Visualização MovieLens

Vasco Sousa, 93049, Tiago Rainho 92984

Visualização da Informação, 2021 (Mestrado em Engenharia informática, Universidade of Aveiro)

Abstract

Neste relatório, são apresentados de forma detalhada todos os passos seguidos para construir um conjunto de visualizações do dataset MovieLens, utilizando a ferramenta de visualização D3.js.

Motivação e objetivos

A principal motivação para a realização deste projeto, consiste no interesse em compreender e detectar padrões nas preferências cinematográficos dos utilizadores que fazem parte do *dataset* em questão.
Como tal, os principais objetivos deste projeto passam por apresentar um conjunto de visualizações que ajudem a perceber esses padrões, assim como fazer uso das mais recentes técnicas de visualização em ambiente *web*, usando a ferramenta d3.js.

Users and the Questions

Caracterização dos utilizadores e o seu contexto

O tipo de utilizador que utilizaria este tipo de métodos de visualização são jovens que querem perceber que género de filmes os seus amigos mais gostam, por exemplo o João Almeida que tem 16 anos e pouco conhecimento das categorias de filmes que existem pois apenas vê filmes de ação.
Outro utilizador é a Ana Silva que tem 39 anos e trabalha como analista, quer investir numa empresa que produza filmes que mulheres empreendedoras gostem.

Perguntas a responder

- Qual o filme de comédia mais bem sucedido?
- Qual dos sexos que sendo cientista gosta mais de drama?
- Que ocupação aumenta a probabilidade de preferir filmes de mistério?
- É mais comum haver filmes de aventura que ao mesmo tempo são de ação ou de romance?

Dataset

O dataset utilizado denomina-se de "Movielens 100k Dataset" [1]. Este é composto por um total de 100,000 avaliações referentes a um total de 9,000 filmes e 600 utilizadores. Este *dataset* é mantido pelo Grouplens, grupo sediado na Universidade de Minnesota. Entre os vários *datasets* construídos por este grupo, esta é a versão indicada como a mais adequada para um uso no âmbito académico.
Cada utilizador avaliou pelo menos um filme, e um filme foi avaliado pelo menos uma vez. Existe uma grande disparidade no número de avaliações, com utilizadores a avaliarem mais de 200 filmes enquanto outros apenas 1, sendo que o mesmo acontece para os filmes, onde certos filmes tem apenas 1 avaliação e outros com mais de 100.
Estes dois fatores, fazem com que o *dataset* não seja balanceado, o que leva a que as possíveis interpretações das visualizações sejam enviesadas, favorecendo os grupos que contêm mais avaliações. Devido a este fator, sempre que possível optou-se por usar percentagens ou outros métodos de normalização da informação em análise.

Soluções de Visualização

Para o desenvolvimento das visualizações começou-se por avaliar a informação mais vantajosa relativa ao dataset em específico. Em cada visualização foram relacionadas as seguintes informações:

Visualização 1:
- Categorias dos filmes
- Classificações atribuídas para cada categoria
- Sexo do utilizador
- Ocupação do utilizador

Visualização 2:
- Categorias dos filmes
- Classificações atribuídas para cada categoria
- Ocupação do utilizador

Visualização 3:
- Filmes
- Categoria dos filmes
- Classificações atribuídas para cada filme

Visualização 4:
- Categorias dos filmes

Primeira visualização

Nesta visualização o objetivo inicial era a representação dos ratings que a amostra de utilizadores deu a uma determinada categoria de filmes, sendo que o protótipo de baixa fidelidade foi desenvolvido em papel e encontra-se na Figura 1:

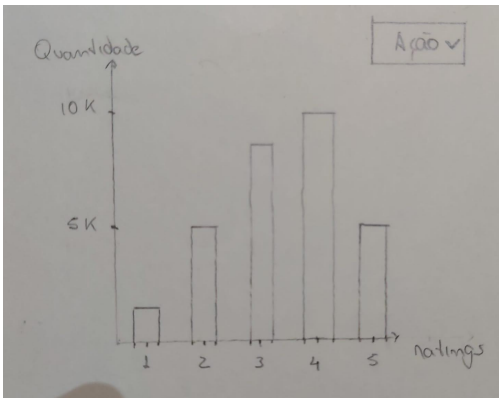

Figura 1- Protótipo inicial equacionado para a Primeira Visualização.

