**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN BÁO CÁO THỰC HÀNH**

KHOA HỆ THỐNG THÔNG TIN Môn: Khai thác dữ liệu

**BÀI THỰC HÀNH SỐ 2**

Giảng viên hướng dẫn: ThS. Nguyễn Hồ Duy Trí

Sinh viên thực hiện: Phạm Thiện Bảo

MSSV: 20521107

Lớp: IS252.N22.HTCL

**Bài làm**

# Thực hành:

## Câu 1: Cho bảng dữ liệu ở một cửa hàng tạp hóa có 6 giao dịch như sau:

|  |  |
| --- | --- |
| **Transaction ID** | **Items** |
| T1 | HotDogs, Buns, Ketchup |
| T2 | HotDogs, Buns |
| T3 | HotDogs, Coke, Chips |
| T4 | Chips, Coke |
| T5 | Chips, Ketchup |
| T6 | HotDogs, Coke, Chips |

Với min\_sup = 33.33% và min\_conf = 60%, sinh viên thực hiện lại các yêu cầu trên.

### a) Tìm tất cả các tập phổ biến từ mẫu dữ liệu trên bằng giải thuật Apriori.

Với min\_sup là 33.33% và tổng số giao dịch là 6

=> Tần số xuất hiện tối thiểu của phần tử để thỏa min\_sup (min\_support\_count) là 2.

Tập các ứng viên 1 phần tử và support count tương ứng với bảng C1 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| HotDogs | 4 |
| Buns | 2 |
| Ketchup | 2 |
| Coke | 3 |
| Chips | 4 |

=> Các tập phổ biến 1 phần tử thỏa min\_support\_count:

L1 = {{HotDogs} ,{Buns} ,{Ketchup} ,{Coke}, {Chips}}

Tập các ứng viên 2 phần tử và support count tương ứng với bảng C2 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| HotDogs, Buns | 2 |
| HotDogs, Ketchup | 1 |
| HotDogs, Coke | 2 |
| HotDogs, Chips | 2 |
| Buns, Ketchup | 1 |
| Buns, Coke | 0 |
| Buns, Chips | 0 |
| Ketchup, Coke | 0 |
| Ketchup, Chips | 1 |
| Coke, , Chips | 3 |

=> Các tập phổ biến 2 phần tử thỏa min\_support\_count:

L2 = {{HotDogs, Buns} ,{HotDogs, Coke} ,{HotDogs, Chips} ,{Coke, Chips}}

Tập các ứng viên 3 phần tử và support count tương ứng với bảng C3 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| HotDogs, Buns, Coke | 0 |
| HotDogs, Buns, Chips | 0 |
| HotDogs, Coke, Chips | 2 |

=> Tập phổ biến 3 phần tử thỏa min\_support\_count:

L3 = {{HotDogs, Coke, Chips}}

Vì L3 chỉ còn 1 tập hợp nên không tiến hành tìm tập ứng cử viên 4 phần tử.

Vậy ta tìm được các tập phổ biến sau thỏa yêu cầu:

L= {{HotDogs} ,{Buns} ,{Ketchup} ,{Coke}, {Chips}, {HotDogs, Buns} ,{HotDogs, Coke} ,{HotDogs, Chips} ,{Coke, Chips}, {HotDogs, Coke, Chips}}

### b) Tìm tất cả các tập phổ biến từ mẫu dữ liệu trên bằng giải thuật FP-growth.

### Bước 1: Xây dựng cây FP

Tập phổ biến 1 phần tử với tần xuất thỏa điều kiện min\_sup và đã được sắp xếp giảm dần

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| HotDogs | 4 |
| Chips | 4 |
| Coke | 3 |
| Buns | 2 |
| Ketchup | 2 |

