không phổ biến.

Các tập con của tập phổ biến cũng là tập phổ biến.

Các luật kết hợp thỏa mãn cả hai ngưỡng độ hỗ trợ cực tiểu (minsupp) và độ tin cậy cực tiểu (minconf) được gọi là luật kết hợp mạnh (strong), tức là $\text{supp}(X \rightarrow Y) \geq \text{minsupp}$ và $\text{conf}(X \rightarrow Y) \geq \text{minconf}$. Người ta thường viết giá trị các độ hỗ trợ và độ tin cậy này giữa 0% và 100% thay cho 0 tới 1.

Nếu độ hỗ trợ cực tiểu minsupp có giá trị cao thì ta sẽ thu được ít tập chỉ mục phổ biến, do vậy sẽ có ít luật hợp lệ phổ biến xuất hiện; ngược lại nếu đặt minsupp thấp thì sẽ xuất hiện nhiều luật hợp lệ hiếm.

Còn đối với độ tin cậy cực tiểu minconf , nếu giá trị minconf cao thì thu được ít luật, nhưng tất cả các luật này "gần như đúng". Còn nếu minconf có giá trị thấp thì ta thu được rất nhiều luật nhưng phần lớn "rất không chắc chắn".

Trong thực tế, người ta thường đặt giá trị minsupp trong khoảng 2 - 10% và minconf trong khoảng 70 - 90%.

Hiện nay, Apriori [4] là thuật toán khai phá luật kết hợp nổi tiếng, sử dụng chiến lược tìm kiếm theo chiều rộng (Breadth-first search) để tính độ hỗ trợ của các tập chỉ mục và tận dụng bổ đề downward closure [4] để tìm ra các tập ứng viên. Apriori rất hiệu quả trong quá trình sinh tập ứng viên do áp dụng sử dụng kĩ thuật cắt tia để tránh phải đánh giá một số tập chỉ mục nhất định mà vẫn bảo đảm tính toàn vẹn.

Apriori là một thuật toán do Rakesh Agrawal, Tomasz Imielinski, Arun Swami đề xuất lần đầu vào năm 1994. Thuật toán Apriori dùng cách tiếp cận lặp, với các tập chỉ mục k -itemsets được dùng để thăm dò các tập $(k+1)$ -itemsets. Đầu tiên, các tập chỉ mục phổ biến 1-itemsets được tìm thấy bằng cách quét cơ sở dữ liệu (CSDL) để đếm số lượng từng item và thu thập những item thỏa mãn độ hỗ trợ cực tiểu, tập kết quả đặt là L_1 . Tiếp theo, L_1 được dùng để tìm L_2 , là các tập chỉ mục phổ biến 2-itemsets, nó lại được dùng để tìm L_3 , và cứ tiếp tục cho tới khi tập chỉ mục phổ biến k -itemsets không thể tìm thấy. Việc tìm kiếm cho mỗi L_k đòi hỏi một lần quét toàn bộ cơ sở dữ liệu.

Đầu vào: CSDL, độ hỗ trợ cực tiểu minsup .

Đầu ra: Tập các chỉ mục phổ biến.

Thuật toán Apriori có độ phức tạp về thời gian là $O(k*(k^2+t*n))$ với k là kích thước tập chỉ mục phổ biến, t là kích thước cơ sở dữ liệu và n là số tập chỉ mục của t . Độ phức tạp về thời gian của thuật toán Apriori là $O(k^3+k*t*n)$.

3. Kết quả và một số thảo luận

Như đã trình bày trước đó, trong nghiên cứu này, chúng tôi tiến hành sử dụng thuật toán tìm ra luật kết hợp Apriori. Kết quả thuật toán được trình bày ở Bảng 1.

Với tất cả các quy tắc mà thuật toán tìm ra thì quy tắc nếu khách mua biscuits và frozen thì tỉ lệ mua Bread and cake là chiếm tới hơn 91% và tổng số tiền giao dịch đều là cao. Do đó, nên đặt các mặt hàng biscuits, frozen, bread và cake cạnh nhau trong cửa hàng.

