Mestrado Engenharia Informática e Multimédia

Aprendizagem e Mineração de Dados

Relatório Trabalho Prático B

Data: January 17, 2022

Docente:
Eng. Artur Ferreira

Realizado por:
Grupo 19
Henrique Martinho - 42363
Mihail Ababii - 46435

Índice de Conteúdos

Lista de Figuras

1 Análise do dataset

A companhia *SoftKnow* é uma companhia de e-Commerce, pelo que pretende adquirir informação sobre os clientes e os seus hábitos. Para tal pretendem utilizar os dados armazenados ao longo do tempo, que contêm os seguintes atributos:

- tracking_record_id - número de identificação de um evento
- date_time - data e hora da ocorrência (evento)
- company - Nome/Sigla da empresa
- tracking_id - Encoder (ASCII/UTF-8)
- meio - meio utilizado para acesso ao produto
- link - URL para acesso ao produto
- referer - URL para acesso ao produto
- user_gui - identificador único do user registado
- campaign_id - identificador da campanha promocional
- product_gui - identificador único do produto, que por vezes pode não ter significado caso o utilizador visite outras páginas que não correspondem a um produto
- ip - ip address de cada user que visitou
- browser - identificador do browser do user que visitou
- session_id - identificador da sessão
- cookie_id - valor do cookie atribuído a cada utilizador

2 Problema "Market-Basket Analysis"

O problema a resolver neste projecto é encontrar um conjunto de regras que associem os diversos produtos, realizando uma analise de market-basket. O objectivo de uma análise market-basket é encontrar grupos de itens (I) que tendem a ocorrer juntos em transacções (T). Assim sendo:

- I = {i1 , i2 , . . . , iM }, onde:
  - I representa um conjunto de M itens;
  - Cada iM representam um item.
- T = t1 , t2 , . . . , tN , onde:
  - T representa um conjunto de N transacções;
  - Cada tN é um conjunto de itens (I);

Neste caso, para se encontrar as relações entre produtos realizando este tipo de analise utilizou-se o product_gui como itens e o cookie_id como transacção, permitindo assim melhor compreender o tipo de produtos que um visitante possa ter bem como a relação entre os diversos produtos

3 PostgreSQL

Para simular uma situação real é pedido que se coloque os dados adquiridos numa base de dados PostgreSQL e manipular os dados, pelo que são criados diversos scripts. Primeiramente, é pedido para apresentar o número total de eventos e o número total de visitantes, pelo que se obteve um total de 415863 eventos, com 263137 visitantes únicos.

```
 totalevents_jan_2012
----------------------
                415863
(1 row)
```

Figure 1: Total de Eventos (Esperado)

```
 totalevents_jan_2012
----------------------
                415863
(1 row)
```

Figure 2: Total de Eventos

```
 totalnumberofvisitors_cookie_id_jan_2012
------------------------------------------
                                   263137
(1 row)
```

Figure 3: Visitantes únicos (E)

```
 totalnumberofvisitors_cookie_id_jan_2012
------------------------------------------
                                   263137
(1 row)
```

Figure 4: Visitantes únicos

Num segundo ponto, as manipulações SQL apresentam maior complexidade, pedindo que se apresente a distribuição dos visitantes por número de sessões participadas e a distribuição dos visitantes pelo número de eventos por sessão.

```
 numberofsessions | numberofsessions
------------------+------------------
           240911 |                1
            15695 |                2
             2529 |                3
             1081 |                4
              692 |                5
              559 |                6
              403 |                7
              308 |                8
              230 |                9
              167 |               10
              141 |               11
              111 |               12
```

Figure 5: Distribuição por sessões (E)

```
 numberofsessions | numberofsessions
------------------+------------------
           240911 |                1
            15695 |                2
             2529 |                3
             1081 |                4
              692 |                5
              559 |                6
              403 |                7
              308 |                8
              230 |                9
              167 |               10
              141 |               11
              111 |               12
```

Figure 6: Distribuição por sessões

numberofeventspersession	numberofvisitors
1	282976
2	11153
3	4507
4	2055
5	1347
6	950
7	720
8	593
9	475
10	431
11	322
12	261
13	263
14	202
15	150

Figure 7: Distribuição por eventos por sessão (E)

numberofeventspersession	numberofvisitors
1	282976
2	11153
3	4507
4	2055
5	1347
6	950
7	720
8	593
9	475
10	431
11	322
12	261
13	263
14	202
15	150

