PESC – Programa de Engenharia de Sistemas e Computação

Clusterização e Regras de Associação utilizando a base de dados do Movie Lens

Aluno: Raul Sena Ferreira
Professor: Geraldo Zimbrão

Relatório do trabalho para a disciplina de Data Mining

Rio de Janeiro, Abril de 2015

Introdução

O relatório trará uma breve informação sobre os experimentos de custerização e regras de associação utilizados sobre a base de dados do movie lens, um site que provê várias bases de dados de diferentes tamanhos contendo informações sobre os filmes que vários usuários assistiram e julgaram utilizando uma pontuação de 0 a 5, onde 0 é ruim e 5 é muito bom, bem como informações do próprios usuários.

Este trabalho fez uso da linguagem Julia por possuir bom desempenho neste tipo de tarefa, além de possuir uma sintaxe simples. O código está hospedado no link:
https://github.com/raulsenaferreira/Systems-Engineering/tree/master/Data%20Mining/Work_1
juntamente com uma cópia deste relatório e todos os arquivos gerados pelo programa.

Procedimentos

REGRAS DE ASSOCIAÇÃO

Primeiro foi extraída a informação dos arquivos do movie lens e em seguida foram estruturadas e colocadas no banco de dados Postgres.

O algoritmo usado foi o Apriori e apesar de simples possui uma série de limitações, entre elas o número de passos que deve ser realizado para revelar os conjuntos finais de regras, que é de: (2^S)-1, onde S é o tamanho do conjunto

Devido a limitação de memória, foi escolhido a estratégia de realizar o processamento em cima de uma quantidade limitada de filmes, ao invés de realizar em cima de toda
os filmes vistos pelos usuários.

Além do mencionado acima, houve a preocupação em se tentar achar regras de associação interessantes, por tanto, a estratégia adotada foi realizar as regras de associação em cima de dados que fossem raros porém com boa confiança.

Depois de definida a estratégia, foram realizadas regras de associação em cima de 3 tarefas:

1) Usuário que assistiu filme X também assistiu Y
2) Usuário que gostou de filme X também gostou de Y
3) Usuário que não gostou de X também não gostou de Y

Para a tarefa 1) foram selecionados os usuários que assistiram os 20 filmes menos assistidos com.

Para a tarefa 2) foram selecionados os 20 filmes mais bem votados pois o contrário não trouxe regras suficientes.

Na tarefa 3) foi adotado o mesmo que para a tarefa 2, mudando apenas as consultas para os filmes que tiveram ratings abaixo de 3.

CLUSTERIZAÇÃO

Para a tarefa de clusterização foi adotada a seguinte estratégia:

Criou-se uma matriz de usuarios x filmes e em seguida aplicou-se o PCA reduzindo a dimensão da matriz para 10.

Em seguida, o algoritmo K-Means foi utilizado para clusterizar a base formando 10 grupos. O algoritmo para depois de 200 iterações ou quando atinge convergência.

Resultados

REGRAS DE ASSOCIAÇÃO:

Foram colocados as regras julgadas mais interessantes, levando em conta o gênero do filme, o restante dos padrões encontrados foram gravados em arquivo e podem ser visualizados e/ou baixados no link do repositorio na pasta “results”.

Alguns padrões interessantes, em um total de 34, encontrados para a regra “Usuário que assistiu filme X também assistiu Y” com suporte de 10%:

1) Assitiu "Damsel in Distress, A (1937)"=>{"Comedy","Musical","Romance"}e também assistiu "Angel on My Shoulder (1946)"=>{"Crime","Drama"}
2) Assistiu "T-Men (1947)"=>{"Film-Noir"} e também assistiu "Terror in a Texas Town (1958)"=>{"Western"}
3) Assistiu "Homage (1995)"=>{"Drama"} e também assistiu "Bird of Prey (1996)"=>{"Action"}
4) Assistiu "Terror in a Texas Town (1958)"=>{"Western"} e também assistiu "Vie est belle, La (Life is Rosey) (1987)"=>{"Comedy","Drama"}
5) Assistiu "Land and Freedom (Tierra y libertad) (1995)"=>{"War"} e também assistiu "Eighth Day, The (1996)"=>{"Drama"}

Alguns padrões interessantes, em um total de 507, encontrados para a regra “Usuário que gostou de filme X também gostou de Y” com suporte de 20%:

1) Gostou de "Raiders of the Lost Ark (1981)"=>{"Action","Adventure"} e também gostou de "Toy Story (1995)"=>{"Animation","Children's","Comedy"}
2) Gostou de "Raiders of the Lost Ark (1981)"=>{"Action","Adventure"}e também gostou de "Contact (1997)"=>{"Drama","Sci-Fi"}
3) Gostou de "Godfather, The (1972)"=>{"Action","Crime","Drama"}e também gostou de "Toy Story (1995)"=>{"Animation","Children's","Comedy"}
4) Gostou de "Star Wars (1977)"=>{"Action","Adventure","Romance","Sci-Fi","War"}e também gostou de "Shawshank Redemption, The (1994)"=>{"Drama"}
5) Gostou de "Pulp Fiction (1994)"=>{"Crime","Drama"}e também gostou de "Princess Bride, The (1987)"=>{"Action","Adventure","Comedy","Romance"}
6) Gostou de "Titanic (1997)"=>{"Action","Drama","Romance"}e também gostou de "Scream (1996)"=>{"Horror","Thriller"}
7) Gostou de "Twelve Monkeys (1995)"=>{"Drama","Sci-Fi"}e também gostou de "Princess Bride, The (1987)"=>{"Action","Adventure","Comedy","Romance"}