Tập phổ biến được sắp xếp theo tần xuất thỏa điều kiện:

|  |  |  |
| --- | --- | --- |
| Transaction ID | Items | Ordered Frequent items |
| T1 | HotDogs, Buns, Ketchup | HotDogs, Buns, Ketchup |
| T2 | HotDogs, Buns | HotDogs, Buns |
| T3 | HotDogs, Coke, Chips | HotDogs, Chips, Coke |
| T4 | Chips, Coke | Chips, Coke |
| T5 | Chips, Ketchup | Chips, Ketchup |
| T6 | HotDogs, Coke, Chips | HotDogs, Chips, Coke |

Từ đó, ta vẽ cây FP như sau:

**Diagram

Description automatically generated**

**Bước 2: Xây dựng cơ sở mẫu điều kiện cho mỗi mục phổ biến**

|  |  |
| --- | --- |
| Item | Conditional Pattern Base |
| Chips | {HotDogs}: 2 |
| Coke | {HotDogs, Chips,}: 2, {Chips}: 1 |
| Buns | {HotDogs}: 2 |
| Ketchup | {HotDogs, Buns}: 1, {Chips}: 1 |

**Bước 3: Xây dựng cây FP điều kiện**

|  |  |  |
| --- | --- | --- |
| Item | Conditional Pattern Base | Conditional FP-Tree (minsup |
| Buns | {HotDogs}: 2 | {{HotDogs}:2} |Buns |
| Chips | {HotDogs}: 2 | {{HotDogs}:2} |Chips |
| Coke | {{HotDogs, Chips}: 2, {Chips}: 1} | {{Chips}:3, {Hotdogs}: 2, {Hotdogs, Chips}: 2} |Coke |
| Ketchup | {{HotDogs, Buns}: 1, {Chips}: 1} | {} |
| HotDogs | {} | {} |

**Bước 4: Xây dựng tập phổ biến**

Ta tìm được các Frequent Pattern tương ứng như sau:

|  |  |  |
| --- | --- | --- |
| Item | Conditional FP-Tree | Frequent Pattern |
| Buns | {{HotDogs}:2} |Buns | {{Buns}:2, {HotDogs, Buns}:2} |
| Chips | {{HotDogs}:2} |Chips | {{Chips}:2, {HotDogs, Chips}:2} |
| Coke | {{Chips}:3, {Hotdogs}: 2, {Hotdogs, Chips}: 2} |Coke | {{Coke}: 3, {Coke, Chip}:3, {Hotdogs, Coke}: 2, {Hotdogs, Chips, Coke}: 2} |
| Ketchup | {} | {{Ketchup}:2} |
| HotDogs | {} | {{HotDogs}:4} |

Lấy những phần tử có 1 phần tử trở lên và thỏa min\_sup ta được các tập hợp sau đây:

L1 = {{HotDogs} ,{Buns} ,{Ketchup} ,{Coke}, {Chips}}

L2 = {{HotDogs, Buns}: 2, {HotDogs, Chips}: 2, {HotDogs, Coke}:2, {Chips, Coke}:3}

L3 = {{HotDogs, Chips, Coke}:2}

Vậy ta tìm được các tập phổ biến sau thỏa yêu cầu:

L= {{HotDogs} ,{Buns} ,{Ketchup} ,{Coke}, {Chips}, {HotDogs, Buns} ,{HotDogs, Coke} ,{HotDogs, Chips} ,{Coke, Chips}, {HotDogs, Coke, Chips}}

### c) Tìm tất cả các luật kết hợp dựa trên các tập phổ biến tìm được ở câu a.

Với min\_conf = 60%

+ Xét tập phổ biến {HotDogs, Buns}: 2 có các tập con không rỗng sau:

{HotDogs}: 4, {Buns}: 2 {HotDogs, Buns}: 2

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| HotDogs 🡪 Buns | 2/4 = 50% |
| Buns 🡪 HotDogs | 2/2 = 100% |

* Luật kết hợp thỏa yêu cầu là: **Buns** 🡪 **HotDogs**

+ Xét tập phổ biến {HotDogs, Chips}: 2 có các tập con không rỗng sau:

