UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Đồ Án KTHP Môn học Khai Phá Dữ Liệu

ÁP DỤNG THUẬT TOÁN ECLAT CHO BỘ DỮ LIỆU GROCERY STORE 2

GVHD: TS. Nguyễn An Tế
Nhóm: 10

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Đồ Án KTHP Môn học Khai Phá Dữ Liệu

THÀNH VIÊN

1. Vũ Nguyễn Thảo Vi
2. Bùi Quốc Việt
3. Nguyễn Quốc Việt
4. Nguyễn Thanh Vy
5. Nguyễn Nhật Thảo Vy

UEH UNIVERSITY
Trường Công nghệ và Thiết kế

Đồ Án KTHP Môn học Khai Phá Dữ Liệu

TỔNG QUAN

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

TỔNG QUAN ĐỀ TÀI

- Tìm hiểu thuật toán ECLAT
- Áp dụng thuật toán ECLAT cho bộ dữ liệu Grocery Store 2
- So sánh với thuật toán Apriori

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

TỔNG QUAN DỮ LIỆU

Bộ dữ liệu Grocery Store 2 bao gồm một danh sách **170** sản phẩm tiêu dùng

```
with open("Groceries 2.csv", 'r') as temp_f:
    col_count = [ len(l.split(",")) for l in temp_f.readlines() ]
column_names = [i for i in range(0, max(col_count))]
df = pd.read_csv("Groceries 2.csv", header=None, delimiter=",",
 ↪  names=column_names)
```

	0	1	2	3	4	5	6	7	8	9	...	22	23	24	25	26	27	28	29	30	31
0	citrus fruit	semi-finished bread	margarine	ready soups	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	tropical fruit	yogurt	coffee	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	whole milk	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	pip fruit	yogurt	cream cheese	meat spreads	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	other vegetables	whole milk	condensed milk	long life bakery product	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

UEH UNIVERSITY
Trường Công nghệ và Thiết kế

Đồ Án KTHP Môn học Khai Phá Dữ Liệu

XÂY DỰNG THUẬT TOÁN ECLAT

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Cho CSDL và *minSup = 40%*

Tid	Items
T1	A,C,D
T2	B,C,E
T3	A,B,C,E
T4	B,E

BƯỚC 1

tìm tidsets của từng item riêng lẻ

Itemset	Tidset
{A}	{T1, T2}
{B}	{T2, T3, T4}
{C}	{T1, T2, T3}
{D}	{T1}
{E}	{T2, T3, T4}

BƯỚC 2

tìm các2-itemsets phải được kết hợp từ 2 item có minFreq ≥ 2

Itemset	Tidset
{A,B}	{T2}
{A,C}	{T1, T2}
{A,E}	{T2}
{B, C}	{T2, T3}
{B, E}	{T2, T3, T4}
{C, E}	{T2, T3}

BƯỚC 3

Tìm 3-itemsets dựa trên việc kết hợp hai cặp 2-itemsets có minFreq ≥ 2

Itemset	Tidset
{B, C, E}	{T2, T3}

BƯỚC 4

Kết thúc tại 3-itemset vì không thể tiếp tục kết hợp

Frequent Itemset
{A, C}
{B, C}
{B, E}
{C, E}
{B, C, E}

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Tạo luật kết hợp mạnh

Frequent Itemset
{A, C}
{B, C}
{B, E}
{C, E}
{B, C, E}

[minSup,minConf] = (40%,70%)

Tạo các luật kết hợp mạnh từ các tập phổ biến

Rule	[Sup, Conf]
A → C	[50%,100%]
C → A	[50%,66.67%]
B → C	[50%,66.67%]
C → B	[50%,66.67%]
B → E	[66.67%,100%]
E → B	[66.67%,100%]
C → E	[50%,66.67%]
E → C	[50%,66.67%]
BC → E	[50%,100%]
BE → C	[50%,66.67%]
CE → B	[50%,100%]
B → CE	[50%,66.67%]
C → BE	[50%,66.67%]
E → BC	[50%,66.67%]

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Đồ Án KTHP Môn học Khai Phá Dữ Liệu

Áp dụng thuật toán ECLAT cho bộ dữ liệu Grocery Store 2

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

- Áp dụng thư viện pyECLAT để phát hiện luật kết hợp, dựa trên mức độ hỗ trợ và độ tin cậy của các item.

