Instituto Superior de Engenharia de Lisboa

ISEL INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Aprendizagem e Mineração de Dados

We-Commerce Data Mining Project

Mestrado em Engenharia Informática de Multimédia

Pedro Gonçalves, 45890

Rodrigo Dias, 45881

Rúben Santos, 49063

Semestre de Inverno, 2021/2022

1. Introdução

A companhia **We-Commerce** captura e armazena todos os dados gerados pelos visitantes que navegam nos seus sites de comércio. Com base num registo de eventos com aproximadamente **420000** entradas, o objetivo deste projeto é transformar todos esses daods em conhecimento (***data – knowledge***).

Procurar-se-á aconselhar a empresa acerca de questões como o tipo de produtos que mais são visualizados ou as relações entre os mesmos. Estes aspetos fornecerão uma dose de conhecimento à companhia **We-Commerce,** que poderá ser utilizada para investir numa estratégia de marketing mais informada, organizada e eficaz.

2. Projeto

2.1. Dados

Maior parte dos dados coletados pela empresa **We-Commerce** são dados provenientes da **Web**. Passa-se à análise de cada atributo do conjunto de dados fornecido:

- **tracking_record_id** – Identificador da transação, cujo propósito é distinguir todas as transações de forma única. Assume valores contínuos e comporta-se como a chave primária da tabela.
- **date_time** – Data em que o utilizador visitou a página Web, ou seja, sempre que um utilizador entra numa página da companhia a data é registada. Assume também valores contínuos.
- **user_gui** – Identificador exclusivo de utilizadores que já estejam inscritos. Este identificador apenas varia entre utilizadores já registados nas páginas da companhia. Como se trata de mais umidentificador, assume valores contínuos.
- **campaign_id** – Identificador da campanha promocional. Sempre que um ou mais produtos estejam em campanha, ela é identificada por este identificador, que mais uma vez, assume valores contínuos.
- **product_gui** – Identificador exclusivo do produto que é visitado por um determinado visitante, numa determinada sessão. Assume valroes contínuos.
- **company** – Nome da empresa que fornece o produto. Assume valores contínuos, que apesar de serem nomes de companhias, não são dados discretizados.
- **link** – ***URL*** da página da Web que foi visitada. Assume valroes contínuos.
- **tracking_id** – ***Encoder*** do ***browser***. Assume valores discretos, visto que os valores possíveis incidem nos vários (mas limitados) ***encoders*** existentes.
- **meio** – Meio pelo qual o visitante navegou. Assume valores discretos (os possíveis meios).
- **ip** – Endereço ***IP*** (Internet Protocol) de o visitante. Assume valores contínuos.
- **brower** – Navegador que o visitante utilizou. Assume valores discretos (browsers existentes).
- **session_id** – Identificador criado quando é iniciada uma nova sessão, podendo ser diferente para o mesmo visitante, indicando que ele visitou em diferentes alturas. Assume valroes contínuos.
- **cookie_id** – Identificador exclusivo global que serve para identificar o visitante. Assume valroes contínuos.

2.2. Regras de Associação entre Produtos

As regras de associação tirarão partido dos dados das transações para encontrar afinidades entre os produtos que são vendidos em simultâneo, assim como um problema "***Market-Basket Analysis***".

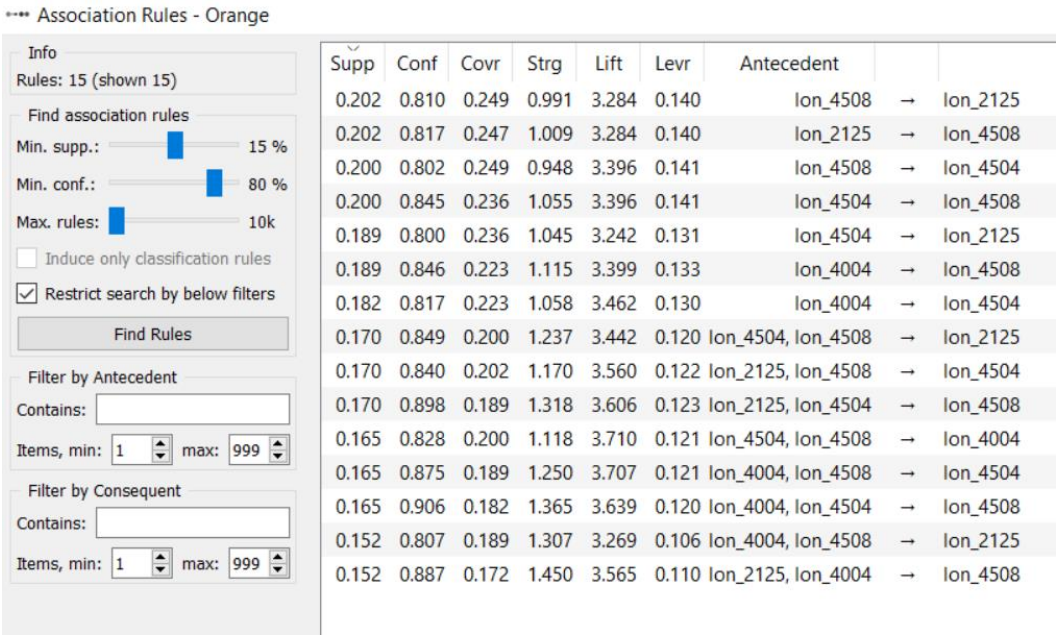
Market Basket Analysis descreve o ***Business Intelligence***, ou seja, são as informações do desempenho passado da empresa que são usadas para ajudar a prever o desempenho futuro da mesma. Isto pode revelar tendências emergentes das quais a empresa pode lucrar no futuro.

