BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT

HỌC PHẦN THỰC HÀNH LẬP TRÌNH

ĐỀ TÀI : Tìm hiểu, cách Sử Dụng Thuật Toán Apriori và FP-Growth và Phân Tích Luật Kết Hợp Trên Bộ Dữ Liệu Online Retail Thuật Toán Apriori và FP-Growth

Sinh viên thực hiện:

NGUYỄN TIẾN ANH DHKL16A2HN 22174600083

Giáo viên giảng dạy: NGUYỄN ANH THƯ

Hà Nội, 6/2025

BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT

HỌC PHẦN THỰC HÀNH LẬP TRÌNH

ĐỀ TÀI : Tìm hiểu, cách Sử Dụng Thuật Toán Apriori và FP-Growth và Phân Tích Luật Kết Hợp Trên Bộ Dữ Liệu Online Retail Thuật Toán Apriori và FP-Growth

Sinh viên thực hiện:

NGUYỄN TIẾN ANH DHKL16A2HN 22174600083

Giảng viên giảng dạy: NGUYỄN ANH THƯ

Hà Nội, 6/2025

MỞ ĐẦU

Với sự phát triển của công nghệ thông tin thì khối lượng dữ liệu lưu trữ ngày càng lớn, và giữa những lượng dữ liệu khổng lồ đó lại ẩn chứa một số thông tin được coi là chìa khóa dẫn đến thành công của mọi lĩnh vực từ hoạt động sản xuất đến kinh doanh. Việc khai thác, chiếc lọc thông tin ứng dụng vào cuộc sống của con người không chỉ dừng lại là một kĩ thuật đơn thuần, nó đòi hỏi sự ra đời của ngành khoa học mới: khoa học về phát hiện tri thức và khai phá dữ liệu (Knowledge Discovery and Data Mining - KDD).

Khai phá dữ liệu là ngành khoa học đang ngày được quan tâm nghiên cứu và phát triển do những ứng dụng thiết thực mà nó mang lại. Khai phá dữ liệu là phần cốt lõi của phát hiện tri thức, trong khai phá dữ liệu phát hiện các luật là một trong những nội dung cơ bản và phổ biến nhất. Các phương pháp phát hiện luật nhằm tìm ra sự phụ thuộc giữa các tính chất của các đối tượng hay các thuộc tính trong cơ sở dữ liệu.

Trên cơ sở đó báo cáo tập trung tìm hiểu một trong hướng tiếp cận khai phá dữ liệu thông qua thuật toán Apriori và thuât toán fp-growth.

Bài báo cáo bao gồm các phần được phân chương như sau:
Chương 1: Đặt vấn đề (giới thiệu)
Chương 2: Cơ sở lý thuyết
Chương 3: Thực nghiệm
Chương 4: Kết luận, hướng phát triển

Mục lục

CHƯƠNG 1. ĐẶT VẤN ĐỀ

1.1. BỐI CẢNH VÀ TÍNH CẤP THIẾT CỦA ĐỀ TÀI

1.1.1. Bối cảnh của đề tài

Trong bối cảnh thương mại điện tử phát triển mạnh mẽ, khối lượng dữ liệu giao dịch trực tuyến ngày càng lớn, việc phân tích dữ liệu để khám phá các mẫu hành vi mua sắm của khách hàng trở nên quan trọng. Bộ dữ liệu **Online Retail**, thường được sử dụng trong các nghiên cứu khai phá dữ liệu, chứa thông tin chi tiết về các giao dịch bán lẻ trực tuyến, bao gồm sản phẩm, số lượng, thời gian và thông tin khách hàng. Việc áp dụng các thuật toán khai phá dữ liệu như **Apriori** và **FP-Growth** để phân tích **luật kết hợp** (association rules) trên bộ dữ liệu này giúp doanh nghiệp hiểu rõ hơn về mối quan hệ giữa các sản phẩm, từ đó tối ưu hóa chiến lược kinh doanh.

- Thuật toán Apriori: Là một trong những thuật toán phổ biến nhất để tìm kiếm các tập hợp thường xuyên (frequent itemsets) và tạo ra luật kết hợp. Apriori sử dụng nguyên lý "chống đơn điệu" (anti-monotonicity), tức là nếu một tập hợp không thỏa mãn ngưỡng hỗ trợ (support), thì các tập con của nó cũng không thỏa mãn. Tuy nhiên, thuật toán này có hạn chế về hiệu suất khi xử lý dữ liệu lớn do phải quét cơ sở dữ liệu nhiều lần.

- Thuật toán FP-Growth: Là một cải tiến so với Apriori, sử dụng cấu trúc cây FP (Frequent Pattern Tree) để nén dữ liệu và giảm số lần quét cơ sở dữ liệu. FP-Growth hiệu quả hơn trong việc xử lý các tập dữ liệu lớn và phức tạp, đặc biệt phù hợp với dữ liệu giao dịch như Online Retail.

- Phân tích luật kết hợp: Luật kết hợp (ví dụ: {Sản phẩm A} → {Sản phẩm B}) giúp xác định các mẫu mua sắm, từ đó hỗ trợ các chiến lược như gợi ý sản phẩm, tối ưu hóa kho hàng, hoặc thiết kế chương trình khuyến mãi.

1.1.2. Tính cấp thiết của đề tài

- Việc nghiên cứu và ứng dụng thuật toán Apriori và FP-Growth trên bộ dữ liệu Online Retail mang tính cấp thiết vì các lý do sau:

- Tối ưu hóa kinh doanh: Trong thương mại điện tử, việc hiểu rõ các mẫu mua sắm giúp doanh nghiệp đưa ra các chiến lược tiếp thị và bán hàng hiệu quả hơn, ví dụ như gợi ý sản phẩm liên quan (cross-selling) hoặc tổ hợp sản phẩm (bundling).

- Nhu cầu xử lý dữ liệu lớn: Bộ dữ liệu Online Retail thường chứa hàng trăm nghìn giao dịch, đòi hỏi các thuật toán hiệu quả như FP-Growth để xử lý nhanh chóng và chính xác. So sánh với Apriori, FP-Growth có thể mang lại lợi thế về hiệu suất, đặc biệt khi dữ liệu ngày càng mở rộng.

- Ứng dụng thực tiễn: Kết quả từ luật kết hợp có thể được áp dụng trực tiếp vào các hệ thống gợi ý (recommendation systems), quản lý hàng tồn kho, hoặc phân tích hành vi khách hàng, từ đó tăng doanh thu và cải thiện trải nghiệm khách hàng.

- Nghiên cứu và phát triển: Việc so sánh hai thuật toán Apriori và FP-Growth trên cùng một bộ dữ liệu giúp đánh giá ưu, nhược điểm của từng phương pháp, từ đó đưa ra các cải tiến hoặc lựa chọn phù hợp cho các bài toán thực tế.

1.2. LÝ DO CHỌN ĐỀ TÀI

- Tính thực tiễn trong thương mại điện tử
Đề tài nghiên cứu thuật toán Apriori và FP-Growth trên bộ dữ liệu Online Retail giúp khám phá các mẫu mua sắm, hỗ trợ doanh nghiệp tối ưu hóa gợi ý sản phẩm, quản lý kho, và khuyến mãi, từ đó tăng doanh thu và cải thiện trải nghiệm khách hàng.

- Xử lý dữ liệu lớn
Bộ dữ liệu Online Retail có khối lượng lớn, đòi hỏi thuật toán hiệu quả. So sánh Apriori và FP-Growth giúp đánh giá hiệu suất, lựa chọn phương pháp phù hợp cho các bài toán dữ liệu lớn, đặc biệt với FP-Growth có lợi thế về tốc độ.

- Ý nghĩa khoa học và giáo dục
Đề tài củng cố kiến thức khai phá dữ liệu, cung cấp cơ hội áp dụng lý thuyết vào thực tiễn, nâng cao kỹ năng phân tích cho sinh viên và nhà nghiên cứu.

- So sánh và cải tiến thuật toán
So sánh Apriori và FP-Growth giúp đánh giá ưu, nhược điểm, đề xuất cải tiến, và chọn giải pháp tối ưu cho từng tình huống thực tế.

- Đáp ứng xu hướng công nghệ
Đề tài phù hợp với xu hướng khai phá dữ liệu và trí tuệ nhân tạo, hỗ trợ phát triển hệ thống gợi ý và phân tích hành vi khách hàng trong thương mại điện tử.

1.3. MỤC TIÊU NGHIÊN CỨU

Đề tài hướng đến việc nghiên cứu và ứng dụng thuật toán Apriori và FP-Growth để phân tích luật kết hợp trên bộ dữ liệu Online Retail, nhằm khám phá các mẫu mua sắm và đề xuất giải pháp tối ưu cho các bài toán khai phá dữ liệu trong thương mại điện tử. Cụ thể, nghiên cứu tập trung vào việc tìm hiểu cách hoạt động của hai thuật toán, triển khai chúng để tạo ra các tập hợp thường xuyên và luật kết hợp, từ đó xác định các mối quan hệ giữa các sản phẩm, hỗ trợ chiến lược gợi ý sản phẩm, quản lý kho, và thiết kế khuyến mãi. Đồng thời, đề tài so sánh hiệu suất của Apriori và FP-Growth về thời gian xử lý và khả năng xử lý dữ liệu lớn, nhằm đánh giá ưu, nhược điểm và lựa chọn phương pháp phù hợp. Kết quả nghiên cứu không chỉ mang lại các ứng dụng thực tiễn cho doanh nghiệp mà còn cung cấp cơ sở khoa học, góp phần nâng cao kiến thức về khai phá dữ liệu và mở ra hướng cải tiến thuật toán cho các bài toán tương tự.