De acordo com as *Heurísticas de Nielsen*, a heurística de consistência e padrões não estava ideal por possuir valores absolutos, por isso o valor foi transformado numa percentagem, no entanto segundo o *feedback* dos utilizadores foi transmitida a ideia de que a visualização passava pouca informação. Com base nesse *feedback* procurou-se a sua extenção através da adição de uma nova relação com outras variáveis, neste caso o sexo e a ocupação dos utilizadores.
A variável ocupação foi adicionada através de um filtro em formato de *textbox* onde o utilizador selecionaria a ocupação de interesse permitindo também selecionar todas [Figura 2].

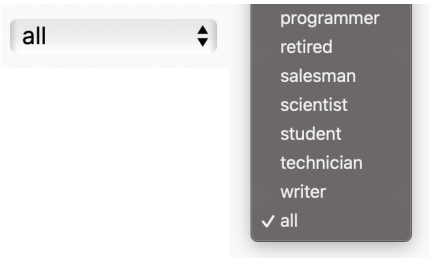

Figura 2- *Textbox* adicionada para a complementação da informação com a seleção da ocupação dos utilizadores.

A variável "sexo" foi adicionada através da divisão da mesma barra entre masculino e feminino de acordo com as suas proporções e ao fazer *hover* com o rato mostraria valores absolutos [Figura 3].

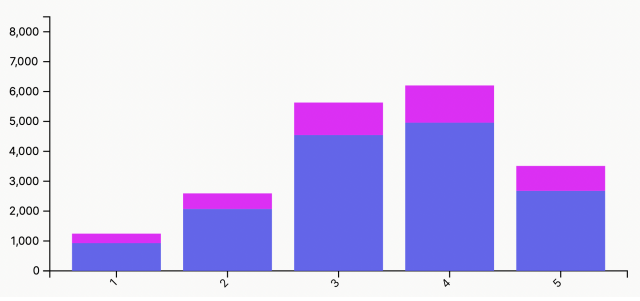

Figura 3- Adição da distinção entre sexo dos utilizadores que atribuiram ratings a cada categoria de filme.

Mais tarde, com base em novo *feedback* de utilizadores, foi transmitida a ideia que para haver uma comparação fiável era necessário os dois sexos estarem à mesma altura inicial. Desta forma procedeu-se à divisão da barra ao separar-se completamente uma da outra e colocando-as lado a lado como se vê na Figura 4.

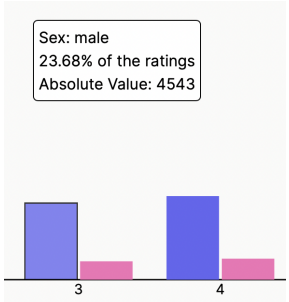

Figura 4- Divisão de cada barra de rating pela distinção entre o sexo dos utilizadores.

O resultado final desta visualização pode ser visto na Figura 5, sendo então possível recolher informações que relacionem os ratings atribuídos a filmes de acordo com as diversas categorias e diferentes ocupações dos utilizadores. Além disso, também é possível distinguir a percentagem de classificações atribuidas e valores absolutos de acordo com o sexo do utilizador, permitindo analisar as preferências gerais de cada sexo e/ou ocupação.

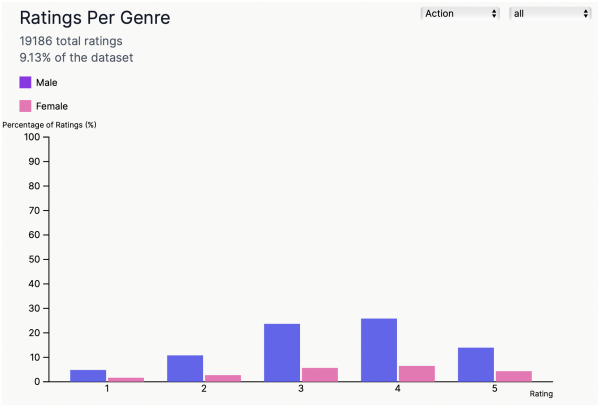

Figura 5- Resultado final da Primeira Visualização

Segunda visualização

Através desta visualização objetivou-se inicialmente relacionar os ratings, as categorias dos filmes e as

ocupações dos utilizadores. O protótipo de baixa fidelidade foi desenvolvido em papel e está presente na Figura 6.

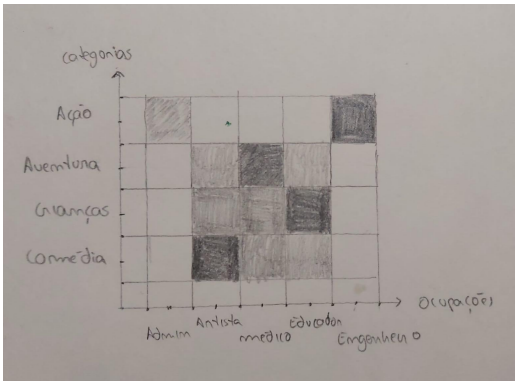

Figura 6- Protótipo inicial equacionado para a Segunda Visualização.

