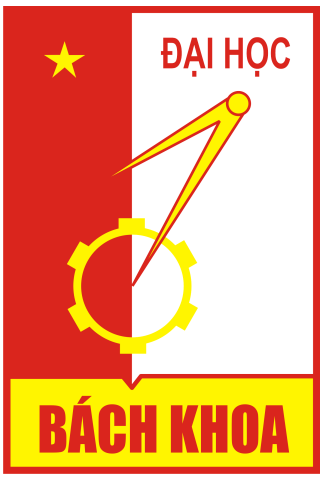
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC



TỐI ƯU GIỎ HÀNG

MÔN HỌC HỆ HỖ TRỢ QUYẾT ĐỊNH

Ngành: HỆ THỐNG THÔNG TIN QUẢN LÝ

Giảng viên: TS. Lê Hải Hà

Sinh viên thực hiện: Dương Thái Huy

Lớp: MI2-02 – K65

HÀ NỘI – 2023

NHẬN XÉT CỦA GIẢNG VIÊN

1. Mục tiêu

(a)

(b)

(c)

2. Nội dung

(a)

(b)

(c)

3. Đánh giá kết quả đạt được

(a)

(b)

(c)

Hà Nội, ngày tháng 8 năm 2023

Giảng viên

TS. Lê Hải Hà

Lời cảm ơn

"Đầu tiên, em xin gửi lời cảm ơn chân thành đến Trường Đại học Bách Khoa Hà Nội đã đưa môn học Hệ hỗ trợ quyết định vào chương trình giảng dạy. Đặc biệt, em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn - TS. Lê Hải Hà đã dạy dỗ, truyền đạt những kiến thức quý báu cho em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp học Hệ hỗ trợ quyết định của thầy, em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý báu, là hành trang để em có thể vững bước sau này.

Bộ môn Hệ hỗ trợ quyết định là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên. Tuy nhiên, do vốn kiến thức còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ. Mặc dù em đã cố gắng hết sức nhưng chắc chắn bài tiểu luận khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, kính mong thầy xem xét và góp ý để bài tiểu luận của em được hoàn thiện hơn.

Em xin chân thành cảm ơn!"

Hà Nội, tháng 08 năm 2023

Sinh viên

Dương Thái Huy

Tóm tắt nội dung Báo cáo

1. Giới thiệu về phân tích giỏ hàng

2. Thuật toán phân tích giỏ hàng

3. Minh họa qua bộ dữ liệu thực tế

4. Đánh giá kết quả

Mục lục

Chương 1 Giới thiệu về Phân tích giỏ hàng **1**

1.1 Phân tích giỏ hàng là gì? 1

1.2 Cách thức hoạt động 3

1.3 Phân loại 3

1.4 Ứng dụng thực tiễn 4

Chương 2 Thuật toán của phân tích giỏ hàng **5**

2.1 Luật kết hợp 5

2.1.1 Các định nghĩa và khái niệm 5

2.2 Thuật toán Apriori 10

2.2.1 Nguyên tắc của Apriori 10

2.2.2 Mô tả thuật toán Apriori 10

2.2.3 Thuật toán Apriori 10

2.2.4 Ví dụ minh họa 11

Chương 3 Áp dụng trên bộ dữ liệu **13**

3.1 Tổng quan về bộ dữ liệu 13

3.2 Áp dụng 13

Chương 4 Đánh giá kết quả **16**

Tài liệu tham khảo **17**

Chương 1

Giới thiệu về Phân tích giỏ hàng

Máy học (Machine Learning) là lĩnh vực đóng góp lợi ích cho ngành công nghiệp bán lẻ một cách rất đặc trưng. Nó hỗ trợ mọi phân khúc trong lĩnh vực bán lẻ, từ dự đoán doanh thu bán hàng cho đến việc xác định tập người dùng. Trong đó, *Phân tích giỏ hàng (Basket Analysis)* là một trong những ứng dụng hàng đầu trong ngành bán lẻ của máy học. Phân tích giỏ hàng thị trường giúp các nhà bán lẻ biết về những đồ vật mà khách hàng thường mua kèm với nhau từ đó các cửa hàng hay website sẽ thiết kế sao cho khách hàng có thể mua những đồ đó dễ dàng hơn. Phân tích giỏ hàng thường chủ yếu được nghiên cứu qua lịch sử hành vi mua hàng. Các công ty cũng đẩy mạnh các chiến lược *bán chéo (cross-selling)* các sản phẩm của họ trên nền tảng trực tuyến. Tuy nhiên, Phân tích giỏ hàng không chỉ được sử dụng trong lĩnh vực bán lẻ mà còn mở rộng ở các lĩnh vực khác như: bảo hiểm và giao dịch tín dụng.