O suporte e a confiança das regras são duas medidas de associação que refletem a utilidade e a certeza das regras de associação descobertas. Normalmente, as regras de associação são consideradas interessantes se satisfazem um limite mínimo de suporte e um limite mínimo de confiança, que são definidos no ***software* Orange**.

O suporte mede a frequência dos produtos da associação, ou seja, mede a quantidade de vezes que os produtos ocorrem juntos numa transação.

A confiança mede a probabilidade de que o antecedente ocorra quando o consequente ocorre.

Em alguns casos, as medidas de suporte e de confiança são muito altas pelo que podem produzir uma regra não muito útil, por isso utiliza-se uma outra medida: o "***lift***" - que indica a força de uma regra. Se a medida "***lift***" for superior a **1**, significa que a regra de associação prevê um bom resultado, mas se for inferior a **1** a regra prevê um resultado não tão significativo.



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.202	0.810	0.249	0.991	3.284	0.140	Ion_4508	→	Ion_2125
0.202	0.817	0.247	1.009	3.284	0.140	Ion_2125	→	Ion_4508
0.200	0.802	0.249	0.948	3.396	0.141	Ion_4508	→	Ion_4504
0.200	0.845	0.236	1.055	3.396	0.141	Ion_4504	→	Ion_4508
0.189	0.800	0.236	1.045	3.242	0.131	Ion_4504	→	Ion_2125
0.189	0.846	0.223	1.115	3.399	0.133	Ion_4004	→	Ion_4508
0.182	0.817	0.223	1.058	3.462	0.130	Ion_4004	→	Ion_4504
0.170	0.849	0.200	1.237	3.442	0.120	Ion_4504, Ion_4508	→	Ion_2125
0.170	0.840	0.202	1.170	3.560	0.122	Ion_2125, Ion_4508	→	Ion_4504
0.170	0.898	0.189	1.318	3.606	0.123	Ion_2125, Ion_4504	→	Ion_4508
0.165	0.828	0.200	1.118	3.710	0.121	Ion_4504, Ion_4508	→	Ion_4004
0.165	0.875	0.189	1.250	3.707	0.121	Ion_4004, Ion_4508	→	Ion_4504
0.165	0.906	0.182	1.365	3.639	0.120	Ion_4004, Ion_4504	→	Ion_4508
0.152	0.807	0.189	1.307	3.269	0.106	Ion_4004, Ion_4508	→	Ion_2125
0.152	0.887	0.172	1.450	3.565	0.110	Ion_2125, Ion_4004	→	Ion_4508

Figura 1 - Antecedentes e Consequentes

Na **figura 1** anterior, observa-se os produtos antecedentes e os produtos consequentes na navegação de visitantes. São produtos que quando ocorrem numa transação (antecedentes), têm maior probabilidade de ocorrer juntamente com outros produtos (consequentes).

2.3. Agregação e Organização dos Dados

Nesta fase, foram reutilizados os procedimentos do **Modelo Prático 7 (MOP7)** onde se criaram ***views*** que agregaram dados específicos para melhor visualizar e analisar o ***Dataset***. Dessa análise, resultou o ficheiro **remove_useless_produts.py**, que serve para remover os produtos que não interessam.

Os ***scripts*** utilizados nesta fase encontram-se na pasta com o nome ***/scripts***.

2.4. Dados mais relevantes

Para gerar um subconjunto de dados mais relevante, criou-se, nesta fase, um ***script*** que mostra todos os eventos gerados pelos visitantes com o número de sessões. Este script é enviado em anexo com o nome de "03_script_CREATE_VIEW".

