Với sự bùng nổ và phát triển của CNTT đã mang lại nhiều hiệu quả đối với khoa học cũng như các hoạt động thực tế, trong đó khai phá dữ liệu là một lĩnh vực mang lại hiệu quả thiết thực cho con người. Khai phá dữ liệu đã giúp người sử dụng thu được những tri thức hữu ích từ những CSDL hoặc các kho dữ liệu khổng lồ khác. CSDL trong các đơn vị, tổ chức kinh doanh, quản lý khoa học chứa đựng nhiều thông tin tiềm ẩn, phong phú và đa dạng, đòi hỏi phải có những phương pháp nhanh, phù hợp chính xác, hiệu quả để lấy được những thông tin bổ ích.

Một trong những nội dung cơ bản nhất trong khai phá dữ liệu và rất phổ biến là phát hiện các luật kết hợp. Phương pháp này nhằm tìm ra các tập phần tử thường xuất hiện đồng thời trong cơ sở dữ liệu và rút ra các luật về ảnh hưởng của một tập phần tử thường xuyên xuất hiện đồng thời trong CSDL và rút ra các luật về ảnh hưởng của một tập phần tử dẫn đến sự xuất hiện của một (hoặc một tập) phần tử khác như thế nào. Bên cạnh đó, nhu cầu khai phá khi dữ liệu gia tăng là rất cần thiết hiện nay bởi kích thước lưu trữ dữ liệu ngày càng nhiều nên đòi hỏi tốc độ xử lý cũng như dung lượng bộ nhớ hệ thống phải đảm bảo. Vì thế, yêu cầu cần có những thuật toán hiệu quả cho việc phát hiện luật kết hợp.

Ngày nay, các phương tiện lưu trữ có dung lượng ngày càng lớn, và các hệ quản trị CSDL cũng phát triển mạnh mẽ, cung cấp cho người dùng khả năng lưu trữ dữ liệu không giới hạn. Dữ liệu tuy nhiều nhưng những giá trị tri thức mà nó chứa đựng lại chưa được sử dụng một cách hiệu quả. Với những thành công trong các công trình nghiên cứu trong lĩnh vực khai phá dữ liệu, con người đã khai thác được những giá trị tri thức từ các dữ liệu được lưu trữ, và sử dụng chúng để giải quyết các bài toán trong nhiều lĩnh vực quan trọng của cuộc sống.

Quá trình khai phá dữ liệu là quá trình trích xuất thông tin, khám phá tri thức có mối tương quan nhất định từ một kho dữ liệu khổng lồ nhằm mục đích dự đoán các xu thế, hành vi trong tương lai, hoặc tìm kiếm tập các thông tin hữu ích mà bình thường không thể nhận diện được.

Tập phổ biến *(Frequent patterns)* là những mẫu (tập món đồ, danh sách con, cấu trúc con) mà thường xuyên xuất hiện trong một tập dữ liệu. Ví dụ, tập các món hàng như sữa và bánh mì, 2 món hàng thường xuyên xuất hiện cùng nhau trong 1 giao dịch thì được gọi là 1 tập phổ biến. Một dãy các hành động, như là đầu tiên mua máy tính, sau đó mua máy ảnh, và mua thẻ nhớ, nếu chúng thường xuyên xuất hiện trong lịch sử mua hàng của CSDL thì được gọi là một tập phổ biến.

Tìm ra các tập phổ biến đóng một vai trò quan trọng trong khai phá luật kết hợp *(associations)*, phân tích tương quan *(correlations)*, và rất nhiều mối quan hệ thú vị khác trong kho lưu trữ dữ liệu. Ngoài ra, nó còn giúp phân loại, gom nhóm dữ liệu và thực thi nhiều tác vụ khai phá dữ liệu khác. Do đó, khai phá tập phổ biến *(frequent pattern mining)* đã trở thành một nhiệm vụ và chủ đề được chú trọng khi nghiên cứu dữ liệu. Trong đề tài này, chúng tôi sẽ giới thiệu về khái niệm tập phổ biến, giải thuật FP Growth, và trình bày làm sao ta có thể ứng dụng giải thuật FP Growth để thực hiện khai phá luật phổ biến các hóa đơn bán hàng.

