ADDETC – Área Departamental de Engenharia Eletrónica e Telecomunicações e de Computadores

MEIM - Mestrado Engenharia informática e multimédia

Aprendizagem e Mineração de Dados

Projeto final B

Turma:

MEIM-11D

Trabalho realizado por:

Grupo 10

Miguel Távora Nº45102

Duarte Domingues Nº45140

Docente:

Artur Ferreira

Data: 17/01/2022

Índice

Índice de ilustrações

1. Introdução

Este projeto baseia-se no processamento da informação e extração de conhecimento de um conjunto de dados recolhidos por um grupo de comércio eletrónico. O grupo é chamado "*We-Commerce*" e tem um conjunto de páginas Web. Ao longo do tempo o grupo tem colecionado um grande volume de dados de visitantes. O grupo captura e guarda cada evento gerado pelos visitantes que navegam pelas páginas Web.

São guardados diferentes atributos para cada evento de um visitante numa das páginas Web, na base de dados. Os atributos mais importantes são os seguintes:

- **cookie_id** – Identificador da *cookie* do visitante. Para cada visita de um visitante pela primeira vez numa das páginas Web é guardada uma *cookie* no seu browser. O valor da *cookie* é um *globally uniqdue identifier* (GUID). Próxima vez que o visitante visita uma das páginas, o GUID da *cookie* é retirado do browser e é guardado na base de dados neste atributo.
- **session_id** – Identificador da sessão. O identificador de sessão é criado sempre que o visitante entra numa das páginas Web. O "session_id" entra na base de dados associado com o "cookie_id". Uma sessão expira passado algum tempo de inatividade do utilizador ou quando o browser fecha, portanto, "cookie_id" pode estar associado a diferentes "session_id".
- **product_gui** – Identificador único de um produto.
- **tracking_record_id -** Cada registo de data é atribuído um único identificador, cada registo representa um evento.

Neste projeto foram abordadas regras de associação para descobrir relações entre produtos visitados, em cada sessão dos visitantes numa das páginas Web, através da análise de dados de um cesto de compras. O objetivo desta análise é identificar padrões de comportamento dos visitantes e identificar os produtos visitados frequentemente em conjunto. A partir desta análise o grupo "*We-Commerce*" pode realizar decisões de negócio vantajosas a partir do conhecimento fornecido.

Para encontrar potencias regras de associação, é utilizado um algoritmo chamado "Algoritmo Apriori", este algoritmo serve para encontrar regras de associação a partir dos conjuntos de itens das transições mais frequentes.

O formato de dados correto para análise de dados de um cesto de compras é o *basket*. Em ficheiros *basket* cada linha é escrita como uma lista de pares nome-valor separados por virgula. De forma a poder realizar a análise de cesto de compras foi necessário converter o conjunto de dados para este formato.

O projeto dividiu-se nas seguintes etapas:

- Análise da base de dados fornecida pelo grupo "*We-Commerce*" e verificação dos significados dos diferentes atributos.
- Geração de um ficheiro que permite associar o identificador da cookie de um visitante e o identificador do produto para um evento numa página Web, de forma a manter apenas os dados mais relevantes do conjunto de dados.
- Processamento do ficheiro mencionado anteriormente, normalizando as "*strings*" que descrevem cada produto, eliminando espaços brancos, acentos e convertendo tudo para letras minúsculas. De seguida transformação do conjunto de dados do ficheiro para o formato *basket*.
- Realização da análise de dados de cesto de compras, utilizando um conjunto de métricas como por exemplo o suporte, confiança e *lift* de forma a encontrar as regras de associação mais fortes.
- Realização de um relatório em Orange que demonstra os resultados e conclusões obtidas na análise de dados de um cestro de compras.
- Realização de um *script* em Python para automaticamente realizar o processamento e conversão dos dados para o formato *basket* e *tab*.