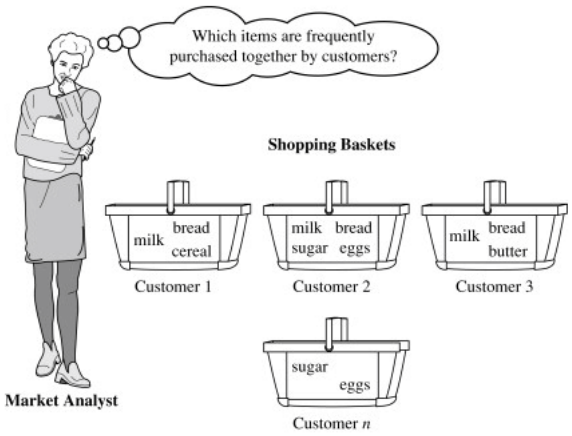
Trong chương 1 này, tôi sẽ trình bày khái niệm , cách thức hoạt động, cách hình thức và ứng dụng thực tiễn của Phân tích giỏ hàng.

1.1 Phân tích giỏ hàng là gì?

Phân tích giỏ hàng là một phương pháp *Khai phá dữ liệu (Data Mining)* nhằm phân tích mô hình của sự đồng thời xảy ra và xác định tính liên kết giữa các sản phẩm được mua đồng thời. Những mô hình phân tích được tận dụng trong các mô hình bán lẻ để hiểu hành vi của khách hàng thông qua việc xác định sự liên quan giữa các sản phẩm mà khác hàng đó mua. Có thể nói đơn giản hơn, Phân tích giỏ hàng sẽ giúp các nhà bán lẻ có thông tin về các sản phẩm thường xuyên được mua từ đó luôn để chúng có mặt trên kệ hàng.

Để tạo ra các mô hình phân tích này cần một lượng dữ liệu lớn mà được liên tục thu tập và lưu trữ. Với việc khai phá dữ liệu có tần suất, ta có thể tìm ra được sự tương quan giữa các sản phẩm thông qua một tập dữ liệu lớn. Việc này có thể giúp ích đáng kể trong quá trình đưa ra quyết định liên quan tới việc phát triển truyền thông, thiết kế thương hiệu hình ảnh và phân tích hành vi mua sắm của khách hàng.
Ta có thể hình dung khái niêm trên qua ví dụ sau:

Nếu một khách hàng đang mua sữa, sẽ có khả năng là thế nào nếu họ cũng mua bánh mì trong lần mua sắm đó? Thông tin này có thể sẽ khiến doanh thu tăng lên nhờ việc giúp các nhà bán lẻ có chiến lược quảng bá sản phẩm phù hợp dựa trên dự đoán, bán chéo và tạo giỏ hàng tối ưu. Ta có thể coi tất cả các sản phẩm trong cửa hàng là một



Hình 1.1: Ảnh minh họa

tập hợp, với mỗi sản phẩm tương ứng với một biến nhị phân ứng với có mặt hoặc không có mặt của sản phẩm đó. Giờ thì mỗi giỏ hàng được biểu diễn bởi một véc-tơ nhị phân của các giá trị ứng với các biến này. Các véc-tơ nhị phân này được phân tích cho mô hình mua sắm đưa ra các thông tin về sản phẩm được mua cùng nhau. Mô hình này thường được thể hiện dưới dạng *Luật kết hợp (association rules)*.

Jeans	Shirt	Jacket	Shoes
1	1	0	1
0	0	0	1
0	1	0	1
0	1	1	1
1	0	1	0
1	0	1	0
1	0	1	0
1	0	0	0
0	1	0	1

Hình 1.2: Ảnh minh họa

1.2 Cách thức hoạt động

1. Thu thập dữ liệu về giao dịch: Thu thập dữ liệu về giao dịch của khác hàng như sản phẩm đã mua trong mỗi giao dịch, ngày giờ giao dịch và các thông tin liên quan.