{HotDogs}: 4, {Chips}: 4, {HotDogs, Chips}: 2

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| HotDogs 🡪 Chips | 2/4 = 50% |
| Chips 🡪 HotDogs | 2/4 = 50% |

* Không tìm được luật kết hợp thỏa yêu cầu

+ Xét tập phổ biến {HotDogs, Coke}: 2 có các tập con không rỗng sau:

{HotDogs}: 4, {Coke}: 3, {HotDogs, Coke}: 2

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| HotDogs 🡪 Coke | 2/4 = 50% |
| Coke 🡪 HotDogs | 2/3 = 66.67% |

* Luật kết hợp thỏa yêu cầu là: **Coke** 🡪 **HotDogs**

+ Xét tập phổ biến {Chips, Coke}: 3 có các tập con không rỗng sau:

{Chips}:4, {Coke}: 3, {Chips, Coke}:3

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| Chips 🡪 Coke | 3/4 =75% |
| Coke 🡪 Chips | 3/3 = 100% |

* Luật kết hợp thỏa yêu cầu là: **Coke** 🡪 **Chips, Chips** 🡪 **Coke**

+ Xét tập phổ biến {HotDogs, Chips, Coke}: 2 có các tập con không rỗng sau:

{HotDogs}: 4, {Coke}: 3, {Chips}: 4, {HotDogs, Coke}: 2, {HotDogs, Chips}: 2, {Chips, Coke}: 3, {HotDogs, Chips, Coke}: 2

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| HotDogs 🡪{Chips, Coke} | 2/4 = 50% |
| Coke 🡪 {HotDogs, Chips} | 2/3 = 66,67% |
| Chips 🡪 {HotDogs, Coke} | 2/4 = 50% |
| {Chips, Coke} 🡪 HotDogs | 2/3 =66,67% |
| {HotDogs, Chips} 🡪 Coke | 2/2 = 100% |
| {HotDogs, Coke} 🡪 Chips | 2/2 = 100% |

* Luật kết hợp thỏa yêu cầu là: **Coke** 🡪 **{HotDogs, Chips}, {Chips, Coke}** 🡪 **HotDogs, {HotDogs, Chips}** 🡪 **Coke, {HotDogs, Coke}** 🡪 **Chips**

## Câu 2: Cho bảng dữ liệu ở một cửa hàng văn phòng phẩm như sau

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TID | KÉO | COMPA | THƯỚC | TẬP TRẮNG | BÚT BI | BÚT MÀU | TẨY |
| T1 |  | x |  | x | x |  |  |
| T2 | x |  | x | x | x |  |  |
| T3 |  | x |  | x | x |  |  |
| T4 | x | x |  | x | x |  |  |
| T5 |  |  | x |  |  |  |  |
| T6 |  |  |  |  | x |  |  |
| T7 |  |  |  | x |  |  |  |
| T8 |  |  |  |  |  |  | x |
| T9 |  |  |  |  |  | x | x |
| T10 |  |  |  |  |  | x |  |

### a) Tìm tất cả các tập phổ biến từ mẫu dữ liệu trên bằng giải thuật Apriori.

Với min\_sup = 30% và tổng số giao dịch là 10 => Tần số xuất hiện tối thiểu của phần tử để thỏa min\_sup là 3.

Tập các ứng viên 1 phần tử và support count tương ứng với bảng C1 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| Kéo | 2 |
| Compa | 3 |
| Thước | 2 |
| Tập trắng | 5 |
| Bút bi | 5 |
| Bút màu | 2 |
| Tẩy | 2 |

=> Các tập phổ biến 1 phần tử thỏa min\_support\_count:

L1 = {{ Compa } ,{ Bút bi } ,{ Tập trắng }}

Tập các ứng viên 2 phần tử và support count tương ứng với bảng C2 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| Compa, Tập trắng | 3 |
| Compa, Bút bi | 3 |
| Bút bi, Tập trắng | 4 |

=> Các tập phổ biến 2 phần tử thỏa min\_support\_count:

L2 = {{Compa, Tập trắng}, {Compa, Bút bi}, {Bút bi, Tập trắng}}

Tập các ứng viên 3 phần tử và support count tương ứng với bảng C3 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| Compa, Tập trắng, Bút bi | 3 |

=> Các tập phổ biến 3 phần tử thỏa min\_support\_count:

L3 = {{Compa; Tập trắng; Bút bi}}

Vì L3 chỉ còn 1 tập hợp nên không tiến hành tìm tập ứng cử viên 4 phần tử.

