CHƯƠNG 2

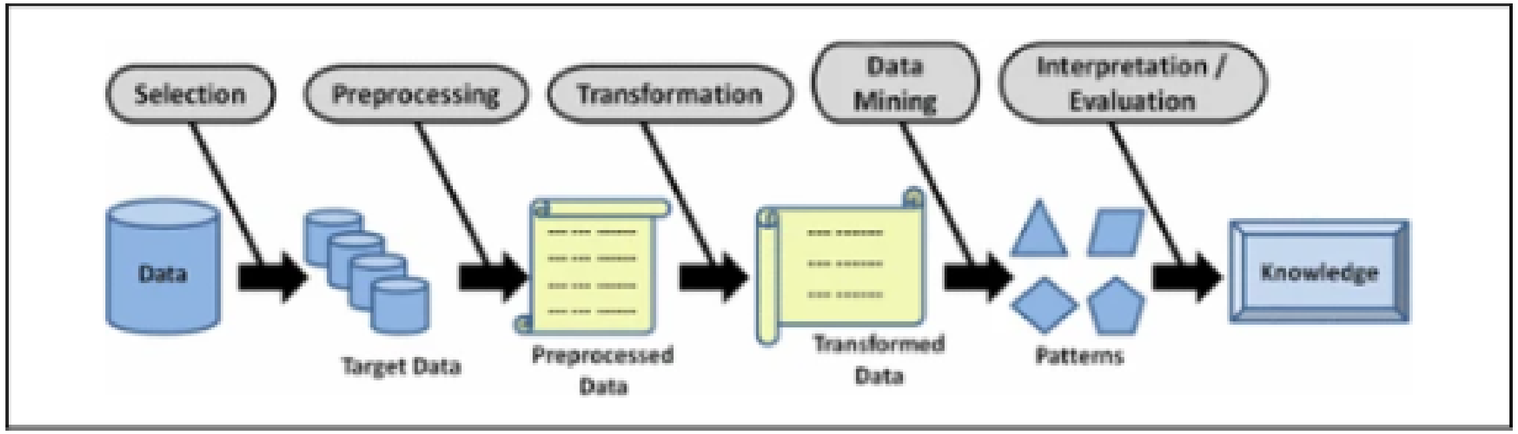
CƠ SỞ LÝ THUYẾT

2.1. MỘT SỐ KHÁI NIỆM CƠ BẢN

2.1.1. Khai phá dữ liệu (Data Mining) là gì?

- Quá trình phân loại các tập dữ liệu lớn
- Xác định xu hướng chung và mối quan hệ giữa các biến
- Là một bước trong quy trình khám phá kiến thức trong cơ sở dữ liệu (KDD)

=> Giải quyết các vấn đề thông qua phân tích dữ liệu.



Quá trình khám phá dữ liệu trong hệ cơ sở dữ liệu (*)

(*) Chee, C., Jaafar, J., Aziz, I. A., Hasan, M. H., & Yeoh, W. (2018). Algorithms for frequent itemset mining: a literature review. Artificial Intelligence Review, 52(4), 2603–2621.

2.1. MỘT SỐ KHÁI NIỆM CƠ BẢN

2.1.2. Tập phổ biến (Frequent Itemset) là gì?

Cho một tập hạng mục X và cơ sở dữ liệu D. Tập X là phổ biến trong D nếu sup(X) >= minsup, với minsup là ngưỡng hỗ trợ tối thiểu do người dùng đặt trước.

Ví dụ: **minsup = 70%.**
Có sup(A) = 66.67% < minsup; sup(C) = 100% > minsup.
Vậy A không phải là tập phổ biến, C là tập phổ biến.

TID	Items
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

2.1. MỘT SỐ KHÁI NIỆM CƠ BẢN

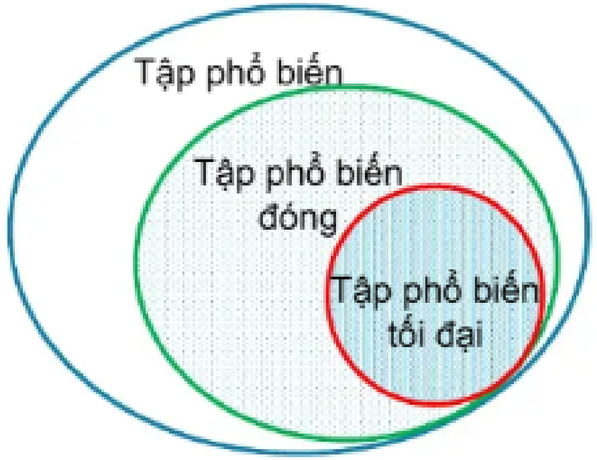
2.1.2. Tập phổ biến (Frequent Itemset) là gì?