2. Tiền xử lý dữ liệu: Làm sạch và tiền xử lí dữ liệu, loại bỏ các thông tin không liên quan, xử lí các giá trị bị thiếu và chuyển dữ liệu thành dạng phù hợp để phân tích.

3. Xác định các tập sản phẩm phổ biến: Sử dụng các thuật toán về *Khai phá luật kết hợp (Association Rule Mining)* như *Apriori* hoặc *FP-Growth* để xác định các tập sản phẩm phổ biến, các tập sản phẩm thường xuyên xuất hiện cùng nhau trong giao dịch.

4. Tính toán chỉ số *support* và *confidence*: Tính 2 chỉ số cho các tập sản phẩm phổ biến nhằm đưa ra sự chắc chắc của việc mua sản phẩm này sau khi đã mua sản phẩm trước đó.

5. Đưa ra luật kết hợp: Đưa ra luật kết hợp dựa trên các tập sản phẩm phổ biến và các chỉ số support và confidence tương ứng. Luật kết hợp chỉ ra sự chắc chắc của việc mua sản phẩm này sau khi đã mua sản phẩm trước đó.

6. Dự đoán kết quả: Dự đoán kết quả của quá trình Phân tích giỏ hàng, xác định sản phẩm nào được mua cùng nhau thường xuyên, tính kết nối giữa các sản phẩm và các thông tin liên quan về hành vi và thị hiếu khách hàng.

7. Ứng dụng: Sử dụng các thông tin lấy được qua quá trình phân tích để đưa ra quyết định cho doanh nghiệp chẳng hạn như gợi ý sản phẩm, tối ưu các sắp đặt hàng và các chiến lược quảng bá sản phẩm.

1.3 Phân loại

1. **Phân tích giỏ hàng mô tả**

   Kiểu phân tích này tìm kiếm các mô hình và sự liên quan trong dữ liệu tồn tại giữa các sản phâm của giỏ hàng. Phân tích giỏ hàng mô tả được sử dụng nhiều nhất để hiểu hành vi của khách hàng. Các nhà bán lẻ có thể bài trí cửa hàng để tăng lợi nhuận nhờ thông tin từ phân tích giỏ hàng mô tả.

2. **Phân tích giỏ hàng dự đoán**

   Phân tích khả năng mua sắm trong tương lai dựa trên mô hình mua sắm trong quá

khứ được gọi là phân tích giỏ hàng dự đoán. Lượng dữ liệu được phân tích sử dụng *thuật toán máy học* tạo nên các dự đoán về các sản phẩm sẽ được mua cùng nhau trong tương lai. Nhà bán lẻ sẽ dựa vào các dữ liệu này để đưa ra quyết định cho các sản phẩm được nhập về, đưa ra mức giá cho các sản phẩm và bài trí chúng trong cửa hàng.

3. **Phân tích giỏ hàng phân loại**

   Kiểu phân tích này sẽ phân tích 2 bộ dữ liệu để xác định sự sai khác giữa chúng. So sánh hành vi của các tập khách hàng khác nhau là mục tiêu của kiểu phân tích này. Các nhà bán lẻ có thể điều hướng hành vi của khách hàng dựa trên việc tùy chỉnh chiến lược quảng bá và doanh thu.

1.4 Ứng dụng thực tiễn

1. **Bán lẻ**

   Phân tích giỏ hàng được thường xuyên sử dụng trong lĩnh vực bán lẻ để kiểm tra mô hình mua sắm của khác hàng để đưa ra quyết định về bày trí sản phẩm, quản lý kho và chiến lược giá.

2. **Thương mại điện tử**

   Phân tích giỏ hàng giúp các nhà bán hàng trực tuyến hiểu về hành vi mua sắm của khác hàng và giúp họ đưa ra các quyết định về gợi ý sản phẩm và chiến dịch quảng bá.

3. **Tài chính**

   Phân tích giỏ hàng có thể được sử dụng để đánh giá hành vi của các nhà đầu tư và dự đoán các giá dản phẩm sẽ được đầu tư trong tương lai.