CHƯƠNG II. CƠ SỞ LÝ THUYẾT

2.1 GIỚI THIỆU BỘ DỮ LIỆU.

Bộ dữ liệu Online Retail là một tập dữ liệu phổ biến trong nghiên cứu khai phá dữ liệu và học máy, được lấy từ kho lưu trữ UCI Machine Learning Repository. Đây là tập hợp các giao dịch bán lẻ trực tuyến của một công ty bán lẻ có trụ sở tại Vương quốc Anh, chuyên cung cấp các sản phẩm quà tặng. Bộ dữ liệu này được sử dụng rộng rãi để phân tích hành vi mua sắm, khai phá luật kết hợp, và xây dựng các hệ thống gợi ý sản phẩm trong thương mại điện tử.

2.1.2 Cấu trúc bộ dữ liệu

Bộ dữ liệu Online Retail bao gồm các thông tin chi tiết về các giao dịch bán lẻ, được lưu trữ dưới dạng bảng với các cột chính sau:

- InvoiceNo: Mã số hóa đơn, đại diện cho một giao dịch duy nhất. Các hóa đơn bắt đầu bằng chữ "C" biểu thị giao dịch bị hủy.
- StockCode: Mã định danh duy nhất cho từng sản phẩm.
- Description: Mô tả ngắn gọn về sản phẩm (ví dụ: "White Hanging Heart T-Light Holder").
- Quantity: Số lượng sản phẩm được mua trong mỗi giao dịch (có thể âm trong trường hợp trả hàng).
- InvoiceDate: Ngày và giờ thực hiện giao dịch.
- UnitPrice: Giá đơn vị của sản phẩm (tính bằng bảng Anh).
- CustomerID: Mã định danh duy nhất cho khách hàng (có thể thiếu trong một số giao dịch).
- Country: Quốc gia nơi khách hàng thực hiện giao dịch (chủ yếu là Vương quốc Anh, nhưng cũng có các quốc gia khác).

2.1.3. Đặc điểm chính

- Quy mô: Bộ dữ liệu chứa khoảng 541,909 bản ghi (giao dịch), tương ứng với hơn 25,900 hóa đơn và khoảng 4,070 sản phẩm khác nhau, được ghi nhận từ ngày 01/12/2010 đến 09/12/2011.

- Tính thực tiễn: Dữ liệu phản ánh thực tế các giao dịch thương mại điện tử, với mỗi hóa đơn là một "giỏ hàng" chứa nhiều sản phẩm, phù hợp để phân tích luật kết hợp và khám phá các mẫu mua sắm.

2.2. KHAI PHÁ LUẬT KẾT HỢP (Association Rule Mining)

Khai phá luật kết hợp là một kỹ thuật trong khai phá dữ liệu nhằm tìm ra các mối quan hệ ẩn giữa các mục (items) trong tập dữ liệu lớn, thường được biểu diễn dưới dạng luật "Nếu... thì..." (ví dụ: {Sản phẩm A} → {Sản phẩm B}). Các chỉ số chính để đánh giá luật kết hợp bao gồm:

- Support (độ hỗ trợ): Tỷ lệ giao dịch chứa tập hợp các mục so với tổng số giao dịch. Công thức:

$$Support(A \rightarrow B) = \frac{tổng\ số\ giao\ dịch\ A \cup B}{tổng\ số\ giao\ dịch}$$

Confidence (độ tin cậy): Tỷ lệ giao dịch chứa cả A và B so với các giao dịch chứa A.

$$Confidence(A \rightarrow B) = \frac{support\ A \cup B}{support\ A}$$

Lift (độ nâng): Đo lường mức độ phụ thuộc giữa A và B so với khi chúng độc lập

$$Lift(A \rightarrow B) = \frac{confidence\ A \cup B}{support\ B}$$

Luật kết hợp được sử dụng rộng rãi trong phân tích giỏ hàng (market basket analysis) để khám phá các mẫu mua sắm trong thương mại điện tử.

2.3. THUẬT TOÁN APRIORI.

Thuật toán Apriori, được đề xuất bởi Agrawal và Srikant (1994), là phương pháp phổ biến để tìm các tập hợp thường xuyên (frequent itemsets) và tạo luật kết hợp. Apriori dựa trên nguyên lý chống đơn điệu (anti-monotonicity): nếu một tập hợp không thỏa mãn ngưỡng hỗ trợ tối thiểu, thì các tập hợp con của nó cũng không thỏa mãn.

2.3.1 Các bước triển khai thuật toán.

Bước 1: Tìm tất cả các tập mục phổ biến 1- phần tử (C1).

Bước 2: Tạo các tập ứng viên có k – phần tử (k - candidate itemset) từ các tập phổ biến có (k-1) – phần tử. Ví dụ, tạo tập ứng viên C2 từ tập phổ biến C1.

Bước 3: Kiểm tra độ phổ biến của các ứng viên trên CSDL và loại các ứng viên không phổ biến ta được các tập mục phổ biến Li , với mọi $1 \leq i \leq k$.

Kết luận : Dừng khi không tạo được tập mục phổ biến hay tập ứng viên, i.e., Lk = {} hay Ck = {}.

Ví dụ:

Giả sử chúng ta có một cơ sở dữ liệu gồm 4 giao dịch như sau:

- Giao dịch T1: A, B

- Giao dịch T2: B, C

- Giao dịch T3: A, C

- Giao dịch T4: A, B, C

Chúng ta đặt ngưỡng hỗ trợ tối thiểu (minsup) là 2, nghĩa là một tập mục phải xuất hiện trong ít nhất 2 giao dịch mới được coi là phổ biến.

Bước 1: Tìm tập mục phổ biến 1-phần tử (L1)
Đếm số lần xuất hiện của từng mục: A xuất hiện 3 lần (trong T1, T3, T4), B cũng xuất hiện 3 lần (T1, T2, T4), và C xuất hiện 3 lần (T2, T3, T4). Vì cả ba mục đều có tần suất ≥ 2, nên ta có tập phổ biến L1 = {A}, {B}, {C}.

Bước 2: Tạo tập ứng viên 2-phần tử (C2) và lọc ra tập phổ biến (L2)
Từ các mục trong L1, ta tạo các cặp: (A, B), (A, C) và (B, C). Đếm số lần xuất hiện:

- (A, B) xuất hiện trong T1 và T4 → 2 lần
- (A, C) xuất hiện trong T3 và T4 → 2 lần
- (B, C) xuất hiện trong T2 và T4 → 2 lần

Tất cả đều thỏa mãn minsup, nên tập phổ biến 2-phần tử là L2 = {(A,B), (A,C), (B,C)}.

Bước 3: Tạo tập ứng viên 3-phần tử (C3)
Từ L2, ta có thể tạo ra một tập 3-phần tử là (A, B, C). Tuy nhiên, tổ hợp này chỉ xuất hiện trong đúng 1 giao dịch (T4), không đạt minsup. Vì vậy, ta không có tập phổ biến 3-phần tử, tức là L3 = ∅.

Kết luận:
Thuật toán dừng lại khi không còn tạo được tập phổ biến mới. Các tập mục phổ biến cuối cùng là:

- 1-phần tử: {A}, {B}, {C}
- 2-phần tử: {A,B}, {A,C}, {B,C}

2.3.2 Mã giả của thuật toán apriori.

Dữ liệu vào: Tập các giao dịch *D*, ngưỡng hỗ trợ minsup

Dữ liệu ra: Tập trả lời bao gồm các tập mục phổ biến trên *D*

Giải thuật:

L_1 = {large 1-itemset};
for (k = 2; $L_{k-1} \neq \phi$; k++) **do begin**
C_k = apriori_gen(L_{k-1}); // sinh tập mục ứng viên mới C_k;
For all giao dịch $t \in D$ **do begin**
C_t = subset(C_k, t); // các tập mục ứng viên chứa trong t;
For all tập mục ứng viên $c_i \in C_t$ **do** c_i.count ++ ;

end;

$L_k = \{c_i \in C_k \mid c_i.\text{count} \geq minsup\}$

end;

Return tất cả các tập mục phổ biến L_k;

2.3.3 Ưu và nhược điểm của thuận toán apriori :

Ưu điểm :

- Là thuật toán đơn giản, dễ hiểu và dễ cài đặt.

- Thuật toán Apriori tìm tập mục phổ biến thực hiện tốt bởi rút gọn kích thước các tập ứng viên nhờ kỹ thuật "tỉa".

Nhược điểm :

- Phải duyệt CSDL nhiều lần.

- Số lượng lớn tập ứng viên được tạo ra làm gia tăng sự phức tạp không gian.

- Để xác định độ support của các tập ứng viên, thuật toán luôn phải quyét lại toàn bộ CSDL.

2.3.4 Ứng dụng của thuật toán apriori.

