KHAI THÁC LUẬT KẾT HỢP VỚI APRIORI_TID

Thuật toán apriori – Tid tương tự như thuật toán apriori nhưng điều khác biệt cơ bản là khi tính số support, thuật toán apriori – Tid không sử dụng lại toàn bộ CSDL của lần duyệt thứ nhất mà kể từ bước thứ hai thuật toán apriori – Tid sử dụng tập C'k. mỗi thành viên của tập C'k có dạng **<Tid, {X_k}>**, trong đó X_k là tập Large k-items có thể có xuất hiện trong giao dịch có mã id là Tid.

Do vậy tập C'1 chính là toàn bộ giao dịch trong cơ sở dữ liệu ban đầu và sau đó với C'k nếu có một giao dịch không chưa sbaats kỳ một tập ứng viên k-items nào , thì giao dịch này sẽ không được nhập vào C'k. Do đó, số lượng nhập vào trong C'k có thể sẽ nhỏ hơn số lượng các giao dịch trong CSDL rất nhiều.

THUẬT TOÁN APRIORI-TID

Bước 1:

Quét tất cả các giao dịch để tìm tất cả các item có độ support lớn hơn min support và đứa vào tập Large 1-items (F_1)

Bước 2:

Đưa toàn bộ các Tid của giao dịch cùng các items vào C'1 dưới dạng <Tid, {X_1}>

Bước 3:

Xây dựng các cặp 2-items từ F_1 đưa vào tập ứng viên C2. Quét tất cả các giao dịch trong C'1 để tìm tất cả các tập large 2-item từ C2 đưa vào C'2 dưới dạng <Tid, {X_2}>, đồng thời đưa các tập Large 2-item ứng viên vào F_2

Bước 4:

Phát sinh luật, xây dựng các k-items từ F_{k-1} đưa vào tập ứng viên C_k. Quét tất cả các giao dịch trong C'$_{k-1}$ để tìm tất cả các tập Large k-item từ C_k và đưa vào C'$_k$ dưới dạng <Tid, {X_k}>, đồng thời đưa các tập Large k-item và F_k. Lặp lại bước 4 cho đến khi hết ứng viên mới (hết luật)

Ví dụ minh họa: Xét CSDL mẫu

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

σ(A) = 4
σ(C) = 6
σ(D) = 4
σ(T) = 4
σ(W) = 5

Với *minSup* = 50% (50*6/100 = 3), ta có
Tập Large 1-item F_1

F_1	
1-item	**Độ phổ biến**
A	4
C	6
D	4
T	4
W	5

Lấy toàn bộ <Tid, {X_1}> đưa vào C'_1

TID	Nội dung
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}}

Ở bước kết từ F_1 trên ta có tập C_2 gồm các cặp 2-item {{AC}, {AD}, {AT},{AW},{CD}{CT},{CW},{DT},{DW},{TW}}
Xác định ứng viên từ C2 khi duyệt Tid trong C'_1 và đưa vào C'_2

TID	Nội dung
1	{{AC},{AT},{AW},{CT},{CW},{TW}}
2	{{AD},{CW},{DW}}
3	{{AC},{AT},{AW},{CT},{CW},{TW}}

4	{{AC}, {AD},{AW},{CD},{CW},{DW}}
5	{{AC}, {AD}, {AT},{AW},{CD}{CT},{CW},{DT},{DW},{TW}}
6	{{CD},{CT},{DT}}

Tính tập Large 2-item ta có F_2

2-item	Độ phổ biến
AC	4
AT	3
AW	4
CD	4
CT	4
CW	5
DW	3
TW	3

Ở bước kết từ F_2 ta có tập C_3 gồm cặp 3-item
{{ACT},{ACW},{ATW},{CDW},{CTW}}
Xác định ứng viên từ C_3 khi duyệt Tid trong C'_2 và đưa vào C'_3

TID	Nội dung
1	{{ACT},{ACW},{ATW},{CTW}}
2	{{CDW}}
3	{{ACT},{ACW},{ATW},{CTW}}
4	{{ACW},{CDW}}
5	{{ACT},{ACW},{ATW},{CDW},{CTW}}
6	{{CDT}}

Tính tập Large 3-ite,, ta có F_3

3-item	Độ phổ biến
ACT	3
ACW	4
ATW	3
CDW	3
CTW	3

Ở bước kết từ F_3 ta có cặp C_4 gồm 4-item {{ACTW}}
Xác định ứng viên từ C_4 khi duyệt Tid trong C'_3 và đưa vào C'_4

TID	Nội dung
1	{{ACTW}}
3	{{ACTW}}
5	{{ACTW}}

Tính tập Large 4-item, ta có F_4

4-item	Độ phổ biến
ACTW	3

Ở bước kết từ F_4 ta có cặp C_5 gồm 5-item là rỗng. Thuật toán kết thúc

---

BÀI TẬP

Bài 1: Cho CSDL giao dịch bên dưới

1. Sử dụng thuật toán Apriori - Tid để tìm các tập phổ biến với minsupp = 22 %

TID	Items
100	M1, M2, M5
200	M2, M4
300	M2, M3
400	M1,M2, M4
500	M1, M3
600	M2, M3
700	M1, M3
800	M1, M2, M3, M5
900	M1, M2, M3

Bài 2:

Cho CSDL giao dịch sau và **minsupp = 60%**. Hãy sử dụng thuật toán **Apriori - Tid** để tìm tất cả các tập phổ biến .

TID	Items
100	K, D, A, B, C, F
200	A, H, C, D
300	C, I , D, E, G, F

400	B, C, H, A, I, D, F ,G
500	F, C, K, E, G