### b) Tìm tất cả các tập phổ biến từ mẫu dữ liệu trên bằng giải thuật FP-growth.

**Bước 1: Xây dựng cây FP**

Tập phổ biến 1 phần tử với tần xuất thỏa điều kiện min\_sup và đã được sắp xếp giảm dần.

|  |  |
| --- | --- |
| Itemset | support count |
| Bút bi | 5 |
| Tập trắng | 5 |
| Compa | 3 |

Tập phổ biến được sắp xếp theo tần xuất thỏa điều kiện:

|  |  |  |
| --- | --- | --- |
| Transaction ID | Items | Ordered frequent items |
| T1 | Compa, Tập trắng, Bút bi | Bút bi, Tập trắng, Compa |
| T2 | Kéo, Thước, Tập trắng, Bút bi | Bút bi, Tập trắng |
| T3 | Compa, Tập trắng, Bút bi | Bút bi, Tập trắng, Compa |
| T4 | Kéo, Compa, Tập trắng, Bút bi | Bút bi, Tập trắng, Compa |
| T5 | Thước |  |
| T6 | Bút bi | Bút bi |
| T7 | Tập trắng | Tập trắng |
| T8 | Tẩy |  |
| T9 | Bút màu, Tẩy |  |
| T10 | Bút màu |  |

Từ đó, ta vẽ cây FP như sau:

Diagram

Description automatically generated

**Bước 2: Xây dựng cơ sở mẫu điều kiện cho mỗi mục phổ biến**

|  |  |
| --- | --- |
| Item | Conditional Pattern Base |
| Compa | {Bút bi, Tập trắng}: 3 |
| Tập trắng | {Bút bi}: 4 |
| Bút bi | {} |

**Bước 3: Xây dựng cây FP điều kiện**

|  |  |  |
| --- | --- | --- |
| Item | Conditional Pattern Base | Conditional FP-Tree (minsup |
| Compa | {Bút bi, Tập trắng}: 3 | {{Bút bi}:3, {Tập trắng}:3, {Bút bi, Trắng}: 3} | Compa |
| Tập trắng | {Bút bi}:4 | {Bút bi}:4 | Tập trắng |
| Bút bi | {} | {} |

**Bước 4: Xây dựng tập phổ biến**

Ta tìm được các Frequent Pattern tương ứng như sau:

|  |  |  |
| --- | --- | --- |
| Item | Conditional FP-Tree | Frequent Pattern |
| Compa | {{Bút bi}: 3, {Tập trắng}: 3, {Bút bi, Trắng}: 3} |Compa | {{Compa}:3, {Compa, Bút bi}: 3, {Compa, Tập trắng}: 3, {Compa, Bút bi, Tập trắng}: 3} |
| Tập trắng | {Bút bi}:4 | Tập trắng | {{Tập trắng}: 4, {Bút bi; Tập trắng}: 4} |
| Bút bi | {} | {{Bút bi}: 5} |

Lấy những phần tử có 1 phần tử trở lên và thỏa min\_sup ta được các tập hợp sau đây:

L1 = {{Compa} ,{ Tập trắng } ,{ Bút bi }}

L2 = {{ Compa; Bút bi }, { Compa: Tập trắng } , { Bút bi, Tập trắng }}

L3 = {{ Compa, Bút bi, Tập trắng }}

Vậy ta tìm được các tập phổ biến sau thỏa yêu cầu:

L= { Compa} ,{ Tập trắng } ,{ Bút bi }, { Compa; Bút bi }, { Compa: Tập trắng } , { Bút bi, Tập trắng },{ Compa, Bút bi, Tập trắng }}

