Khai thác dữ liệu và ứng dụng

BT01: Khai thác tập phổ biến & Luật kết hợp

GVLT:	Thầy Lê Hoài Bắc	
GVTH:	Thầy Nguyễn Tiến Huy	
Sinh viên:	Nguyễn Phan Mạnh Hùng	1312727
	La Ngọc Thùy An	1312716

Mục lục

1 Bài 1: Apriori

1.1 Câu 1

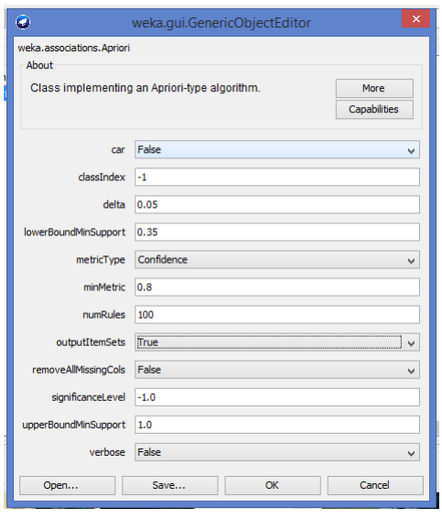
1.1.1 Tập phổ biến

Kích thước	ID	Tập phổ biến	Support
1	1	Bread	4
	2	Peanuts	4
	3	Milk	6
	4	Fruit	6
	5	Jam	5
	6	Soda	6
	7	Chips	4
2	8	Bread Milk	3
	9	Bread Jam	4
	10	Bread Soda	3
	11	Bread Chips	3
	12	Peanuts Milk	3
	13	Peanuts Fruit	4
	14	Milk Fruit	5
	15	Milk Jam	4
	16	Milk Soda	5
	17	Milk Chips	3

Kích thước	ID	Tập phổ biến	Support
2	18	Fruit Jam	3
	19	Fruit Soda	4
	20	Jam Soda	4
	21	Jam Chips	3
	22	Soda Chips	4
3	23	Bread Milk Jam	3
	24	Bread Jam Soda	3
	25	Bread Jam Chips	3
	26	Bread Soda Chips	3
	27	Peanuts Milk Fruit	3
	28	Milk Fruit Jam	3
	29	Milk Fruit Soda	4
	30	Milk Jam Soda	3
	31	Milk Soda Chips	3
	32	Jam Soda Chips	3
4	33	Bread Jam Soda Chips	3

1.1.2 Thiết lập tham số - Tập phổ biến sinh bởi Weka

Hình 1: Màn hình thiết lập tham số



Hình 2: Màn hình kết quả tập phổ biến

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Large Itemsets L(1):
Bread=TRUE 4
Peanuts=TRUE 4
Milk=TRUE 6
Fruit=TRUE 6
Jam=TRUE 5
Soda=TRUE 6
Chips=TRUE 4
```

```
Size of set of large itemsets L(2): 15

Large Itemsets L(2):
Bread=TRUE Milk=TRUE 3
Bread=TRUE Jam=TRUE 4
Bread=TRUE Soda=TRUE 3
Bread=TRUE Chips=TRUE 3
Peanuts=TRUE Milk=TRUE 3
Peanuts=TRUE Fruit=TRUE 4
Milk=TRUE Fruit=TRUE 5
Milk=TRUE Jam=TRUE 4
Milk=TRUE Soda=TRUE 5
Milk=TRUE Chips=TRUE 3
Fruit=TRUE Jam=TRUE 3
Fruit=TRUE Soda=TRUE 4
Jam=TRUE Soda=TRUE 4
Jam=TRUE Chips=TRUE 3
Soda=TRUE Chips=TRUE 4
```

```
Size of set of large itemsets L(3): 10

Large Itemsets L(3):
Bread=TRUE Milk=TRUE Jam=TRUE 3
Bread=TRUE Jam=TRUE Soda=TRUE 3
Bread=TRUE Jam=TRUE Chips=TRUE 3
Bread=TRUE Soda=TRUE Chips=TRUE 3
Peanuts=TRUE Milk=TRUE Fruit=TRUE 3
Milk=TRUE Fruit=TRUE Jam=TRUE 3
Milk=TRUE Fruit=TRUE Soda=TRUE 4
Milk=TRUE Jam=TRUE Soda=TRUE 3
Milk=TRUE Soda=TRUE Chips=TRUE 3
Jam=TRUE Soda=TRUE Chips=TRUE 3