2.5. Geração do Conjunto de Dados

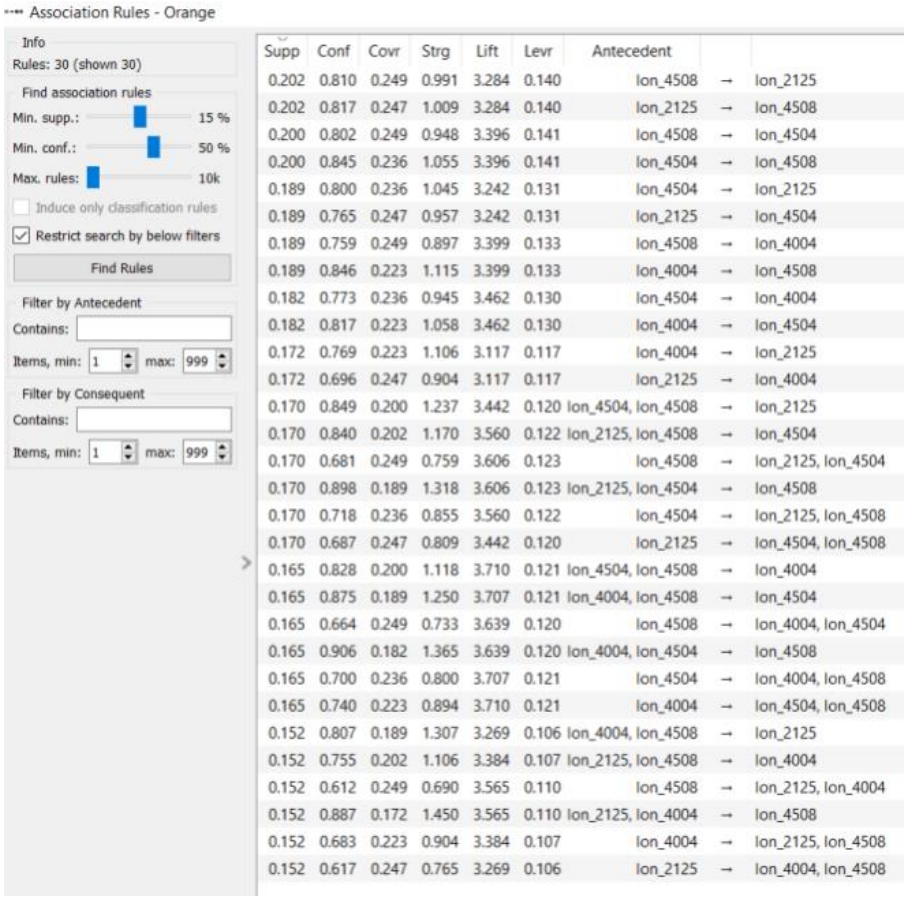
Para gerar um ***dataset*** com as transações registadas que foram filtradas de acordo com o critério definido, foi desenvolvido o ***script*** de nome "03_script_CREATE_VIEW".

2.6. Processamento dos Dados

Em primeiro lugar é necessário realizar um processo de normalização das ***Strings*** que descrevem cada atributo, como por exemplo eliminar os espaços, acentos, e colocar todas as letras em minúsculas. Em segundo lugar, foi necessário, através de um script Python (de nome "_goPy_transform_v02.py"), gerar o ficheiro ". basket".

2.7. Marketing Decisions

Ao analisar a tabela da **figura 2** das regras de associação do **Orange** podemos verificar que no topo, os produtos na coluna **consequent** são os que têm maior probabilidade de serem visitados após o visitante ter visualizado os produtos na coluna **antecedent**. Logo, os produtos no topo destas colunas devem continuar juntos pois contêm mais visualizações, que se traduzirão em mais lucro para a empresa.



Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		
0.202	0.810	0.249	0.991	3.284	0.140	Ion_4508	→	Ion_2125
0.202	0.817	0.247	1.009	3.284	0.140	Ion_2125	→	Ion_4508
0.200	0.802	0.249	0.948	3.396	0.141	Ion_4508	→	Ion_4504
0.200	0.845	0.236	1.055	3.396	0.141	Ion_4504	→	Ion_4508
0.189	0.800	0.236	1.045	3.242	0.131	Ion_4504	→	Ion_2125
0.189	0.765	0.247	0.957	3.242	0.131	Ion_2125	→	Ion_4504
0.189	0.759	0.249	0.897	3.399	0.133	Ion_4508	→	Ion_4004
0.189	0.846	0.223	1.115	3.399	0.133	Ion_4004	→	Ion_4508
0.182	0.773	0.236	0.945	3.462	0.130	Ion_4504	→	Ion_4004
0.182	0.817	0.223	1.058	3.462	0.130	Ion_4004	→	Ion_4504
0.172	0.769	0.223	1.106	3.117	0.117	Ion_4004	→	Ion_2125
0.172	0.696	0.247	0.904	3.117	0.117	Ion_2125	→	Ion_4004
0.170	0.849	0.200	1.237	3.442	0.120	Ion_4504, Ion_4508	→	Ion_2125
0.170	0.840	0.202	1.170	3.560	0.122	Ion_2125, Ion_4508	→	Ion_4504
0.170	0.681	0.249	0.759	3.606	0.123	Ion_4508	→	Ion_2125, Ion_4504
0.170	0.898	0.189	1.318	3.606	0.123	Ion_2125, Ion_4504	→	Ion_4508
0.170	0.718	0.236	0.855	3.560	0.122	Ion_4504	→	Ion_2125, Ion_4508
0.170	0.687	0.247	0.809	3.442	0.120	Ion_2125	→	Ion_4504, Ion_4508
0.165	0.828	0.200	1.118	3.710	0.121	Ion_4504, Ion_4508	→	Ion_4004
0.165	0.875	0.189	1.250	3.707	0.121	Ion_4004, Ion_4508	→	Ion_4504
0.165	0.664	0.249	0.733	3.639	0.120	Ion_4508	→	Ion_4004, Ion_4504
0.165	0.906	0.182	1.365	3.639	0.120	Ion_4004, Ion_4504	→	Ion_4508
0.165	0.700	0.236	0.800	3.707	0.121	Ion_4504	→	Ion_4004, Ion_4508
0.165	0.740	0.223	0.894	3.710	0.121	Ion_4004	→	Ion_4504, Ion_4508
0.152	0.807	0.189	1.307	3.269	0.106	Ion_4004, Ion_4508	→	Ion_2125
0.152	0.755	0.202	1.106	3.384	0.107	Ion_2125, Ion_4508	→	Ion_4004
0.152	0.612	0.249	0.690	3.565	0.110	Ion_4508	→	Ion_2125, Ion_4004
0.152	0.887	0.172	1.450	3.565	0.110	Ion_2125, Ion_4004	→	Ion_4508
0.152	0.683	0.223	0.904	3.384	0.107	Ion_4004	→	Ion_2125, Ion_4508
0.152	0.617	0.247	0.765	3.269	0.106	Ion_2125	→	Ion_4004, Ion_4508

Figura 2 - Tabela de regras de associação

- **Que produtos são raramente visitados, mas normalmente são visitados sequencialmente?**

Ao analisar a tabela das regras de associação do **Orange** da **figura 2**, podemos verificar que os produtos que normalmente não aparecem na coluna **antecedente**, mas que aparecem frequentemente na coluna **consequent**, são produtos que são mais frequentemente visitados como consequência da visita de outro produto. Isto é, normalmente, o interesse do visitante por esse produto é despertado por um outro produto antecedente.

- **Quais são os produtos mais visitados?**

Para deduzir acerca dos produtos que mais foram visitados, gerou-se a ***view*** da **figura 3**. Esta ***view*** ilustra todas as visitas aos produtos de todas as transações registadas. Desta forma, conclui-se que os produtos mais procurados foram, em geral, pertencentes à categoria do calçado, da roupa, da tecnologia e do divertimento.

	product_gui character varying	product_count bigint
1	Botas	6391
2	Botins	6294
3	Pumps e Open Toes	6001
4	Tecnologia	4682
5	Outlet	4582
6	Divertimento	3855
7	Botas Rasas e Cavaleiro	2941
8	Sabrinas e Mocassins	2665
9	Spy	2608
10	Estilo	2287
11	Botas de Salto	2063
12	Green 21	1999
13	Primavera/Verão	1938
14	Robots	1870
15	Segurança	1860

Figura 3 - Produtos mais procurados

2.8. Relatório de Análise do Mercado da We-Commerce

Para gerar um relatório gráfico do problema "***Market-Basket Analysis***" no **Orange**, recorreu-se ao ficheiro resultante dos capítulos anteriores, originando o relatório "market-basket analysis.report" enviado em anexo.

3. Conclusão

A conversão de dados para conhecimento é um processo muito útil, que permite organizar e analisar um conjunto de dados com tamanho significativo, de modo a tirar conclusões acerca do mesmo. Recorrendo a tecnologias como Python, SQL e Orange, concluiu-se que a mineração de dados, hoje em dia, se revela cada vez mais um processo importante em empresas que armazenam grandes quantias de dados.