- Chuyển đổi dữ liệu giao dịch sang dạng nhị phân trong DataFrame để tối ưu hóa phân tích và tính toán.

```python
from pyECLAT import ECLAT
eclat = ECLAT(data=df)

# Chuyển về binary dataframe
binary_df = eclat.df_bin
print("Binary DataFrame:")
print(binary_df[:5])
```

```
Binary DataFrame:
   dish cleaner  bathroom cleaner  brandy  jam  UHT-milk  liquor  pork  \
0             0                 0       0    0         0       0     0
1             0                 0       0    0         0       0     0
2             0                 0       0    0         0       0     0
3             0                 0       0    0         0       0     0
4             0                 0       0    0         0       0     0

   nut snack  kitchen utensil  zwieback  ...  instant coffee  flower (seeds)  \
0          0                0         0  ...               0               0
1          0                0         0  ...               0               0
2          0                0         0  ...               0               0
3          0                0         0  ...               0               0
4          0                0         0  ...               0               0
```

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Áp dụng thuật toán Eclat với ngưỡng giá trị minSup được chỉ định.

```
#Áp dụng thuật toán ECLAT tìm frequent itemsets với minSup = 0.03
get_ECLAT_indexes, get_ECLAT_supports = eclat.fit_all(min_support=0.03,
                                                      min_combination=1,
                                                      separator=', ')
#Áp dụng thuật toán ECLAT tìm frequent itemsets với minSup = 0.03
get_ECLAT_indexes, get_ECLAT_supports = eclat.fit_all(min_support=0.03,
                                                      min_combination=1,
                                                      separator=', ')
```

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Xem các tập phổ biến

```
# Tạo dataframe chứa frequent itemsets
frequent_itemsets_eclat = pd.DataFrame({'support':
    get_ECLAT_supports.values(),'itemsets': get_ECLAT_supports.keys()})
frequent_itemsets_eclat = frequent_itemsets_eclat.sort_values(by='support',
    ascending=False)
frequent_itemsets_eclat
```

	support	itemsets
12	0.074835	other vegetables, whole milk
8	0.056634	rolls/buns, whole milk
14	0.056024	yogurt, whole milk
4	0.048907	root vegetables, whole milk
3	0.047382	root vegetables, other vegetables
11	0.043416	other vegetables, yogurt
5	0.042603	rolls/buns, other vegetables
2	0.042298	tropical fruit, whole milk
18	0.040061	whole milk, soda
9	0.038332	rolls/buns, soda
1	0.035892	tropical fruit, other vegetables
6	0.034367	rolls/buns, yogurt
15	0.034367	bottled water, whole milk
17	0.033249	whole milk, pastry
13	0.032740	other vegetables, soda
16	0.032232	whole milk, whipped/sour cream
7	0.030605	rolls/buns, sausage
10	0.030503	citrus fruit, whole milk
0	0.030097	pip fruit, whole milk

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Kết quả chỉ số Support

	antecedents	consequents	support
0	(whole milk)	(other vegetables)	0.074835
1	(other vegetables)	(whole milk)	0.074835
2	(whole milk)	(rolls/buns)	0.056634
3	(rolls/buns)	(whole milk)	0.056634
4	(whole milk)	(yogurt)	0.056024
5	(yogurt)	(whole milk)	0.056024
6	(root vegetables)	(whole milk)	0.048907
7	(whole milk)	(root vegetables)	0.048907
9	(other vegetables)	(root vegetables)	0.047382
8	(root vegetables)	(other vegetables)	0.047382
10	(other vegetables)	(yogurt)	0.043416
11	(yogurt)	(other vegetables)	0.043416
12	(rolls/buns)	(other vegetables)	0.042603
13	(other vegetables)	(rolls/buns)	0.042603
14	(whole milk)	(tropical fruit)	0.042298
15	(tropical fruit)	(whole milk)	0.042298
16	(whole milk)	(soda)	0.040061
17	(soda)	(whole milk)	0.040061
19	(soda)	(rolls/buns)	0.038332
18	(rolls/buns)	(soda)	0.038332
21	(tropical fruit)	(other vegetables)	0.035892
20	(other vegetables)	(tropical fruit)	0.035892