### c) Tìm tất cả các luật kết hợp dựa trên các tập phổ biến tìm được ở câu a.

Với min\_conf = 80%

+ Xét tập phổ biến {Compa; Bút bi}: 3 có các tập con không rỗng sau:

{Compa}:3, {Bút bi}: 5, {Compa; Bút bi}:3

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| Compa 🡪 Bút bi | 3/3 = 100% |
| Bút bi 🡪 Compa | 3/5 = 60% |

* Luật kết hợp thỏa yêu cầu là: **Compa** 🡪 **Bút bi**

+ Xét tập phổ biến {Compa; Tập trắng}: 3 có các tập con không rỗng sau:

{Compa}:3, {Tập trắng}: 5, {Compa; Tập trắng}:3

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| Compa 🡪 Tập trắng | 3/3 = 100% |
| Tập trắng 🡪 Compa | 3/5 = 60% |

* Luật kết hợp thỏa yêu cầu là: **Compa** 🡪 **Tập trắng**

+ Xét tập phổ biến {Bút bi; Tập trắng}:4 có các tập con không rỗng sau:

{Bút bi}: 5, {Tập trắng}: 5, {Bút bi, Tập trắng}: 4

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| Bút bi🡪 Tập trắng | 4/5 = 80% |
| Tập trắng 🡪 Bút bi | 4/5 = 80% |

* Luật kết hợp thỏa yêu cầu là: **Bút bi** 🡪 **Tập trắng, Tập trắng** 🡪 **Bút bi**

+ Xét tập phổ biến {Compa; Bút bi; Tập trắng}: 3 có các tập con không rỗng sau:

{Compa}: 3, {Bút bi}: 5, {Tập trắng}: 5, {Compa, Bút bi}: 3, {Compa, Tập trắng}: 3; {Bút bi, Tập trắng}: 4, {Compa; Bút bi, Tập trắng}: 3

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| Compa 🡪 {Bút bi; Tập trắng} | 3/3 = 100% |
| Bút bi 🡪 {Tập trắng, Compa} | 3/5 = 60% |
| Tập trắng 🡪 {Bút bi, Compa} | 3/5 = 60% |
| {Bút bi, Tập trắng} 🡪 Compa | 3/4 =75% |
| {Tập trắng, Compa} 🡪 Bút bi | 3/3 = 100% |
| {Bút bi, Compa} 🡪 Tập trắng | 3/3 = 100% |

* Luật kết hợp thỏa yêu cầu là: **Compa** 🡪 **{Bút bi, Tập trắng}, {Tập trắng, Compa}** 🡪 **Bút bi, {Bút bi, Compa}** 🡪 **Tập trắng**

## Câu 3: CSDL về Nhân viên được cho trong bảng sau:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | GT | Tuổi (T) | NL | LGD | Thu nhập (TN) | Thăng chức (TC) |
| 1 | Nữ | 20..25 | Giỏi | Rồi | Rất cao | Có |
| 2 | Nam | 20..25 | Khá | Chưa | Khá | Không |
| 3 | Nữ | 26..30 | Giỏi | Chưa | Khá | Có |
| 4 | Nữ | 31..40 | T.Bình | Chưa | T.Bình | Có |
| 5 | Nam | 26..30 | T.Bình | Rồi | Rất cao | Không |
| 6 | Nữ | 26..30 | Khá | Chưa | Cao | Không |
| 7 | Nữ | 31..40 | Khá | Chưa | T.Bình | Không |
| 8 | Nam | 26..30 | Khá | Rồi | Cao | Có |
| 9 | Nữ | >40 | Giỏi | Rồi | T.Bình | Không |
| 10 | Nữ | 26..30 | Giỏi | Chưa | Khá | Có |

Cho B= {Tuổi, Năng lực làm việc, Thăng chức}. Hãy tìm tất cả các luật kết hợp có về phải chỉ gồm thuộc tính Thăng chức (TC) thỏa ngưỡng minsup = 30% và minconf = 80%.