Tập phổ biến đóng (closed frequent itemset)

- Một tập mục X được gọi là đóng (closed) nếu không có tập cha nào của X có cùng độ hỗ trợ với nó
- Ký hiệu: FCI

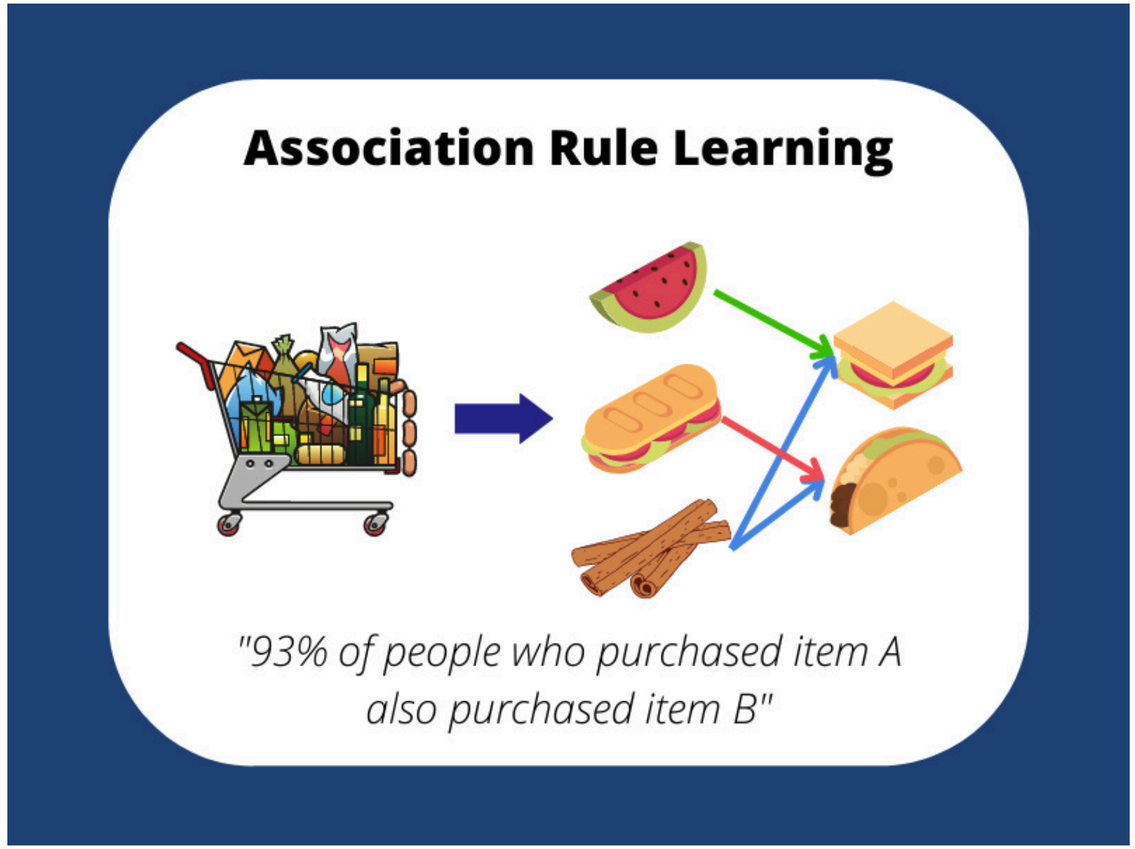
Tập phổ biến lớn nhất (maximally frequent itemset)

- Một tập mục X được gọi là đóng (closed) nếu không có tập cha nào của X có cùng độ hỗ trợ với nó
- Ký hiệu: FCI



2.1. MỘT SỐ KHÁI NIỆM CƠ BẢN

2.1.3. Luật kết hợp (Association Rules) là gì?



Luật kết hợp

- Tìm ra các mối tương quan, các mẫu thường gặp, các liên kết hoặc các cấu trúc ngẫu nhiên giữa tập hợp các mục
- Ứng dụng ở nhiều lĩnh vực
- Sử dụng nhiều trong kinh doanh

