Khoa Công Nghệ Thông Tin

Trường Đại học Công Nghệ - ĐHQGHN

KHAI PHÁ MẪU PHỔ BIẾN, KẾT HỢP VÀ TƯƠNG QUAN

TS. Trần Trọng Hiếu

Email: hieutt@vnu.edu.vn

HUS - 07/2021

Nội dung

- **Khái niệm cơ bản**
- Các phương pháp khai phá tập mục phổ biến
- Các phương pháp đánh giá mẫu
- Tổng kết

Phân tích mẫu phổ biến

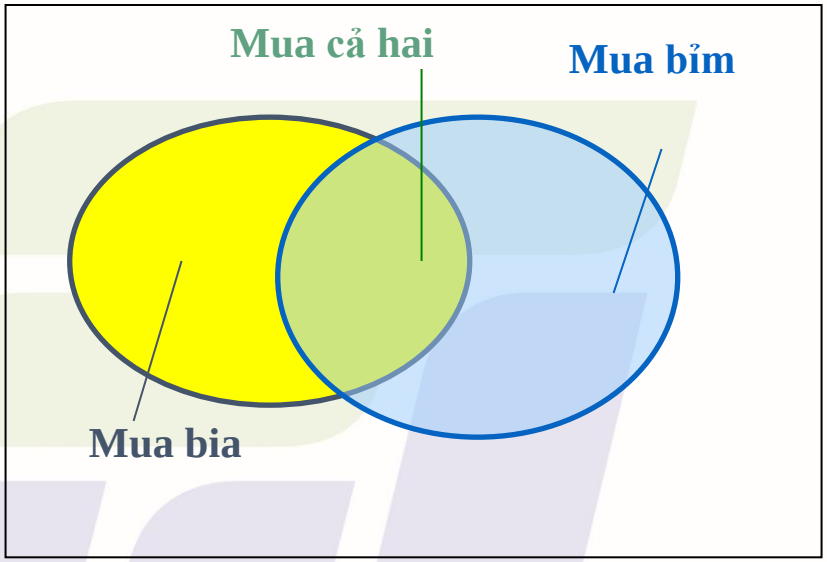
- Mẫu phổ biến: một mẫu (một tập các mục, các dãy con, các cấu trúc con,…) mà xuất hiện một cách liên tục trong một tập dữ liệu.
- Được đề xuất bởi Agrawal, Imielinski, and Swami [AIS93] khi xem xét về các tập mục phổ biến và khai phá luật kết hợp
- Động lực: Tìm kiếm sự đều đặn vốn có trong dữ liệu
  - Những sản phẩm nào thường được mua cùng nhau?
  - Sau khi mua PC sẽ mua gì tiếp theo?
  - Những loại DNA nào nhạy cảm với loại thuốc mới này?
  - Chúng ta có thể tự động phân loại tài liệu web không?
- Các ứng dụng
  - Phân tích dữ liệu giỏ hàng, tiếp thị chéo, thiết kế danh mục, phân tích chiến dịch bán hàng, phân tích nhật ký web và phân tích chuỗi DNA.

Phân tích mẫu phổ biến

- Mẫu phổ biến: Thuộc tính nội tại và quan trọng của bộ dữ liệu
- Nền tảng cho nhiều nhiệm vụ khai thác dữ liệu thiết yếu
  - Phân tích liên kết, tương quan và quan hệ nhân quả
  - Các mẫu tuần tự hay có cấu trúc
  - Phân tích mẫu trong dữ liệu không gian, đa phương tiện, chuỗi thời gian và luồng
  - Phân lớp: phân biệt, phân tích mẫu phổ biến
  - Phân tích cụm: Phân cụm dựa trên mẫu phổ biến
  - Kho dữ liệu
  - ....

Khái niệm cơ bản: Mẫu phổ biến

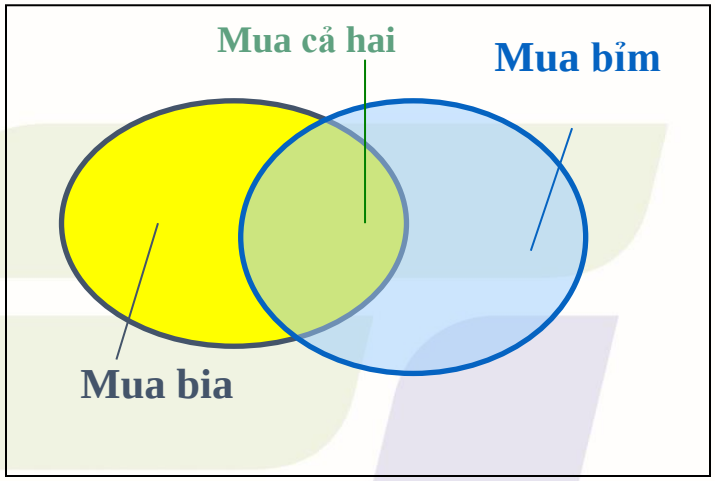
Tid	Các mục hang được mua
10	Bia, Hoa quả, Bỉm
20	Bia, Cà phê, Bỉm
30	Bia, Bỉm, Trứng
40	Hoa quả, Trứng, Sữa
50	Hoa quả, Cà phê, Bỉm, Trứng, Sữa



- ***Tập mục (Tập mục)***: Một tập hợp của các mục
- ***K-tập mục (k-Tập mục)***: Tập mục có đúng ***k*** mục. VD: X = {x_1, …, x_k}
- ***Hỗ trợ (support) tuyệt đối*** hoặc ***số lượng hỗ trợ*** của tập mục X: Số lần xuất hiện của X.
- ***hỗ trợ tương đối*** là xác suất mà một giao dịch chứa X.
- Tập mục X là ***phổ biến*** nếu hỗ trợ của X không dưới ngưỡng *minsup.*

Khái niệm cơ bản: luật kết hợp

Tid	Items bought
10	Bia, Hoa quả, Bỉm
20	Bia, Cà phê, Bỉm
30	Bia, Bỉm, Trứng
40	Hoa quả, Trứng, Sữa
50	Hoa quả, Cà phê, Bỉm, Trứng, Sữa



- Tìm tất cả các luật X → Y với *độ hỗ trợ* và *độ tin cậy* nhỏ nhất
  - ***Độ hỗ trợ (support)*** *s* là xác suất mà một giao dịch có chứa X ∪ Y
  - ***Độ tin cậy (confidence)*** *c là* xác suất mà một giao dịch có X thì cũng có Y

Giả sử minsup = 50%, minconf = 50%

Mẫu phổ biến: Bia:3, Hoa quả:3, Bỉm:4, Trứng:3, {Bia, Bỉm}:3

- Các luật kết hợp:
  - *Bia* → *Bỉm* (60%, 100%)
  - *Bỉm* → *Bia* (60%, 75%)