Size of set of large itemsets L(4): 1

Large Itemsets L(4):
Bread=TRUE Jam=TRUE Soda=TRUE Chips=TRUE 3
```

1.2 Câu 2

1.2.1 Luật kết hợp

Rule ID	Set ID	Rule			Confidence
1	8	Bread 4	→	Jam 4	1
2		Jam 5	→	Bread 4	0.8
3	13	Peanuts 4	→	Fruit 4	1
4	22	Chips 4	→	Soda 4	1
5	14	Fruit 6	→	Milk 5	0.83
6		Milk 6	→	Fruit 5	0.83
7	16	Soda 6	→	Milk 5	0.83
8		Milk 6	→	Soda 5	0.83
9	15	Jam 5	→	Milk 4	0.8
10	20	Jam 5	→	Soda 4	0.8
11	19	Fruit Soda 4	→	Milk 4	1
12	23	Bread Milk 3	→	Jam 3	1
13	24	Bread Soda 3	→	Jam 3	1
14	25	Jam Chips 3	→	Bread 3	1
15		Bread Chips 3	→	Jam 3	1
16	26	Bread Chips 3	→	Soda 3	1
17		Bread Soda 3	→	Chips 3	1
18	27	Peanuts Milk 3	→	Fruit 3	1
19	28	Fruit Jam 3	→	Milk 3	1
20	31	Milk Chips 3	→	Soda 3	1
21	32	Jam Chips 3	→	Soda 3	1
22	29	Milk Soda 5	→	Fruit 4	0.8
23		Milk Fruit 5	→	Soda 4	0.8
24	33	Jam Soda Chips 3	→	Bread 3	1
25		Bread Soda Chips 3	→	Jam 3	1
26		Bread Jam Chips 3	→	Soda 3	1
27		Bread Jam Soda 3	→	Chips 3	1
28		Jam Chips 3	→	Bread Soda 3	1
29		Bread Chips 3	→	Jam Soda 3	1
30		Bread Soda 3	→	Jam Chips 3	1

1.2.2 Kết quả luật kết hợp

Hình 3: Kết quả luật kết hợp bằng Weka

Associator output

```
Best rules found:

 1. Bread=TRUE 4 ==> Jam=TRUE 4    conf:(1)
 2. Peanuts=TRUE 4 ==> Fruit=TRUE 4    conf:(1)
 3. Chips=TRUE 4 ==> Soda=TRUE 4    conf:(1)
 4. Fruit=TRUE Soda=TRUE 4 ==> Milk=TRUE 4    conf:(1)
 5. Bread=TRUE Milk=TRUE 3 ==> Jam=TRUE 3    conf:(1)
 6. Bread=TRUE Soda=TRUE 3 ==> Jam=TRUE 3    conf:(1)
 7. Jam=TRUE Chips=TRUE 3 ==> Bread=TRUE 3    conf:(1)
 8. Bread=TRUE Chips=TRUE 3 ==> Jam=TRUE 3    conf:(1)
 9. Bread=TRUE Chips=TRUE 3 ==> Soda=TRUE 3    conf:(1)
10. Bread=TRUE Soda=TRUE 3 ==> Chips=TRUE 3    conf:(1)
11. Peanuts=TRUE Milk=TRUE 3 ==> Fruit=TRUE 3    conf:(1)
12. Fruit=TRUE Jam=TRUE 3 ==> Milk=TRUE 3    conf:(1)
13. Milk=TRUE Chips=TRUE 3 ==> Soda=TRUE 3    conf:(1)
14. Jam=TRUE Chips=TRUE 3 ==> Soda=TRUE 3    conf:(1)
15. Jam=TRUE Soda=TRUE Chips=TRUE 3 ==> Bread=TRUE 3    conf:(1)
16. Bread=TRUE Soda=TRUE Chips=TRUE 3 ==> Jam=TRUE 3    conf:(1)
17. Bread=TRUE Jam=TRUE Chips=TRUE 3 ==> Soda=TRUE 3    conf:(1)
18. Bread=TRUE Jam=TRUE Soda=TRUE 3 ==> Chips=TRUE 3    conf:(1)
19. Jam=TRUE Chips=TRUE 3 ==> Bread=TRUE Soda=TRUE 3    conf:(1)
20. Bread=TRUE Chips=TRUE 3 ==> Jam=TRUE Soda=TRUE 3    conf:(1)
21. Bread=TRUE Soda=TRUE 3 ==> Jam=TRUE Chips=TRUE 3    conf:(1)
22. Fruit=TRUE 6 ==> Milk=TRUE 5    conf:(0.83)
23. Milk=TRUE 6 ==> Fruit=TRUE 5    conf:(0.83)
24. Soda=TRUE 6 ==> Milk=TRUE 5    conf:(0.83)
25. Milk=TRUE 6 ==> Soda=TRUE 5    conf:(0.83)
26. Jam=TRUE 5 ==> Bread=TRUE 4    conf:(0.8)
27. Jam=TRUE 5 ==> Milk=TRUE 4    conf:(0.8)
28. Jam=TRUE 5 ==> Soda=TRUE 4    conf:(0.8)
29. Milk=TRUE Soda=TRUE 5 ==> Fruit=TRUE 4    conf:(0.8)
30. Milk=TRUE Fruit=TRUE 5 ==> Soda=TRUE 4    conf:(0.8)
```

2 Bài 2: FP-Growth

2.1 Câu 1

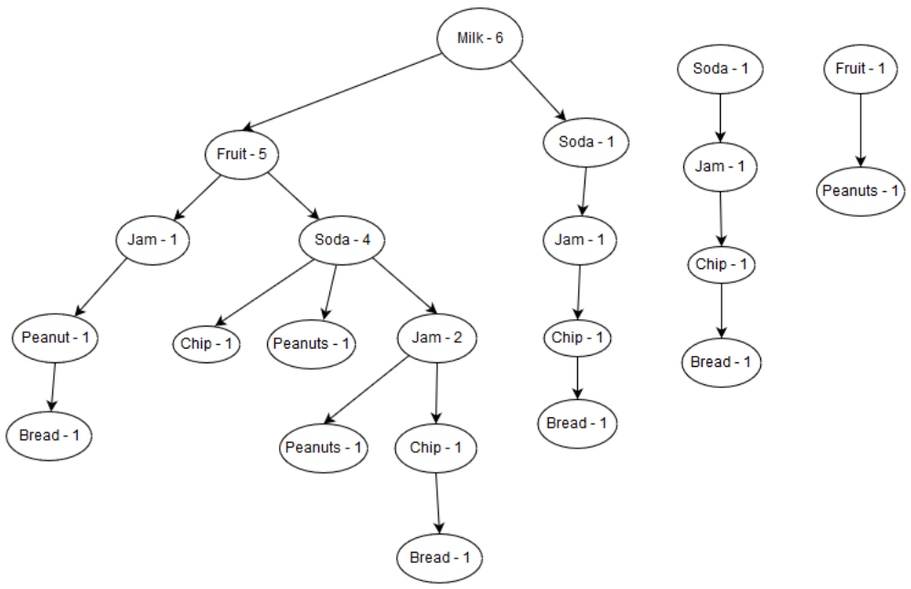
Bước 1: Duyệt cơ sở dữ liệu, chọn ra các item phổ biến và sắp xếp giảm dần. Ta được:

Milk	Fruit	Soda	Jam	Chips	Peanuts	Bread
6	6	6	5	4	4	4

Bước 2: Duyệt cơ sở dữ liệu lần hai. Với mỗi transaction, ta sắp xếp các item theo thứ tự độ phổ biến giảm dần trong bảng trên. Ta được:

Tid	Items					
1	Milk	Fuit	Jam	Peanuts	Bread	
2	Milk	Fruit	Soda	Jam	Chips	Bread
3	Soda	Jam	Chips	Bread		
4	Milk	Fruit	Soda	Jam	Peanuts	
5	Milk	Soda	Jam	Chips	Bread	
6	Milk	Fruit	Soda	Chips		
7	Milk	Fruit	Soda	Peanuts		
8	Fruit	Peanuts				

Bước 3: Xây dựng cây FP-growth



Bước 4: Xây dựng các cơ sở dữ liệu điều kiện và sắp xếp theo thứ tự giảm dần của support, đồng thời loại bỏ các item không thỏa min support.

Bảng 2: CSDL điều kiện - Bread

Bread					
	Jam - 1	Milk - 1	~~Peanuts~~	~~Fruit~~	
	Jam - 1	Milk - 1	Chips - 1	Soda - 1	~~Fruit~~
	Jam - 1	Milk - 1	Chips - 1	Soda - 1	
	Jam - 1	Chips - 1	Soda - 1		

Bảng 3: CSDL điều kiện - Peanuts