2. Desenvolvimento

Processamento do conjunto de dados

De forma a poder realizar a análise de dados de cesto de compras é necessário inicialmente agrupar os eventos realizados por um visitante com um determinado identificador de *cookie*, associando "cookie_id" com "product_gui". Desta forma é possível analisar os produtos que ocorrem em conjunto em cada transição do visitante numa das páginas Web. Contudo, de forma a manter apenas os dados mais relevantes selecionaram-se apenas os visitantes com um número de sessões entre 5 e 30. Ao realizar esta filtração está-se apenas a considerar a informação dos visitantes mais importantes. Limitando o número máximo de sessões de um visitante a 30, impede considerar demasiado os dados de um único visitante, impossibilitando também evitar considerar informação de possíveis "*bots*" ou administradores que podem visitar as páginas Web numa quantidade muita elevada de vezes. Isto poderia ser prejudicável para a análise dos dados.

Para agrupar o conjunto de dados da maneira mencionada foi necessário criar primeiro um conjunto de *views* auxiliares. As *views* auxiliares realizadas mais importantes foram as seguintes:

- **v_cookie_session_number_of_events(cookie_id,session_id,number_of_events_per_session)** – Agrega visitantes com as sessões realizadas e obtém o número total de eventos.
- **v_cookie_number_of_sessions(cookie_id, number_of_sessions, number_of_events)** – Agrega um visitante com o número de sessões e o número de eventos realizados.
- **v_number_of_cookies_number_of_sessions(number_of_cookies, number_of_sessions)** – Agrega para o número de sessões o número de visitantes, por exemplo o número de visitantes que só realizaram uma sessão é 240911.
- **v_export(cookie_id, session_id, product_gui) –** Esta *view* irá ser utilizada para exportar o conjunto de dados para o ficheiro. Esta *view* agrupa o visitante com as sessões realizadas e os produtos, obtendo então os eventos realizados por um visitante. Porém são apenas considerados os visitantes que tenham um número de sessões entre 5 e 30.

Por fim, apôs ter os dados preparados foi necessário exportá-los para um ficheiro de texto. No ficheiro de texto cada linha equivale a um evento do visitante, associando o identificador da *cookie* do visitante ao identificador do produto (cookie_id, product_gui) num evento.

Transformação dos dados para o formato basket

O formato correto de dados para análise de dados de um cesto de compras é o *basket*. Neste tipo de ficheiros, cada linha é escrita como uma lista de pares nome-valor separados por virgula. Os dados no ficheiro *basket* são uma coleção de todos os produtos presentes numa transação, com o número de ocorrências associado. Pelas razões mencionadas foi necessário converter os dados para o formato *basket*.

Inicialmente começou-se por normalizar as "*strings*" que descrevem cada produto, eliminando espaços brancos, acentos e convertendo tudo para letras minúsculas. De seguida passou-se para o processo de geração do ficheiro *basket*.

O processo de geração do ficheiro *basket* seguiu as seguintes etapas:

1. Organizar os dados num formato que associa cada transação com os produtos e o número de ocorrências dos respetivos produtos. Para isto, cada transação corresponde a um dicionário. Em cada dicionário a chave é o identificador da transição ("cookie_id") e os valores são pares (produto, ocorrência) dos produtos.
2. De seguida é necessário escrever a informação para um ficheiro *basket*. Para cada transação é escrita uma linha no ficheiro *basket*, com todos os produtos que ocorreram na transação, separados por vírgulas. No caso de o produto ter ocorrência superior a 1, escreve-se à frente do nome do produto ("=n") em que n é o número de ocorrências.

Além da geração do ficheiro *basket*, foi também gerado um ficheiro ".tab" com a matriz esparsa com a taxa de ocorrência de cada produto em cada transação. O número de linhas é igual ao número de transações e o número de colunas equivale ao número de diferentes produtos.

Por fim foi realizado um script Python que permite automatizar todo este processo, sendo apenas necessário correr este ficheiro para obter o ficheiro "tab" e o ficheiro "basket" a partir do conjunto de dados inicial.