4. **Viễn thông**

   Nhờ vào Phân tích giỏ hàng, ta có thể đánh giá được đâu là dịch vụ tốt để cung cấp nhằm cung cấp trải nghiệm tốt nhất cho khách hàng.

5. **Sản xuất**

   Nhờ vào Phân tích giỏ hàng, ta có thể đánh giá được đâu là sản phẩm cần sản xuất và nguyên liệu nào cần được thuê trong quá trình sản xuất từ đó có thể tối ưu chi phí và gia tăng hiệu quả.

Chương 2

Thuật toán của phân tích giỏ hàng

Như đã đề cập ở phần **1.2**, Phân tích giỏ hàng dựa trên Luật kết hợp. Trong khai phá Luật kết hợp, ta có thể sử dụng nhiều thuật toán khác nhau như: thuật toán Apriori, thuật toán AIS, thuật toán SETM,... Nhưng phổ biến nhất vẫn là thuật toán Apriori. Vì vậy trong chương này tôi sẽ trình bày tập trung vào Luật kết hợp và thuật toán Apriori.

2.1 Luật kết hợp

2.1.1 Các định nghĩa và khái niệm

Tập mục, Giao dịch và Cở sở dữ liệu giao dịch

1. **Tập mục**

   Gọi $I = \{x_1, x_2, x_3, ..., x_n\}$ là tập n mục. Mỗi tập $X \subseteq I$ được gọi là một tập mục (itemset).

   Nếu X có k mục (tức $|X| = k$) thì X được gọi là tập mục k phần tử.

   Ví dụ:

   - Tập tất cả các mặt hàng thực phẩm trong siêu thị: I = {sữa, trứng, đường, bánh mỳ, mật ong, mứt, bơ, thịt bò, giá, . . . }.
   - Tập tất cả các bộ phim: I = {pearl harbor, fast and furious 7, fifty shades of grey, spectre, . . . }.

2. **Giao dịch**

   Ký hiệu $D = T_1, T_2, ..., T_m$ là cơ sở dữ liệu gồm m giao dịch (transaction). Mỗi giao dịch $T_i \in D$ là một tập mục, tức $T_i \subseteq I$.

Ví dụ:

Tập tất cả các mục $I = A, B, C, D, E$. Cơ sở dữ liệu giao dịch $D = T_1, T_2, T_3, T_4, T_5, T_6$ trong đó:

- $T_1 = \{A, B, D, E\}$
- $T_2 = \{B, C, E\}$
- $T_3 = \{A, B, D, E\}$
- $T_4 = \{A, B, C, E\}$
- $T_5 = \{A, B, C, D, E\}$
- $T_6 = \{B, C, D\}$

Tập mục I là các sản phẩm trong siệu thị, Cơ sở giao dịch là những đơn mua của khách hàng.

- $T_1 = \{$sữa, trứng, đường, bánh mỳ$\}$
- $T_2 = \{$sữa, mật ong, mứt, bơ$\}$
- $T_3 = \{$trứng, mì tôm, thịt bò, giá$\}$

Tập phổ biến, Luật kết hợp

1. **Tập phổ biến (frequent itemset)**

   Cho tập mục $X(\subseteq I)$

   - Độ hỗ trợ của X , kí hiệu là $supp(X, D)$, là số lượng giao dịch trong D chứa tập X:
   $$supp(X, D) = |\{T|(T \subseteq D) \cap (X \subseteq T)\}|$$
   - Độ hỗ trợ tương đối của X, kí hiệu là $rsupp(X, D)$ là số phần trăm các giao dịch trong D chứa X:
   $$rsupp(X, D) = \frac{supp(X, D)}{|D|}$$
   - Tập mục X được gọi là ***tập phổ biến*** trong cơ sở giao dịch D nếu $supp(X, D) \geq minsupp$, với minsupp là một *ngưỡng độ hỗ trợ tối thiểu (minimum support threshold) do người dùng định nghĩa.*
   - F là kí hiệu của tất cả các tập phổ biến
     $F^{(k)}$ là kí hiệu của tập các tập phổ biến có độ dài k

VD: Các tập phổ biến (với *minsupp = 3*) từ cơ sở dữ liệu D (tức số lần xuất hiện của tập trong 6 giao dịch ≥ 3) $D = \{T_1, T_2, T_3, T_4, T_5, T_6\}$ trong đó:

