Análise de Redes de Interesses em Filmes

Díaz, Juan S. B.[1], **Fadel, Samuel G.**[1] **Vinco e Silva, Fábio S.**[1],

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)
Campinas – SP – Brazil

Resumo. *A análise de redes tem sido aplicada para modelar e entender diversos fenômenos em diferentes áreas do conhecimento. Este trabalho investiga os dados do "Netflix Prize", por meio de análises de redes, buscando entender os interesses por filmes, partindo da hipótese de que os gêneros de filmes tem forte influência nestes interesses. Uma nova metodologia, onde as avaliações são agregadas por gênero, é proposta e avaliada comparativamente à filtragem colaborativa. Os resultados mostram que a nova metodologia capturou diversas comunidades dentro da rede de usuários sendo analisada, ilustrando interesses bem definidos para usuários mais experientes. Comparativamente, a filtragem colaborativa tende a capturar comunidades menos isoladas e mais homogêneas.*

1. Introdução

O estudo de redes vem se tornando cada vez mais importante para compreender diversos fenômenos em diferentes áreas, como: logística em redes de transportes; velocidade e alcance de divulgação de informações e notícias na internet; estrutura de mercados financeiros; espalhamento de epidemias; padrões comportamentais em redes sociais, dentre várias outras. O avanço da capacidade de processamento computacional e maior poder de armazenamento possibilitaram a análise de grandes bases de dados, o que, de certa forma, colaborou para aumentar o interesse no estudo de redes, pois muitas delas são oriundas de enorme quantidade de dados.

Considerando o vasto domínio de aplicação de redes, a proposta desse trabalho é criar uma rede de interesses em filmes, ou seja, uma rede onde os usuários se conectam quando eles avaliam os mesmos filmes de maneira similar. Para criá-la, a base de dados utilizada foi a referente ao "Netflix Prize", descrita na próxima seção.

Buscou-se validar a seguinte hipótese nas análises das redes: os interesses dos usuários refletem interesses por gêneros de filmes. Com isso, busca-se saber, por exemplo, relações entre a experiência do usuário em termos de números de filmes assistidos e a forma como se distribuem as avaliações deste nos diversos gêneros.

2. Trabalhos relacionados

Lançado em Outubro de 2006, o Netflix Prize [Bennett and Lanning 2007] foi um desafio onde Netflix disponibilizou um conjunto de dados de mais de 100 milhões de avaliações (com as datas) de aproximadamente 480.000 usuários escolhidos ao acaso e mantendo a informação dos usuários anônima, com uma amostra de aproximadamente de 18.000 filmes. Estes dados foram coletados entre Outubro de 1998 e Outubro de 2005. As avaliações são números inteiros no intervalo fechado $[1, 5]$. O objetivo do desafio foi desenvolver um sistema de recomendação para melhorar a precisão em ao menos um 10% do algoritmo Cinematch usado por Netflix.

Trabalhos anteriores foram realizados no sentido de unir a análise de grandes redes em que se relacionam produtos com usuários, utilizadas em sistemas de recomendação. A filtragem colaborativa [Herlocker et al. 2004] (do inglês *collaborative filtering*), parte da uma rede bipartida para computar a similaridade entre usuários e é um dos métodos mais utilizados para construir sistemas de recomendação. Entretanto, nesta, calcula-se a similaridade entre usuários baseando-se em seu histórico de interesses, dado pela rede bipartida original. A partir desta similaridade, computa-se uma pontuação que relaciona itens com usuários, recomendando-se os itens mais bem pontuados para cada usuário. Uma das vantagens deste método está no uso de informações tanto globais (o quanto um produto é, no geral, interessante para usuários) quanto locais (o quanto um produto é interessante para usuários parecidos).

Também relacionado a sistemas de recomendação, em [Zhou et al. 2007], os autores descrevem um método para construir uma rede de interesses também a partir de uma rede bipartida de produtos e usuários. O objetivo do método é construir uma rede de produtos em que o peso das arestas refletem o quanto um produto influencia na recomendação de outro. Combinando a rede de produtos com as informações da rede bipartida original, obtém-se uma pontuação para o quão interessante é um produto para um usuário. Em comparação à filtragem colaborativa, a pontuação calculada por meio deste método reflete melhor os interesses dos usuários.