Mẫu đóng và mẫu cực đại

- Một mẫu dài chứa một tổ hợp của các đoạn mẫu con. VD: mẫu $\{a_1, \ldots, a_{100}\}$ chứa $\binom{100}{1} + \binom{100}{2} + \ldots + \binom{100}{100} = 2^{100} - 1 = 1.27*10^{30}$ đoạn mẫu con!
- Giải pháp: Khai phá theo các mẫu đóng và mẫu cực đại
- Một tập mục X là đóng nếu X là phổ biến và không tồn tại tập mục Y sao cho: Y ↄ X và Y có cùng độ hỗ trợ với X (đề xuất bởi Pasquier, et al. @ ICDT'99)
- Một tập mục X là mẫu cực đại nếu X là phổ biến và không tồn tại tập mục phổ biến Y sao cho: Y ↄ X (đề xuất bởi Bayardo @ SIGMOD'98)
- Mẫu đóng là một dạng nén không mất mát của các mẫu phổ biến
  - Làm giảm số các mẫu và luật

Mẫu đóng và mẫu cực đại

- **Bài tập:** Cho DB = {<a_1, …, a_{100}>, < a_1, …, a_{50}>}
  - Min_sup = 1.
- Tìm tập các tập mục đóng?
  - <a_1, …, a_{100}>: 1
  - < a_1, …, a_{50}>: 2
- Tìm tất cả các tập các mẫu cực đại?
  - <a_1, …, a_{100}>: 1
- Tìm tất cả các mẫu có thể?
  - !!

Độ phức tạp tính toán

- Có bao nhiêu tập mục có khả năng được tạo trong trường hợp xấu nhất?
  - Số lượng các tập mục phổ mục sẽ được tạo tương ứng với ngưỡng minsup
  - Khi minsup thấp, số tập mục phổ biến có thể là một số mũ
  - Trường hợp xấu nhất: M^N trong đó M: số các mục riêng biệt và N: độ dài tối đa của giao dịch
- Trường hợp độ phức tạp xấu nhất so với xác suất dự kiến
  - Ví dụ. Giả sử Walmart có 10^4 loại sản phẩm
  - Cơ hội chọn một sản phẩm 10^{-4}
  - Cơ hội chọn bộ 10 sản phẩm cụ thể: ~ 10^{-40}
  - Cơ hội mà bộ 10 sản phẩm cụ thể này phổ biến xuất hiện 10^3 lần trong 10^9 giao dịch là bao nhiêu?

Nội dung

- Khái niệm cơ bản
- **Các phương pháp khai phá tập mục phổ biến**
- Các phương pháp đánh giá mẫu
- Tổng kết

Các phương pháp

- **Apriori**
- FPGrowth
- ECLAT

Bao đóng đi xuống và các phương pháp

- Thuộc tính ***bao đóng đi xuống*** của các mẫu phổ biến:
  - *Bất kỳ tập con nào của một tập mục phổ biến nào cũng là tập phổ biến.*
  - VD: Nếu {Bia, Bỉm, Hoa quả} là phổ biến, thì {Bia, Bỉm} cũng vậy vì mọi giao dịch có {Bia, Bỉm, Hoa quả} cũng chứa {Bia, Bỉm}
- Các phương pháp khai thác có thể mở rộng: Ba cách tiếp cận chính
  - *Apriori* (Agrawal & Srikant @ VLDB'94)
  - *Tăng trưởng mẫu phổ biến* (FPgrowth — Han, Pei & Yin @ SIGMOD'00)
  - *Phương pháp định dạng dữ liệu dọc* (Charm — Zaki & Hsiao @ SDM'02)

Apriori: Phương pháp sinh-và-kiểm-tra ứng viên

- Nguyên tắc tỉa nhánh Apriori: Nếu có bất kỳ một tập mục nào mà không phổ biến thì mọi tập cha của nó cũng không là phổ biến (Agrawal & Srikant @VLDB'94, Mannila, et al. @ KDD' 94)
- Phương thức:
  - Khởi đầu, quét CSDL một lần đề lấy được các 1-tập mục phổ biến
  - Sinh các (k+1)-tập mục ứng viên từ các k-tập mục
  - Kiểm tra các ứng viên theo CSDL
  - Kết thúc khi không còn ứng viên nào được sinh ra

Ví dụ

$Sup_{min} = 2$

CSDL

Tid	Các mục
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

Lần quét 1

C_1

Tập mục	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Tập mục	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Tập mục
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

Lần quét 2

C_2

Tập mục	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Tập mục	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Tập mục	
{A,B, C}	1
{B, C, E}	2

Lần quét 3

L_3

Tập mục	sup
{B, C, E}	2

Thuật toán Apriori

C_k: Các k-Tập mục ứng viên

L_k : k-tập mục phổ biến

L_1 = {các mục phổ biến};

for (k = 1; L_k !=∅; k++) **do begin**

C_{k+1} = Các ứng viên được sinh ra từ L_k;

for each t in CSDL do

Tăng số đếm của tất cả các ứng viên trong C_{k+1} mà t chứa;

L_{k+1} = Các ứng viên trong C_{k+1} có *hỗ trợ ≥ min_support*

end

return $\cup_k L_k$;