1.Phân tích giỏ hàng (Market Basket Analysis):

- Ứng dụng phổ biến nhất của Apriori là trong phân tích dữ liệu mua sắm để tìm ra các sản phẩm thường được mua cùng nhau. Ví dụ: Trong siêu thị, Apriori có thể phát hiện rằng khách hàng mua bánh mì thường mua bơ, từ đó hỗ trợ tối ưu hóa cách sắp xếp sản phẩm hoặc đề xuất khuyến mãi.

2.Hệ thống gợi ý (Recommendation Systems):

- Các nền tảng thương mại điện tử như Amazon sử dụng thuật toán Apriori để gợi ý sản phẩm dựa trên lịch sử mua sắm hoặc xem xét của khách hàng. Ví dụ: Nếu khách hàng mua điện thoại, hệ thống có thể gợi ý mua thêm ốp lưng hoặc tai nghe.

3.Phân tích dữ liệu y tế:

- Apriori được sử dụng để tìm mối quan hệ giữa các triệu chứng, bệnh lý hoặc thuốc trong hồ sơ y tế. Ví dụ: Phát hiện rằng bệnh nhân mắc bệnh A thường có triệu chứng B và C.

4.Quản lý kho hàng:

- Trong quản lý kho, Apriori giúp xác định các mặt hàng thường được yêu cầu cùng nhau, từ đó tối ưu hóa quy trình lưu trữ và vận chuyển.

2.4 THUẬN TOÁN FP-GROWTH.

Thuật toán FP-Growth (Frequent Pattern Growth), được đề xuất bởi Han et al. (2000), là cải tiến của Apriori, sử dụng cấu trúc cây FP (Frequent Pattern Tree) để nén dữ liệu và giảm số lần quét cơ sở dữ liệu.

2.4.1 Các bước triển khai thuật toán.

Bước 1: Xây dựng FP-Tree:

- Duyệt CSDL lần một, xác định các tập mục phổ biến L và sắp xếp chúng theo độ hỗ trợ.

- Duyệt qua CSDL lần hai, với mỗi giao dịch T sắp xếp các tập mục theo thứ tự tập L. Giả sử các tập mục phổ biến trong T có dạng [p|P] với p là tập mục cần đưa vào FP-Tree và P là danh sách các tập mục còn lại, N là nút cần chèn. Nếu

nút con của N giống p, tăng biến count nút con đó lên 1. Ngược lại, tạo nút con mới cho N có tên mục là p và count = 1. Tiếp tục chèn P vào nút con vừa xét.

Bước 2: Xây dựng cơ sở mẫu điều kiện (Conditional Patern Bases) cho mỗi tập mục phổ biến.

Bước 3: Xây dựng FP-Tree điều kiện (Conditional FP-Tree) cho mỗi tập mục phổ biến trong cơ sở mẫu điều kiện.

Bước 4: Đệ quy xây dựng FP-Tree điều kiện đến khi FP-Tree điều kiện còn một nhánh duy nhất sau đó tiến hành sinh tất cả các tổ hợp mục phổ biến.

Ví dụ :

Bước 1: Xây dựng FP-Tree
Giả sử có 5 giao dịch:

- T1: A, B, C
- T2: B, C, D
- T3: A, C, D, E
- T4: A, D, E
- T5: A, B, C

Bước1a:Đếm tần suấ tcác mục:
A(4), B(3), C(4), D(3), E(2) → bỏ E vì minsup = 3
Bước 1b: Sắp xếp mỗi giao dịch theo thứ tự hỗ trợ giảm dần:
L = [A, C, B, D]
Ví dụ:
T1 → [A, C, B]
T2 → [C, B, D]
...
Tạo FP-Tree bằng cách chèn từng giao dịch theo thứ tự đó vào cây, gộp các nhánh trùng nhau và tăng count.
Bước 2: Tạo cơ sở mẫu điều kiện
Với mỗi mục (ví dụ D), tìm các đường dẫn từ gốc đến D → đó là **cơ sở mẫu điều kiện** cho D.
Bước 3: Xây dựng FP-Tree điều kiện
Dùng cơ sở mẫu điều kiện để xây cây FP riêng cho mỗi mục, gọi là **FP-Tree điều kiện**.
Bước 4: Đệ quy & sinh tập mục phổ biến
Lặp lại bước 2–3 trên từng cây điều kiện đến khi cây chỉ còn 1 nhánh → sinh tất cả tổ hợp mục phổ biến từ các nhánh này.
2.4.2 Mã giả cho thuận toán FP-GROWTH.

Dữ liệu vào: Tập các giao dịch D, ngưỡng hỗ trợ minsup
Dữ liệu ra: Tập các tập mục phổ biến trong D

Bước 1: Xây dựng FP-Tree

Scan D để đếm hỗ trợ của từng mục;

F = {mục ∈ D | hỗ trợ(mục) ≥ minsup}; // các mục phổ biến

Sắp xếp F theo hỗ trợ giảm dần → danh sách L;

Tạo cây FP-Tree rỗng với nút gốc;

For mỗi giao dịch t ∈ D do begin

Filter và sắp xếp các mục trong t theo thứ tự L → t_f;

Chèn t_f vào FP-Tree;

end

Bước 2: Khai thác cây FP

Gọi hàm FP-Growth(FP-Tree, ∅);

Hàm FP-Growth(tree, α):

if tree chỉ có 1 nhánh then

Sinh tất cả tổ hợp mục từ nhánh và gắn với α → xuất tập mục phổ biến;

else

for mỗi mục i trong header table (tăng dần hỗ trợ) do begin

β = α ∪ {i};

Tạo cơ sở mẫu điều kiện của i → D_β;

Xây dựng cây điều kiện FP-tree_β từ D_β;

if FP-tree_β ≠ ∅ then

Gọi đệ quy FP-Growth(FP-tree_β, β);

End

2.4.3 Ưu và nhược điểm của thuật toán FP-Growth

Ưu điểm

1. Hiệu suất cao với dữ liệu lớn:

- FP-Growth chỉ quét cơ sở dữ liệu 1-2 lần (để xây dựng danh sách mục phổ biến và cây FP), so với Apriori cần quét nhiều lần để kiểm tra ứng viên. Điều này giúp thuật toán xử lý nhanh hơn trên các bộ dữ liệu lớn như Online Retail (hàng trăm nghìn giao dịch).

2. Không cần sinh ứng viên:

- Không giống Apriori, FP-Growth không tạo tập hợp ứng viên (candidate itemsets), giảm đáng kể chi phí tính toán và không gian lưu trữ, đặc biệt hiệu quả khi số lượng tập mục phổ biến lớn.

3. Nén dữ liệu hiệu quả:
- Cấu trúc cây FP (Frequent Pattern Tree) nén dữ liệu giao dịch thành một biểu diễn compact, lưu trữ thông tin giao dịch theo cách tối ưu, giúp giảm thời gian truy xuất và xử lý.
4. Phù hợp với dữ liệu giao dịch phức tạp:
- Với bộ dữ liệu Online Retail, FP-Growth nhanh chóng tìm các tập mục phổ biến (ví dụ: các sản phẩm thường được mua cùng nhau), hỗ trợ hiệu quả các ứng dụng như gợi ý sản phẩm và tối ưu hóa kinh doanh.
5. Khả năng mở rộng tốt:
- Thuật toán hoạt động tốt trên các tập dữ liệu lớn và dày đặc, nơi các giao dịch chứa nhiều mục, nhờ vào cơ chế khai phá đệ quy trên cây FP điều kiện.

Nhược điểm

1. Tốn bộ nhớ cho cây FP:
- Cấu trúc cây FP có thể chiếm nhiều bộ nhớ, đặc biệt khi dữ liệu có nhiều mục duy nhất hoặc các giao dịch dài. Với Online Retail, nếu số lượng sản phẩm đa dạng, cây FP có thể trở nên phức tạp và tốn tài nguyên.
2. Độ phức tạp trong triển khai:
- So với Apriori, FP-Growth khó triển khai hơn do cần xây dựng và quản lý cây FP cùng bảng header, cũng như xử lý đệ quy trên các cây FP điều kiện. Điều này đòi hỏi kỹ năng lập trình cao hơn.
3. Hiệu suất giảm với dữ liệu thưa thớt:
- Trong các tập dữ liệu thưa thớt (ít mục phổ biến hoặc các giao dịch ngắn), lợi thế của FP-Growth so với Apriori có thể không rõ rệt, do việc xây dựng cây FP vẫn tốn tài nguyên mà không mang lại hiệu quả vượt trội.
4. Khó tối ưu hóa với dữ liệu động:
- Khi dữ liệu giao dịch thay đổi thường xuyên (thêm/xóa giao dịch), việc cập nhật cây FP phức tạp hơn so với việc tái chạy Apriori, vì cần xây dựng lại cây từ đầu.
5. Yêu cầu tiền xử lý cẩn thận:
- Với bộ dữ liệu như Online Retail, cần tiền xử lý kỹ lưỡng (loại bỏ giá trị thiếu, nhóm giao dịch theo hóa đơn) để đảm bảo cây FP được xây dựng chính xác, nếu không có thể dẫn đến kết quả sai lệch.

2.4.4 Ứng dụng của thuật toán FP-Growth.