Peanuts				
	Fruit - 1	Milk - 1	~~Jam~~	
	Fruit - 1	Milk - 1	~~Soda~~	
	Fruit - 1	Milk - 1	~~Soda~~	~~Jam~~
	Fruit - 1			

Bảng 4: CSDL điều kiện - Chips

Chips				
	Soda - 1	Milk - 1	~~Fruit~~	
	Soda - 1	Milk - 1	Jam - 1	~~Fruit~~
	Soda - 1	Milk - 1	Jam - 1	
	Soda - 1	Jam - 1		

Bảng 5: CSDL điều kiện - Jam

Jam			
	Milk - 1	Fruit - 1	
	Milk - 2	Soda - 2	Fruit - 2
	Milk - 1	Soda - 1	
	Soda - 1		

Bảng 6: CSDL điều kiện - Soda

Soda		
	Milk - 4	Fruit - 4
	Milk - 1	

Bảng 7: CSDL điều kiện - Fruit

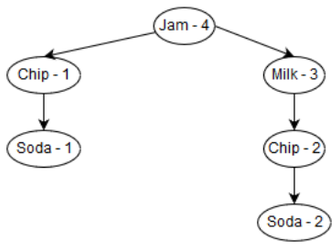
Fruit	
	Milk - 5

Ở đây ta được các tập phổ biến (ứng với mỗi điều kiện):

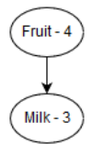
- Bread
- Peanuts
- Chips
- Jam
- Soda
- Fruit
- Milk

Bước 5: Xây dựng các FP-trees tương ứng:

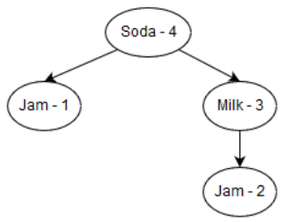
Hình 4: Bread - FP tree



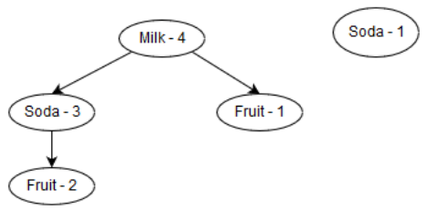
Hình 5: Peanuts - FP tree



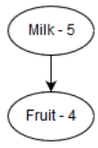
Hình 6: Chips - FP tree



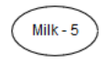
Hình 7: Jam - FP tree



Hình 8: Soda - FP tree



Hình 9: Fruit - FP tree



Ta thấy các cây điều kiện Peanuts, Soda, và Fruit suy biến, nên ta tiến hành khai thác các tập phổ biến:

- Peanuts Fruit
- Peanuts Milk
- Peanuts Fruit Milk
- Soda Fruit
- Soda Milk
- Soda Fruit Milk
- Fruit Milk

Sau đó ta lại tiếp tục đệ quy xây dựng các cơ sở dữ liệu điều kiện và cây FP cho tới khi các cây suy biến để khai thác tập phổ biến.
Cụ thể:

Điều kiện	Transactions	FP-tree	Tập phổ biến
Bread Soda	Chips - 1, Jam - 1 Chips - 2, Jam - 2, ~~Milk~~	(Chips - 3 → Jam - 3)	Bread Soda Bread Soda Chips Bread Soda Jam Bread Soda Chips Jam
Bread Chips	Jam 1 Jam - 2, ~~Milk~~	(Jam - 3)	Bread Chips Bread Chips Jam
Bread Milk	Jam - 3	(Jam - 3)	Bread Milk Bread Milk Jam
Bread Jam	∅	∅	Bread Jam
Chips Jam	Soda - 1 Soda - 2, ~~Milk~~	(Soda - 3)	Chips Jam Chips Jam Soda
Chips Milk	Soda - 3	(Soda - 3)	Chips Milk Chips Milk Soda
Chips Soda	∅	∅	Chips Soda
Jam Fruit	Milk - 2, ~~Soda~~ Milk - 1	(Milk-3)	Jam Fruit Jam Fruit Milk
Jam Soda	Milk - 3	(Milk - 3)	Jam Soda Jam Soda Milk
Jam Milk	∅	∅	Jam Milk

2.2 Câu 2

Các tập phổ biến được sinh ra bởi thuật toán Apriori theo thứ tự kích thước từ nhỏ tới lớn. Trong khi đó, nếu dùng FP-growth thì các tập kích thước lớn có thể được sinh ra trước.
Tuy vậy, các tập phổ biến được sinh ra bởi 2 thuật là giống nhau.

3 Bài 3: Các độ đo lý thú

3.1 Câu 1

	B, C:	itemset
	$conf\,(B \to C)$:	độ đo **confidence** của luật $B \to C$
Quy ước:	$lift\,(B, C)$:	độ đo **lift** của luật $B \to C$ và C $\to B$ (do có cách tính tương tự)
	$conv\,(B \to C)$:	độ đo **conviction** của luật $B \to C$