Chủ đề khai phá tập phổ biến thực sự là rất rộng lớn. Trong quy mô đề tài, chúng tôi tập trung bàn về một phương pháp cụ thể của khai phá tập phổ biến ứng dụng thuật toán FP Growth. Chúng tôi sẽ trả lời cho các câu hỏi: Tập phổ biến là gì? Nó có ý nghĩa, vai trò như thế nào? Vì sao chúng ta phải thực hiện khai phá tập phổ biến? Giải thuật FP Growth thực thi như thế nào? Ứng dụng giải thuật FP Growth để giải quyết bài toán tìm tập phổ biến các hóa đơn bán hàng? Tính hiệu quả của giải thuật FP Growth và các biện pháp cải thiện nếu có?

Khai phá tập phổ biến thực hiện tìm kiếm, nghiên cứu các mối quan hệ trong một hệ CSDL. Một trong những ứng dụng cụ thể nhất của công việc khai phá tập phổ biến đó là Phân tích giỏ thị trường *(Market Basket Analysis)*

**Market Basket Analysis**

Khai phá các tập món hàng dẫn đến một khám phá thú vị về sự kết hợp, tương hợp giữa các món hàng trong những bộ dữ liệu giao dịch hoặc quan hệ khổng lồ. Với số lượng lớn dữ liệu liên tục được thu thập và lưu trữ, nhiều công việc kinh doanh, các ngành công nghiệp ngày càng muốn khai phá những mẫu từ CSDL. Những khám phá này có thể hỗ trợ trong nhiều nghiệp vụ kinh doanh như đưa ra quyết định, thiết kế catalog, bán hàng chéo, bán hàng gia tăng, phân tích hành vi người mua hàng,… Ví dụ điển hình là phân tích giỏ thị trường. Quy trình này phân tích thói quen mua hàng của khách bằng các tìm ra các tập món hàng thường xuyên được mua, sự kết hợp giữa các món hàng khác nhau mà người dùng cho vào giỏ hàng của họ. Sự khám phá đó có thể hỗ trợ người bán hàng trong các chiến dịch marketing bằng cách tìm sự thật ngầm hiểu *(insight)*. Ví dụ, khi một khách hàng đang tìm mua sữa, có thể bao nhiêu phần trăm khách hàng đó sẽ mua thêm bánh mì trong cùng một lần tìm kiếm món hàng trên giá. Tìm hiểu về những tri thức, thông tin này có thể giúp người bán hàng gia tăng doanh số bằng cách đưa ra chiến lược marketing thích hợp, những cách bố trí vị trí của các món hàng trên giá để tiện lợi cho người mua.

Applications: Basket data analysis, cross-marketing, catalog design, loss-leader analysis, clustering, classification, recommendation systems, etc

**Một giỏ thị trường là một nhóm các sản phẩm được thiết kế để theo dõi hiệu suất của một phân khúc thị trường cụ thể**. Giỏ thị trường kinh tế tập trung vào các chỉ số giá tiêu dùng (CPI), theo dõi hàng hóa tiêu dùng khác nhau và sử dụng mức giá của họ để cung cấp một ước tính lạm phát. Tuy nhiên, các nhà đầu tư, một giỏ thị trường liên quan đến chứng khoán và tài chính là ý tưởng chính đằng sau các quỹ chỉ số