(*) Nguồn ảnh: Arora, A. (2022, June 30). What is Association Rule Learning? Machine Learning Interview Questions. Analytics Arora

2.1. MỘT SỐ KHÁI NIỆM CƠ BẢN

2.1.3. Luật kết hợp (Association Rules) là gì?

Độ hỗ trợ

- Độ hỗ trợ của một tập hợp X trong cơ sở dữ liệu D là tỷ số giữa các bản ghi T ⊆ D có chứa tập X và tổng số bản ghi trong D
- Ký hiệu: support(X) hay supp(X)

$$S_0 = \frac{|\{T \subset D : Y \supset X\}|}{|D|}$$

- Độ hỗ trợ của một luật kết hợp XY là tỷ lệ giữa số lượng các bản ghi chứa tập hợp X ∪ Y, so với tổng số các bản ghi trong D
- Ký hiệu: supp(X->Y).

$$\text{Supp}(X \rightarrow Y) = \frac{|\{T \subset D : T \supseteq X \cup Y\}|}{|D|}$$

Độ tin cậy

- Độ tin cậy của một luật kết hợp X->Y là tỷ lệ giữa số lượng các bản ghi trong D chứa X ∪ Y với số bản ghi trong D có chứa tập hợp X.
- Ký hiệu: conf(r)
- Độ hỗ trợ và độ tin cậy có xác suất:

  supp(X->Y) = P(X ∪ Y)

  conf(X->Y) = P(X/Y) = supp(X ∪ Y)/supp(X)

2.2. BÀI TOÁN KHAI THÁC TẬP PHỔ BIẾN

2.2.1. Định nghĩa bài toán & Ứng dụng thực tế

ĐỊNH NGHĨA BÀI TOÁN

KHAI THÁC TẬP PHỔ BIẾN (FREQUENT ITEMSET)

- Vấn đề cốt lõi trong lĩnh vực khai phá dữ liệu
- Xác định các tập mẫu, liên kết, tương quan hoặc cấu trúc nhân quả có tần suất xuất hiện cao trong tập hợp các đối tượng từ các kho thông tin dữ liệu

ỨNG DỤNG

ỨNG DỤNG TRONG CÁC LĨNH VỰC

- Nổi bật nhất là **Basket Data Analysis** – phân tích giỏ hàng
- Các lĩnh vực: tiếp thị chéo, thiết kế danh mục sản phẩm, phân tích thua lỗ, phân cụm, phân loại, hệ thống khuyến nghị, v.v.

(*) Nguồn ảnh: Bot Penguin - How is Market Basket Analysis Conducted & what are its trends

2.2. BÀI TOÁN KHAI THÁC TẬP PHỔ BIẾN

2.2.2.THÁCH THỨC TRONG KHAI PHÁ TẬP PHỔ BIẾN

- **Số lượng luật tạo ra lớn**
- **Hạn chế trong việc phát hiện các mối quan hệ phức tạp**
- **Tốn kém về mặt tính toán**
- **Cần xác định ngưỡng hỗ trợ, độ tin cậy tối thiểu**

