

Extração de Regras de Associação no Consumo de Filmes



MARIA EDUARDA MACIEL
RENATA CAVALHEIRO DA SILVA

Introdução

Contexto

- O streaming dos filmes cada vez mais comum
 - Objetivo de construir um perfil de consumo de seus usuários
 - Relevância da funcionalidade levou ao desenvolvimento de sistemas de recomendação
 - Análise geral ao invés de apenas a individual, considerando o cenário geral de consumo
-



Técnicas de mineração de
dados e de regras de
associação

Algoritmos para mineração de regras de associação:
Partition, FP-Growth, Eclat e Apriori

Introdução

Fundamentação Teórica

REGRAS DE ASSOCIAÇÃO

Dado um conjunto de transações, sendo cada transação um conjunto de itens e sendo X e Y dois conjuntos de itens, uma regra de associação seria $X \rightarrow Y$.

SUORTE E CONFIANÇA

Suporte é a frequência com a qual são localizados padrões específicos entre os dados da base. A confiança mede a força das regras, ou seja, a quantidade percentual de um determinado $X \rightarrow Y$.

MINSUP E MINCONF

minsup: é a parcela de transações que satisfaz a união dos itens do antecedente com o consequente;
minconf: define que no mínimo $c\%$ das transações que satisfazem o antecedente de uma regra, também satisfaçam o consequente dessa regra.

Métodos

SELEÇÃO E PRÉ-PROCESSAMENTO DA BASE DE DADOS

- Primeira etapa: selecionar a base de dados que seria utilizada para o experimento
- Base escolhida: "*Large Movie Dataset*", disponível no *Kaggle*
- A base possui os dados de mais de 7.000 usuários, mais de 200.000 títulos - sendo desses, 58.958 filmes diferentes
- Para esse estudo, foram utilizados os dados de consumo de 5.000 usuários, levando em consideração apenas os títulos assistidos

ESCOLHA DA TÉCNICA E DO ALGORITMO

- O algoritmo aplicado nesse estudo foi o *Apriori*
- Implementado em Python, utilizando a biblioteca *Apyori*, disponível no *Python Package Index (PyPI)*
- O *Apriori* funciona encontrando todos os conjuntos de itens frequentes de uma base de dados, e utiliza funções para criar os candidatos e eliminar os itens que não são frequentes dentro dos conjuntos
- Foram retirados do conjunto os usuários com o identificador (id) maior do que 5000
- Nova tabela de dados utilizando 4 passos

Tabela utilizada

PASSOS

1. Criação de uma lista para adicionar todos os filmes assistidos por cada usuário
2. Iteração na base de dados de zero até 5.000, adicionando os filmes de acordo com o id do usuário
3. Uma lista para adicionar cada lista de filmes de cada usuário
4. Transformação dessa lista de listas em uma tabela de dados em que a primeira coluna corresponde ao id do usuário e as colunas subsequentes contém cada uma o título de um filme

PARÂMETROS DO ALGORITMO

- min_support: 0,1; o que corresponde à análise de 500 usuários que assistiram determinado filme dentre os 5.000, chamado de suporte (ou seja, $500/5000 = 0,1$);
 - min_confidence: 0,4; é a medida da proporção dos usuários que assistiram ao filme X e que também assistiram ao filme Y. A confiança entre dois itens (filme X e filme Y) é definida como o número total de usuários que assistiram tanto o filme X quanto o Y, sobre o número total de usuários que assistiram X. Isto é: $\frac{\text{número de usuários que assistiram X e Y}}{\text{número de usuários que assistiram X}}$
 - min_lift: 2; é a relação entre a confiança e o suporte, dado por: $\frac{\text{confiança}(X \rightarrow Y)}{\text{suporte}(Y)}$
 - min_length: 10; é o número mínimo de filmes que devem estar na lista de filmes assistidos de cada usuário
-

Resultados



COM OS PARÂMETROS

Confiança dos filmes assistidos de pelo menos 40%, o que indica que de todas as pessoas que assistiram ao filme X, o resultado deve mostrar pelo menos 40% que também assistiram ao filme Y; e número de filmes no mínimo 10 para evitar resultados apenas com sequências (por exemplo, se fosse selecionado três como min_length, teríamos uma sequência óbvia de filmes nas regras como: assistir o primeiro “Senhor dos Anéis” implica em assistir também o segundo “Senhor dos Anéis”)

APÓS A EXECUÇÃO DO ALGORITMO

A estrutura de exibição dessa regra mostra primeiramente quais foram os dois itens associados e o valor de suporte



