Programa de Engenharia de Sistemas e Computação - PESC/Coppe/UFRJ
Aluno: Christian da Silva Cabral Cardozo **DRE:** 117029985
Data Mining - Professor Geraldo Zimbrão

Extração de Regras de Associação no 100k MovieLens Dataset

Introdução

Este relatório tem por objetivo descrever o desenvolvimento do trabalho proposto na disciplina Data Mining. Usando o dataset 100k disponibilizado pelo MovieLens [1], foi proposto extrair regras de associação entre os filmes avaliados por usuários.

Dataset

O dataset 100K do MovieLens é uma base de dados que possui as avaliações que usuários deram para filmes vistos. Para isso, foram utilizadas as avaliações de 943 usuários distintos em 1682 filmes distintos, gerando 100 mil avaliações que compõem o dataset. Esta base contém outras informações sobre os usuários e os filmes, mas basicamente as informações necessárias para desenvolver o problema são: id do usuário, id e nome do filme, e, lista de avaliações de cada usuário, onde ter avaliado um filme significa ter assistido o mesmo. Cada usuário avaliou no mínimo 20 filmes e a média de quantidade de filmes avaliados por usuário é igual a 106.

Pré-processamento

O pré-processamento desta base é simples e consiste nos dois passos seguintes:

1. Extrair nomes dos filmes: utilizando o arquivo “u.item”, são extraídos id e nome do filme, que são as duas primeiras colunas respectivamente.
2. Extrair itemset dos usuários: utilizando o arquivo “u.data” são extraídas as avaliações feitas por usuários. O resultado é salvo em um arquivo CSV, onde cada i-ésima linha representa a lista de id de filmes que foram avaliados pelo i-ésimo usuário.

Extração de Regras: Algoritmo Apriori

Para a mineração das regras de associação foi utilizado o algoritmo apriori. O algoritmo recebe como entrada a lista de itens de cada usuário e de acordo com um suporte e confidência mínimos computa e fornece as regras de associação. O código

do algoritmo utilizado em Python foi desenvolvido por terceiros e está disponibilizado no GitHub [2].

Experimentos e Resultados

Os experimentos foram executados utilizando os dados e o algoritmo das seções anteriores. Dois parâmetros podem ser passados como entrada do algoritmo: suporte mínimo e confidência mínima. A seguir, temos os experimentos realizados variando tais parâmetros.

1. Suporte = 0.5; Confidência = 0.8

Quem assistiu	Também assiste	Confidência
Return of the Jedi (1983)	Star Wars (1977)	0.947
Star Wars (1977)	Return of the Jedi (1983)	0.823

2. Suporte = 0.35; Confidência = 0.9

Quem assistiu	Também assiste	Confidência
Raiders of the Lost Ark (1981), Return of the Jedi (1983)	Star Wars (1977)	0.985
Toy Story (1995), Return of the Jedi (1983)	Star Wars (1977)	0.979
Return of the Jedi (1983)	Star Wars (1977)	0.947
Empire Strikes Back, The (1980)	Star Wars (1977)	0.940
Raiders of the Lost Ark (1981)	Star Wars (1977)	0.905

3. Suporte = 0.3; Confidência = 0.925

Quem assistiu	Também assiste	Confidência
Raiders of the Lost Ark (1981),Return of the Jedi (1983),Empire Strikes Back, The (1980)	Star Wars (1977)	0.997
Return of the Jedi (1983),Empire Strikes Back, The (1980)	Star Wars (1977)	0.994
Return of the Jedi (1983),Pulp Fiction (1994)	Star Wars (1977)	0.986
Raiders of the Lost Ark (1981),Return of the Jedi (1983)	Star Wars (1977)	0.985
Toy Story (1995),Return of the Jedi (1983)	Star Wars (1977)	0.979

Silence of the Lambs, The (1991),Return of the Jedi (1983)	Star Wars (1977)	0.976
Return of the Jedi (1983),Twelve Monkeys (1995)	Star Wars (1977)	0.976
Return of the Jedi (1983),Godfather, The (1972)	Star Wars (1977)	0.971
Return of the Jedi (1983),Star Trek: First Contact (1996)	Star Wars (1977)	0.97
Raiders of the Lost Ark (1981),Empire Strikes Back, The (1980)	Star Wars (1977)	0.966
Independence Day (ID4) (1996),Return of the Jedi (1983)	Star Wars (1977)	0.965
Contact (1997),Return of the Jedi (1983)	Star Wars (1977)	0.956
Fargo (1996),Return of the Jedi (1983)	Star Wars (1977)	0.954
Return of the Jedi (1983)	Star Wars (1977)	0.947
Empire Strikes Back, The (1980)	Star Wars (1977)	0.94
Silence of the Lambs, The (1991),Raiders of the Lost Ark (1981)	Star Wars (1977)	0.936
Return of the Jedi (1983),Star Wars (1977),Empire Strikes Back, The (1980)	Raiders of the Lost Ark (1981)	0.933
Return of the Jedi (1983),Empire Strikes Back, The (1980)	Raiders of the Lost Ark (1981)	0.931
Raiders of the Lost Ark (1981),Star Wars (1977),Empire Strikes Back, The (1980)	Return of the Jedi (1983)	0.93
Return of the Jedi (1983),Empire Strikes Back, The (1980)	Raiders of the Lost Ark (1981),Star Wars (1977)	0.927