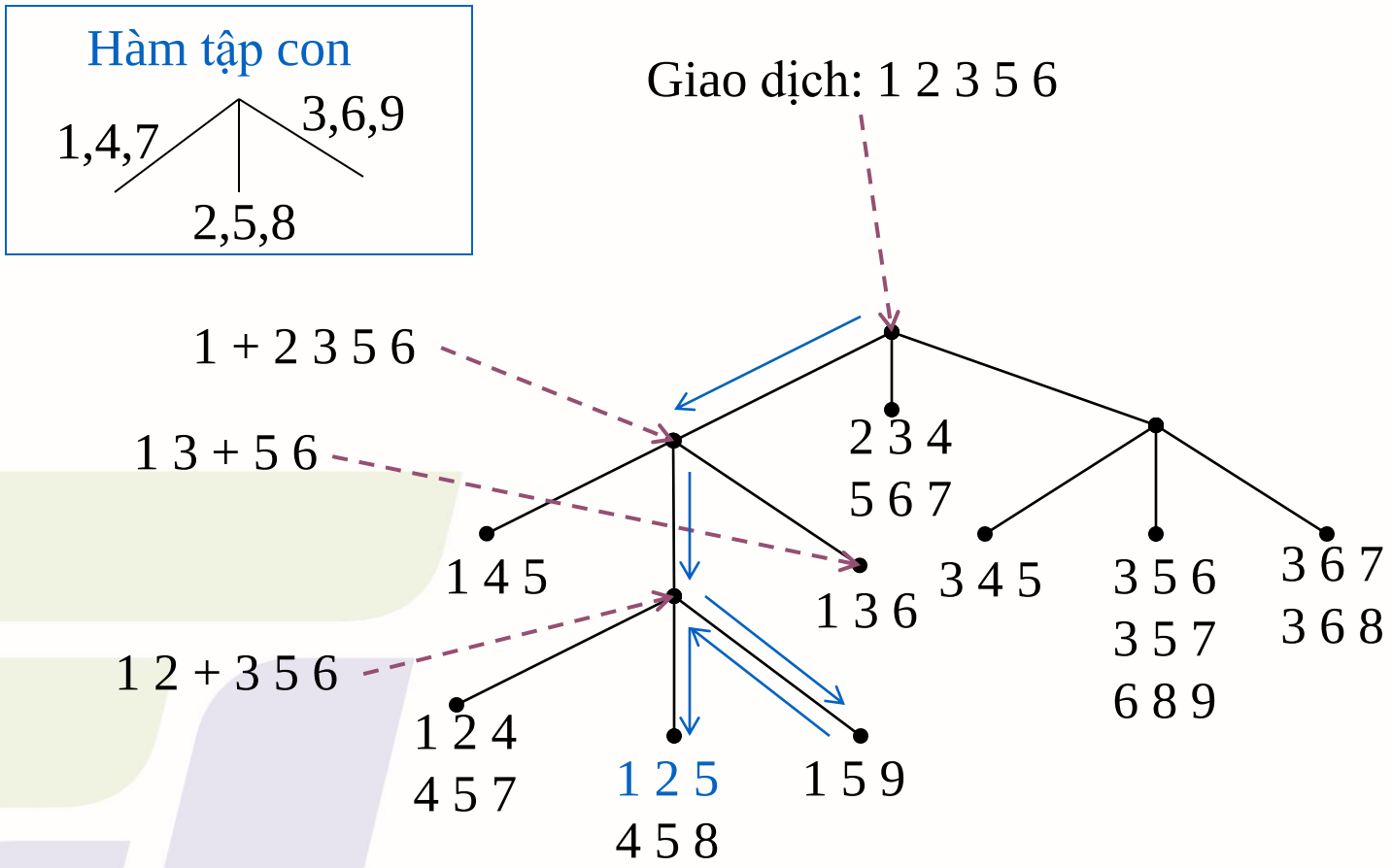
Cách thức thi của Apriori

- Cách sinh các ứng viên:
  - Bước 1: Ghép các cặp trong L_k để sinh ra các (k+1)-tập mục
  - Bước 2: Tỉa nhánh
- Ví dụ:
  - L_3={*abc, abd, acd, ace, bcd*}
  - Ghép cặp: L_3*L_3
    - *abcd* từ *abc* và *abd*
    - *acde* từ *acd* và *ace*
  - Tỉa nhánh:
    - *acde* bị loại bỏ vì *ade* không có trong L_3
  - C_4 = {*abcd*}

Đếm số hỗ trợ của các ứng viên

- *Những vấn đề gặp phải khi đếm số hỗ trợ:*
  - *Tổng số ứng viên có thể rất lớn*
  - *Một giao dịch có thể chứa nhiều ứng cử viên*
- *Phương pháp:*
  - *Các tập mục ứng cử viên được lưu trữ trong một cây băm*
  - *Nút lá của cây băm chứa một danh sách các tập mục và số đếm của chúng*
  - *Nút bên trong chứa một bảng băm*
  - *Hàm tập con: tìm tất cả các ứng cử viên có trong một giao dịch*

Đếm số hỗ trợ bằng cây băm



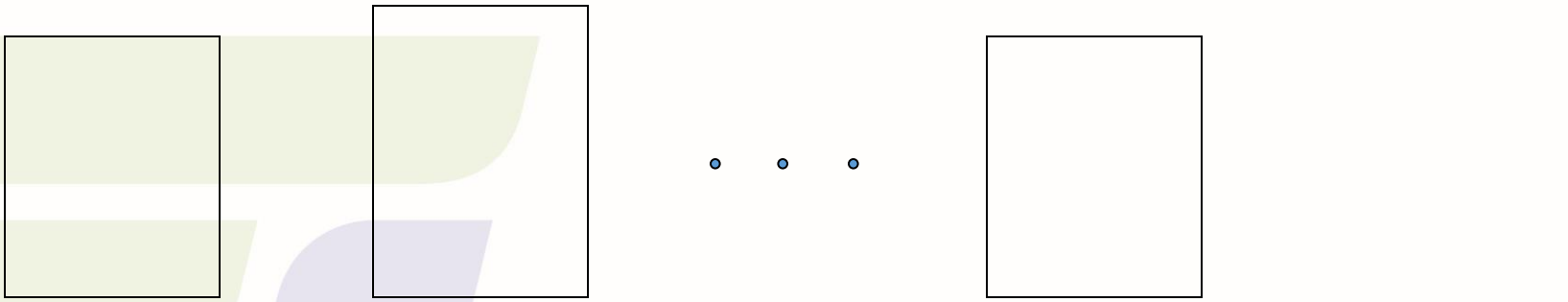
Các cải tiến của phương pháp Apriori

- Những thách thức lớn về tính toán
  - Nhiều lần quét cơ sở dữ liệu giao dịch
  - Số lượng lớn các ứng cử viên
  - Khối lượng công việc nhàm chán về đếm số hỗ trợ cho các ứng viên
- Cải tiến Apriori: ý tưởng chung
  - Giảm số lượt quét cơ sở dữ liệu giao dịch
  - Thu hẹp số lượng ứng viên
  - Tạo điều kiện hỗ trợ đếm số lượng ứng viên

Phân hoạch: Quét CSDL hai lần

- Bất kỳ tập mục nào có khả năng là phổ biến trong CSDL thì cũng phải phổ biến trong ít nhất một trong các phân vùng của CSDL.
  - Quét 1: phân hoạch CSDL và tìm các mẫu phổ biến cục bộ
  - Quét 2: hợp nhất các mẫu phổ biến toàn cục.

(A. Savasere, E. Omiecinski and S. Navathe, VLDB'95)



DHP: Giảm số các ứng viên

- Một *k*-tập mục mà có số đếm theo thùng băm nằm dưới ngưỡng sẽ không thể là phổ biến
  - Các ứng viên: a, b, c, d, e
  - Mục băm
    - {ab, ad, ae}
    - {bd, be, de}
    - …
  - 1-Tập mục phổ biến: a, b, d, e
  - ab không là 2-tập mục ứng viên nếu tổng số đếm của {ab, ad, ae} nằm dưới ngưỡng hỗ trợ

count	Tập mucs
35	{ab, ad, ae}
88	{bd, be, de}
. . .	. . .
102	{yz, qs, wt}

Hash Table

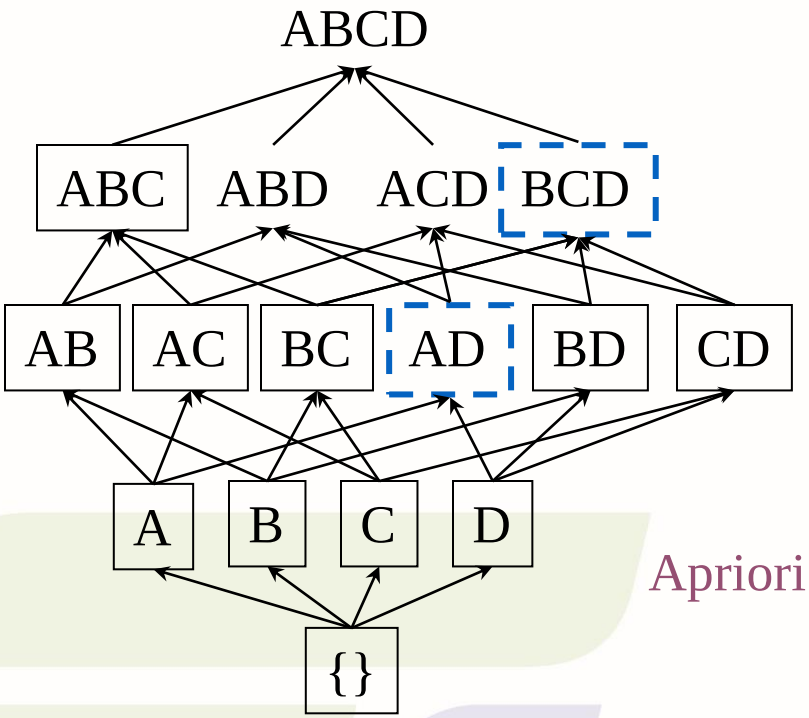
(J. Park, M. Chen, and P. Yu. An effective hash-based algorithm for mining association rules. SIGMOD'95)