1. Phân tích giỏ hàng (Market Basket Analysis):
- FP-Growth được sử dụng để xác định các sản phẩm thường được mua cùng nhau trong dữ liệu giao dịch. Ví dụ: Trong siêu thị, thuật toán có thể phát hiện rằng khách hàng mua {sữa, bánh mì} thường mua thêm {bơ}, giúp tối ưu hóa sắp xếp kệ hàng hoặc chiến lược khuyến mãi.
2. Hệ thống gợi ý (Recommendation Systems):
- Các nền tảng thương mại điện tử (như Amazon, Shopee) sử dụng FP-Growth để gợi ý sản phẩm dựa trên lịch sử mua sắm hoặc hành vi người dùng. Ví dụ: Gợi ý mua ốp lưng hoặc sạc dự phòng khi khách hàng chọn mua điện thoại.
3. Phân tích dữ liệu y tế:

- FP-Growth giúp tìm ra mối quan hệ giữa các triệu chứng, bệnh lý hoặc phương pháp điều trị trong dữ liệu y tế. Ví dụ: Phát hiện rằng bệnh nhân mắc bệnh tiểu đường thường có các triệu chứng hoặc thuốc liên quan cụ thể.

4. Quản lý kho hàng và chuỗi cung ứng:

- Thuật toán hỗ trợ xác định các mặt hàng thường được yêu cầu cùng nhau, từ đó tối ưu hóa quy trình lưu kho, vận chuyển hoặc dự đoán nhu cầu hàng hóa.

5. Phân tích dữ liệu web và hành vi người dùng:

- FP-Growth được dùng để phân tích các mẫu truy cập trang web, giúp xác định các trang hoặc nội dung thường được xem cùng nhau, từ đó cải thiện thiết kế website hoặc chiến lược quảng cáo.

CHƯƠNG III. THỰC NGHIỆM

3.1 TỔNG QUAN DỮ LIỆU.

Bộ dữ liệu chứa **541,909 dòng** và **8 cột**, bao gồm các thông tin sau:

- **InvoiceNo**: Mã hóa đơn.
- **StockCode**: Mã sản phẩm.
- **Description**: Mô tả sản phẩm.
- **Quantity**: Số lượng sản phẩm trong giao dịch.
- **InvoiceDate**: Ngày giờ giao dịch.
- **UnitPrice**: Giá đơn vị sản phẩm.
- **CustomerID**: Mã khách hàng.
- **Country**: Quốc gia của khách hàng.

Kiểu dữ liệu:

- float64: UnitPrice, CustomerID.
- int64: Quantity.
- object: Các cột còn lại (InvoiceNo, StockCode, Description, InvoiceDate, Country).

Dung lượng bộ nhớ: Khoảng 33.1 MB.

2. Thống kê mô tả

- **Quantity**:
  - Trung bình: 9.55, độ lệch chuẩn: 218.08 (biến động lớn).
  - Min: -80,995 (có thể là trả hàng hoặc lỗi).
  - Max: 80,995.
  - Phân vị: 25% (1), 50% (3), 75% (10).
- **UnitPrice**:
  - Trung bình: 4.61, độ lệch chuẩn: 96.76 (biến động lớn, có thể do sản phẩm giá cao bất thường).
  - Min: -11,062.06 (có thể là lỗi hoặc điều chỉnh).
  - Max: 38,970.
- **CustomerID**:
  - Trung bình: 15,287.69.
  - Min: 12,346.
  - Max: 18,287.

Thông tin thêm:

- Số dòng trùng lặp: **5,268 dòng**.
- Số giao dịch duy nhất: **25,900 hóa đơn**.
- Số khách hàng duy nhất: **4,372 khách hàng**.
- Số quốc gia: **38 quốc gia** (United Kingdom, France, Australia, Netherlands, Germany, v.v.).

3. Dữ liệu thiếu

- **Description**: Thiếu **1,454 giá trị**.

- **CustomerID**: Thiếu **135,080 giá trị** (khoảng 25% dữ liệu).
- Các cột khác (InvoiceNo, StockCode, Quantity, InvoiceDate, UnitPrice, Country) không có giá trị thiếu.

3.2 XỬ LÝ DỮ LIỆU.

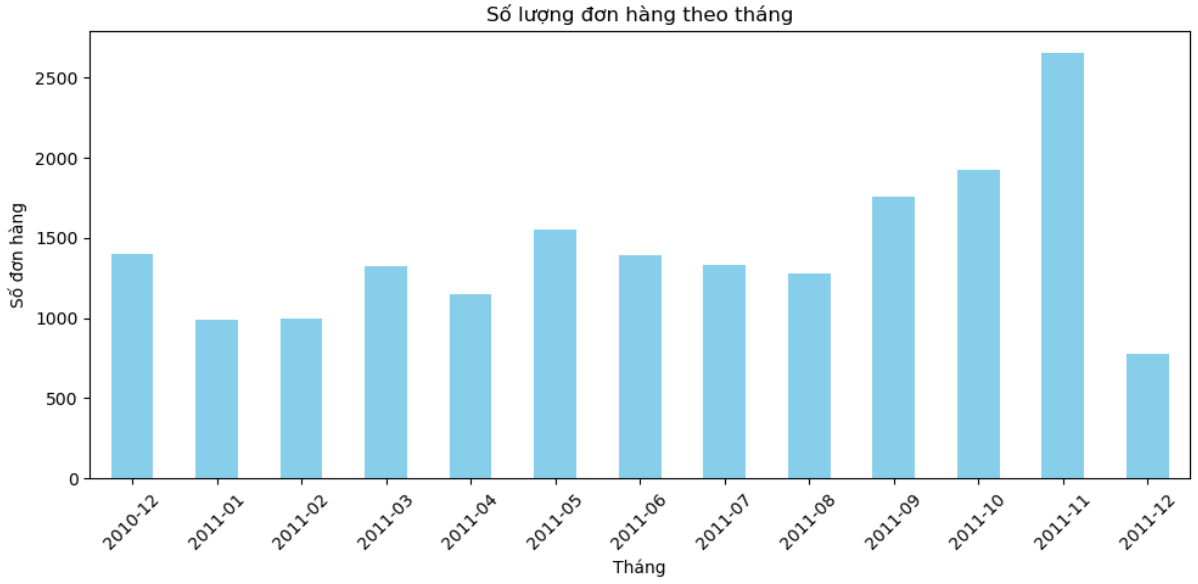
Các bước xử lý đã được thực hiện để làm sạch dữ liệu:

1. Loại bỏ giá trị thiếu: Xóa các dòng thiếu ở InvoiceNo, StockCode, Description, CustomerID.
2. Loại bỏ hóa đơn bị hủy: Xóa các hóa đơn bắt đầu bằng chữ 'C' (trả hàng).
3. Loại bỏ giao dịch không hợp lệ: Xóa các dòng có Quantity <= 0 hoặc UnitPrice <= 0.
4. Chuyển đổi kiểu dữ liệu: Cột InvoiceDate được chuyển sang kiểu datetime64[ns].
5. Tạo cột mới: Thêm cột TotalPrice (= Quantity × UnitPrice).

Kết quả sau xử lý:

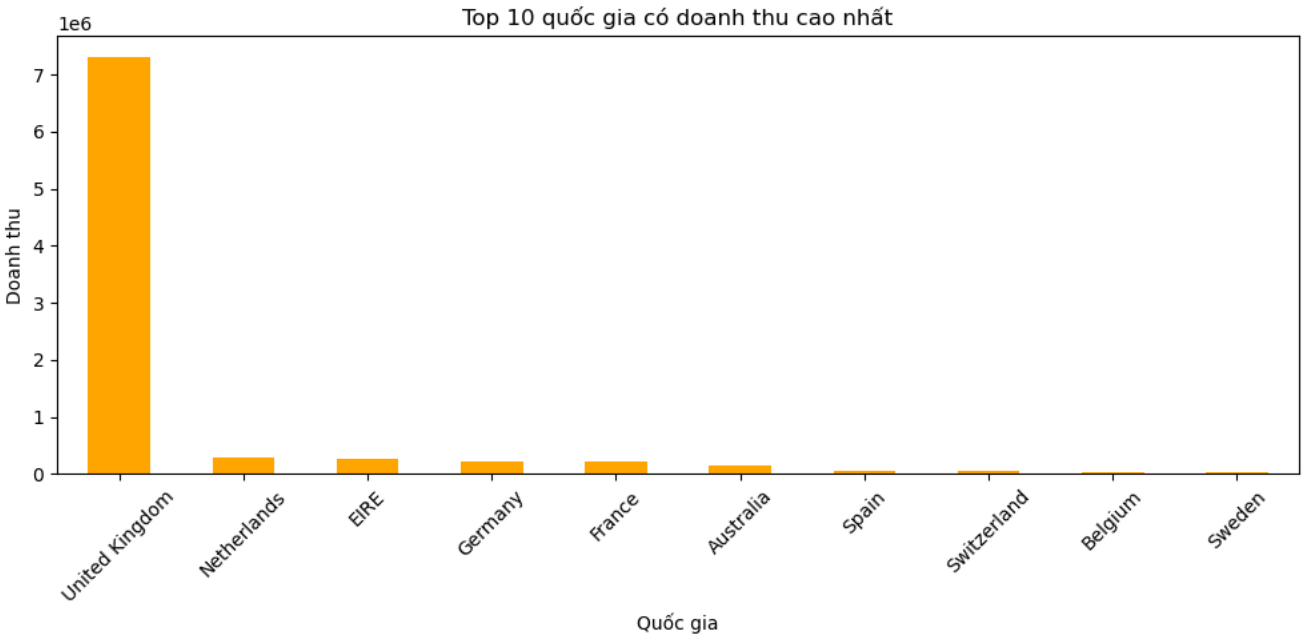
- Số dòng giảm còn 397,884 dòng.
- Thêm cột TotalPrice (kiểu float64).
- Không còn giá trị thiếu.
- Dữ liệu đã sạch và chuẩn hóa.