Tendo em conta as *Heurísticas de Nielsen*, a única que não está em correspondência é a de Ajuda e Documentação pois não tem legenda de cores, dificultando assim a interpretação para os utilizadores que na sua primeira abordagem ainda não interagiram com a informação disponível na visualização com o *hover* do rato. Face ao feedback recolhido dos utilizadores, não houve comentários negativos relativos à falta de legenda, pelo contrário, houve quem considerasse ter sido um aspeto positivo pois incentiva a interação com a visualização, no entanto esta permite uma interpretação mais eficaz, pelo que foi adicionada [Figura 7].

Figura 7- Legenda adicionada.

O resultado final desta visualização pode ser visto na Figura 8, sendo então possível ter uma perceção geral sobre as preferências de cada ocupação. Após a conclusão desta visualização são notórias algumas particularidades, como por exemplo em média, os filmes de horror serem os que têm piores ratings em todas as ocupações excepto para desempregados onde esta categoria é a melhor classificada. Além disso, os profissionais de saúde distinguem-se pela atribuição dos piores ratings independentemente das categorias.

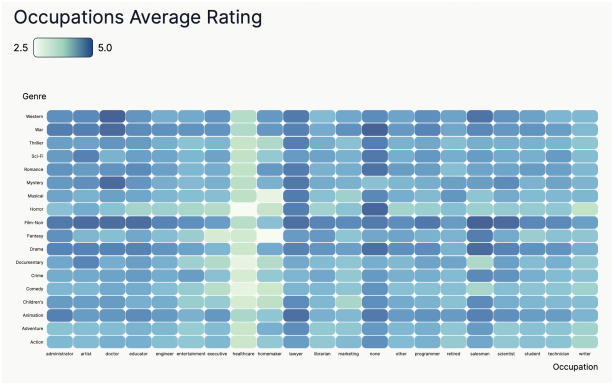

Figura 8- Resultado final da Segunda Visualização

Terceira visualização

Esta visualização tem como objetivo inicial analisar os melhores filmes de cada categoria presentes no *dataset* tendo como base a média dos ratings atribuídos. O primeiro protótipo feito em papel é o da Figura 9.

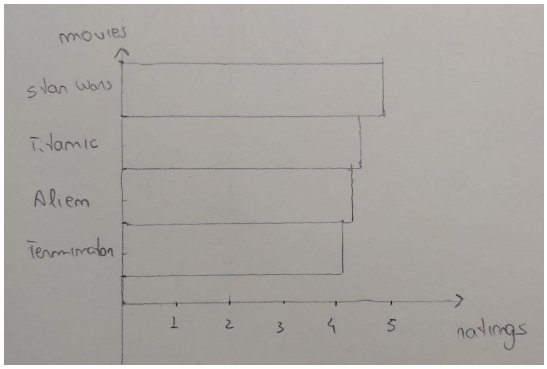

Figura 9- Protótipo inicial equacionado para a Terceira Visualização.

De acordo com as *Heurísticas de Nielsen*, concluiu-se que na heurística de liberdade e controlo do utilizador, falta fornecer ao mesmo a possibilidade de poder usar um número maior ou menor dos filmes que quer analisar. Nos testes de usabilidade esta também foi a maior crítica pelos utilizadores, sendo que estes demonstraram vontade de ter a possibilidade de também analisar os filmes com piores ratings.

Primeiramente passou-se ao aumento do conjunto dos filmes em análise ao permitir que os utilizadores possam escolher quantos filmes querem ver simultaneamente. Posteriormente implementou-se um botão onde é possível selecionar se se pretende analisar os *n* melhores ou piores filmes [Figura 10].

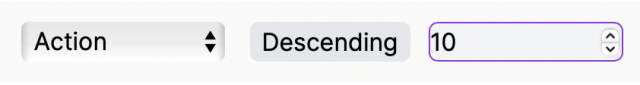

Figura 10- Adições que procuraram corresponder às questões impostas através do *feedback* do utilizador. À direita encontra-se um campo onde se introduz a quantidade de filmes desejáveis para ter simultaneamente em análise e centralmente encontra-se o botão que selecionado, permite diferenciar se se pretende a apresentação dos *n* piores ou melhores filmes.