Figure 8: Distribuição por eventos por sessão

É realizado uma query para obter todos os utilizadores que estiveram presentes em 18 sessões distintas. Com isto podemos confirmar que a base de dados foi correctamente preenchida, pois os dados obtidos através de queries correspondem ao pretendido, estando esta pronta para ser utilizada.

cookie_id	c1	c2
1b70f82c-d17c-447b-ac8c-4d16d3dd3ffd	18	24
21c87837-c3ea-4482-952a-fa1eef21f3d1	18	18
22a10aed-6455-444b-bb81-ed661de64337	18	75
325437cd-27fa-4f01-a95e-8ddad81c0a46	18	28
4af72963-62c2-490a-9295-685d3e28afe5	18	36
530df755-b04e-4ec8-ae9d-18f59537a652	18	28
5867b140-1d63-4078-8758-6917fc0f4dd6	18	25
62ae0655-bd69-480e-af74-3858178b5d19	18	34
ec0cb166-38e1-48dc-963a-e54b592e3328	18	21
fd4e8075-f0f2-4aec-9001-ce238b9e0e08	18	28
fe695227-41b7-4757-8881-4a13c7966788	18	165

(11 rows)

Figure 9: Utilizadores em 18 sessões (E)

cookie_id	c1	c2
1b70f82c-d17c-447b-ac8c-4d16d3dd3ffd	18	24
21c87837-c3ea-4482-952a-fa1eef21f3d1	18	18
22a10aed-6455-444b-bb81-ed661de64337	18	75
325437cd-27fa-4f01-a95e-8ddad81c0a46	18	28
4af72963-62c2-490a-9295-685d3e28afe5	18	36
530df755-b04e-4ec8-ae9d-18f59537a652	18	28
5867b140-1d63-4078-8758-6917fc0f4dd6	18	25
62ae0655-bd69-480e-af74-3858178b5d19	18	34
ec0cb166-38e1-48dc-963a-e54b592e3328	18	21
fd4e8075-f0f2-4aec-9001-ce238b9e0e08	18	28
fe695227-41b7-4757-8881-4a13c7966788	18	165

(11 rows)

Figure 10: Utilizadores em 18 sessões

Devido ao elevado número de dados, pretende-se utilizar apenas uma porção dos mesmos tendo-se seleccionado apenas os eventos criados por utilizadores que cujo número de eventos atendido seja entre 5 e 30. Como o objectivo deste projecto é encontrar regras de correlação entre os produtos, decidiu-se que apenas alguns dos atributos são relevantes para o problema, sendo estes o tracking_record_id, o user_gui, o campaign_id, o product_gui, a company, o link, o session_id e o cookie_id.

4 *Support & Confidence*

Support - mede a frequência com que a coleção de itens em uma associação ocorre em conjunto como uma percentagem de todas as transações

Confidence - mede a probabilidade de que uma coleção de itens ocorra quando uma coleção de itens ocorreu.

Lift - O *lift* indica a força de uma regra de acordo com a co-ocorrência aleatória do antecedente e do consequente, dado o seu apoio individual. O lift ainda fornece informações sobre a melhoria, ou seja, o aumento na probabilidade do consequente dado o antecedente.

Neste ponto do projecto é necessário realizar um programa que apresente os valores de *support* e de *confidence* cujo output seja no máximo *"maxR"* regras. Após realizarmos o programa podemos observar que foram encontrados apenas 2 conjuntos de valores de *support* mínimos e *confidence* mínimos, para os quais é possível obter *"maxR"* regras. Idealmente seria necessário seleccionar o conjunto que tivesse o menor *support* e a maior *confidence*, pelo que neste caso seria 6% *support* e 1% *confidence*.

```
number os rules (maxR):5
Suporte =  0.07  | Confiança =  0.01
Regras:
(['divertimento'], ['tecnologia'], 672, 0.7593220338983051)
(['tecnologia'], ['divertimento'], 672, 0.6170798898071626)

Suporte =  0.06  | Confiança =  0.01
Regras:
(['display.category*homepage'], ['tecnologia'], 573, 0.2064864864864865)
(['tecnologia'], ['display.category*homepage'], 573, 0.5261707988980716)
(['divertimento'], ['tecnologia'], 672, 0.7593220338983051)
(['tecnologia'], ['divertimento'], 672, 0.6170798898071626)
```

Figure 11: Output Exercício7

5 Market-Basket Report

Com o intuito de obter um relatório de acordo com a análise *market-basket* realizada para enviar à We-Commerce foi utilizado *Orange DM*. Desta forma foi possível obter o relatório com as regras de associação através da inserção de valores de suporte e de confiança.

