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

# 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.

# Projeto

## 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.

## 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.

Uma imagem com mesa

Descrição gerada automaticamente

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).

## 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***.

## 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”.

## 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”.

## 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”.

## 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.

Uma imagem com mesa

Descrição gerada automaticamente

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.

Uma imagem com mesa

Descrição gerada automaticamente

Figura 3 - Produtos mais procurados

## 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.

# 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.