O resultado final desta visualização pode ser visto na Figura 11, sendo então possível recolher informações que relacionem os *n* melhores/piores filmes de uma determinada categoria

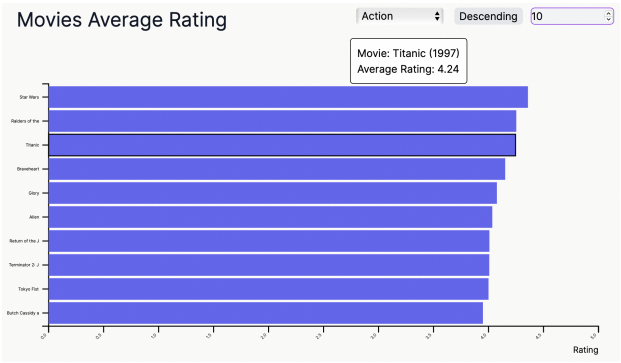

Figura 11- Resultado final da Terceira Visualização

Quarta visualização

Na última visualização procurou-se analisar as relações intra categorias mais frequentes presentes nos filmes, para tal foram consideradas várias tipologias de apresentação de informação sendo que o primeiro protótipo encontra-se na Figura 12.

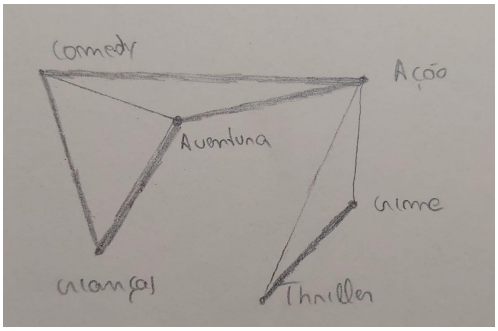

Figura 12- Protótipo inicial equacionado para a Quarta Visualização.

Este não seguiu para a fase de desenvolvimento tendo em consideração a presença de pontos dispersos, pelo que não havia uma forma consistente de demonstração desta informação sem o recurso a técnicas de algoritmos avançados, pelo que se equacionou manter os pontos mas apenas segundo uma organização diferente. Esta encontra-se presente na Figura 13 e usa o facto de estar numa disposição circular para manter conectividade direta entre todas as categorias, no entanto tornou-se evidente que nas extremidades as ligações ficam demasiado juntas para se fazer uma clara distinção e consequentemente análise, pelo que não progrediu para a fase seguinte (desenvolvimento).

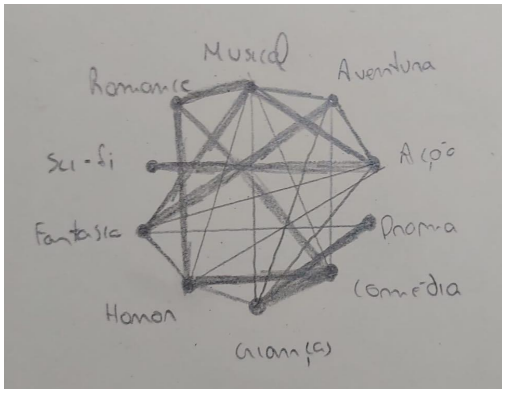

Figura 13- Protótipo secundário equacionado para a Quarta Visualização.

Após pesquisa sobre métodos de visualização que pudessem corresponder a todos os requisitos, foram considerados dois métodos: o *Sankey* (Figura 14) e o *Chord*. O primeiro foi excluído uma vez que a representação de informação exigia a introdução individual de cada categoria de filmes, ou seja, sempre que fosse necessário alterar a principal categoria em estudo, seria preciso ter um filtro para a sua seleção, o que resultaria numa perda de visão geral sobre as relações entre todas as categorias. Esta característica foi essencial para a seleção do Chord como visualização para esta informação, pois além de uma representação individual rica, também apresenta uma visão geral bastante representativa.

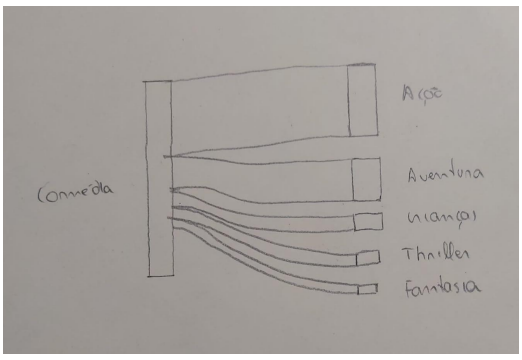

Figura 14- Protótipo secundário equacionado para a Quarta Visualização (Sankey).