Em referência à relação entre gêneros preferidos dos usuários e o tipo de avaliação dada aos filmes desses gêneros, um estudo anterior [Moon et al. 2010], que também usou a base de dados disponibilizada pelo Netflix Prize, buscou uma abordagem diferente da aqui proposta. Utilizando técnicas estatísticas, os autores constataram a existência da relação descrita acima para os usuários que veem filmes com mais frequência. Essa relação tem o perfil aproximado de um "U" em um gráfico de avaliações por proporção de gênero. A conclusão que eles chegaram é que usuários que veem vários gêneros avaliam positivamente os seus filmes preferidos dentro de cada gênero e as avaliações tendem a cair depois disso. Conforme eles vão limitando os gêneros vistos, as avaliações tendem a subir novamente, alcançando o valor máximo quando a maioria dos filmes vistos são de um ou dois gêneros preferidos. Os autores alegam que isso ocorre devido à expertise de gênero que esses usuários adquirem, que os auxilia na escolha de melhores filmes para serem vistos.

3. Construindo e analisando a rede de usuários

Nesta seção, descrevemos os dois métodos utilizados para construir a rede de usuários que reflete os interesses em filmes, a partir dos dados do "Netflix Prize".

Ao construir e analisar as redes em questão, procuramos entender como os usuários compartilham interesses de filmes. Essencialmente, buscamos observar a *formação de comunidades*, definida em [Boccaletti et al. 2006] como grupos de nós que possuem maiores densidades de arestas dentre si, do que entre outros grupos. Para isso, propomos uma nova abordagem para definir interesses em comum e investigamos como esta se compara à filtragem colaborativa, uma abordagem mais estabelecida para tal tarefa. Sumarizando, as abordagens são:

1. Avaliações agregadas por gênero, uma nova abordagem onde cada usuário é descrito como a avaliação média dada em cada um dos 22 diferentes gêneros de filme

categorizados;
2. Filtragem colaborativa, para comparar a abordagem proposta a uma bem estabelecida na literatura.

3.1. Pré-processamento

Neste projeto foi usada a API do OMDB, que fornece dados básicos de filmes a partir do título ou do identificador IMDB. Obtivemos os gêneros de 12.440 filmes, aproximadamente o 70% dos filmes totais disponíveis no conjunto de dados.

Alguns dos motivos pelos quais os gêneros não puderam ser obtidos em sua totalidade foram:

- O uso não padronizado de algumas palavras como "&" e "and" nos títulos no conjunto de dados do Netflix Prize e na OMDB
  - Netflix Prize Dataset: "Lilo and Stitch"
  - OMDB API: "Lilo & Stitch"
- Alguns títulos de vídeos não são considerados filmes pela OMDB
- O uso inapropriado de descrições que não devem ser incluídas no titulo
  - Netflix Prize Dataset: "Lord of the Rings: The Return of the King: Extended Edition: Bonus Material"
  - OMDB API: "The Lord of the Rings: The Return of the King"

3.2. Construção da rede de usuários via avaliações por gênero

Nesta abordagem, partimos da hipótese de que existe uma relação entre preferências por gêneros de filmes e a avaliação dada para o filme, sendo esta maior quando há um casamento entre a preferência pessoal e o gênero do filme. Em outras palavras, buscamos construir uma rede de usuários em que usuários com avaliações altas de certos gêneros e baixas de outro sejam conectados. Utilizamos os dados de gênero coletados no pré-processamento para construir um descritor de usuário que consiste em um vetor de 22 dimensões, sendo uma dimensão para cada um dos 22 gêneros. Em cada dimensão, atribuímos a média de todas as avaliações dadas pelo usuário para um certo filme, resultando em um valor entre 0 e 5.

Em seguida, a rede é gerada a partir de um critério simples de conectividade: dois usuários estão conectados se a dissimilaridade entre eles for menor do que um certo limiar, ou seja, eles são considerados "próximos" ou "parecidos" segundo o descritor. Para medir a dissimilaridade entre os usuários, utilizamos a distância Euclidiana. Já o limiar, definido empiricamente, é de 15% da distância média. O limiar foi definido assim para que o número de conexões não fosse muito grande e nem muito pequeno.