CÁC THUẬT TOÁN KHAI PHÁ CÁC TẬP PHỔ BIẾN

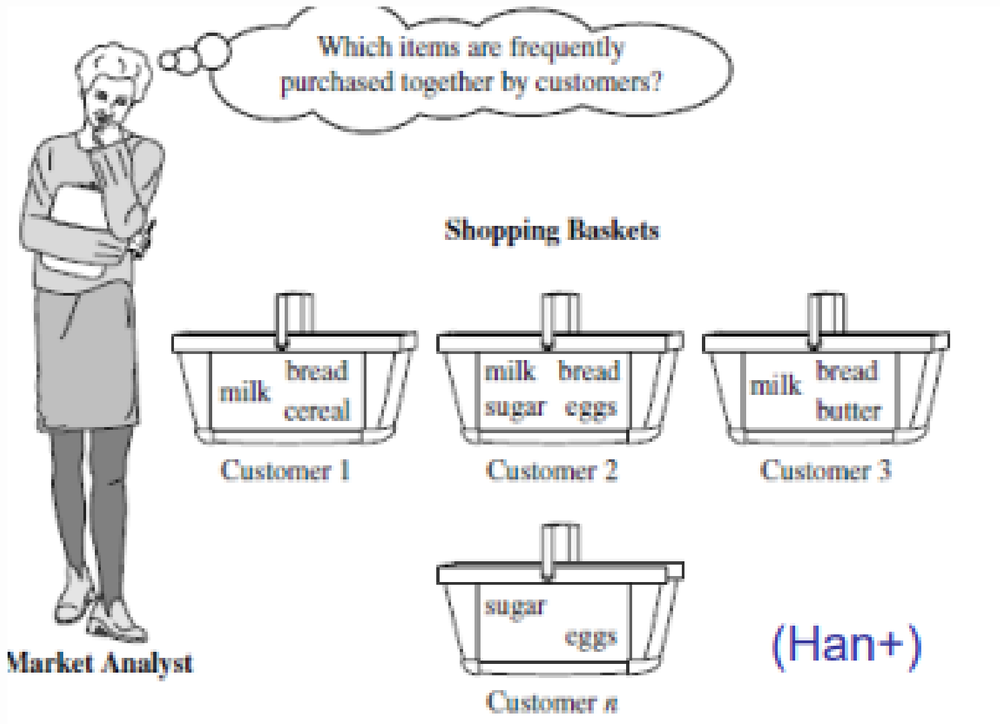
APRIORI

FP-Growth

ECLAT

APRIORI

Apriori là thuật toán khai phá tập mục thường xuyên do R. Agrawal và R. Srikant đề xuất vào năm 1993. Thuật toán Apriori còn là nền tảng cho việc phát triển nhiều thuật toán khai phá tập mục thường xuyên khác về sau.



Thuật toán Apriori (Agrawal & Srikant, 94) tìm tất cả X phổ biến bằng level-wise search: dùng k-itemset để tìm (k+1)-itemsets

Tạo L1 = { frequent 1-itemset }

Mở rộng L1 để tìm L2 = { frequent 2-itemset }

Mở rộng L2 để tìm L3 = { frequent 3-itemset } . . .

Tiếp tục cho đến khi không tìm được k-itemset nào.

Tid	Items
1	bánh, chocolate, sữa
2	bia, chocolate
3	bia, chocolate, hamburger, sữa
4	bánh, hamburger, sữa
5	chocolate, hamburger, sữa

minSup = 60% → minFreq = 0.6 x 5 = 3

1-itemsets

Itemset	freq.
bánh	2
bia	2
chocolate	4
hamburger	3
sữa	4

2-itemsets

Itemset	freq.
chocolate, hamburger	2
chocolate, sữa	3
hamburger, sữa	3

3-itemsets

∅

2 tập phổ biến:

X_1 = {chocolate, sữa}, X_2 = {hamburger, sữa}

Ts. Nguyễn An Tế (2024) Chương 4: Mẫu phổ biến & Luật kết hợp

FP-Growth

Thuật toán FP-growth được xây dựng với 3 kỹ thuật chính:

(1) Nén dữ liệu thích hợp vào một cấu trúc cây gọi là cây FP-tree. Chỉ có các 1-tập mục (1-item) ở trong cây và các nút của cây được sắp xếp để các nút xuất hiện thường xuyên hơn có thể dễ dàng chia sẻ với các nút xuất hiện ít hơn.

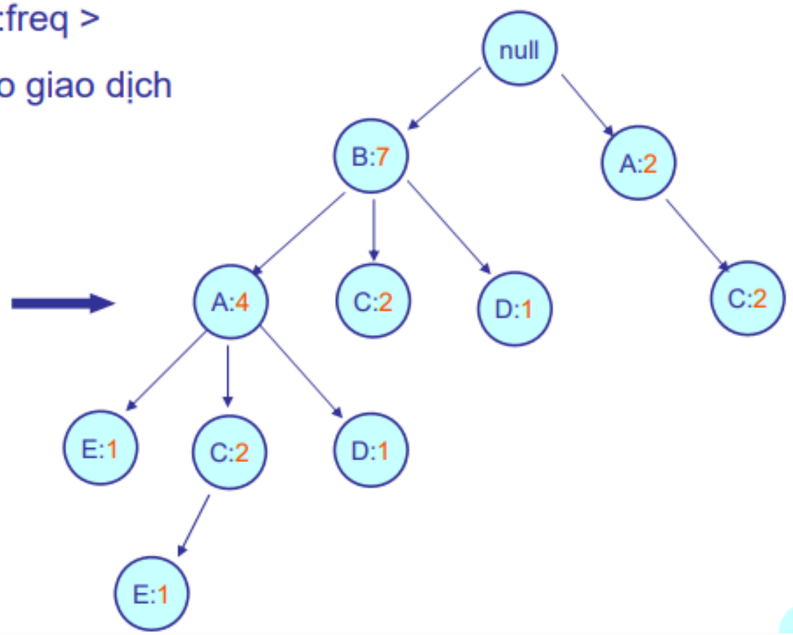
(2) Thực hiện phương pháp khai phá phát triển (growth) từng đoạn dựa trên cây FP-tree gọi là phương pháp FP-growth.