- $T_1 = \{A, B, D, E\}$
- $T_2 = \{B, C, E\}$
- $T_3 = \{A, B, D, E\}$
- $T_4 = \{A, B, C, E\}$
- $T_5 = \{A, B, C, D, E\}$
- $T_6 = \{B, C, D\}$

Ta có tập các tập phổ biến (với *minsupp=1*) là:
F = {A, B, C, D, E, AB, AD, AE, BC, BD, BE, CE, DE, ABD, ABE, ADE, BCE, BDE, ABDE}

- $F_{(1)} = \{A, B, C, D, E\}$
- $F_{(2)} = \{AB, AD, AE, BC, BD, BE, CE, DE\}$
- $F_{(3)} = \{ABE, ABD, ADE, BCE, BDE\}$
- $F_{(4)} = \{ABDE\}$

2. Luật kết hợp

Luật kết hợp là mối quan hệ giữa các tập thuộc tính trong cơ sở dữ liệu. Luật kết hợp là phương tiện hữu ích để khám phá các mối liên kết trong dữ liệu.
Một luật kết hợp là một mệnh đề kéo theo có dạng $X \to Y$, trong đó $X, Y \subseteq I$, thỏa mãn điều kiện $X \cap Y = \emptyset$. Các tập hợp X và Y được gọi là các tập hợp thuộc tính (itemset). Tập X gọi là nguyên nhân, tập Y gọi là hệ quả.

$$X \to Y | X \subseteq I, Y \subseteq I, X \cap Y = \emptyset$$

Ta có thể hiểu đơn giản rằng khi khách hàng mua sắm nhóm sản phẩm X thì sẽ có khả năng dùng sản phẩm Y với 1 xác suất nào đấy.
Có 2 độ đo quan trọng đối với luật kết hợp: Độ hỗ trợ (support) và độ tin cậy (confidence). Bên cạnh đó còn các độ đo như Lift và Conviction sẽ được trình bày ở dưới.
Ta có ví dụ sau: Có 7 giao dịch của 1 cửa hàng quần áo như bảng sau:

Giao dịch (Transaction)	Sản phẩm (Item)
t1	{T-shirt, Trousers, Belt}
t2	{T-shirt, Jacket}
t3	{Jacket, Gloves}
t4	{T-shirt, Trousers, Jacket}
t5	{T-shirt, Trousers, Sneakers, Jacket, Beltt}
t6	{Trousers, Sneakers, Belt}
t7	{Trousers, Belt, Sneakers}

- Đặt các sản phẩm (item) như sau: $I = \{i_1, i_2, ..., i_k\}$.
  Tương ứng:
  $I = \{T - shirt, Trousers, Belt, Jacket, Gloves, Sneakers\}$
- Giao dịch (transaction): $T = \{t_1, t_2, ..., t_n\}$. Ví dụ:
  $t_1 = \{T - shirt, Trousers, Belt\}$
  $\Rightarrow$ Luật kết hợp: $\{T - shirt, Trousers\} \Rightarrow \{Belt\}$

2.1 Độ hỗ trợ (Support)

Độ hỗ trợ là tần suất xuất hiện của nhóm sản phẩm X và Y trong tổng số các giỏ hàng. Hay số lần X và Y cùng 1 giỏ hàng chia tổng số giỏ hàng:

$$supp(X \Rightarrow Y) = \frac{|X \cup Y|}{n}$$

Ví dụ:

$supp(T - shirt \Rightarrow Trousers) = \frac{3}{7} = 43\%$

$supp(Trousers \Rightarrow Belt) = \frac{4}{7} = 57\%$

$supp(\{T - shirt, Trousers\} \Rightarrow \{Belt\}) = \frac{2}{7} = 28\%$

2.2 Độ tin cậy (Confidence)

Độ tin cậy là là tỷ lệ % số lần xuất hiện Y trong những giỏ hàng có nhóm sản phẩm X

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Ví dụ: Trousers xuất hiện $\frac{5}{7}$ giỏ hàng, Trousers và Belt đồng thời xuất hiện $\frac{4}{7}$ giỏ hàng. Khi đó

$conf(Trousers \Rightarrow Belt) = \frac{4/7}{5/7} = 80\%$

Tương tự với nhóm khác:

$$conf(T - shirt \Rightarrow Belt) = \frac{2/7}{4/7} = 50\%$$

$$conf(\{T - shirt, Trousers\} \Rightarrow \{Belt\}) = \frac{2/7}{3/7} = 66\%$$

2.3 Lift

Lift là tỷ lệ giữa số lần xuất hiện đồng thời nhóm X và Y chia cho số lần xuất hiện X và số lần xuất hiện Y.