Bảng 1. Kết quả thuật toán Apriori

	Antecedents	Consequents	Support	Confidence	lift
788	Fruit, frozen food, biscuits, total = high	Bread and cake	0,03	0,92	1,27
760	Fruit, baking needs, biscuits, total = high	Bread and cake	0,03	0,92	1,27
705	Fruit, baking needs, frozen foods, total = high	Bread and cake	0,03	0,92	1,27
746	Fruit, vegetables, biscuits, total = high	Bread and cake	0,03	0,92	1,27
779	Party snack foods, total = high	Bread and cake	0,04	0,91	1,27
725	Vegetables, frozen foods, biscuits, total = high	Bread and cake	0,03	0,91	1,26
701	Vegetables, baking needs, biscuits, total = high	Bread and cake	0,03	0,91	1,26
866	Fruit, biscuits, total = high	Bread and cake	0,04	0,91	1,26
757	Fruit vegetables, frozen foods, total = high	Bread and cake	0,03	0,91	1,26
877	Fruit, frozen foods, total = high	Bread and cake	0,04	0,91	1,26

4. Kết luận

Qua kết quả trên ta thấy, thuật toán Apriori hỗ trợ rất tốt trong việc tìm ra các quy luật liên kết giữa các sản phẩm trong một cửa hàng.

Theo kết quả phân tích cho thấy, nếu khách mua biscuits và frozen thì tỉ lệ khách quyết định mua bread và cake trên 91%. Do đó, cửa hàng nên đặt các mặt hàng này cạnh nhau để khách hàng thuận lợi trong việc mua hàng.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] E. H. Hermaliani *et al*, "Data Mining Technique to Determine the Pattern of Fruits Sales & Supplies Using Apriori Algorithm," *Journal of Physics: conference series*, vol. 1641, 2020, Art. no. 012070.
- [2] J. Silva *et al*, "Association Rules Extraction for Customer Segmentation in the SMEs Sector Using the Apriori Algorithm," *International Workshop on Web Search and Data Mining (WSDM)*, April 29 - May 02, 2019, Leuven, Belgium.
- [3] M. Kavitha and S. Subbaiah, "Association Rule Mining using Apriori Algorithm for Extracting Product Sales Patterns in Groceries," *Int. J. Eng. Res. Technol.*, vol. 08, no. 03, pp. 1-4, 2020.
- [4] I. R. V. Srinivasa Kumar, R. Renganathan, and C. Vijaya Banu, "Consumer Buying Pattern Analysis using Apriori Association Rule," *International Journal of Pure and Applied Mathematics*, vol. 119, no. 7, pp. 2341-2349, 2018.
- [5] N. Verma, D. Malhotra, and S. Jatinder, "Big data analytics for retail industry using MapReduce-Apriori framework," *J. Manag. Anal.*, vol. 7, pp. 424-442, 2020.
- [6] P. Yazgan, *Association Rules And Market Basket Analysis: A Case Study In Retail Sector*, Istanbul Commerce University, 2016.
- [7] Y. Kurnia, Y. Isharianto, Y. C. Giap, A. Hermawan, and Riki, "Study of application of data mining market basket analysis for knowing sales pattern (association of items) at the O! Fish restaurant using apriori algorithm," *1st International Conference on Advance and Scientific Innovation (ICASI) - IOP Conf. Series: Journal of Physics: Conf. Series 1175*, 2019, pp. 1-6.
- [8] R. Husna, R. Lestari, and Y. Hendra, "Inventory model of goods availability with apriori algorithm," *ICOMSET, IOP Conf. Series: Journal of Physics: Conf. Series 1317*, vol. 2018, pp. 1-8, 2018.
- [9] V. Singh and K. Kumar, "Data Mining and Knowledge Management," *Int. Res. J. Eng. Technol.*, vol. 4, no. 2, pp. 200-206, 2017.
- [10] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*, Elsevier, vol. 2, 2011.
- [11] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 9, no. 2, pp. 447-459, 2018.
- [12] F. M. Ba-Alwi and H. M. Hintaya, "Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach," *Int. J. Sci. Eng. Res.*, vol. 4, no. 8, p. 64, 2013.
- [13] U. Fayyad, P. G. Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *American Association for Artificial Intelligence Magazine*, vol. 17, pp. 36-54, 1996.
- [14] F. Ba-Alwi, "Discovery of novel association rules based on genetic algorithms," *Br. J. Math. Comput. Sci.*, vol. 4, no. 23, p. 17, 2014.

- [15] M. Z. Susac and Adela, "Discovering market basket patterns using hierarchical association rules," *Croatian Operational Research Review*, 2015, pp. 475-487.
- [16] A. K. Shrivastav and R. N. Panda, "Implementation of Apriori Algorithm using WEKA," *KIET International Journal of Intelligent Computing and Informatics*, vol. 1, no. 1, pp. 12-15, January 2021.
- [17] Kaggle Inc, "Kaggle: Your Machine Learning and Data Science Community", 2021. <https://www.kaggle.com/>. [Accessed June 20, 2021].