Ý nghĩa - Giải thích

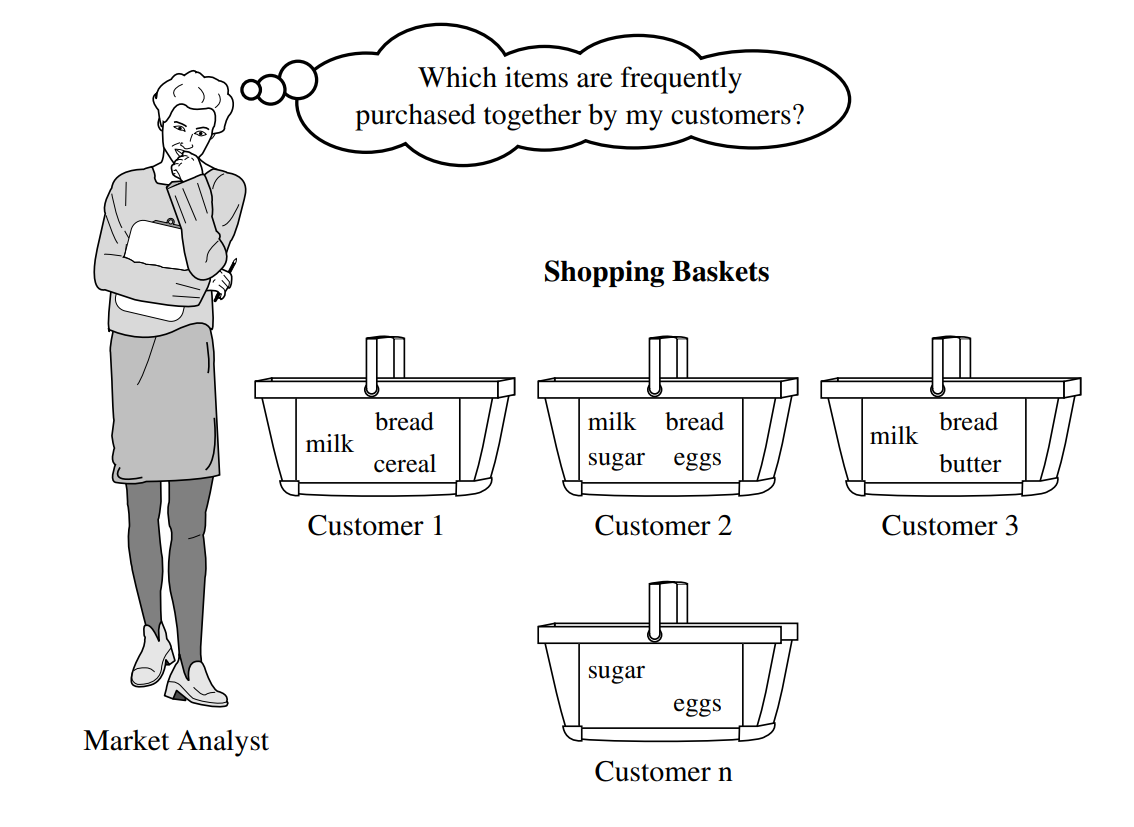
Market Basket

Phân tích giỏ thị trường là một kỹ thuật khai thác dữ liệu được các nhà bán lẻ sử dụng để tăng doanh số bằng cách hiểu rõ hơn các mô hình mua hàng của khách hàng. Nó liên quan đến việc phân tích các tập dữ liệu lớn như lịch sử mua hàng, để tiết lộ các nhóm sản phẩm và các sản phẩm có khả năng được mua cùng nhau.

## Ví dụ

Let’s look at an example of how market basket analysis can be useful.

Suppose, as manager of an AllElectronics branch, you would like to learn more about the buying habits of your customers. Specifically, you wonder, “Which groups or sets of items are customers likely to purchase on a given trip to the store?” To answer your question, market basket analysis may be performed on the retail data of customer transactions at your store. You can then use the results to plan marketing or advertising strategies, or in the design of a new catalog. For instance, market basket analysis may help you design different store layouts. In one strategy, items that are frequently purchased together can be placed in proximity in order to further encourage the sale of such items together. If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items. In an alternative strategy, placing hardware and software at opposite ends of the store may entice customers who purchase such items to pick up other items along the way. For instance, after deciding on an expensive computer, a customer may observe security systems for sale while heading toward the software display to purchase antivirus software and may decide to purchase a home security system as well. Market basket analysis can also help retailers plan which items to put on sale at reduced prices. If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers as well as computers.



## Mục tiêu

Với những gì đã đề cập ở trên, đề tài sẽ tập trung vào những vấn đề sau:

* Thực hiện khảo sát, phân tích và triển khai giải thuật FP Growth bằng ngôn ngữ Java.
* Khảo sát tập dữ liệu hóa đơn bán lẻ của UK, France, Portual, Sweden, tìm các tập món hàng phổ biến bằng giải thuật FP Growth đã triển khai.
* Thực hiện so sánh kết quả của chương trình tự triển khai với kết quả từ công cụ Weka.
* Đánh giá kết quả đã đạt.

## Phạm vi

Đề tài sử dụng dữ liệu từ

Giải thuật sử dụng là FP Growth tự triển khai, thực hiện tìm các tập hạng mục phổ biến.

Những vấn đề nâng cao hơn có liên quan như tìm luật kết hợp, luật tương hợp sẽ không nằm trong phạm vi của assignment 2 này.