⇒ **Giá trị của Lift càng lớn thì sự kết hợp giữa X và Y càng chặt**

$$lift(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)supp(Y)}$$

Ví dụ:

$$lift(T - shirt \Rightarrow Trousers) = \frac{3/7}{(4/7)(5/7)} = 1.05$$

$$supp(Trousers \Rightarrow Belt) = \frac{4/7}{(5/7)(4/7)} = 1.4$$

$$supp(\{T - shirt, Trousers\} \Rightarrow \{Belt\}) = \frac{2/7}{(3/7)(4/7)} = 1.17$$

2.4 Conviction

Conviction được định nghĩa như sau:

$$conv(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)}$$

⇒ Chỉ số này được hiểu là khả năng X xảy ra mà không có Y

Tóm lại:

- Support dùng để đánh giá số lần xuất hiện của luật $X \Rightarrow Y$ trong tổng số các giỏ hàng
- Confidence dùng để đánh giá khả năng xuất hiện Y trong những giỏ hàng có nhóm sản phẩm X
- Lift đo lường mức độ chặt chẽ trong sự kết hợp của luật $X \Rightarrow Y$ (càng lớn càng chặt)
- Conviction đo lường khả năng xảy ra X mà không có Y (càng nhỏ càng tốt)

2.2 Thuật toán Apriori

Bài toán đặt ra là:

1) *Tìm tất cả các tập mục phổ biến với minsupp nào đó.*

2) Sử dụng các tập mục phổ biến để sinh ra các luật kết hợp với độ tin cậy minconf nào đó.

2.2.1 Nguyên tắc của Apriori

- Mọi tập con của một tập phổ biến đều phổ biến
- Mọi tập mẹ của một tập không phổ biến đều không phổ biến

2.2.2 Mô tả thuật toán Apriori

Bước 1: Đếm số support cho mỗi tập gồm một phần tử và xem chúng như một *tập phổ biến*. Support của chúng là minsupp.

Bước 2: Với mỗi*tập phổ biến* bổ sung các sản phẩm vào và tạo một *tập phổ biến* mới, tập này được gọi là tập ứng viên (Candidate itemset - C). Đếm số support cho mỗi tập C trên cơ sở dữ liệu, từ đó quyết định tập C nào là *tập phổ biến* thực sự và ta dùng làm hạt giống cho bước kế tiếp.

Bước 3: Lặp lại bước 2 cho đến khi không còn tìm thấy 1 tập *tập phổ biến* nào nữa.

2.2.3 Thuật toán Apriori

Input: Tập các giao dịch D, ngưỡng support tối thiểu minsup
Output: F- tập mục phổ biến trong D
Phương pháp:

```
1: procedure APRIORI(𝔻 = {T1, T2, ..., Tm}, 𝕀 = {x1, x2, ..., xn}, minsup)
2:     Khởi tạo tập các tập phổ biến: 𝔽 ← ∅;
3:     𝔽(1) ← FindFrequent1Itemsets(𝔻, 𝕀, minsup);
4:     for (k = 2; 𝔽(k−1) ≠ ∅; k++) do
5:         ℂ(k) ← AprioriGen(𝔽(k−1));
6:         for (each transaction T ∈ 𝔻) do
7:             ℂT ← SubsetsOfT(ℂ(k), T);
8:             for (each C ∈ ℂT) do
9:                 C.count++;
10:            end for
11:        end for
12:        𝔽(k) ← {C ∈ ℂ(k) | C.count ≥ minsup};
13:    end for
14:    𝔽 ← 𝔽(1) ∪ 𝔽(2) ∪ ··· ∪ 𝔽(k);
15:    return 𝔽;
16: end procedure
```

Hình 2.1: Thuật toán Apriori

2.2.4 Ví dụ minh họa

Minh họa 1: Cho một ví dụ tập các giao dịch từ các hóa đơn mua hàng như sau:

Tid	Các món hàng được mua (Item)
1	{b, m, t, y}
2	{b, m}
3	{p, s, t}
4	{a, b, c, d}
5	{a, b}
6	{e, t, y}
7	{a, b, m}

Cho Min Support = 30%, Min Confidence = 60%

Tính *tập phổ biến 1 sản phẩm*, ta có $F^{(1)}$:

Tập sản phẩm	**Số lần xuất hiện**
{a}	3
{b}	5
{m}	3
{t}	3

Ở bước trên từ $F^{(1)}$ ta có tập C_2 gồm các cặp 2 sản phẩm:

{{a, b}, {a, m}, {a,t}, {b,m}, {b,t}, {m,t}}

Tính *tập phổ biến 2 sản phẩm*, ta có $F^{(2)}$:

Tập sản phẩm	**Số lần xuất hiện**
{a,b}	3
{a,m}	1
{a,t}	0
{b,m}	3
{b,t}	1
{m,t}	1

Chỉ lấy các cặp 2 sản phẩm có Support $\geq$ Minsupp (= 30%) gồm: {a,b} và {b,m}.

Phát sinh luật:

a $\rightarrow$ b có độ Confidence 3/3 = 100%

b $\rightarrow$ a có độ Confidence 3/5 = 60%

b $\rightarrow$ m có độ Confidence 3/5 = 60%

m $\rightarrow$ b có độ Confidence 3/3 = 100%

Ở bước lược bỏ, ta có $F^{(2)}$ = {{a, b}, {b,m}}

Ở bước cuối, từ $F^{(2)}$ ta có tập C_3 gồm các cặp 3 sản phẩm là {$\emptyset$}

Thuật toán kết thúc.

Chương 3

Áp dụng trên bộ dữ liệu

3.1 Tổng quan về bộ dữ liệu

Trong phần này tôi sẽ áp dụng kỹ thuật Phân tích giỏ hàng lên bộ dữ liệu chứa 7501 giao dịch mua đồ của khách hàng (file csv `https://www.kaggle.com/datasets/devchauhan1/market-basket-optimisationcsv`).

7487	red wine	tomato sauce	spaghetti	chocolate	olive oil	french fries	salt	asparagus
7488	soup	milk						
7489	eggs	whole wheat rice						
7490	brownies							
7491	herb & pepper	spaghetti	low fat yogurt					
7492	herb & pepper							
7493	chocolate	escalope						
7494	burgers	salmon	pancakes	french fries	frozen smoothie	fresh bread	mint	
7495	turkey	burgers	dessert wine	shrimp	pasta	tomatoes	pepper	milk
7496	pancakes	light mayo						
7497	butter	light mayo	fresh bread					
7498	burgers	frozen vegetables	eggs	french fries	magazines	green tea		
7499	chicken							
7500	escalope	green tea						
7501	eggs	frozen smoothie	yogurt cake	low fat yogurt				

Hình 3.1: Bộ dữ liệu về Phân tích giỏ hàng

Đây là bộ dữ liệu tốt để thực hiện việc Phân tích giỏ hàng. Từ bộ dữ liệu ta có thể cho ra dữ liệu giúp hiệu quả cho chiến lược kinh doanh.

3.2 Áp dụng

Nhập vào ra các thư viện cần dùng

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import os
for dirname, _, filenames in os.walk('Market_Basket_Optimisation.csv'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

Sử dụng thuật toán Apriori từ thư viện có sẵn

```
    !pip install apyori
```

Đọc file và tiền xử lí dữ liệu

```
    df = pd.read_csv("/Market_Basket_Optimisation.csv")
    df.head()
    df.info()
    transactions = []
    for i in range(0, 7501):transactions.append([str(dataset.values[i,j]) for j in range(0,
    20)])
```