Análise de dados de cesto de compras

Finalmente, apôs os dados estarem no formato correto realizou-se a análise de dados de cesto de compras. A análise de dados de cesto de compras teve como objetivo encontrar padrões entre os diferentes produtos nas transações e em comportamentos dos visitantes. Este tipo de análise baseia-se na aplicação e identificação de regras de associação. Regras de associação são usadas em dados de transações, e têm o intuito de identificar relações e conexões entre os diferentes produtos do conjunto de dados.

Um exemplo de uma regra de associação do conjunto de dados pode ser por exemplo:

```
display.product*ljv_0459*helicopterotelecomandadoatravesiphone3/4/4s, eandroid  →  ipodtouch, ipad1/2
```

Como se pode observar visitantes que visitaram na transação produtos relacionados com iphone, neste caso um helicóptero telecomandado através iphone e android eletrónico pesquisaram também por produtos relacionados com o iphone, neste caso, ipodtouch e ipad.

Entretanto, para uma regra de associação ser útil para o cliente é necessário recolher evidência suficiente que a regra irá se útil e lucrativa para a companhia. Portanto a força das regras é medida através de um conjunto de diferentes métricas.

As métricas principais e as que foram tidas em conta nesta análise foram:

- Suporte – Mede quão frequentemente a coleção de itens numa associação, ocorrem juntos como uma percentagem de uma transação. Regras com maior suporte são preferíveis, pois são mais prováveis de ser aplicáveis a um grande número de futuras transações.
- Confiança – Probabilidade de ocorrer os itens no lado direito da regra, sabendo que os itens do lado da esquerda ocorreram. Quanto maior for a confiança, maior é a probabilidade de o item do lado direito da regra ocorrer, logo há um maior retorno.
- *Lift* – Divisão da confiança pela confiança esperada, confiança esperada corresponde à percentagem do número de ocorrências dos itens do lado direito em todas as transações (suporte). O *Lift* indica a força da associação entre produtos no lado esquerdo e lado direito da regra, se o valor da *lift* for inferior a 1 os itens têm uma correlação negativa, portanto quanto maior for o seu valor, melhor.

A análise de cesto de compras foi realizada na Orange API, a partir dos *Association Rules e Frequent Itemsets widgets*. A partir do Orange é possível obter os conjuntos de itens e as diferentes regras de associação, consoante um valor mínimo para o suporte e a confiança.

Inicialmente começou-se por analisar o conjunto de dados. O conjunto de dados tem 1282 instâncias e 2125 *features*. A primeira etapa da análise foi procurar os produtos que são visitados mais frequentemente, respetivamente os 15 produtos mais frequentes. Os produtos mais frequentes são os produtos com maior suporte.

O item mais frequente foi a *homepage*, o que seria de esperar visto que em páginas Web é normal que os visitantes passem muitas vezes por esta página. Outros produtos que apareceram muitas vezes, foram itens relacionados com botas, por exemplo, botins e sabrinas, isto é um indicador que o grupo "We-Commerce" tem uma grande parte do seu negócio no comércio de botas.

Itemsets	Support	%
display.category*homepage	325	25.35
lon_4508	154	12.01
botins	147	11.47
lon_2125	147	11.47
botas	135	10.53
pumpseopentoes	134	10.45
lon_4504	131	10.22
lon_4004	127	9.906
outlet	115	8.97
lon_4013	104	8.112
intro*fbpage*link1	104	8.112
lon_4108	96	7.488
lon_2119	93	7.254
display.product*lon_4508*botaempelecastanhamogno-luisonofre	90	7.02
display.product*lon_4504*botasemcamurcaverdeelacolateral-luisonofre	89	6.942
tecnologia	86	6.708
display.product*lon_2125*botasluisonofreempeletaupeesaltosemmadeira	76	5.928
display.product*lon_4004*pumpsempelecrococastanho-luisonofre	76	5.928
lon_4401	75	5.85
sabrinasemocassins	75	5.85

Figura 1 - 20 *itemsets* mais frequentes

Outra área que oferece um bom suporte é tecnologia, que pode ser uma área também de possível interesse para o grupo.

tecnologia	86	6.708
display.category*homepage	57	4.446
green21	46	3.588
robots	43	3.354
spy	42	3.276
seguranca	42	3.276
stockoff	37	2.886

Figura 2 - *itemset* tecnologia

Itens que aparecem em conjunto com tecnologia, são por exemplo, green21, robots, espionagem e segurança. O que é um indicador dos visitantes na área de tecnologia mostrarem interesse em áreas da informática.