Como mencionado anteriormente, o conjunto original continha 480 mil usuários. Este número elevado de nós fez com que a abordagem aqui descrita se tornasse computacionalmente inviável para ser realizada. Por isso, para realização dos experimentos, foram escolhidos 50 mil usuários aleatoriamente, de tal forma que a distribuição de grau dos nós tivesse as mesmas características dos dados originais.

3.3. Construção da rede de usuários via filtragem colaborativa

Nesta abordagem para a construção da rede de usuários, utilizamos a filtragem colaborativa para determinar a similaridade entre os usuários, conforme descrito em [Zhou et al. 2007]. Sejam $F = \{f_1, f_2, \ldots, f_m\}$ e $U = \{u_1, u_2, \ldots, u_n\}$ os conjuntos de filmes e usuários, respectivamente. Seja também $a_{ij} = 1$ quando f_i foi avaliado pelo usuário u_j e $a_{ij} = 0$, caso contrário. Nestes casos, assumimos que os valores a_{ij} representam a rede bipartida de filmes e usuários. Para computar a similaridade s_{ij} entre dois usuários u_i, u_j, fazemos

$$s_{ij} = \frac{\sum_{l=1}^{m} a_{li} a_{lj}}{\min\{k(u_i), k(u_j)\}}, \tag{1}$$

para todos os usuários sendo considerados, onde $k(u_i)$ representa o grau de u_i.

O segundo passo na construção da rede de usuários é semelhante ao método anterior. É importante notar que a similaridade, calculada conforme a Equação (1), sempre respeita $0 \leq s_{ij} \leq 1$, para quaisquer i e j. Com base nesta observação, agora que possuímos as similaridades entre usuários, podemos transformar esta em uma medida de distância fazendo $d_{ij} = 1 - s_{ij}$.

Em seguida, utilizando os mesmos critérios para conectividade da rede descrita anteriormente, construímos a rede de usuários conectando dois usuários u_i e u_j sempre que d_{ij} for menor do que um certo limiar, estabelecido empiricamente. Para tornar as redes mais comparativas, utilizamos o mesmo critério utilizado na rede anterior, ou seja, o limiar adotado é de 15% da distância média.

Assim como na abordagem anterior, o custo computacional do método se tornou um limitante para o uso dos 480 mil usuários originais. Entretanto, a filtragem colaborativa é ainda mais cara computacionalmente do que a abordagem descrita anteriormente. Como consequência, foram escolhidos aleatoriamente 5 mil usuários para a realização dos experimentos, também com o cuidado de manter a distribuição de graus semelhante à distribuição original.

4. Resultados e discussão

Inicialmente, foram feitas análises dos graus da rede com os dados originais. A Figura 1 representa a distribuição de frequência dos graus dos filmes, onde o grau de cada filme é determinado usando o número de avaliações feitas pelos usuários para cada filme. A Figura 2 representa o histograma de frequência nos graus dos usuários, onde o grau de cada usuário é calculado usando o número total de avaliações feitas por cada usuário para os filmes.

Tanto na Figura 1 quanto na Figura 2 é possível observar um comportamento assintótico que segue uma *power-law*, esta característica permite classificar como uma rede *scale-free* a rede obtida de conectar os filmes e os usuários, usando as avaliações dos filmes como arestas.

Utilizamos o software Gephi [Bastian et al. 2009] para visualizar as redes de interesses. O layout de posicionamento dos nós escolhido foi o Yifan Hu, pois, dentre todos os algoritmos de distribuição testados, ele apresentou melhor qualidade nos resultados com um desempenho computacional mais competitivo. O algoritmo Force Atlas, para a rede de 50 mil nós, executou em torno de 4 horas e mesmo assim apresentou mudanças