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Kết quả chỉ số Confidence

	antecedents	consequents	support	confidence
5	(yogurt)	(whole milk)	0.056024	0.401603
7	(root vegetables)	(whole milk)	0.048907	0.448694
9	(root vegetables)	(other vegetables)	0.047382	0.434701
15	(tropical fruit)	(whole milk)	0.042298	0.403101
31	(whipped/sour cream)	(whole milk)	0.032232	0.449645

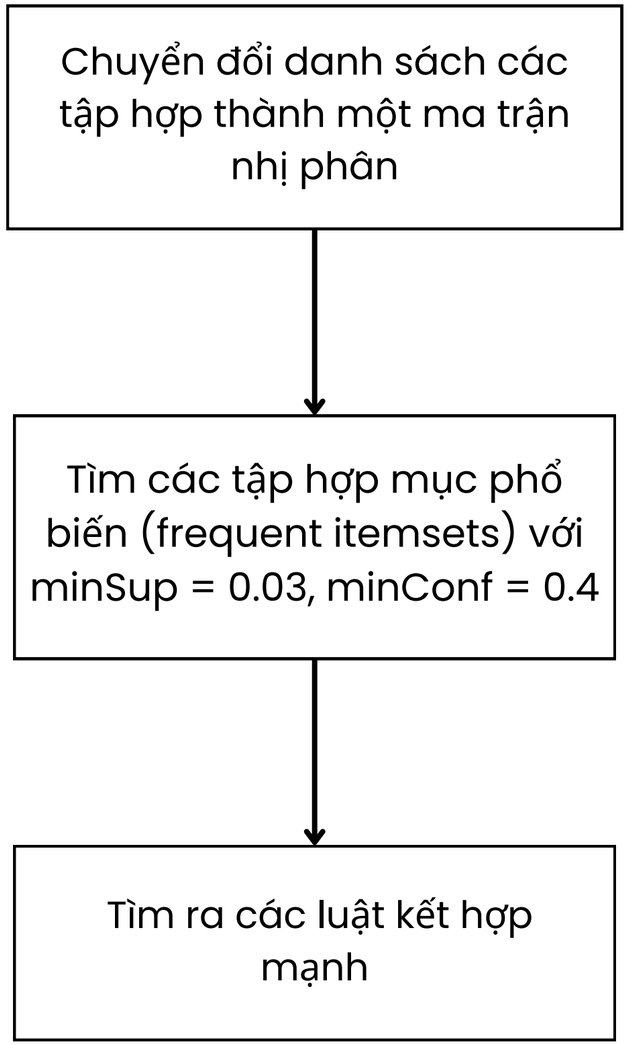
UEH UNIVERSITY
Trường Công nghệ và Thiết kế

Đồ Án KTHP Môn học Khai Phá Dữ Liệu

ỨNG DỤNG VỚI THUẬT TOÁN APRIORI

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

THUẬT TOÁN APRIORI



Itemsets phổ biến

	support	itemsets
50	0.074835	(whole milk, other vegetables)
56	0.056634	(whole milk, rolls/buns)
62	0.056024	(whole milk, yogurt)
58	0.048907	(whole milk, root vegetables)
47	0.047382	(other vegetables, root vegetables)
51	0.043416	(other vegetables, yogurt)
46	0.042603	(other vegetables, rolls/buns)
60	0.042298	(whole milk, tropical fruit)
59	0.040061	(whole milk, soda)
55	0.038332	(rolls/buns, soda)
49	0.035892	(other vegetables, tropical fruit)
57	0.034367	(rolls/buns, yogurt)
44	0.034367	(whole milk, bottled water)
52	0.033249	(pastry, whole milk)
48	0.032740	(other vegetables, soda)
61	0.032232	(whole milk, whipped/sour cream)
54	0.030605	(sausage, rolls/buns)
45	0.030503	(whole milk, citrus fruit)
53	0.030097	(whole milk, pip fruit)