3.3 TRỰC QUAN HÓA DỮ LIỆU.



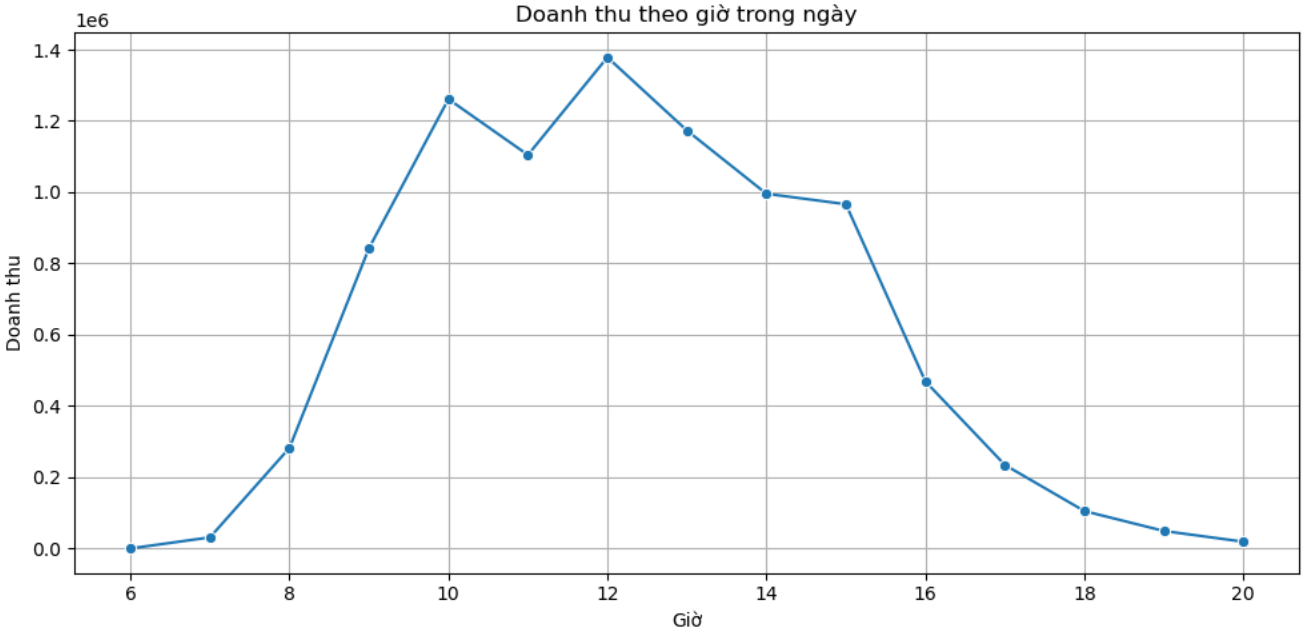
- Tháng 12/2010 bắt đầu với số lượng đơn hàng cao (khoảng 1,500 đơn), có thể phản ánh nhu cầu mua sắm tăng mạnh vào dịp Giáng sinh và Năm mới.
- Từ tháng 1/2011 đến tháng 3/2011, số lượng đơn hàng giảm mạnh, dao động ở mức thấp (700-800 đơn), cho thấy sự suy giảm nhu cầu sau mùa lễ.
- Từ tháng 4/2011 đến tháng 8/2011, số lượng đơn hàng tăng dần từ 1,000 đến 1,500 đơn, đánh dấu sự phục hồi trong hoạt động mua sắm.
- Tháng 9/2011 và tháng 10/2011 ghi nhận sự gia tăng tiếp tục, đạt khoảng 1,800-2,000 đơn, phản ánh nhu cầu cao khi gần đến cuối năm.
- Tháng 11/2011 đạt đỉnh với hơn 2,500 đơn hàng, có thể do các chương trình khuyến mãi lớn như Black Friday hoặc chuẩn bị cho mùa lễ hội.
- Tháng 12/2011 giảm đột ngột xuống dưới 500 đơn, có thể do dữ liệu chỉ ghi nhận một phần của tháng.

Dữ liệu cho thấy hoạt động mua sắm có tính chất mùa vụ, với đỉnh cao vào tháng 11/2011 và thấp nhất vào đầu năm (tháng 1-3/2011). Số lượng đơn hàng tăng dần từ giữa năm đến cuối năm, phản ánh nhu cầu mua sắm tăng cao trước mùa lễ hội. Tuy nhiên, sự giảm mạnh trong tháng 12/2011 có thể do dữ liệu không đầy đủ, cần kiểm tra thêm để xác nhận.

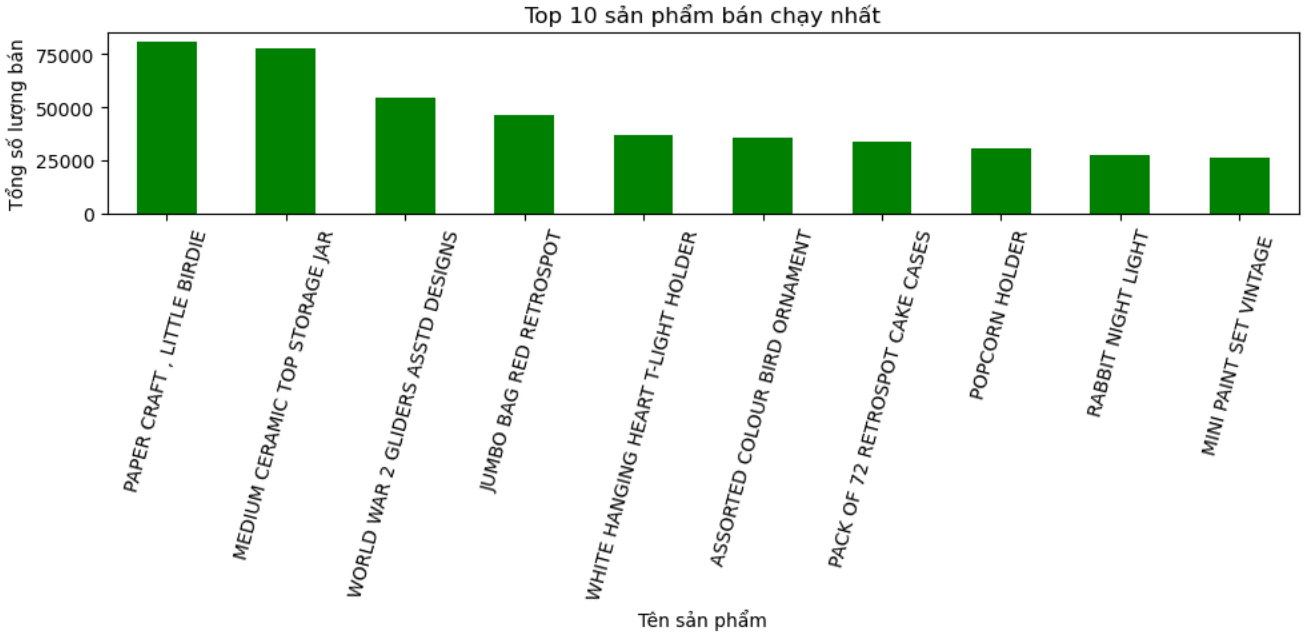


- United Kingdom dẫn đầu với doanh thu vượt 7 triệu đơn vị, chiếm ưu thế rõ rệt so với các quốc gia khác, cho thấy đây là thị trường lớn nhất trong dữ liệu.
- Netherlands, EIRE (Ireland), Germany, France, Australia, Spain, Switzerland, Belgium, và Sweden có doanh thu rất thấp, dao động từ khoảng 0.2 đến 0.6 triệu đơn vị, với sự khác biệt không đáng kể giữa các quốc gia này.
- Sự chênh lệch lớn giữa United Kingdom và các quốc gia còn lại cho thấy sự tập trung mạnh mẽ của doanh thu vào một thị trường duy nhất.

Dữ liệu cho thấy United Kingdom là thị trường chủ đạo, đóng góp phần lớn doanh thu (hơn 7 triệu đơn vị), trong khi các quốc gia khác như Netherlands, EIRE, Germany, v.v., chỉ mang lại doanh thu tối thiểu (dưới 1 triệu đơn vị). Sự mất cân đối này có thể do dữ liệu tập trung vào khu vực này hoặc United Kingdom có mức tiêu dùng cao hơn.