	$leve\,(B \to C)$:	độ đo **leverage** của luật $B \to C$

3.1.1 Confidence

$$conf\,(B \to C) = \frac{sup\,(B \cup C)}{sup\,(B)} \tag{1}$$

Ý nghĩa: thể hiện tỉ lệ các transaction chứa itemset B thì chứa cả itemset C, hay có thể hiểu như tỉ lệ luật dự đoán chính xác.

3.1.2 Lift

$$lift\,(B, C) = \frac{conf\,(B \to C)}{sup\,(C)} = \frac{sup\,(B \cup C)}{sup\,(B) \times sup\,(C)} \tag{2}$$

Ý nghĩa: Thể hiện sự độc lập giữa itemset B và itemset C. Nếu lift = 1, thì 2 itemsets này độc lập. Nếu lift < 1 thì nó thể hiện tính negative correlated, tức sự xuất hiện của itemset này ảnh hưởng ít tới sự xuất hiện của itemset kia. Nếu lift > 1 thì sự xuất hiện của itemset này ảnh hưởng nhiều tới sự xuất hiện của itemset kia, hay còn gọi là positive correlated.

3.1.3 Conviction

$$conv\,(B \to C) = \frac{1 - sup\,(C)}{1 - conf\,(B \to C)} \tag{3}$$

Ý nghĩa: thể hiện tỉ lệ itemset B xuất hiện mà không có C. Nói cách khác, thể hiện tần số luật B $\to$ C dự đoán sai.

3.1.4 Leverage

$$leve(B \to C) = sup\,(B \cup C) - sup\,(B) \times sup\,(C) \tag{4}$$

3.2 Câu 2

Rule Id	Confidence	Lift	Leverage	Conviction
1	1.00	1.60	0.19	∞
2	0.80	1.60	0.19	2.50
3	1.00	1.33	0.13	∞
4	1.00	1.33	0.13	∞
5	0.83	1.11	0.06	1.50
6	0.83	1.11	0.06	1.50
7	0.83	1.11	0.06	1.50
8	0.83	1.11	0.06	1.50
9	0.80	1.07	0.03	1.25
10	0.80	1.07	0.03	1.25
11	1.00	1.33	0.13	∞
12	1.00	1.60	0.14	∞
13	1.00	1.60	0.14	∞
14	1.00	2.00	0.19	∞
15	1.00	1.60	0.14	∞

Rule Id	Confidence	Lift	Leverage	Conviction
16	1.00	1.33	0.09	∞
17	1.00	2.00	0.19	∞
18	1.00	1.33	0.09	∞
19	1.00	1.33	0.09	∞
20	1.00	1.33	0.09	∞
21	1.00	1.33	0.09	∞
22	0.80	1.07	0.03	1.25
23	0.80	1.07	0.03	1.25
24	1.00	2.00	0.19	∞
25	1.00	1.60	0.14	∞
26	1.00	1.33	0.09	∞
27	1.00	2.00	0.19	∞
28	1.00	2.67	0.23	∞
29	1.00	2.00	0.19	∞
30	1.00	2.67	0.23	∞

3.3 Câu 3

Hình 10: Giá trị Confidence, Lift, Leverage, Conviction tính bằng Weka

```
1. Bread=TRUE Soda=TRUE 3 ==> Jam=TRUE Chips=TRUE 3    conf:(1) lift:(2.67) lev:(0.23) [1] < conv:(1.88)>
2. Jam=TRUE Chips=TRUE 3 ==> Bread=TRUE Soda=TRUE 3    conf:(1) lift:(2.67) lev:(0.23) [1] < conv:(1.88)>
3. Bread=TRUE 4 ==> Jam=TRUE 4    conf:(1) lift:(1.6) lev:(0.19) [1] < conv:(1.5)>
4. Jam=TRUE Chips=TRUE 3 ==> Bread=TRUE 3    conf:(1) lift:(2) lev:(0.19) [1] < conv:(1.5)>
5. Bread=TRUE Soda=TRUE 3 ==> Chips=TRUE 3    conf:(1) lift:(2) lev:(0.19) [1] < conv:(1.5)>
```

3.4 Câu 4

Có sự khác biệt trong cách tính bằng công thức thông thường và trong Weka ở cột Conviction. Có thể chương trình đã thêm vào các hệ số làm trơn nhằm khử trường hợp không xác định khi $confidence = 1$ nhằm giữ được mối liên hệ giữa các itemset cho việc tính toán sau này.