Bảng dữ liệu của B:

|  |  |  |  |
| --- | --- | --- | --- |
|  | T | NL | TC |
| 1 | 20..25 | Giỏi | Có |
| 2 | 20..25 | Khá | Không |
| 3 | 26..30 | Giỏi | Có |
| 4 | 31..40 | T.Bình | Có |
| 5 | 26..30 | T.Bình | Không |
| 6 | 26..30 | Khá | Không |
| 7 | 31..40 | Khá | Không |
| 8 | 26..30 | Khá | Có |
| 9 | >40 | Giỏi | Không |
| 10 | 26..30 | Giỏi | Có |

### Bước 1: Tìm tất cả các tập phổ biến từ mẫu dữ liệu trên bằng giải thuật Apriori.

Với min\_sup = 30% và tổng số giao dịch là 10 => Tần số xuất hiện tối thiểu của phần tử để thỏa min\_sup là 3.

Tập các ứng viên 1 phần tử và support count tương ứng với bảng C1 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| 20..25 | 2 |
| 26..30 | 5 |
| 31..40 | 2 |
| >40 | 1 |
| Giỏi | 4 |
| Khá | 4 |
| Trung bình | 2 |
| Có | 5 |
| Không | 5 |

=> Các tập phổ biến 1 phần tử thỏa min\_support\_count:

L1 = {{ 26..30} ,{ Giỏi } ,{ Khá },{ Có },{ Không }}

Tập các ứng viên 2 phần tử và support count tương ứng với bảng C2 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| 26..30, Giỏi | 2 |
| 26..30, Khá | 2 |
| 26..30, Có | 3 |
| 26..30, Không | 2 |
| Giỏi, Có | 3 |
| Giỏi, Không | 1 |
| Khá, Có | 1 |
| Khá, Không | 3 |

=> Các tập phổ biến 2 phần tử thỏa min\_support\_count:

L2 = {{26..30, Có}, {Giỏi, Có}, {Khá, Không}}

Tập các ứng viên 3 phần tử và support count tương ứng với bảng C3 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| 26..30, Giỏi, Có | 2 |

=> Các tập phổ biến 3 phần tử thỏa min\_support\_count: L3 =

Vậy ta tìm được các tập phổ biến sau thỏa yêu cầu:

L= {{26..30, Có}, {Giỏi, Có}, {Khá, Không}, { 26..30} ,{ Giỏi } ,{ Khá },{ Có },{ Không }}

### Bước 2: Tìm tất cả các luật kết hợp có vế phải chỉ gồm thuộc tính Thăng chức (TC) thỏa minconf = 80%.

Với min\_conf = 80%

Xét tất cả luật kết hợp có vế phải là thuộc tính Thăng chức

+ Xét tập phổ biến {26..30, Có} có các tập con không rỗng sau:

{26..30}: 5, {Có}: 5, {26..30, Có}: 3

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| 26..30 🡪 Có | 3/5= 60% |

* Luật kết hợp không thỏa yêu cầu min\_conf.

+ Xét tập phổ biến {Giỏi, Có} có các tập con không rỗng sau:

{Giỏi}: 4, {Có}: 5, {Giỏi, Có}: 3

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| Giỏi 🡪 Có | 3/4= 75% |

* Luật kết hợp không thỏa yêu cầu min\_conf.

+ Xét tập phổ biến {Khá, Không} có các tập con không rỗng sau:

{Khá}: 4, {Không}: 5, {Khá, Không}: 3

Tính confidence của các luật kết hợp:

|  |  |
| --- | --- |
| Khá 🡪 Không | 3/4= 75% |

* Luật kết hợp không thỏa yêu cầu min\_conf.