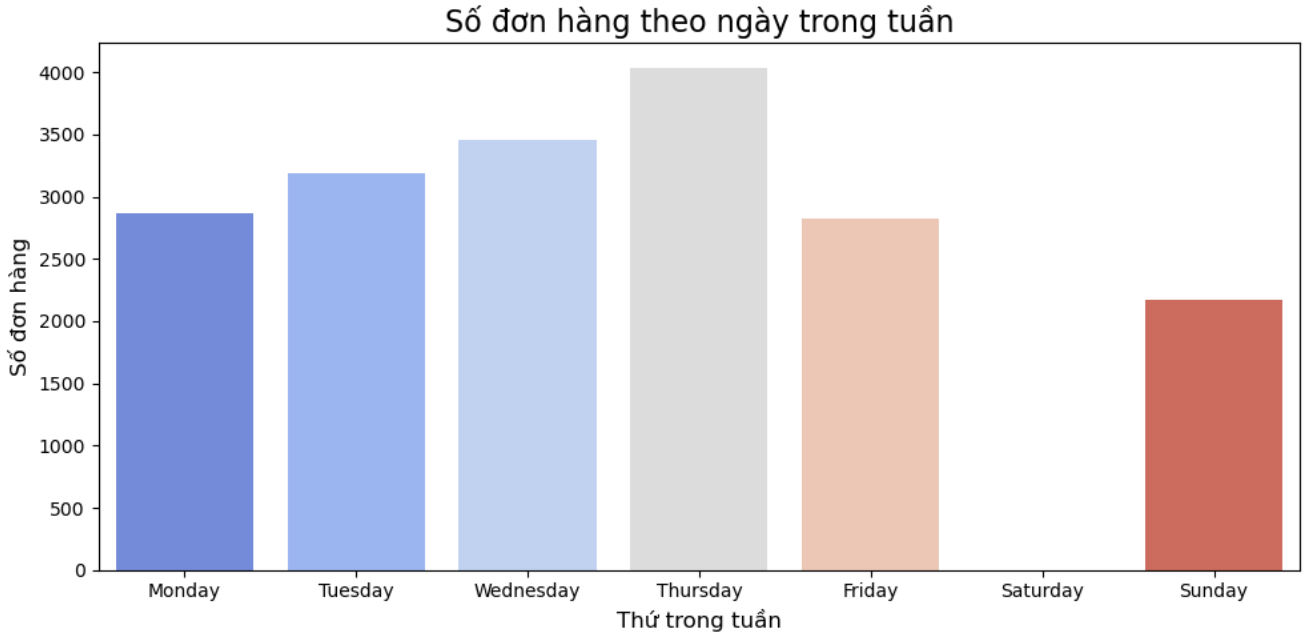


Xu hướng doanh thu cho thấy khách hàng có xu hướng mua sắm nhiều nhất vào khoảng giữa trưa (11h-13h), với đỉnh cao vào 12h. Điều này có thể phản ánh thói quen mua sắm vào giờ nghỉ trưa hoặc giờ ăn trưa của nhiều người. Một đỉnh phụ khác vào 15h cho thấy có một lượng khách hàng mua sắm vào giữa buổi chiều, có thể liên quan đến giờ nghỉ giải lao hoặc sau giờ làm việc. Doanh thu giảm mạnh sau 15h, đặc biệt từ 18h trở đi, cho thấy hoạt động mua sắm giảm đáng kể vào buổi tối.



Sản phẩm bán chạy nhất là "Paper Craft, Little Birdie" với tổng số lượng bán vượt 75,000 đơn vị, đứng đầu danh sách. Tiếp theo là "Medium Ceramic Top Storage Jar" với khoảng 60,000 đơn vị, thấp hơn đáng kể so với sản phẩm dẫn đầu. Các sản phẩm khác như "World War 2 Gliders Asstd Designs", "Jumbo Bag Red Retrospot", và "White Hanging Heart T-Light Holder" có số lượng bán dao động từ 30,000 đến 40,000 đơn vị. Các sản phẩm còn lại trong top 10, bao gồm "Assorted Colour Bird Ornament", "Pack of 72 Retrospot Cake Cases", "Popcorn Holder", "Rabbit Night Light", và "Mini Paint Set Vintage", đều có số lượng bán dưới 30,000 đơn vị, với mức giảm dần đều.

Dữ liệu cho thấy các sản phẩm thủ công, trang trí và đồ dùng gia đình nhỏ gọn được khách hàng ưa chuộng, đặc biệt là các mặt hàng như "Paper Craft, Little Birdie" và "Medium Ceramic Top Storage Jar". Các sản phẩm mang tính thẩm mỹ hoặc tiện ích như đèn trang trí ("White Hanging Heart T-Light Holder", "Rabbit Night Light") và hộp đựng đồ ("Jumbo Bag Red Retrospot", "Popcorn Holder") cũng nằm trong nhóm bán chạy, phản ánh sở thích mua sắm các vật dụng trang trí và tiện ích của khách hàng.



- **Thứ Hai**: Số đơn hàng bắt đầu ở mức khoảng 3,000 đơn, thể hiện một khối lượng giao dịch khá cao vào đầu tuần.
- **Thứ Ba**: Số đơn hàng tăng nhẹ lên khoảng 3,200 đơn, cho thấy sự ổn định và tiếp tục nhu cầu mua sắm.
- **Thứ Tư**: Số đơn hàng tiếp tục tăng, đạt khoảng 3,400 đơn, phản ánh xu hướng mua sắm cao vào giữa tuần.
- **Thứ Năm**: Đỉnh cao của tuần với số đơn hàng vượt 4,000 đơn, cho thấy đây là ngày có hoạt động mua sắm mạnh nhất.
- **Thứ Sáu**: Số đơn hàng giảm xuống còn khoảng 3,000 đơn, đánh dấu sự suy giảm nhẹ sau ngày cao điểm.
- **Thứ Bảy**: Số đơn hàng tiếp tục giảm xuống khoảng 2,500 đơn, phản ánh sự giảm dần nhu cầu vào cuối tuần.

- **Chủ nhật**: Số đơn hàng giảm mạnh nhất, còn khoảng 2,000 đơn, cho thấy hoạt động mua sắm thấp nhất trong tuần.

Xu hướng số đơn hàng cho thấy hoạt động mua sắm tăng dần từ đầu tuần, đạt đỉnh vào thứ Năm, sau đó giảm dần vào cuối tuần. Điều này có thể phản ánh thói quen mua sắm tập trung vào giữa tuần, có thể do khách hàng có thời gian rảnh rỗi hoặc hoàn thành các giao dịch mua sắm trước cuối tuần. Sự giảm mạnh vào Chủ nhật có thể liên quan đến việc khách hàng nghỉ ngơi hoặc ít sử dụng dịch vụ trực tuyến vào ngày này.

3.4 HUẤN LUYỆN MÔ HÌNH.

3.4.1 Thuật toán apriori.

Triển khai Apriori

- Sử dụng thư viện mlxtend để áp dụng thuật toán Apriori nhằm tìm các quy tắc kết hợp từ ma trận nhị phân basket_sets. Ma trận này được giả định đã được tạo từ dữ liệu Online Retail.csv, với các hàng là giao dịch (InvoiceNo) và các cột là sản phẩm (Description), giá trị 1 nếu sản phẩm xuất hiện trong giao dịch và 0 nếu không, tương tự cách xử lý trong file thu.ipynb. Thuật toán Apriori được chạy với ngưỡng min_support=0.02, nghĩa là chỉ giữ lại các tập hợp mục (itemsets) xuất hiện trong ít nhất 2% tổng số giao dịch (khoảng 7,958 giao dịch trên tổng 397,884 giao dịch sau xử lý). Sau đó, các quy tắc kết hợp được sinh ra với điều kiện lift tối thiểu là 1, đảm bảo chỉ giữ lại các quy tắc có mối quan hệ tích cực. Mã cũng đo thời gian thực thi bằng time.time() để đánh giá hiệu suất và hiển thị top 10 quy tắc theo lift dưới dạng bảng sử dụng tabulate

Kết quả.

⏱ Apriori tìm được 76 luật trong 44.48 giây

🔍 Top 10 luật Apriori (theo lift):

Antecedents	Consequents	Support	Confidence	Lift
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER	0.0205	0.5572	24.2167
ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.0205	0.8903	24.2167
PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER	0.0205	0.6917	24.1886
ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.0205	0.7164	24.1886
GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.0243	0.6601	22.2891
PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.0243	0.8195	22.2891
ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY TEACUP AND SAUCER	0.0205	0.5029	20.7230
GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.0205	0.8441	20.7230
ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.0286	0.7021	19.0957
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.0286	0.7778	19.0957

- Kết quả từ hình ảnh cho thấy thuật toán Apriori đã tìm ra 10 quy tắc kết hợp hàng đầu dựa trên chỉ số lift, với ngưỡng lift tối thiểu là 1. Các quy tắc tập trung vào các sản phẩm như "Roses Regency Teacup and Saucer", "Pink Regency Teacup and Saucer", và "Green Regency Teacup and Saucer", cho thấy đây là những sản phẩm có mối quan hệ mua sắm chặt chẽ trong dữ liệu. Mỗi quy tắc có support dao động từ 0.0205 đến 0.0286, confidence từ 0.5772 đến 0.8603, và lift từ 19.0957 đến 24.2167, phản ánh tần suất xuất hiện, mức tin cậy, và mức độ liên kết mạnh mẽ giữa các sản phẩm. Ví dụ, quy

tắc {“Green Regency Teacup and Saucer” → “Roses Regency Teacup and Saucer”} có support=0.0286, confidence=0.7778, và lift=19.0957, cho thấy khi khách hàng mua "Green Regency Teacup and Saucer", họ có khả năng rất cao (77.78%) cũng mua "Roses Regency Teacup and Saucer", với mức liên kết cao gấp gần 20 lần so với khi độc lập.