Lấy mẫu cho các mẫu phổ biến

- Chọn một mẫu (sample) từ cơ sở dữ liệu gốc, khai phá các mẫu phổ biến trong mẫu bằng cách sử dụng Apriori
  - Quét cơ sở dữ liệu một lần để xác minh tập mục phổ biến được tìm thấy trong mẫu, chỉ các đường bao của bao đóng của các mẫu phổ biến được kiểm tra.
    - Ví dụ: kiểm tra abcd thay vì ab, ac,…, v.v.
  - Quét lại cơ sở dữ liệu để tìm các mẫu phổ biến bị bỏ sót

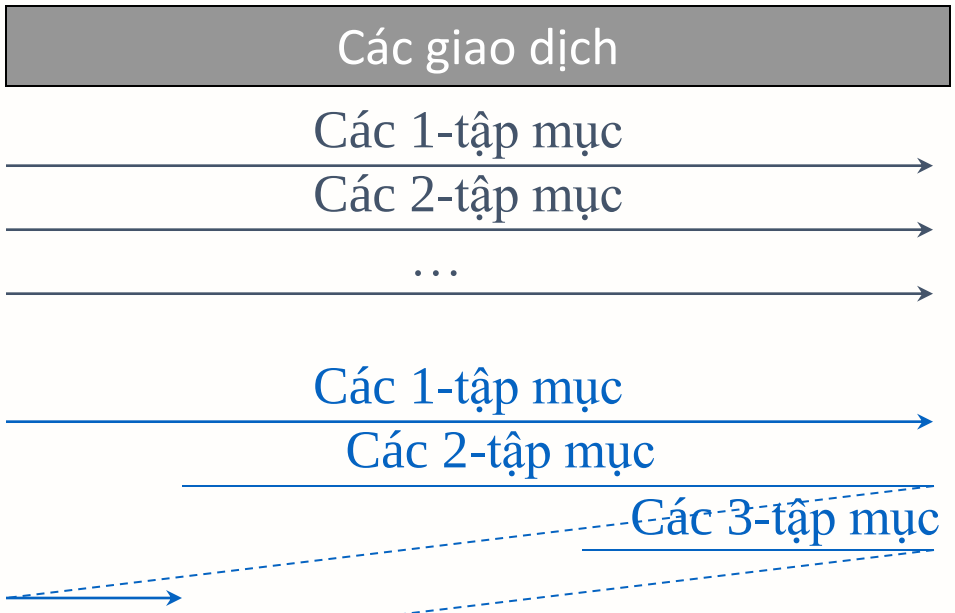
(H. Toivonen. Sampling large databases for association rules. In VLDB'96)

DIC: Giảm số lần quét CSDL



Tập mục

- Khi cả A và D được xác định là phổ biến, việc đếm AD sẽ bắt đầu
- Khi tất cả các tập con có độ dài 2 của BCD được xác định là phổ biến, việc đếm BCD bắt đầu



Apriori

DIC

S. Brin R. Motwani, J. Ullman, and S. Tsur. Dynamic Tập mục counting and implication rules for market basket data. *SIGMOD'97*

Các phương pháp

- Apriori
- **FPGrowth**
- ECLAT

Phương pháp mở rộng mẫu

- Những điểm nghẽn của cách tiếp cận Apriori
  - Tìm kiếm theo chiều rộng (tức là theo cấp độ)
  - Tạo ứng viên và kiểm tra
    - Thường tạo ra một số lượng lớn các ứng cử viên
- Phương pháp tiếp cận FPGrowth (J. Han, J. Pei và Y. Yin, SIGMOD ’00)
  - Tìm kiếm theo chiều sâu
  - Tránh tạo ứng viên

Phương pháp mở rộng mẫu

- Triết lý chính: Phát triển các mẫu dài từ các mẫu ngắn chỉ sử dụng các mục phổ biến địa phương
  - “Abc” là một mẫu phổ biến
  - Lấy tất cả các giao dịch có “abc”, tức là chiếu CSDL trên abc: CSDL | abc
  - “D” là một mục phổ biến cục bộ trong CSDL | abc → abcd là một biến phổ mẫu

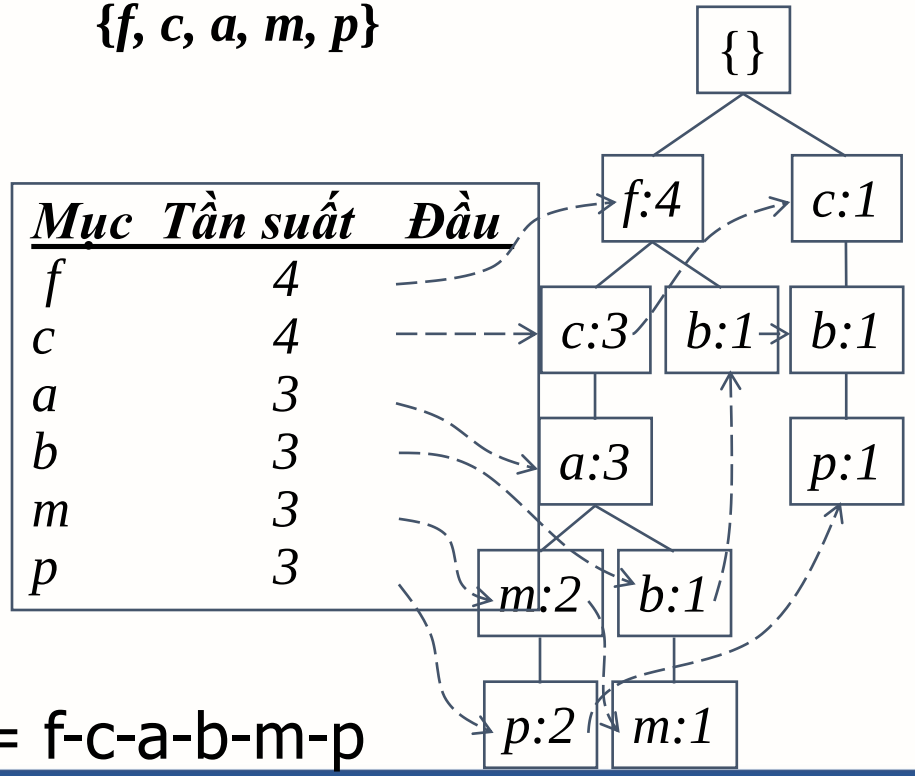
Xây dựng cây FP từ CSDL

TID	*Mục hang đã mua*	*Các mục phổ biến*
100	{*f, a, c, d, g, i, m, p*}	{*f, c, a, m, p*}
200	{*a, b, c, f, l, m, o*}	{*f, c, a, b, m*}
300	{*b, f, h, j, o, w*}	{*f, b*}
400	{*b, c, k, s, p*}	{*c, b, p*}
500	{*a, f, c, e, l, p, m, n*}	{*f, c, a, m, p*}