## Câu 4: Cho bảng dữ liệu ở một công ty có các giao dịch như sau:

|  |  |
| --- | --- |
| TID | Items |
| T1 | A, B, C, D |
| T2 | A, B, C |
| T3 | A, B, C |
| T4 | A, B, D |
| T5 | A, B |
| T6 | A, C, D |
| T7 | A, D |
| T8 | B, C, D |

Với min\_support\_count > 1, sinh viên hãy:

### a) Tìm tất cả các tập phổ biến từ mẫu dữ liệu trên bằng giải thuật Apriori.

Tập các ứng viên 1 phần tử và support count tương ứng với bảng C1 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| A | 7 |
| B | 6 |
| C | 5 |
| D | 5 |

=> Các tập phổ biến 1 phần tử thỏa min\_support\_count:

L1 = {{ A} ,{ B} ,{ C },{ D }}

Tập các ứng viên 2 phần tử và support count tương ứng với bảng C2 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| A, B | 5 |
| A, C | 4 |
| A, D | 3 |
| B, C | 4 |
| B, D | 3 |
| C, D | 3 |

=> Các tập phổ biến 2 phần tử thỏa min\_support\_count:

L2 = {{A, B}, {A, C}, {A, D}, {B, C}, {B, D}, {C, D}}

Tập các ứng viên 3 phần tử và support count tương ứng với bảng C3 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| A, B, C | 3 |
| A, B, D | 2 |
| A, C, D | 2 |
| B, C, D | 2 |

=> Các tập phổ biến 3 phần tử thỏa min\_support\_count:

L3 = {{A, B, C}, {A, B, D}, {A, C, D}, {B, C, D}}

Tập các ứng viên 4 phần tử và support count tương ứng với bảng C4 dưới đây:

|  |  |
| --- | --- |
| **Itemset** | **support count** |
| A, B, C, D | 1 |

Vì tập phổ biến 4 phần tử không thỏa min\_support\_count => L4 =

Vậy ta tìm được các tập phổ biến sau thỏa yêu cầu:

L = {{ A } ,{ B } ,{ C },{ D }, {A, B}, {A, C}, {A, D}, {B, C}, {B, D}, {C, D}, {A, B, C}, {A, B, D}, {A, C, D}, {B, C, D}}

### b) Tìm tất cả các tập phổ biến từ mẫu dữ liệu trên bằng giải thuật FP-growth.

**Bước 1: Xây dựng cây FP**

Tập phổ biến 1 phần tử với tần xuất thỏa điều kiện min\_sup và đã được sắp xếp giảm dần.

|  |  |
| --- | --- |
| Itemset | support count |
| A | 7 |
| B | 6 |
| C | 5 |
| D | 5 |

Tập phổ biến được sắp xếp theo tần xuất thỏa điều kiện:

|  |  |  |
| --- | --- | --- |
| Transaction ID | Items | Ordered frequent items |
| T1 | A, B, C, D | A, B, C, D |
| T2 | A, B, C | A, B, C |
| T3 | A, B, C | A, B, C |
| T4 | A, B, D | A, B, D |
| T5 | A, B | A, B |
| T6 | A, C, D | A, C, D |
| T7 | A, D | A, D |
| T8 | B, C, D | B, C, D |

Từ đó, ta vẽ cây FP như sau:

Diagram

Description automatically generated

**Bước 2: Xây dựng cơ sở mẫu điều kiện cho mỗi mục phổ biến**

|  |  |
| --- | --- |
| Item | Conditional Pattern Base |
| D | {A, B, C}: 1, {A}: 1, {A, B}: 1, {A, C}: 1, {B, C}: 1 |
| C | {A, B}: 3, {A}: 1, {B}: 1 |
| B | {A}: 5 |
| A | {} |

**Bước 3: Xây dựng cây FP điều kiện**

|  |  |  |
| --- | --- | --- |
| Item | Conditional Pattern Base | Conditional FP-Tree (minsup |
| D | {A, B, C}: 1, {A}: 1, {A, B}: 1, {A, C}: 1, {B, C}: 1 | {{A}: 4, {B}: 3, {C}: 3, {AB}: 2, {A, C}: 2, {B, C}: 2} | D |
| C | {A, B}: 3, {A}: 1, {B}: 1 | {{A}: 4, {B}: 4, {A, B}: 3} | C |
| B | {A}: 5 | {{A}:5} | B |
| A | {} | {} |