(3) Kỹ thuật tìm kiếm được dùng ở đây là dựa vào sự phân chia, "chia để trị", phân rã nhiệm vụ khai phá thành các nhiệm vụ nhỏ hơn.

Cây mẫu phổ biến (*FP-Tree*): xây dựng từ CSDL D

- nút gốc: null
- nút trong < {item}:freq >
- nhánh: liên kết tạo giao dịch

Tid	Items
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C



ECLAT

Thuật toán ECLAT (Equivalence Class Clustering and Bottom-Up Lattice Traversal) được đề xuất bởi Zaki sử dụng cấu trúc IT - Tree (TIDset Itemset-tree)

Chỉ một lần quét CSDL là một tiếp cận hiện đại, tiết kiệm thời gian xử lý và có thể áp dụng khai thác FI (Frequent Itemset - Tập mục thường xuyên) trên nhiều loại CSDL một cách hiệu quả.

Áp dụng tính chất bao đóng giảm trên IT - Tree sẽ cắt nhánh được tất cả các nhánh chắc chắn không chứa FI và các tập mục trên IT - Tree chính là FI cần khai thác.

THUẬT TOÁN ECLAT

Đầu vào: CSDL và một ngưỡng minSup

Đầu ra: Các tập phổ biến thỏa minSup

```
ECLAT()
    [∅] = {i∈I | σ(i) ≥ minSup}
    ENUMERATE_FREQUENT([∅])
ENUMERATE_FREQUENT([P])
  for all l_i ∈ [P] do
    [P_i] = ∅
    for all l_j ∈ [P] with j > i do
      X = l_i ∪ l_j
      T = t(l_i) ∩ t(l_j)
      if |T| ≥ minSup then
        [P_i] = [P_i] ∪ {X×T}
    ENUMERATE_FREQUENT([P_i])
```

Trong đó $t(X) = \{y \in T \mid X$ xuất hiện trong giao dịch $y\}$ được gọi là Tidset của X.

B1. Quét toàn bộ CSDL để xác định tập giao dịch (TIDset) của các mục. Chọn các mục có support thỏa ngưỡng minSup (1-itemset).

B2. Chèn 1-itemset vào mức 1 của IT - Tree.

B3. Mỗi nút ở mức k - 1 kết hợp với các nút có cùng nút cha với nó tạo ra các nút ở mức k nếu support của các nút này thỏa ngưỡng minSup.

B4. Lặp lại B3 cho đến khi không thể tạo thêm nút mới trên IT - Tree.

B5. Duyệt cây IT - Tree để lấy ra các tập phổ biến (tất cả các nút trên IT - Tree đều là tập phổ biến).

Xét CSDL mẫu $\Rightarrow$ định dạng dữ liệu dọc

Mã giao dịch	Nội dung giao dịch
1	*A, C, T, W*
2	*C, D, W*
3	*A, C, T, W*
4	*A, C, D, W*
5	*A, C, D, T, W*
6	*C, D, T*

$\Rightarrow$

Mã danh mục	Các giao dịch chứa danh mục
A	1, 3, 4, 5
C	1, 2, 3, 4, 5, 6
D	2, 4, 5, 6
T	1, 3, 5, 6
W	1, 2, 3, 4, 5

$$t(A) = 1345;\ t(AD) = t(A) \cap t(D) = 1345 \cap 2456 = 45$$

IT TREE VỚI MIN-SUP = 50%