1. Quét CSDL một lần, tìm 1-tập mục phổ biến.
2. Sắp xếp các mục phổ biến theo thứ tự giảm dần của tần suất, danh sách f-list,
3. Quét lại CSDL, xây dựng Cây FP.

min_support = 3

Mục	*Tần suất*	*Đầu*
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	



f-list = f-c-a-b-m-p

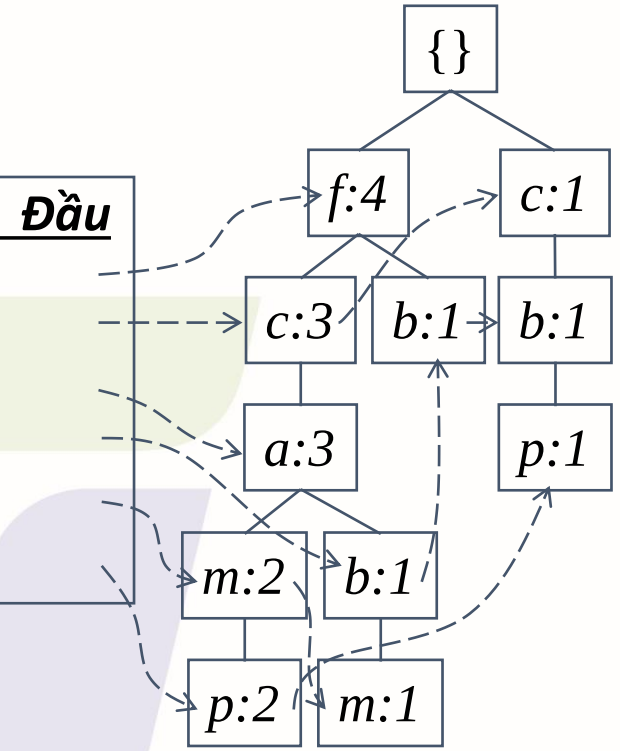
Phân hoạch các mẫu và CSDL

- Các mẫu phổ biến có thể được phân hoạch thành các tập con theo f-list
  - F-list = f-c-a-b-m-p
  - Các mẫu có chứa p
  - Các mẫu có m nhưng không có p
  - …
  - Các mẫu có c nhưng không có a hay b, m, p
  - Mẫu f
- Tính đầy đủ và không dư thừa

Tìm các mẫu có p từ CSDL điều kiện p

- Bắt đầu từ bảng các mục phổ biến trong cây FP
- Duyệt cây FP bằng cách theo liên kết của từng mục phổ biến p
- Tích lũy tất cả các *đường dẫn tiền tố đã được biến đổi* của mục p để tạo thành cơ sở mẫu có điều kiện của p.

Mục	Tần suất	Đầu
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	



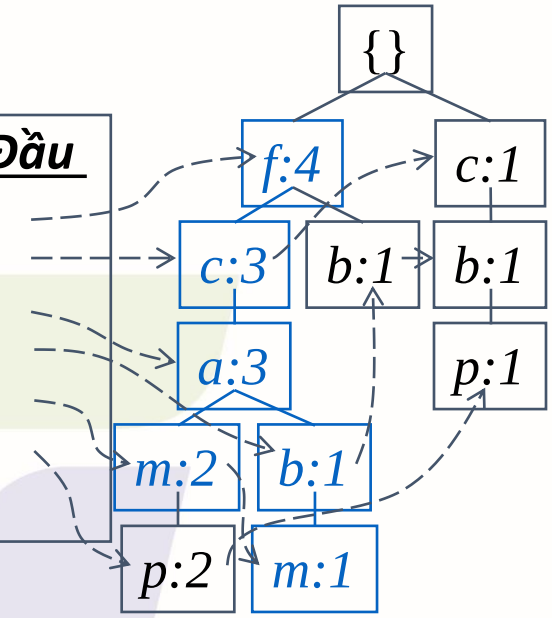
Các cơ sở mẫu có điều kiện

Mục	cơ sở mẫu
c	f:3
a	fc:3
b	fca:1, f:1, c:1
m	fca:2, fcab:1
p	fcam:2, cb:1

Xây dựng cây FP điều kiện từ các CSDL điều kiện

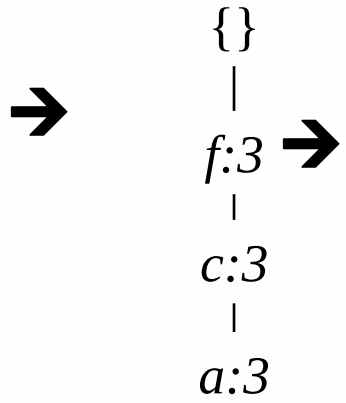
- Với mỗi cơ sở mẫu
  - Tích lũy số đếm cho mỗi mục trong cơ sở
  - Xây dựng cây FP cho các mục phổ biến của cơ sở mẫu

Mục	Tần suất	Đầu
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	



Cơ sở mẫu theo điều kiện m

fca:2, fcab:1



Tất cả các mẫu phổ biến liên quan tới *m*

m,

fm, cm, am,

fcm, fam, cam,

fcam

Cây FP theo điều kiện m

Đệ quy: Khai thác từng cây FP có điều kiện

{}
\
f:3
|
c:3
|
a:3

Cây FP có điều kiện m

Cơ sở mẫu có điều kiện của "am": (fc:3)

{}
|
f:3
|
c:3

Cây FP có điều kiện am

Cơ sở mẫu có điều kiện của "cm": (f:3)

{}
|
f:3

Cây FP có điều kiện cm

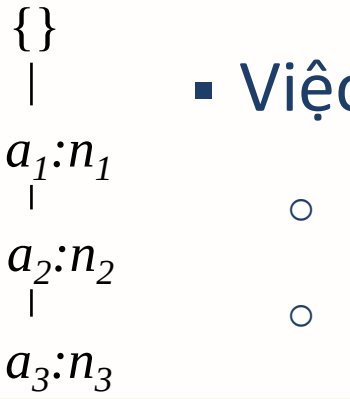
Cơ sở mẫu có điều kiện của "cam": (f:3)