Màn hình sẽ hiện ra các sản phẩm và tần suất xuất hiện

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7500 entries, 0 to 7499
Data columns (total 20 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   shrimp             7500 non-null   object
 1   almonds            5746 non-null   object
 2   avocado            4388 non-null   object
 3   vegetables mix     3344 non-null   object
 4   green grapes       2528 non-null   object
 5   whole weat flour   1863 non-null   object
 6   yams               1368 non-null   object
 7   cottage cheese     980 non-null    object
 8   energy drink       653 non-null    object
 9   tomato juice       394 non-null    object
 10  low fat yogurt     255 non-null    object
 11  green tea          153 non-null    object
 12  honey              86 non-null     object
 13  salad              46 non-null     object
 14  mineral water      24 non-null     object
 15  salmon             7 non-null      object
 16  antioxydant juice  3 non-null      object
 17  frozen smoothie    3 non-null      object
 18  spinach            2 non-null      object
 19  olive oil          0 non-null      float64
dtypes: float64(1), object(19)
memory usage: 1.1+ MB
```

Thực hiện thuật toán Apriori với độ hỗ trợ tối thiểu (minsupp=0.003, minconf=0.2, minlift=3, minlength và max length = 2 thể hiện luật kết hợp giữa 2 sản phẩm).

```
from apyori import apriori
rules = apriori(transactions = transactions, min_support = 0.003, min_confidence = 0.2,
    min_lift = 3, min_length = 2, max_length = 2)
```

Đưa ra kết quả

```
    results = list(rules)
    results
    def inspect(results):
    lhs         = [tuple(result[2][0][0])[0] for result in results]
    rhs         = [tuple(result[2][0][1])[0] for result in results]
    supports    = [result[1] for result in results]
    confidences = [result[2][0][2] for result in results]
    lifts       = [result[2][0][3] for result in results]
    return list(zip(lhs, rhs, supports, confidences, lifts))
resultsinDataFrame = pd.DataFrame(inspect(results), columns = ['Left Hand Side', 'Right Hand
     Side', 'Support', 'Confidence', 'Lift'])
resultsinDataFrame
```

	Left Hand Side	Right Hand Side	Support	Confidence	Lift
0	light cream	chicken	0.004533	0.290598	4.843951
1	mushroom cream sauce	escalope	0.005733	0.300699	3.790833
2	pasta	escalope	0.005866	0.372881	4.700812
3	fromage blanc	honey	0.003333	0.245098	5.164271
4	herb & pepper	ground beef	0.015998	0.323450	3.291994
5	tomato sauce	ground beef	0.005333	0.377358	3.840659
6	light cream	olive oil	0.003200	0.205128	3.114710
7	whole wheat pasta	olive oil	0.007999	0.271493	4.122410
8	pasta	shrimp	0.005066	0.322034	4.506672

Ta có thể chọn ra 10 cặp sản phẩm có chỉ số lift lớn nhất để tối ưu giỏ hàng.

```
    resultsinDataFrame.nlargest(n = 10, columns = 'Lift')
```

	Left Hand Side	Right Hand Side	Support	Confidence	Lift
3	fromage blanc	honey	0.003333	0.245098	5.164271
0	light cream	chicken	0.004533	0.290598	4.843951
2	pasta	escalope	0.005866	0.372881	4.700812
8	pasta	shrimp	0.005066	0.322034	4.506672
7	whole wheat pasta	olive oil	0.007999	0.271493	4.122410
5	tomato sauce	ground beef	0.005333	0.377358	3.840659
1	mushroom cream sauce	escalope	0.005733	0.300699	3.790833
4	herb & pepper	ground beef	0.015998	0.323450	3.291994
6	light cream	olive oil	0.003200	0.205128	3.114710

Chương 4

Đánh giá kết quả

Kết quả đã đưa ra thành công các cặp sản phẩm có hệ số lift cao nhất từ đó có thể áp dụng vào chiến lược kinh doanh. Thuật toán Apriori đơn giản, dễ hiểu dễ cài đặt. Tuy nhiên, Apriori có các nhược điểm:

- Duyệt CSDL nhiều lần.
- Tập ứng viên sinh ra rất lớn $2^n - 1$
- Việc tính độ phổ biến nhiều.

Tài liệu tham khảo

Tiếng Việt

[1] https://viblo.asia/p/khai-pha-mau-pho-bien-va-luat-ket-hop-gGJ59QAa5X2

[2] https://rpubs.com/nguyenngocbinhneu/basket_analysis

Tiếng Anh

[3] https://www.turing.com/kb/market-basket-analysis#terminologies-used-in-market-basket-analysis

[4] https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/#What_Is_Market_Basket_Analysis?

[5] Link bộ dữ liệu
https://www.kaggle.com/datasets/devchauhan1/market-basket-optimisationcsv