Resultados

```
Association 1 is: RelationRecord(items=frozenset({'2001: A Space Odyssey (1968)', 'Alien (1979)'}), support=0.10702140428085617, ordered_statistics=[OrderedStatistic(items_base=frozenset({'2001: A Space Odyssey (1968)'}), items_add=frozenset({'Alien (1979)'}), confidence=0.6430288461538463, lift=2.9654070128441674), OrderedStatistic(items_base=frozenset({'Alien (1979)'}), items_add=frozenset({'2001: A Space Odyssey (1968)'}), confidence=0.4935424354243543, lift=2.9654070128441674)])
```

```
-----
Association 2 is: RelationRecord(items=frozenset({'2001: A Space Odyssey (1968)', 'Back to the Future (1985)'}), support=0.11002200440088018, ordered_statistics=[OrderedStatistic(items_base=frozenset({'2001: A Space Odyssey (1968)'}), items_add=frozenset({'Back to the Future (1985)'}), confidence=0.6610576923076924, lift=2.2480458529565674)])
```

```
-----
Association 3 is: RelationRecord(items=frozenset({'2001: A Space Odyssey (1968)', 'Blade Runner (1982)'}), support=0.11302260452090418, ordered_statistics=[OrderedStatistic(items_base=frozenset({'2001: A Space Odyssey (1968)'}), items_add=frozenset({'Blade Runner (1982)'}), confidence=0.6790865384615385, lift=3.0945794036182597), OrderedStatistic(items_base=frozenset({'Blade Runner (1982)'}), items_add=frozenset({'2001: A Space Odyssey (1968)'}), confidence=0.5150410209662716, lift=3.0945794036182592)])
```

```
-----
Association 4 is: RelationRecord(items=frozenset({'2001: A Space Odyssey (1968)', 'Godfather, The (1972)'}), support=0.10802160432086418, ordered_statistics=[OrderedStatistic(items_base=frozenset({'2001: A Space Odyssey (1968)'}), items_add=frozenset({'Godfather, The (1972)'}), confidence=0.6490384615384616, lift=2.1054790845105575)])
```

Resultados

Regra	Suporte	Antecedente	Consequente	Confiança	Lift
Regra de associação 1	0.1070	2001: A Space Odyssey (1968)	Alien (1979)	0.6430	2.9654
		Alien (1979)	2001: A Space Odyssey (1968)	0.4935	2.9654
Regra de associação 2	0.1100	2001: A Space Odyssey (1968)	Back to the Future (1985)	0.6610	2.2480
	0.1130	2001: A Space Odyssey (1968)	Blade Runner (1982)	0.6790	3.0945
Regra de associação 3		Blade Runner (1982)	2001: A Space Odyssey (1968)	0.5150	3.0945
Regra de associação 4	0.1080	2001: A Space Odyssey (1968)	Godfather, The (1972)	0.6490	2.1054
Regra de associação 5	0.1138	2001: A Space Odyssey (1968)	Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)	0.6838	2.0657

Resultados

Regra de associação 6	0.1218	2001: A Space Odyssey (1968)	Star Wars: Episode V - The Empire Strikes Back (1980)	0.7319	2.0957
Regra de associação 7	0.1122	2001: A Space Odyssey (1968)	Star Wars: Episode VI - Return of the Jedi (1983)	0.6742	2.0453
Regra de associação 8	0.1012	2001: A Space Odyssey (1968)	Terminator, The (1984)	0.6081	2.6739
		Terminator, The (1984)	2001: A Space Odyssey (1968)	0.4450	2.6739
Regra de associação 9	0.1076	Ace Ventura: Pet Detective (1994)	Ace Ventura: When Nature Calls (1995)	0.4487	3.4141
		Ace Ventura: When Nature Calls (1995))	Ace Ventura: Pet Detective (1994)	0.8188	3.4141
Regra de associação 10	0.1504	Aladdin (1992)	Ace Ventura: Pet Detective (1994)	0.6271	2.3241
		Ace Ventura: Pet Detective (1994)	Aladdin (1992)	0.5574	2.3241

Conclusão



Recomendar um filme aleatório pode não ser uma boa ideia para manter os usuários ativos na plataforma, já que o filme pode não ser interessante para quem está assistindo



Sistemas de recomendação podem ser úteis para manter os clientes fiéis e interessados, que irão enxergar na plataforma que assinam um bom catálogo com filmes que atendem aos seus gostos



Pode-se criar uma lista com algumas sugestões do que recomendar para o usuário. Por exemplo, na regra que mostra que 64,3% dos usuários que assistiram "2001: Uma Odisseia no Espaço" também assistiram "Alien", pode-se inferir que recomendar "Alien" para aqueles usuários que acabaram de assistir "2001: Uma Odisseia no Espaço" tem uma grande chance de lhes parecer interessante

Obrigada!

Referências

Hipp, J., Güntzer, U. e Nakhaeizadeh, G. (2000) "Algorithms for Association Rule Mining – A General Survey and Comparison." ACM SIGKDD Explorations Newsletter, 58–64.

Romão, W., Niederauer, C., Martins, A., Tcholakian, A., Pacheco, R. e Barcia, R. (1999) "Extração De Regras De Associação em C&t: O Algoritmo Apriori". Programa de Pós-Graduação em Engenharia de Produção - PPGEP Universidade Federal de Santa Catarina, Centro Tecnológico, Florianópolis, Brasil.

Phorasm, P. e Yu, L. (2017) "Movies recommendation system using collaborative filtering and k-means". International Journal of Advanced Computer Research, Vol 7(29).

Bourreau, M. e Gaudin, G. (2018) "Streaming Platform and Strategic Recommendation Bias". CESifo Working Paper No. 7390, Munique, Alemanha