Alguns padrões interessantes, em um total de 23, encontrados para a regra “Usuário que não gostou de X também não gostou de Y” com suporte de 20%:

1) Não gostou de "Liar Liar (1997)"=>{"Comedy"} e também não gostou de "English Patient, The (1996)"=>{"Drama","Romance","War"}
2) Não gostou de "Liar Liar (1997)"=>{"Comedy"} e também não gostou de "Scream (1996)"=>{"Horror","Thriller"}
3) Não gostou de "Independence Day (ID4) (1996)"=>{"Action","Sci-Fi","War"} e também não gostou de "Scream (1996)"=>{"Horror","Thriller"}

CLUSTERIZAÇÃO:

A clusterização dos filmes visou formar gruposde filmes que foram assistidos pelos mesmos grupos de usuários. O arquivo com todos os filmes em seus respectivos grupos resultantes também estão no repositório, no link:
https://github.com/raulsenaferreira/Systems-Engineering/tree/master/Data%20Mining/Work_1

Alguns grupos ficaram com filmes bem semelhantes como é era de se esperar, mas também continham alguns resultados inusitados, como Babe, o porquinho no mesmo grupo de Star Wars e Pulp Fiction do Tarantino.

Também resultados interessantes puderam ser observados, como os filmes da franquia Star Trek e Star Wars que pertenciam a grupos diferentes, indicando que as franquias apesar de estarem no mesmo gênero não parecem compartilhar do mesmo público.

Um grupo maior ficou com o resto que não pôde ser classificado.

Seguem alguns filmes e seus respectivos grupos:

Grupo 1: Until the End of the World (Bis ans Ende der Welt) (1991), Waiting for Guffman (1996), I Shot Andy Warhol (1996), Basquiat (1996), Anaconda (1997), Shiloh (1997), Tie Me Up! Tie Me Down! (1990), Die xue shuang xiong (Killer, The) (1989), Gaslight (1944), 8 1/2 (1963), Fast, Cheap & Out of Control (1997), "Fathers' Day (1997)

Grupo 2: GoldenEye (1995), Get Shorty (1995), Seven (Se7en) (1995), Crimson Tide (1995), Professional, The (1994), Stargate (1994), Star Trek VI: The Undiscovered Country (1991), Star Trek: The Wrath of Khan (1982), Star Trek III: The Search for Spock (1984), Star Trek IV: The Voyage Home (1986)

Grupo 3: Jungle2Jungle (1997), Devil's Own, The (1997), George of the Jungle (1997), Event Horizon (1997), Mimic (1997), Chasing Amy (1997), Full Monty, The (1997), Gattaca (1997), Starship Troopers (1997), Good Will Hunting (1997)

Grupo 4: Aladdin (1992), Babe (1995), Usual Suspects, The (1995), Braveheart (1995), 2001: A Space Odyssey (1968), Apollo 13 (1995), Star Wars (1977), Empire Strikes Back, The (1980), Return of the Jedi (1983), Pulp Fiction (1994)

Grupo 5: Secret Garden, The (1993), Return of Martin Guerre, The (Retour de Martin Guerre, Le) (1982), Tin Drum, The (Blechtrommel, Die) (1979), Cook the Thief His Wife & Her Lover, The (1989), Rosencrantz and Guildenstern Are Dead (1990), Paris, Texas (1984), Crucible, The (1996), Persuasion (1995), Little Women (1994), Barcelona (1994)

Grupo 6: Taxi Driver (1976, Breakfast at Tiffany's (1961), Gone with the Wind (1939), Citizen Kane (1941), Mr. Smith Goes to Washington (1939), 20,000 Leagues Under the Sea (1954), Sleeper (1973), On Golden Pond (1981), Return of the Pink Panther, The (1974), Good, The Bad and The Ugly, The (1966)

Grupo 7: Contact (1997), English Patient, The (1996), Scream (1996), Liar Liar (1997), Air Force One (1997), Titanic (1997), Conspiracy Theory (1997), Saint, The (1997)

Grupo 8: Toy Story (1995), Twelve Monkeys (1995), Dead Man Walking (1995), Mr. Holland's Opus (1995), Rumble in the Bronx (1995), Birdcage, The (1996), Fargo (1996), Truth About Cats & Dogs, The (1996), Rock, The (1996), Twister (1996)

Grupo 9: Batman Forever (1995), Desperado (1995), Net, The (1995), Strange Days (1995), Clerks

(1994), Dolores Claiborne (1994), Ed Wood (1994), I.Q. (1994), Legends of the Fall (1994), Natural Born Killers (1994), Outbreak (1995)

Grupo 10: Mortal Kombat (1995), Pocahontas (1995), Things to Do in Denver when You're Dead (1995), Species (1995), Walk in the Clouds, A (1995), Mary Shelley's Frankenstein (1994), Quick and the Dead, The (1995), Wes Craven's New Nightmare (1994), Wolf (1994), Wyatt Earp (1994), Another Stakeout (1993)

Referências

https://github.com/JuliaDB/DBI.jl
https://github.com/iamed2/PostgreSQL.jl
https://github.com/jonathanewerner/julia-apriori-algorithm
https://github.com/JuliaStats/Clustering.jl
http://en.wikipedia.org/wiki/Principal_component_analysis
https://github.com/raulsenaferreira/Systems-Engineering/tree/master/Data%20Mining