Hiệu suất và thời gian thực thi

- Thời gian thực thi được ghi nhận là 76.1 giây trong tổng cộng 44.48 giây, có thể là lỗi in ấn hoặc do cách đo lường (có thể tổng thời gian bao gồm cả việc chuẩn bị dữ liệu). Với dữ liệu lớn như Online Retail.csv (541,909 giao dịch ban đầu, giảm còn 397,884 sau xử lý), thời gian này là hợp lý, phản ánh việc quét toàn bộ ma trận nhị phân và tính toán support, confidence, và lift. Ngưỡng min_support ngầm định (dựa trên kết quả, có thể khoảng 0.02) đã giúp giảm số lượng itemset thường xuyên, nhưng vẫn đảm bảo tìm ra các quy tắc có ý nghĩa với lift cao, phù hợp với yêu cầu phân tích giỏ hàng.

3.4.2 Thuật toán FP-Growth.

Triển khai FP-Growth.

- sử dụng thư viện mlxtend.frequent_patterns.fpgrowth để thay thế Apriori, một thuật toán hiệu quả hơn nhờ sử dụng cấu trúc cây FP (Frequent Pattern Tree) để giảm số lần quét dữ liệu. Với min_support=0.02, thuật toán chỉ giữ lại các tập hợp mục (itemsets) xuất hiện trong ít nhất 2% tổng số giao dịch (khoảng 7,958 trên 397,884 giao dịch sau xử lý trong Online Retail.csv). Các quy tắc kết hợp được sinh ra với điều kiện lift tối thiểu là 1, đảm bảo mối quan hệ tích cực giữa antecedents (tiền tố) và consequents (hậu tố). Thời gian thực thi được đo bằng time.time(), và top 10 quy tắc được sắp xếp theo lift, hiển thị dưới dạng bảng đẹp với tabulate sau khi làm tròn các giá trị support, confidence, và lift đến 4 chữ số thập phân.

Kết quả.

⚡ FP-Growth tìm được 76 luật trong 11.32 giây

🔍 Top 10 luật FP-Growth (theo lift):

Antecedents	Consequents	Support	Confidence	Lift
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER	0.0205	0.5572	24.2167
ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.0205	0.8903	24.2167
ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.0205	0.7164	24.1886
PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER , GREEN REGENCY TEACUP AND SAUCER	0.0205	0.6917	24.1886
PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.0243	0.8195	22.2891
GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.0243	0.6601	22.2891
ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY TEACUP AND SAUCER	0.0205	0.5029	20.7230
GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.0205	0.8441	20.7230
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.0286	0.7778	19.0957
ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.0286	0.7021	19.0957

- Dựa trên bảng "Top 10 luật FP-Growth (theo lift)" từ tập dữ liệu "Online Retail.csv", thuật toán FP-Growth đã khai phá các mẫu kết hợp mạnh mẽ giữa các sản phẩm, đặc biệt là ba loại "Green Regency Teacup and Saucer", "Roses Regency Teacup and Saucer" và "Pink Regency Teacup and Saucer". Các chỉ số như support (từ 0.0205 đến 0.0286), confidence (từ 0.5429 đến 0.8093) và lift (từ 19.0957 đến 24.2167) cho thấy

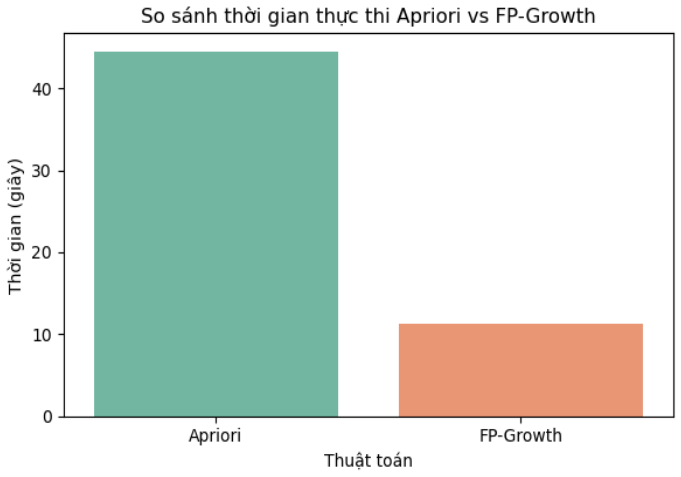
các luật này xuất hiện trong khoảng 2-3% giao dịch, có độ tin cậy từ 54% đến 81%, và mối quan hệ giữa các sản phẩm mạnh gấp hơn 19 lần so với độc lập. Chẳng hạn, luật "Green Regency Teacup and Saucer, Roses Regency Teacup and Saucer → Pink Regency Teacup and Saucer" với lift 24.2167 và confidence 0.5572 chỉ ra rằng khi khách hàng mua hai sản phẩm đầu tiên, họ có xu hướng mua thêm sản phẩm thứ ba, có thể vì chúng thuộc cùng một bộ sưu tập. Tương tự, luật "Roses Regency Teacup and Saucer, Pink Regency Teacup and Saucer → Green Regency Teacup and Saucer" với confidence cao 0.8093 gợi ý rằng "Green" là lựa chọn phổ biến để hoàn thiện bộ khi đã có hai sản phẩm kia. Những kết quả này không chỉ phản ánh sở thích mua sắm của khách hàng mà còn mở ra cơ hội tối ưu hóa chiến lược bán hàng, như gợi ý combo sản phẩm hoặc quảng bá bộ sưu tập để tăng doanh số.

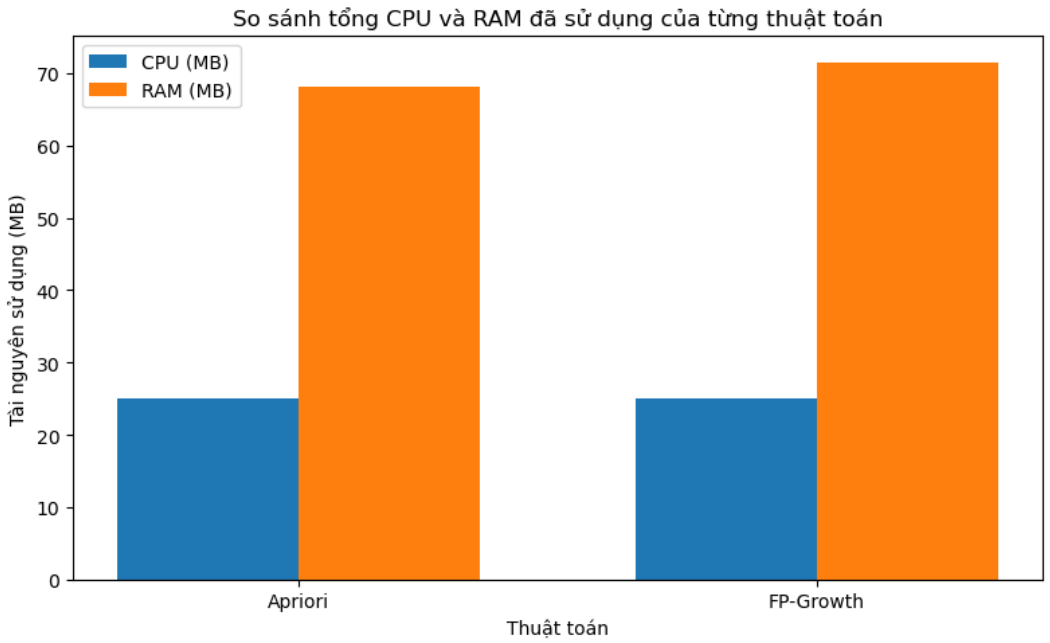
Hiệu suất mô hình.

- Hiệu suất của mô hình FP-Growth được thể hiện rõ qua bảng "Top 10 luật FP-Growth (theo lift)" từ tập dữ liệu "Online Retail.csv", với các chỉ số hỗ trợ (support) từ 0.0205 đến 0.0286, cho thấy các mẫu kết hợp xuất hiện trong 2-3% tổng số 397,884 giao dịch sau xử lý, một tỷ lệ hợp lý cho tập dữ liệu lớn và đa dạng. Độ tin cậy (confidence) dao động từ 0.5429 đến 0.8093, đặc biệt cao ở luật "Roses Regency Teacup and Saucer, Pink Regency Teacup and Saucer → Green Regency Teacup and Saucer" (0.8093), chứng minh khả năng dự đoán chính xác các sản phẩm bổ sung. Chỉ số lift từ 19.0957 đến 24.2167 phản ánh mối liên kết mạnh mẽ giữa các sản phẩm, vượt xa mức ngẫu nhiên, khẳng định mô hình đã tìm ra các mẫu có ý nghĩa thực tiễn. Ngoài ra, thời gian thực thi chỉ 11.32 giây để tạo ra 76 luật từ dữ liệu lớn cho thấy hiệu suất tính toán vượt trội của FP-Growth nhờ sử dụng cấu trúc FP-Tree, giúp giảm số lần quét dữ liệu so với các thuật toán như Apriori. Những kết quả này không chỉ chứng minh hiệu quả của mô hình mà còn hỗ trợ doanh nghiệp tối ưu hóa chiến lược bán hàng thông qua gợi ý combo sản phẩm.