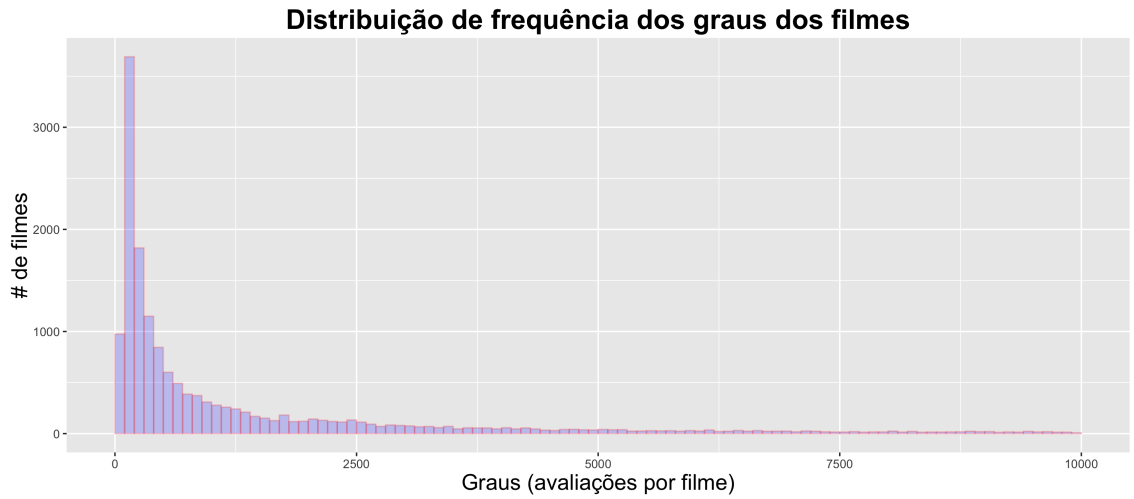

Figura 1. Distribuição de frequência dos graus dos filmes para os dados completos do "Netflix Prize".

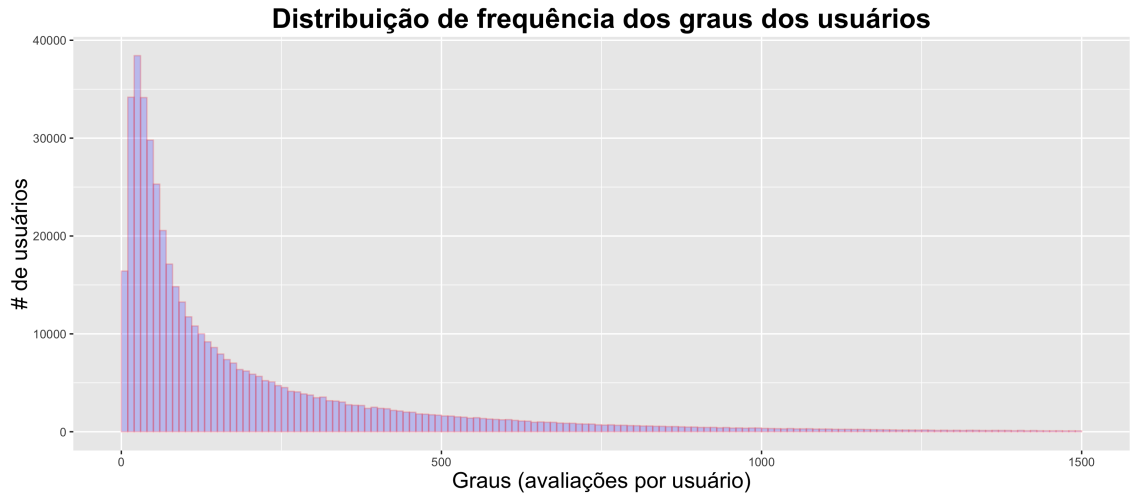

Figura 2. Distribuição de frequência dos graus dos usuários para os dados completos do "Netflix Prize".

pouco significativas no posicionamento dos nós. Por outro lado, com o Yifan Hu, em questão de minutos era possível perceber o efeito de agrupamentos.

4.1. Rede de usuários via médias de avaliações por gênero

Conforme mencionado anteriormente, a rede de usuários construída pela primeira abordagem, possui 50 mil usuários. A Figura 3 mostra uma visualização da rede. Para ajudar na análise, foram computadas as modularidades de cada nó, afim de identificar a formação de comunidades dentro da rede. As cores de cada nó indicam a comunidade a que pertencem. Ainda na figura, é possível notar a formação de diversas comunidades, em que seus números estão indicados em cima de cada uma.