Os resultados obtidos podem ser observados na imagem abaixo.

	Antecedent	Consequent	Support	Confidence	Coverage	Strength	Lift	Leverage
1	divertimento	tecnologia	0.083	0.759	0.11	1.231	5.623	0.069
2	tecnologia	divertimento	0.083	0.617	0.135	0.813	5.623	0.069
3	display.categor...	tecnologia	0.071	0.206	0.344	0.392	1.529	0.025
4	tecnologia	display.categor...	0.071	0.526	0.135	2.548	1.529	0.025
5	botins	pumpseopentoes	0.065	0.617	0.106	0.979	5.97	0.054
6	pumpseopentoes	botins	0.065	0.631	0.103	1.022	5.97	0.054
7	display.categor...	botas	0.064	0.187	0.344	0.332	1.643	0.025
8	botas	display.categor...	0.064	0.565	0.114	3.016	1.643	0.025

Figure 12: Regras de Associação

Através do suporte podemos concluir que o divertimento e a tecnologia são a combinação mais frequente. E com o confidence podemos concluir que a probabilidade de um cliente procurar produtos de tecnologia após procurar produtos de divertimento é ligeiramente superior ao inverso. Podemos também observar que para valores de confidence mais elevados o valor do lift também aumenta.

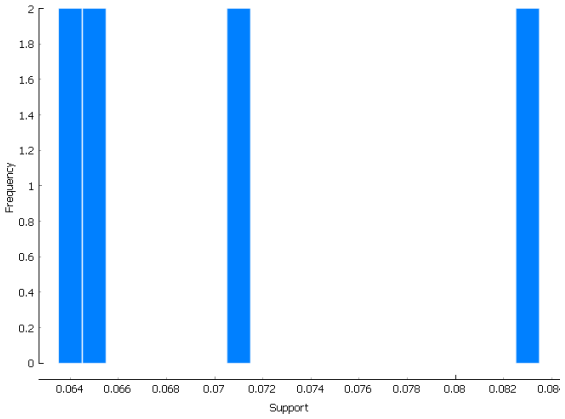

Figure 13: Gráfico de Distribuição de *Support*

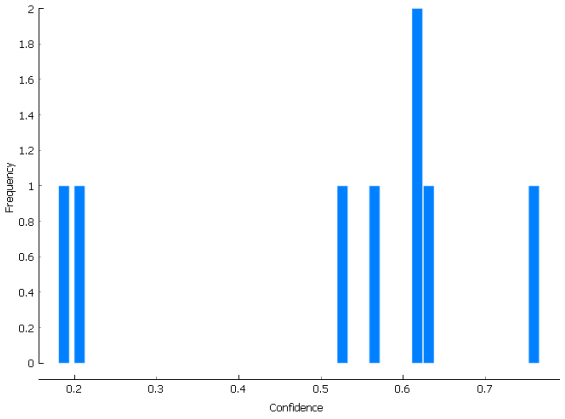

Figure 14: Gráfico de Distribuição de *Confidence*

6 Conclusões

A realização do trabalho prático da unidade curricular de Aprendizagem e Mineração de Dados teve como objetivo consolidar os conteúdos lecionados ao longo do semestre e aplicar este conhecimento num problema *market-basket analysis*.

Com este trabalho o grupo conseguiu, através da utilização de uma base de dados *PostgreSQL*, algoritmos em *Python* e do *Orange* para visualização e confirmação de dados e adquirir as regras de correlação entre os produtos.

Assim sendo, foi possível melhorar as capacidades funcionais de trabalho com o *PostgreSQL*, realizado queries consideravelmente mais complexas que em trabalhos anteriores, utilizado funções SQL como o *GROUP BY* e o *ORDER BY*. Com a necessidade de criar um programa *Python* foi também possível trabalhar com as tecnologias disponibilizadas pelo *Orange DM* como o *OneHot*, *frequent_itemsets* e *association_rules* e por consequente melhor compreende algumas métricas como o *Support* e o *confidence*.

Deste modo, o grupo atingiu os objectivos com sucesso realizando o trabalho na sua totalidade.

7 Reference

https://github.com/biolab/orange3-associate/blob/master/orangecontrib/associate/fpgrowth.py