De seguida foram abordadas as regras de associação. De forma a cortar as regras que ocorrem menos vezes que não oferecem muito valor, utilizou-se um valor minimo de suporte de 2%. Tem-se o objetivo das regras descobertas serem verdadeiras na maior parte dos casos, então escolheu-se um valor de confiança relativamento alto, de 60%. Estes parâmetros, permitem cortar um grande número de regras que não teriam muito interesse para o grupo, removendo tempo de computação desnecessário para correr os algoritmos e permitindo focar a análise apenas em regras mais fortes. A partir destes parametros há uma boa hipotese de serem geradas regras relevantes.

Defeniu-se o número de itens minimos a 2 e máximos a 4 no antecendente e 1 item minimo e 4 máximos no consequente. Um valor de itens no consequente superior a 4 poderia causar confusão a entender a utilidade da regra. A partir destes parâmetros são geradas 2534 regras.

De modo a poder limitar o número de regras, para ter apenas as regras mais relevantes, defeniu-se o número máximo de regras a 10000 como o valor minimo possivel no Orange, porém isto é indeferente, pois como foi mencionado anteriormente o número de regras obtidas foi só 2534.

Min. suporte	Min. confiança	Max. regras
2.00%	60%	10000

Figura 3 - Parâmetros utilizados no *Association Rules*

Apesar de existirem um número elevado de regras, para ser possivel observar cada uma individualmente, podemos observar as 8 regras com maior *lift*.

Antecendente	Consequente	Suporte	Confiança	*Lift*
lon_2125, display.product*lon_4504* botasemcamurcaverdeelacolateral-luisonofre	lon_4504, display.product*lon_2125* botasluisonofreempeletaupeesaltosemmadeira	0.023	0.935	39.976
lon_4504, display.product*lon_2125* botasluisonofreempeletaupeesaltosemmadeira	lon_2125, display.product*lon_4504* botasemcamurcaverdeelacolateral-luisonofre	0.023	0.967	39.976
estilo, divertimento, seguranca	jogos, robots	0.02	0.867	38.313
jogos, robots	estilo, divertimento, seguranca	0.02	0.897	38.313
estilo, divertimento, tecnologia, seguranca	jogos, robots	0.02	0.867	38.313
estilo, divertimento, seguranca	jogos, tecnologia, robots	0.02	0.867	38.313
jogos, tecnologia, robots	estilo, divertimento, seguranca	0.02	0.897	38.313
jogos, robots	estilo, divertimento, tecnologia, seguranca	0.02	0.897	38.313

Figura 4 - 5 regras com maior *lift*

Estas regras mostram ter um sentido intuitivo, as duas primeiras regras parecem representar itens relacionados com a transação de itens relacionados com botas, estas regras podem ser uteis para o "We-Commerce" ter noção que itens os visitantes costumam visitar em conjunto.

De seguida tencionou-se encontrar as regras que são raramente visitadas, mas são muitas vezes visitadas em conjunto. Para isto, escolheu-se regras com baixo suporte e alto valor de confiança. Para o antecendente defeniu-se o número de itens máximos a 3 e apenas 1 item máximo para o consequente. As 5 melhores regras obtidas foram as seguintes:

Antecendente	Consequente	Suporte	Confiança	*Lift*
jogos, robots, green21	divertimento	0.02	1	18.314
estilo, divertimento, rc	tecnologia	0.02	1	14.907
estilo, robots, spy	tecnologia	0.02	1	14.907
jogos, robots, green21	tecnologia	0.02	1	14.907
seguranca, green21, display.category*homepage	tecnologia	0.02	1	14.907

Figura 5 - Regras baixo suporte / alta confiança

Para facilitar o processo de identificação regras uteis foi realizado um gráfico que ilustra a relação entre as diferentes métricas, a cor dos pontos representa o *lift*.