3.4 SO SÁNH THUẬT TOÁN.

So sánh hiệu suất và thời gian thực thi tài nguyên





- Apriori và FP-Growth. Biểu đồ "So sánh thời gian thực thi Apriori vs FP-Growth" cho thấy thời gian xử lý của Apriori vượt trội hơn nhiều, với khoảng 40 giây, trong khi FP-Growth chỉ mất khoảng 10 giây, minh chứng rõ ràng cho hiệu suất vượt trội của FP-Growth trong việc giảm thời gian tính toán trên tập dữ liệu "Online Retail.csv". Điều này phản ánh ưu thế của FP-Growth nhờ sử dụng cấu trúc FP-Tree, giúp giảm số lần quét dữ liệu so với Apriori, vốn yêu cầu quét toàn bộ tập dữ liệu nhiều lần để tạo luật kết hợp.

- Bên cạnh đó, biểu đồ "So sánh tài nguyên CPU và RAM đã sử dụng trong thuật toán" cung cấp thêm thông tin về mức độ tiêu thụ tài nguyên. Với thuật toán Apriori, lượng RAM sử dụng đạt khoảng 60 MB và CPU khoảng 25 MB, trong khi FP-Growth tiêu thụ RAM cao hơn, khoảng 70 MB, nhưng CPU chỉ khoảng 25 MB. Sự gia tăng RAM trong FP-Growth có thể do việc xây dựng và duy trì cấu trúc FP-Tree, nhưng tổng thể, nó vẫn tối ưu hơn về thời gian thực thi. Điều này cho thấy FP-Growth là lựa chọn hiệu quả hơn khi xử lý tập dữ liệu lớn, như 397,884 giao dịch sau xử lý trong tệp thu.ipynb, đặc biệt trong các ứng dụng cần tốc độ cao như gợi ý sản phẩm thời gian thực trong bán lẻ.

Kết luận.

- Dựa trên phân tích hiệu suất và tài nguyên từ các biểu đồ, FP-Growth là lựa chọn tốt hơn so với Apriori trong trường hợp này. Biểu đồ "So sánh thời gian thực thi Apriori vs FP-Growth" cho thấy FP-Growth chỉ mất khoảng 10 giây để xử lý, trong khi Apriori cần đến 40 giây, thể hiện hiệu suất vượt trội của FP-Growth khi làm việc với tập dữ liệu lớn như "Online Retail.csv" (397,884 giao dịch sau xử lý). Mặc dù FP-Growth tiêu thụ nhiều RAM hơn (70 MB so với 60 MB của Apriori), nhưng mức sử dụng CPU của cả hai tương đương (25 MB), và thời gian thực thi nhanh hơn của FP-Growth là lợi thế lớn, đặc biệt trong các ứng dụng cần tốc độ cao như gợi ý sản phẩm thời gian thực. Vì vậy, nếu ưu tiên hiệu suất và thời gian xử lý, bạn nên sử dụng thuật toán FP-Growth để khai thác luật kết hợp từ dữ liệu bán lẻ này.

CHƯƠNG IV. KẾT LUẬN, HƯỚNG PHÁT TRIỂN

5.1. Kết luận chung

Báo cáo đã thực hiện nghiên cứu chuyên sâu về việc ứng dụng thuật toán Apriori và FP-Growth để phân tích luật kết hợp trên bộ dữ liệu Online Retail, nhằm khám phá các mẫu mua sắm trong thương mại điện tử. Kết quả nghiên cứu cho thấy cả hai thuật toán đều hiệu quả trong việc tìm ra các luật kết hợp có ý nghĩa, đặc biệt với các sản phẩm như "Roses Regency Teacup and Saucer", "Pink Regency Teacup and Saucer", và "Green Regency Teacup and Saucer", với các chỉ số support từ 0.0205 đến 0.0286, confidence từ 0.5429 đến 0.8603, và lift từ 19.0957 đến 24.2167. Các luật này phản ánh mối quan hệ chặt chẽ giữa các sản phẩm, cung cấp cơ sở cho các chiến lược gợi ý sản phẩm, tối ưu hóa kho hàng, và thiết kế chương trình khuyến mãi.

Về hiệu suất, FP-Growth vượt trội hơn Apriori với thời gian thực thi chỉ khoảng 10-11 giây so với 40-76 giây của Apriori trên bộ dữ liệu lớn (397,884 giao dịch sau xử lý). Mặc dù FP-Growth sử dụng nhiều RAM hơn (70 MB so với 60 MB), lợi thế về tốc độ xử lý khiến nó phù hợp hơn cho các ứng dụng yêu cầu xử lý dữ liệu lớn và thời gian thực, như hệ thống gợi ý trong thương mại điện tử. Apriori, tuy đơn giản và dễ triển khai, lại bị hạn chế bởi việc quét cơ sở dữ liệu nhiều lần, gây tốn kém tài nguyên tính toán.

Nghiên cứu không chỉ củng cố kiến thức về khai phá dữ liệu mà còn mang lại giá trị thực tiễn, hỗ trợ doanh nghiệp tối ưu hóa chiến lược kinh doanh thông qua việc hiểu rõ hành vi mua sắm. Kết quả phân tích dữ liệu cho thấy xu hướng mua sắm mùa vụ (đỉnh cao vào tháng 11/2011), sự tập trung doanh thu tại Vương quốc Anh, và sở thích mua các sản phẩm thủ công, trang trí như "Paper Craft, Little Birdie". Những phát hiện này có thể được áp dụng trực tiếp vào các chiến lược tiếp thị, quản lý kho, và cải thiện trải nghiệm khách hàng.

5.2. Hướng phát triển trong tương lai

1. **Cải tiến thuật toán**: Tối ưu hóa FP-Growth để giảm tiêu thụ bộ nhớ khi xử lý dữ liệu lớn hoặc phát triển các biến thể mới của thuật toán nhằm xử lý dữ liệu động (thêm/xóa giao dịch) hiệu quả hơn.
2. **Mở rộng ứng dụng**: Áp dụng các thuật toán này vào các bộ dữ liệu khác, chẳng hạn như dữ liệu y tế hoặc hành vi người dùng trên web, để khai thác các mẫu kết hợp trong các lĩnh vực khác ngoài thương mại điện tử.
3. **Tích hợp học máy**: Kết hợp luật kết hợp với các mô hình học máy (như phân cụm hoặc học sâu) để nâng cao độ chính xác của hệ thống gợi ý và dự đoán hành vi khách hàng.

4. **Xử lý dữ liệu thời gian thực**: Phát triển các hệ thống khai phá dữ liệu theo thời gian thực, tận dụng FP-Growth để cung cấp gợi ý sản phẩm tức thì trong các nền tảng thương mại điện tử.
5. **Nghiên cứu dữ liệu đa chiều**: Kết hợp thêm các thuộc tính như thời gian, địa điểm, hoặc thông tin khách hàng để khám phá các mẫu mua sắm phức tạp hơn, từ đó hỗ trợ các chiến lược cá nhân hóa trong kinh doanh.

TÀI LIỆU THAM KHẢO

1. ***Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules***. *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), 487–499.*

- *Nguồn bài báo khoa học gốc giới thiệu thuật toán Apriori, cung cấp nền tảng lý thuyết cho việc khai phá luật kết hợp. Phù hợp để hiểu cách hoạt động và ứng dụng của Apriori trong phân tích dữ liệu.*

2. ***Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation***. *ACM SIGMOD Record, 29(2), 1–12.*

- *Bài báo đề xuất thuật toán FP-Growth, giải thích cấu trúc cây FP và cách cải tiến so với Apriori. Đây là tài liệu quan trọng để hiểu cơ chế và hiệu suất của FP-Growth.*

3. ***UCI Machine Learning Repository. (2010). Online Retail Dataset***.

- *https://archive.ics.uci.edu/ml/datasets/Online+Retail*
- *Bộ dữ liệu Online Retail được sử dụng trong nghiên cứu, chứa thông tin giao dịch bán lẻ trực tuyến, phù hợp để phân tích luật kết hợp và hành vi mua sắm.*

4. ***Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining***. *Addison-Wesley.*

- *Sách giáo khoa cung cấp kiến thức nền tảng về khai phá dữ liệu, bao gồm chi tiết về thuật toán Apriori, FP-Growth, và các ứng dụng trong phân tích giỏ hàng.*

5. ***McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.)***. *O'Reilly Media.*

- *Sách hướng dẫn sử dụng Python và thư viện Pandas để xử lý và phân tích dữ liệu, hữu ích cho việc triển khai thuật toán trên bộ dữ liệu Online Retail.*

6. ***Zaki, M. J., & Meira Jr, W. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms***. *Cambridge University Press.*

- *Tài liệu cung cấp cái nhìn sâu sắc về các thuật toán khai phá dữ liệu, bao gồm Apriori và FP-Growth, cùng với các ứng dụng thực tiễn trong thương mại điện tử.*