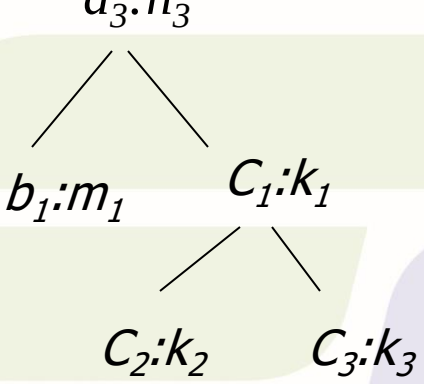
{}
|
f:3

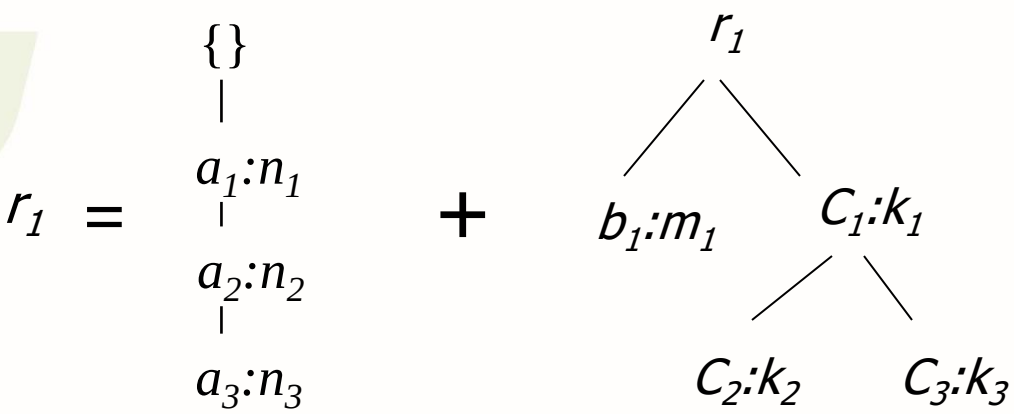
Cây FP có điều kiện cam

Đường dẫn tiền tố đơn trong cây FP

- Giả sử cây FP (có điều kiện) T có một đường dẫn tiền tố P được chia sẻ
- Việc khai phá có thể được chia thành hai phần
  - Giảm đường dẫn tiền tố đơn thành một nút
  - Kết hợp các kết quả khai thác của hai phần







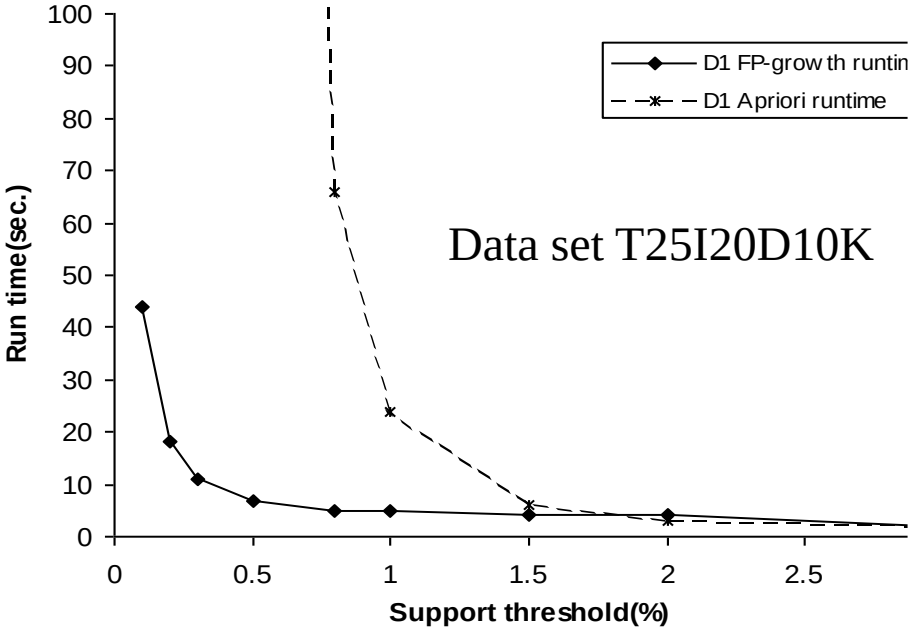
Những lợi ích của cấu trúc cây FP

- Tính đầy đủ
  - Lưu giữ thông tin đầy đủ cho khai phá mẫu phổ biến
  - Không bao giờ phá vỡ một mẫu dài của bất kỳ giao dịch nào
- Tính nhỏ gọn
  - Giảm thông tin không liên quan — các mục không phổ biến không còn nữa
  - Các mục theo thứ tự giảm dần của tần suất : càng xuất hiện phổ biến thì càng có nhiều khả năng được chia sẻ
  - Không bao giờ lớn hơn cơ sở dữ liệu gốc (không tính liên kết nút và trường đếm)

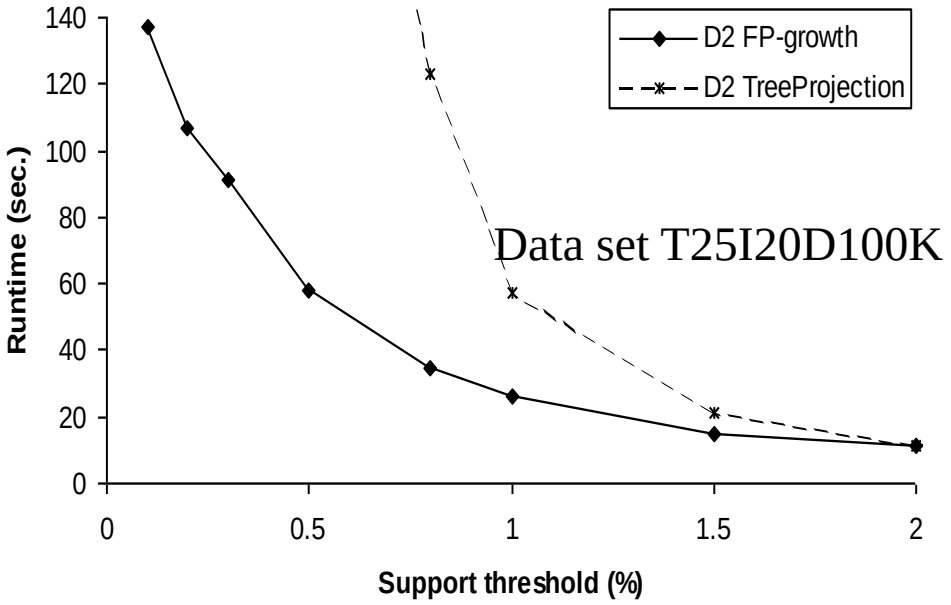
Phương pháp khai phá mẫu mở rộng

- Ý tưởng: Mẫu phổ biến mở rộng
  - Đệ quy tăng các mẫu phổ biến bởi các mẫu và phân hoạch CSDL
- Phương thức:
  - Đối với mỗi mục phổ biến, xây dựng cơ sở mẫu có điều kiện và sau đó là cây FP có điều kiện
  - Lặp lại quy trình trên mỗi cây FP có điều kiện mới được tạo
  - Cho đến khi cây FP kết quả trống hoặc nó chỉ chứa một đường dẫn — đường dẫn duy nhất sẽ tạo ra tất cả các kết hợp của các đường dẫn con của nó, mỗi đường dẫn trong số đó là một mẫu phổ biến