4. Suporte = 0.239; Confidência = 1

Quem assistiu	Também assiste	Confidência
Toy Story (1995),Return of the Jedi (1983),Empire Strikes Back, The (1980)	Star Wars (1977)	1

5. Suporte = 0.2; Confidência = 0.95

Quem assistiu	Também assiste	Confidência
Raiders of the Lost Ark (1981),Return of the Jedi (1983),Princess Bride, The (1987),Empire Strikes Back, The (1980)	Star Wars (1977)	1

E.T. the Extra-Terrestrial (1982),Return of the Jedi (1983),Empire Strikes Back, The (1980)	Star Wars (1977)	1
Toy Story (1995),Return of the Jedi (1983),Empire Strikes Back, The (1980)	Star Wars (1977)	1

Conforme diminuímos o valor do suporte, as possibilidades de conjuntos de filmes que estão dentro do limite de confidência aumenta. Aumentando o valor de confidência, restringimos o resultado para que tenha menos itens. A partir de suportes menores do que 0.24 começamos a obter regras com confidência igual a 1. A maioria das regras extraídas têm como resultado de “Também assiste” o filme Star Wars. É uma característica aceitável, já que é um filme bastante assistido e popular.

Obviamente, quando mineramos dados estamos buscando informações não triviais que sejam úteis. Como foi mencionado, é fácil concluir que qualquer indivíduo tenha assistido Star Wars. Por isso, mais alguns experimentos foram feitos em busca de uma relação que não fosse tão óbvia e que fosse mais interessante. Com o suporte igual a 0.18 e confidência 0.97, podemos encontrar algumas regras que finalmente não possuem o filme Star Wars na sua definição:

Quem assistiu	Também assiste	Confidência
Rock, The (1996),Broken Arrow (1996),Mission: Impossible (1996)	Independence Day (ID4) (1996)	0.978
Mission: Impossible (1996),Eraser (1996)	Independence Day (ID4) (1996)	0.973
Mission: Impossible (1996),Broken Arrow (1996)	Independence Day (ID4) (1996)	0.972

Se continuarmos variando os valores de suporte e confidência, podemos encontrar outras regras interessantes. Como neste caso, em que temos suporte igual a 0.2 e confidência de 0.9:

Quem assistiu	Também assiste	Confidência
Return of the Jedi (1983),Monty Python and the Holy Grail (1974),Star Wars (1977),Indiana Jones and the Last Crusade (1989)	Back to the Future (1985)	0.937
Seven (Se7en) (1995),Star Wars (1977)	Pulp Fiction (1994)	0.914
Jaws (1975),Pulp Fiction (1994)	Silence of the Lambs, The (1991)	0.912
Empire Strikes Back, The (1980),Twelve Monkeys (1995)	Pulp Fiction (1994)	0.906
Silence of the Lambs, The (1991),Usual Suspects, The (1995)	Fargo (1996)	0.901

Por fim, executamos um último experimento buscando variações além de Star Wars, com suporte igual 0.15 e confidência igual a 0.98. Algumas das regras obtidas são:

Quem assistiu	Também assiste	Confidência
Mission: Impossible (1996),Broken Arrow (1996),Eraser (1996)	Independence Day (ID4) (1996)	0.986
Star Trek IV: The Voyage Home (1986),Star Trek III: The Search for Spock (1984)	Star Trek: The Wrath of Khan (1982)	0.986
Jurassic Park (1993),Aliens (1986),Star Wars (1977),Terminator 2: Judgment Day (1991)	Alien (1979)	0.953

Código

O projeto foi desenvolvido utilizando a linguagem Python. Todo o código desenvolvido para a realização deste relatório está disponível online e pode ser acessado através do GitHub [3].

Referências

[1] MovieLens 100K Dataset - Group Lens , disponível em https://grouplens.org/datasets/movielens/100k/, acessado em 28/03/2017.

[2] Python Implementation of Apriori Algorithm for finding Frequent sets and Association Rules, disponível em https://github.com/asaini/Apriori, acessado em 28/03/2017.

[3] Apriori 100K MovieLens, disponível em https://github.com/chriiscardozo/apriori-100k-ml/tree/master, acessado em 31/03/2017.