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

THUẬT TOÁN APRIORI | LUẬT KẾT HỢP MẠNH

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	conviction
4	(yogurt)	(whole milk)	0.139502	0.255516	0.056024	0.401603	1.571735	1.244132
1	(root vegetables)	(whole milk)	0.108998	0.255516	0.048907	0.448694	1.756031	1.350401
0	(root vegetables)	(other vegetables)	0.108998	0.193493	0.047382	0.434701	2.246605	1.426693
2	(tropical fruit)	(whole milk)	0.104931	0.255516	0.042298	0.403101	1.577595	1.247252
3	(whipped/sour cream)	(whole milk)	0.071683	0.255516	0.032232	0.449645	1.759754	1.352735

UEH UNIVERSITY
Trường Công nghệ và Thiết kế

Đồ Án KTHP Môn học Khai Phá Dữ Liệu

ĐỘ ĐO LIFT

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

$$\text{lift}(X \Rightarrow Y) = \frac{\sup(X \Rightarrow Y)}{\sup(X) \cdot \sup(Y)} = \frac{\text{conf}(X \Rightarrow Y)}{\sup(Y)}$$

	antecedents	consequents	support	confidence	lift
5	(yogurt)	(whole milk)	0.056024	0.401603	1.571735
7	(root vegetables)	(whole milk)	0.048907	0.448694	1.756031
9	(root vegetables)	(other vegetables)	0.047382	0.434701	2.246605
15	(tropical fruit)	(whole milk)	0.042298	0.403101	1.577595
31	(whipped/sour cream)	(whole milk)	0.032232	0.449645	1.759754

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Đồ Án KTHP Môn học Khai Phá Dữ Liệu

ĐỘ ĐO CHI-SQUARE

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

	$\mathbf{Y}$	$\bar{\mathbf{Y}}$	
$\mathbf{X}$	$O_1(E_1)$	$O_2(E_2)$	$O_1 + O_2$
$\bar{\mathbf{X}}$	$O_3(E_3)$	$O_4(E_4)$	$O_1 + O_2$
	$O_1 + O_3$	$O_2 + O_4$	

	antecedents	consequents	Chi2	pvalue	freq	expected_freq
21	(yogurt)	(whole milk)	178	1.333683e-40	551	351
13	(root vegetables)	(whole milk)	235	5.124682e-53	481	274
15	(root vegetables)	(other vegetables)	447	3.519862e-99	466	207
23	(tropical fruit)	(whole milk)	131	2.292893e-30	416	264
25	(whipped/sour cream)	(whole milk)	149	2.397312e-34	317	180

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

Đồ Án KTHP Môn học Khai Phá Dữ Liệu

SO SÁNH ĐIỂM GIỐNG VÀ KHÁC NHAU GIỮA APRIORI VÀ ECLAT

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

GIỐNG NHAU

- ECLAT và Apriori tìm quan hệ kết hợp trong dữ liệu mà không phụ thuộc vào thứ tự xuất hiện của mặt hàng.
- Cả hai sử dụng minSup để lọc các tập ứng viên, xác định tập item phổ biến.
- Sử dụng bước cắt tỉa để loại bỏ tập item không phổ biến, dựa trên nguyên tắc tập cha không phổ biến.

UEH UNIVERSITY | Trường Công nghệ và Thiết kế

KHÁC NHAU

Tiêu chí so sánh	ECLAT	APRIORI
Cơ sở dữ liệu	Theo cấu trục bảng dọc	Theo cấu trúc bảng ngang
Thời gian thực thi	Về lý thuyết, thời gian thực thi nhanh hơn vì chỉ cần duyệt qua CSDL 1 lần.	Thời gian thực thi không tối ưu vì liên tục lặp lại việc duyệt qua toàn bộ CSDL.
Tài nguyên sử dụng	Tốn bộ nhớ hơn để lưu trữ nếu số lượng giao dịch lớn	Tốn bộ nhớ để lưu các tập itemset trong mỗi lần duyệt CSDL

UEH UNIVERSITY
Trường Công nghệ và Thiết kế

THANK YOU