Hiệu suất của FPGrowth trong các tập dữ liệu lớn



FP-Growth vs. Apriori



FP-Growth vs. Tree-Projection

Ưu điểm của tiếp cận mẫu mở rộng

- Chia để trị:
  - Phân rã cả nhiệm vụ khai phá và DB theo mẫu phổ biến thu được
  - Dẫn đến việc tìm kiếm tập trung vào các cơ sở dữ liệu nhỏ hơn
- Những yếu tố khác
  - Không sinh ứng viên, không kiểm tra ứng viên
  - Cơ sở dữ liệu được nén: cấu trúc cây FP
  - Không quét lặp lại toàn bộ cơ sở dữ liệu
  - Các hoạt động cơ bản: đếm các mục phổ biến cục bộ và xây dựng cây FP con, không tìm kiếm và đối sánh mẫu
- Một triển khai mã nguồn mở tốt và cải tiến từ FPGrowth
  - FPGrowth + (Grahne và J. Zhu, FIMI'03)

Các cải tiến

- AFOPT (Liu và cộng sự @ KDD'03)
  - Phương pháp "đẩy sang phải" để khai phá cây mẫu phổ biến cô đọng
- Carpenter (Pan, et al. @ KDD'03)
  - Khai thác tập dữ liệu với ít hàng nhưng rất nhiều cột
  - Xây dựng cây liêt kê theo hàng để khai phá hiệu quả
- FPgrowth + (Grahne và Zhu, FIMI'03)
  - Sử dụng hiệu quả Prefix-Trees trong khai thác các tập mục phổ biến
- TD-Close (Liu, et al, SDM'06)

Extension of Pattern Growth Mining Methodology

- Khai thác tập mục phổ biến đóng và các mẫu cực đại
  - CLOSET (DMKD'00), FPclose và FPMax (Grahne & Zhu, Fimi'03)
- Khai thác các mẫu tuần tự
  - PrefixSpan (ICDE'01), CloSpan (SDM'03), BIDE (ICDE'04)
- Khai phá các mẫu biểu đồ
  - gSpan (ICDM'02), CloseGraph (KDD'03)
- Khai phá dựa trên ràng buộc của các mẫu phổ biến
  - Các ràng buộc có thể chuyển đổi (ICDE'01), gPrune (PAKDD'03)
- Tính toán dữ liệu khối với các độ đo phức tạp
  - H-tree, H-cubing và Star-cubing (SIGMOD'01, VLDB'03)
- Phân cụm dựa trên mẫu mở rộng
  - MaPle (Pei, et al., ICDM'03)
- Phân lớp dựa trên mẫu mở rộng
  - Khai phá các mẫu bổ biến và phân biệt (Cheng, et al, ICDE'07)

Các phương pháp

- Apriori
- FPGrowth
- **ECLAT**

ECLAT: Khai phá bởi khám phá định dạng dữ liệu dọc

- Định dạng dọc: t(AB) = {T_{11}, T_{25}, …}
  - tid-list: danh sách các id của các giao dịch có chứa một tập mục
- Dẫn ra các mẫu phổ biến dựa vào các phép giao dọc
  - t(X) = t(Y): X và Y luôn luôn xảy ra cùng nhau
  - t(X) $\subset$ t(Y): Giao dịch có X cũng luôn có Y
- Dùng phép diffset để tăng tốc độ khai phá
  - Chỉ theo dõi sự khác biệt của các tid
  - t(X) = {T_1, T_2, T_3}, t(XY) = {T_1, T_3}
  - Diffset (XY, X) = {T_2}
- Eclat (Zaki et al. @KDD'97)
- Mining Closed patterns using vertical format: CHARM (Zaki & Hsiao@SDM'02)

Nội dung

- Khái niệm cơ bản
- Các phương pháp khai phá tập mục phổ biến
- **Các phương pháp đánh giá mẫu**
- Tổng kết

Đo lường sự thú vị: Tương quan (Lift)

- *chơi bóng rổ* $\Rightarrow$ *ăn ngũ cốc* [40%, 66.7%] gây hiểu nhầm
  - Tỷ lệ sinh viên ăn ngũ cốc là 75%> 66,7%.
- *chơi bóng rổ* $\Rightarrow$ *không ăn ngũ cốc* [20%, 33.3%] chính xác hơn, mặc dù với sự hỗ trợ và độ tin cậy thấp hơn
- Đo lường các sự kiện phụ thuộc / tương quan : lift

$$lift = \frac{P(A \cup B)}{P(A)P(B)}$$

$$lift(B,C) = \frac{2000/5000}{3000/5000 * 3750/5000} = 0.89$$

$$lift(B,\neg C) = \frac{1000/5000}{3000/5000 * 1250/5000} = 1.33$$

	Bóng rổ	Không bóng rổ	Tổng (hàng)
Ngũ cốc	2000	1750	3750
Không ngũ cốc	1000	250	1250
Tổng (cột)	3000	2000	5000

lift và χ^2 có là các độ đo tương quan tốt?

- *"Mua Hoa quả $\Rightarrow$ Mua Sữa* [1%, 80%]" là nhầm lẫn nếu 85% khách hàng mua Sữa
- Hỗ trợ và độ tin cậy không tốt để chỉ ra các mối tương quan
- Hơn 20 độ đo độ thú vị đã được đề xuất (see Tan, Kumar, Sritastava @KDD'02)
- Những độ đo nào là tốt?

symbol	measure	range	formula
ϕ	ϕ-coefficient	-1 . . . 1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Q	Yule's Q	-1 . . . 1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Y	Yule's Y	-1 . . . 1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
k	Cohen's	-1 . . . 1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
PS	Piatetsky-Shapiro's	-0.25 . . . 0.25	$P(A,B)-P(A)P(B)$
F	Certainty factor	-1 . . . 1	$\max(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)})$
AV	added value	-0.5 . . . 1	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
K	Klosgen's Q	-0.33 . . . 0.38	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$
g	Goodman-kruskal's	0 . . . 1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-max_k P(B_k)}$
M	Mutual Information	0 . . . 1	$\frac{\Sigma_i\Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\Sigma_i P(A_i)\log P(A_i)\log P(A_i), -\Sigma_i P(B_i)\log P(B_i)\log P(B_i))}$
J	J-Measure	0 . . . 1	$\max(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}))$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})$
G	Gini index	0 . . . 1	$\max(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A}[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B}[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]-P(A)^2-P(\overline{A})^2)$
s	support	0 . . . 1	$P(A,B)$
c	confidence	0 . . . 1	$max(P(B\|A), P(A\|B))$
L	Laplace	0 . . . 1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
IS	Cosine	0 . . . 1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
γ	coherence(Jaccard)	0 . . . 1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
α	all_confidence	0 . . . 1	$\frac{P(A,B)}{\max(P(A),P(B))}$
o	odds ratio	0 . . . ∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
V	Conviction	0.5 . . . ∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
λ	lift	0 . . . ∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective strength	0 . . . ∞	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
χ^2	χ^2	0 . . . ∞	$\Sigma_i \frac{(P(A_i)-E_i)^2}{E_i}$

Các độ đo bất biến null

Table 6: Properties of interestingness measures. Note that none of the measures satisfies all the properties.