De acordo com as *Heurísticas de Nielsen*, concluiu-se que na heurística de Estética e Design Minimalista está no limite entre poder ser considerado um problema ou não, pois é uma representação que apresenta muita informação agregada para um *dataset* como o usado. Neste método de visualização, quanto maior a extensão do *dataset*, mais problemático será o cumprimento desta heurística uma vez que não tolera a presença de um grande número de objetos. Todavia optou-se pela sua utilização por se considerar que o número de objetos diferentes está dentro do limiar tolerável.
Nos testes de usabilidade, os utilizadores relataram que no início a interpretação é ligeiramente difícil, no entanto após análise concluíram que em determinados aspetos se torna mais acessível, como por exemplo quando se quer analisar a relação direta entre duas categorias, onde selecionando uma ou outra, obtém-se a mesma resposta. Em complementação deste *feedback*, os utilizadores relataram não gostar de ter que estar sempre a passar o rato em cima das fatias dos círculos para saber a que género correspondia [Figura 15]. Por esta razão procedeu-se à adição do nome de cada categoria perto de cada pedaço de círculo correspondente.

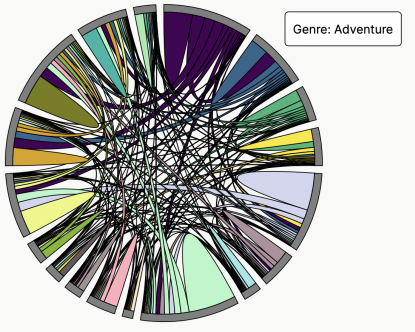

Figura 15- Apresentação anterior do Nome da Categoria, que só aparecia com a passagem do rato por cima da fatia.

O resultado final desta visualização pode ser visto na Figura 16 onde se evidencia uma visão geral, para sobressair informação mais específica, pode-se passar o rato em cima da fatia de círculo destacando-a como apresenta a Figura 17.
É de notar por exemplo que a maioria das ligações da categoria de "Ação" é com a de "Aventura" assim como a maioria das ligações com a categoria de "Animação" é com a de "Crianças", algo previsível no entanto com este método torna-se bastante evidente.

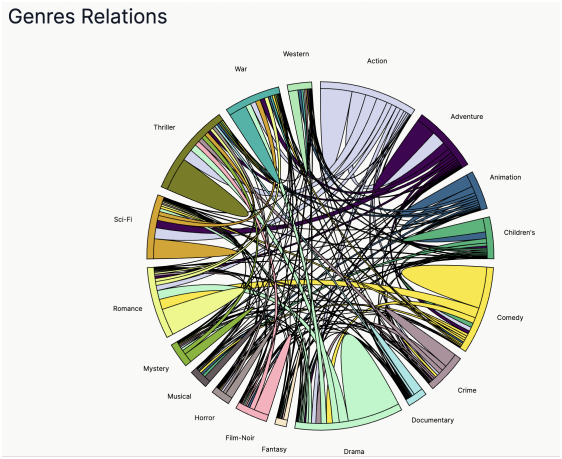

Figura 16- Resultado final da Quarta Visualização- Visão Geral das Categorias

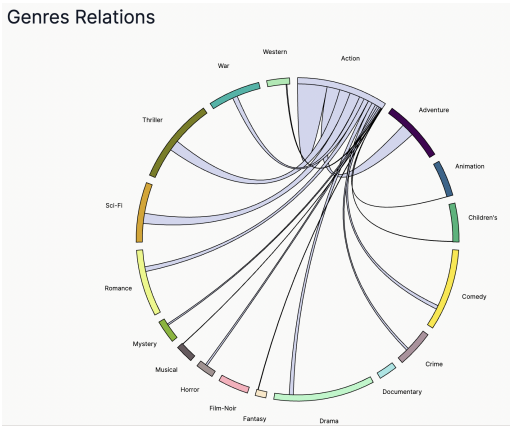

Figura 17- Resultado final da Quarta Visualização- Visão Individual da Categoria "Ação"

Conclusões e Trabalho Futuro

Em suma, os objetivos a que o grupo se propôs inicialmente foram atingidos. É possível através das 4 visualizações construídas, detetar padrões presentes no dataset utilizado.
Contudo, como mencionado anteriormente, o dataset utilizado não é balanceado, o que leva a que os grupos que avaliaram mais vezes, tenham uma posição dominante, enviesados os resultados.
Por outro lado, já na parte técnica do trabalho, as visualizações não são totalmente responsivas, o que faz com que em certos tamanhos de ecrã, certos elementos das mesmas, apareçam em posições incorretas.
Um trabalho futuro passaria por balancear o dataset e tornar as visualizações completamente responsivas. Para além disso, poderia ser também interessante, construir mais algumas visualizações que combinadas com um sistema de recomendação, permitissem os utilizadores receber recomendações de filmes para ver.

References

[1] https://grouplens.org/datasets/movielens/

[2] https://www.data-to-viz.com