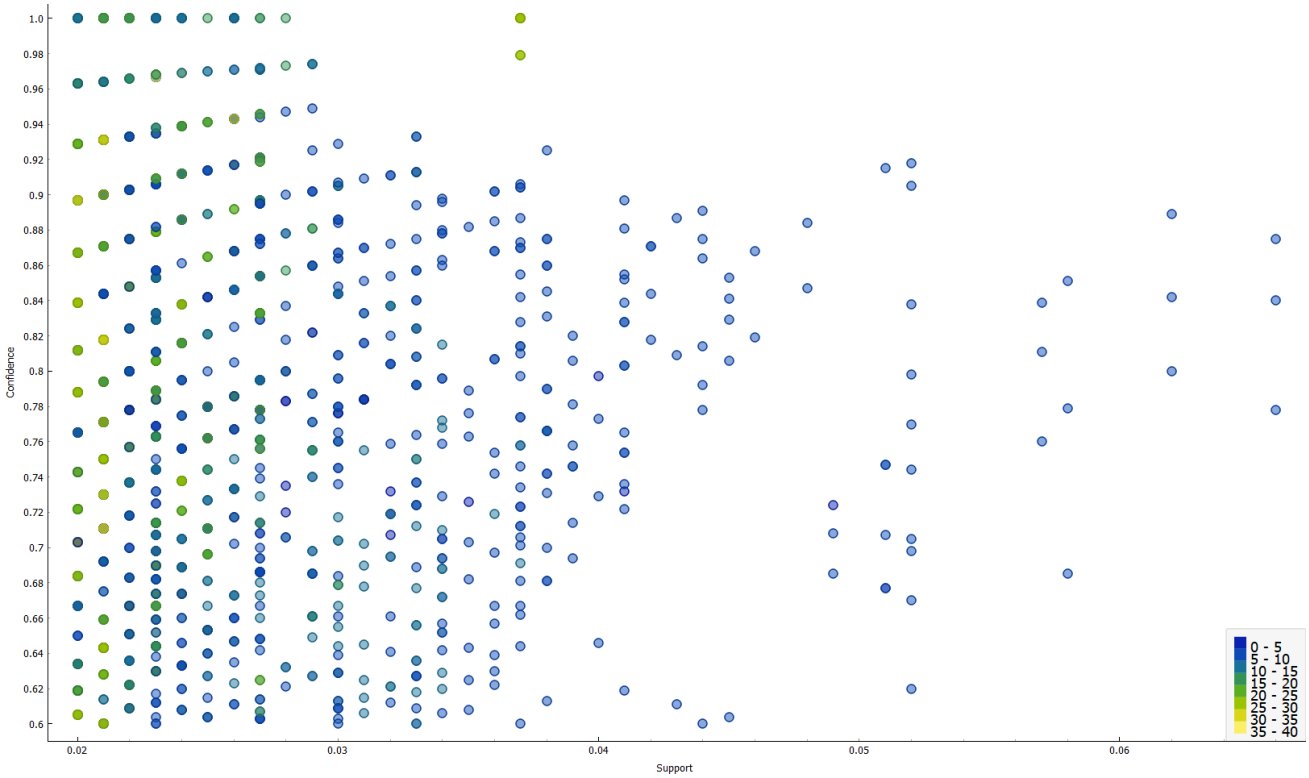

Figura 6 - *Scatter plot* das regras em relação com as métricas

As regras ideias vão ser aquelas que se apresentam o quanto mais próximo possivel do canto superior direito, onde a confiança e o suporte é maximo. No neste gráfico pode-se passar por cima de cada ponto e vizualisar a regra em questão, o que é bastante eficiente para encontrar regras uteis. Como se pode observar na seguinte figura:

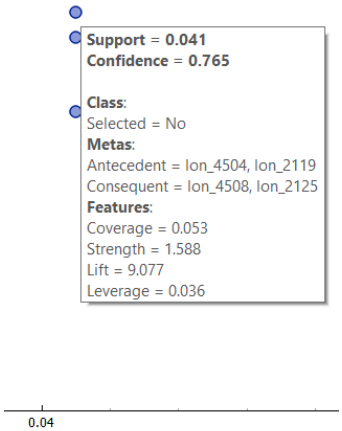

Figura 7 - Exemplo da visualização de uma regra no gráfico

Também foi importante verificar as regras com maior suporte, quando o antecendete contém a página principal. Desta forma é possível verificar os comportamentos dos visitantes ao entrarem na página Web e os itens que lhes despertam maior interesse.

Antecendente	Consequente	Suporte	Confiança	*Lift*
botins, display.category*homepage	pumpseopentoes	0.051	0.707	6.759
display.category*homepage, pumpseopentoes	botins	0.051	0.747	6.516
botins, display.category*homepage	botas	0.049	0.685	6.503
display.category*homepage, botas	botins	0.049	0.708	6.173
display.category*homepage, outlet	botas	0.04	0.646	6.131
botins, display.category*homepage, botas	pumpseopentoes	0.037	0.746	7.137
botins, display.category*homepage, pumpseopentoes	botas	0.037	0.723	6.867
display.category*homepage, pumpseopentoes, botas	botins	0.037	0.904	7.883
display.category*homepage, sabrinasemocassins	botins	0.032	0.804	7.011
display.category*homepage, sabrinasemocassins	pumpseopentoes	0.032	0.804	7.691

Figura 8 - Regras com maior suporte com a página principal como antecedente

Como se pode observar, associado com a página principal está muitas vezes botins e o item sabrinas e mocassins, um tipo de botas, que leva os visitantes a visitarem itens de botas. Isto desmonstra que é algo que este item é algo que captiva a atenção dos visitantes. A partir destas descobertas o grupo "We-Commerce" pode realizar estratégias de *marketing* para alcançar melhor o seu cliente alvo.

Por fim analisou-se as relações de um para um entre itens que ocorrem mais vezes (1 item -> 1 item). Para isto procorou-se as regras com maior suporte.

Antecendente	Consequente	Suporte	Confiança	*Lift*
lon_2125	lon_4508	0.084	0.735	6.116
lon_4508	lon_2125	0.084	0.701	6.116
lon_4504	lon_4508	0.078	0.763	6.355
lon_4508	lon_4504	0.078	0.649	6.355
pumpseopentoes	botins	0.075	0.716	6.391
botins	pumpseopentoes	0.075	0.653	6.391
lon_4504	lon_2125	0.075	0.733	6.248
lon_2125	lon_4504	0.075	0.653	6.248

Figura 9 - Regras com maior suporte (1->1)

3. Conclusões

Em suma neste projeto foram utilizadas e consolidadas diferentes técnicas de processamento de dados, implementação de algoritmos, extração de padrões e conhecimento de regras de associação.

A partir da análise de cesto de compras foi possível obter um conjunto de informação util, relacionada com o comportamento e padrões de utilização dos diferentes visitantes nas páginas Web do grupo “We-Commerce”. Descobriu-se os itens visitados mais frequentemente, regras de associação fortes entre alguns produtos, itens que devem ser mantidos juntos nas páginas Web e comportamentos frequentes dos visitantes ao entrarem nas páginas Web. Pela análise, é possivel também notar que a área de produtos com maior interesse dos visitantes são: botas, tecnologia e divertimento.

A análise de cesto de compras é uma ferramenta util para encontrar relações entre produtos que ocorrem em transações de visitantes. Este tipo de análise é muito util para qualquer empresa ou forma de comércio electrónico, sendo que um bom entendimento das necessidades e preferências do consumidor é chave para um negócio de sucesso.

4. Bibliografia

Paulo Trigo Silva, Aprendizagem e Mineração de Dados,
b06en_algorithmsSearchForAssociationRules_v01.pdf
Market Basket Analysis: Understanding Customer Behaviour:
https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/

A Comprehensive Guide on Market Basket Analysis:

https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/