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
ϕ	ϕ-coefficient	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Goodman-Kruskal's	$0 \cdots 1$	Yes	No	No	Yes	No	No*	Yes	No
α	odds ratio	$0 \cdots 1 \cdots \infty$	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	$0 \cdots 1$	Yes	Yes	Yes	No**	No	No*	Yes	No
J	J-Measure	$0 \cdots 1$	Yes	No	No	No**	No	No	No	No
G	Gini index	$0 \cdots 1$	Yes	No	No	No**	No	No*	Yes	No
s	Support	$0 \cdots 1$	No	Yes	No	Yes	No	No	No	No
c	Confidence	$0 \cdots 1$	No	Yes	No	No**	No	No	No	Yes
L	Laplace	$0 \cdots 1$	No	Yes	No	No**	No	No	No	No
V	Conviction	$0.5 \cdots 1 \cdots \infty$	No	Yes	No	No**	No	No	Yes	No
I	Interest	$0 \cdots 1 \cdots \infty$	Yes*	Yes	Yes	Yes	No	No	No	No
IS	Cosine	$0 \cdots \sqrt{P(A,B)} \cdots 1$	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	$-0.25 \cdots 0 \cdots 0.25$	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	$-1 \cdots 0 \cdots 1$	Yes	Yes	Yes	No**	No	No	Yes	No
AV	Added value	$-0.5 \cdots 0 \cdots 1$	Yes	Yes	Yes	No**	No	No	No	No
S	Collective strength	$0 \cdots 1 \cdots \infty$	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	$0 \cdots 1$	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$(\frac{2}{\sqrt{3}} - 1)^{1/2}[2 - \sqrt{3} - \frac{1}{\sqrt{3}}] \cdots 0 \cdots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No**	No	No	No	No

where: P1: $O(\mathbf{M}) = 0$ if $det(\mathbf{M}) = 0$, *i.e.*, whenever A and B are statistically independent.
P2: $O(\mathbf{M_2}) > O(\mathbf{M_1})$ if $\mathbf{M_2} = \mathbf{M_1} + [k \ -k; \ -k \ k]$.
P3: $O(\mathbf{M_2}) < O(\mathbf{M_1})$ if $\mathbf{M_2} = \mathbf{M_1} + [0 \ k; \ 0 \ -k]$ or $\mathbf{M_2} = \mathbf{M_1} + [0 \ 0; \ k \ -k]$.
O1: Property 1: Symmetry under variable permutation.
O2: Property 2: Row and Column scaling invariance.
O3: Property 3: Antisymmetry under row or column permutation.
O3': Property 4: Inversion invariance.
O4: Property 5: Null invariance.
Yes*: Yes if measure is normalized.
No*: Symmetry under row or column permutation.
No**: No unless the measure is symmetrized by taking $\max(M(A,B), M(B,A))$.

So sánh các độ đo độ thú vị

- Bất biến null là quan trọng để phân tích tương quan
- Lift và χ^2 không là bất biến null
- 5 độ đo bất biến null

	Sữa	Không Sữa	Tổng (hàng)
Cà phê	m, c	~m, c	c
Không Cà phê	m, ~c	~m, ~c	~c
Tổng(cột)	m	~m	Σ

Measure	Definition	Range	Null-Invariant
$\chi^2(a,b)$	$\sum_{i,j=0,1} \frac{(e(a_i,b_j)-o(a_i,b_j))^2}{e(a_i,b_j)}$	$[0,\infty]$	No
$Lift(a,b)$	$\frac{P(ab)}{P(a)P(b)}$	$[0,\infty]$	No
$AllConf(a,b)$	$\frac{sup(ab)}{max\{sup(a),sup(b)\}}$	$[0,1]$	Yes
$Coherence(a,b)$	$\frac{sup(ab)}{sup(a)+sup(b)-sup(ab)}$	$[0,1]$	Yes
$Cosine(a,b)$	$\frac{sup(ab)}{\sqrt{sup(a)sup(b)}}$	$[0,1]$	Yes
$Kulc(a,b)$	$\frac{sup(ab)}{2}(\frac{1}{sup(a)}+\frac{1}{sup(b)})$	$[0,1]$	Yes
$MaxConf$(a,b)	$max\{\frac{sup(ab)}{sup(a)},\frac{sup(ab)}{sup(b)}\}$	$[0,1]$	Yes

Table 3. Interestingness measure definitions.

Giao dịch null đối với m và c

Độ đo Kulczynski(1927)

Bất biến null

Data set	mc	$\overline{mc}$	$m\overline{c}$	$\overline{mc}$	χ^2	$Lift$	$AllConf$	$Coherence$	$Cosine$	$Kulc$	$MaxConf$
D_1	10,000	1,000	1,000	100,000	90557	9.26	0.91	0.83	0.91	0.91	0.91
D_2	10,000	1,000	1,000	100	0	1	0.91	0.83	0.91	0.91	0.91
D_3	100	1,000	1,000	100,000	670	8.44	0.09	0.05	0.09	0.09	0.09
D_4	1,000	1,000	1,000	100,000	24740	25.75	0.5	0.33	0.5	0.5	0.5
D_5	1,000	100	10,000	100,000	8173	9.18	0.09	0.09	0.29	0.5	0.91
D_6	1,000	10	100,000	100,000	965	1.97	0.01	0.01	0.10	0.5	0.99

Table 2. Example data sets.

Subtle: They disagree

Độ đo bất biến null nào tốt hơn?

- Tỷ lệ mất cân bằng (IR - Imbalance Ratio): đo lường sự mất cân bằng của hai tập mục A và B trong luật kéo theo

$$IR(A,B) = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$

- Kulczynski và IR cùng nhau trình bày một bức tranh rõ ràng cho cả ba tập dữ liệu từ D_4 đến D_6
  - D_4 cân bằng & trung tính
  - D_5 không cân bằng & trung tính
  - D_6 rất không cân bằng & trung tính

Data	mc	$\overline{m}c$	$m\overline{c}$	$\overline{mc}$	*all_conf.*	*max_conf.*	*Kulc.*	*cosine*	IR
D_1	10,000	1,000	1,000	100,000	0.91	0.91	0.91	0.91	0.0
D_2	10,000	1,000	1,000	100	0.91	0.91	0.91	0.91	0.0
D_3	100	1,000	1,000	100,000	0.09	0.09	0.09	0.09	0.0
D_4	1,000	1,000	1,000	100,000	0.5	0.5	0.5	0.5	0.0
D_5	1,000	100	10,000	100,000	0.09	0.91	0.5	0.29	0.89
D_6	1,000	10	100,000	100,000	0.01	0.99	0.5	0.10	0.99

Nội dung

- Khái niệm cơ bản
- Các phương pháp khai phá tập mục phổ biến
- Which Patterns Are Interesting?—Các phương pháp đánh giá mẫu
- **Tổng kết**

Tổng kết

- Khái niệm cơ bản: các luật kết hợp, Khung làm việc hỗ trợ - tin cậy, mẫu đóng và mẫu cực đại
- Các phương pháp khai phá mẫu phổ biến có thể mở rộng
  - Apriori (Sinh và kiểm tra các ứng viên)
  - Dựa trên phép chiếu (FPgrowth, CLOSET+, ...)
  - Tiếp cận định dạng dọc (ECLAT, CHARM, ...)
- Các phương pháp đánh giá mẫu