**Bước 4: Xây dựng tập phổ biến**

Ta tìm được các Frequent Pattern tương ứng như sau:

|  |  |  |
| --- | --- | --- |
| Item | Conditional FP-Tree | Frequent Pattern |
| D | {{A}: 4, {B}: 3, {C}: 3, {AB}: 2, {A, C}: 2, {B, C}: 2} | D | {{D}: 5, {A, D}: 4, {B, D}: 3, {C, D}: 3, {A, B, D}: 2, {A, C, D}: 2, {B, C, D}: 2} |
| C | {{A}: 4, {B}: 4, {A, B}: 3} | C | {{C}: 5, {A, C}: 4, {B, C}: 4, {A, B, C}: 3} |
| B | {{A}:5} | B | {{B}: 6, {A, B}: 5} |
| A | {} | {A: 7} |

Lấy những phần tử có 1 phần tử trở lên và thỏa min\_sup ta được các tập hợp sau đây:

L1 = {{A} ,{ B } ,{ C }, {D}}

L2 = {{A, B}, {A, C}, {A, D}, {B, C}, {B, D}, {C, D}}

L3 = {{A, B, C}, {A, B, D}, {A, C, D}, {B, C, D}}

Vậy ta tìm được các tập phổ biến sau thỏa yêu cầu:

L = {{ A } ,{ B } ,{ C },{ D }, {A, B}, {A, C}, {A, D}, {B, C}, {B, D}, {C, D}, {A, B, C}, {A, B, D}, {A, C, D}, {B, C, D}}

# Bài tập thêm:

## 3. Tìm hiểu thuật toán: Apriori+, FPMax

Thuật toán Apriori là một trong những thuật toán khai phá luật kết hợp phổ biến nhất trong dữ liệu lớn. Nó được sử dụng để tìm ra các mẫu phổ biến (frequent patterns) trong tập dữ liệu. Cụ thể, thuật toán này sử dụng một phương pháp đệ quy để liệt kê tất cả các mẫu con (subsets) của các mục (items) trong tập dữ liệu. Tuy nhiên, với các tập dữ liệu lớn, thuật toán này sẽ gặp phải vấn đề về thời gian và không gian tính toán, do đó Apriori+ và FPMax ra đời để giải quyết vấn đề này.

Thuật toán Apriori+ được cải tiến từ Apriori bằng cách sử dụng kỹ thuật lưu trữ (hash tree) để xác định mẫu phổ biến. Thay vì tạo tất cả các cặp mẫu con có thể có, Apriori+ chỉ tạo các cặp mẫu con nếu chúng có ít nhất một phần tử chung. Điều này giúp giảm số lượng cặp mẫu con cần xét và làm tăng tốc độ thực hiện.

FPMax là một thuật toán khác để khai phá mẫu phổ biến trong tập dữ liệu lớn. Khác với Apriori, FPMax tập trung vào các mẫu phổ biến cực đại (maximal frequent patterns) thay vì các mẫu phổ biến. Mẫu phổ biến cực đại là các mẫu không thể được mở rộng thêm mà vẫn là mẫu phổ biến. Thuật toán FPMax sử dụng một cây sufﬁx để liệt kê tất cả các mẫu phổ biến cực đại. Điều này giúp giảm thời gian và không gian tính toán so với Apriori khi tìm kiếm các mẫu phổ biến cực đại.

Tóm lại, Apriori+ và FPMax là hai thuật toán khai phá luật kết hợp cải tiến được sử dụng để tìm kiếm các mẫu phổ biến và mẫu phổ biến cực đại trong tập dữ liệu lớn. Cả hai đều giúp giảm thời gian và không gian tính toán so với Apriori và là hai phương pháp hiệu quả để khai phá dữ liệu.