A fim de investigar esta formação, coletamos algumas informações sobre as cinco maiores comunidades. Na Tabela 1, vemos informações sobre o número de usuários e a média de avaliações por comunidade. Observa-se que o número médio de avaliações e a

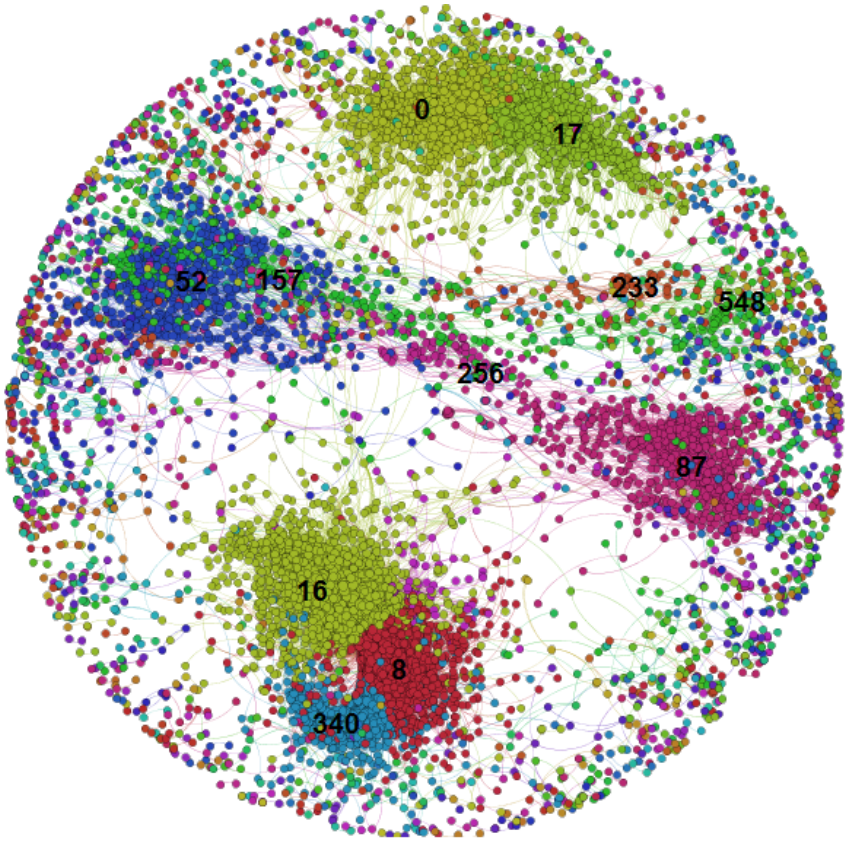

Figura 3. Visualização da rede de 50 mil usuários, com posicionamento dos nós computado pelo algoritmo Yifan Hu. As onze maiores comunidades estão com seu respectivo número indicado.

quantidade de usuários não necessariamente refletem o quão próximas estão as comunidades na visualização. Mais ainda, um fato surpreendente é que, mesmo unindo usuários pelos gostos parecidos, nota-se que a maior parte dos usuários (mais de 75%) não faz parte das cinco maiores comunidades, se distribuindo entre as 729 totais. Olhando o número de avaliações por usuário globalmente e, tendo em mente a distribuição original (Figura 2), podemos concluir que a experiência dos usuários torna os gostos por filmes bem definidos, levando à formação das comunidades.

Na Figura 4, estão desenhadas em um gráfico de paralelas coordenadas as avaliações médias de cada uma das cinco maiores comunidades. Nota-se que a formação de comunidades observável na visualização também se mostra no gráfico quando observamos que as comunidades 0 e 17 são as únicas das cinco que avaliaram bem filmes do gênero "Film-Noir". Outro fato interessante é que a comunidade 340 se destaca por ser a menor destas, cerca de duas vezes menor do que a próxima em ordem crescente. Aliado a isto tem a menor média de avaliações por usuário dentre as maiores comunidades, mas mesmo assim, a maior média de avaliação pra todos os gêneros, excluindo "Film-Noir". Conforme o estudo conduzido em [Moon et al. 2010], espera-se que usuários mais experientes sejam mais duros nas avaliações dos filmes. Com isso, podemos concluir que as comunidades 340 e 8 provavelmente estão próximas do limiar de experiência para o fenômeno observado por [Moon et al. 2010].

Tabela 1. Informações sobre as cinco maiores comunidades encontradas na rede de 50 mil usuários. Em destaque, as mesmas informações considerando todos os usuários da rede.

Comunidade	Avaliações por usuário	Usuários
16	1234,42	2949
8	1089,77	2283
0	1752,35	2192
17	1430,44	1907
340	1032,08	956
Total	**467,26**	**50000**

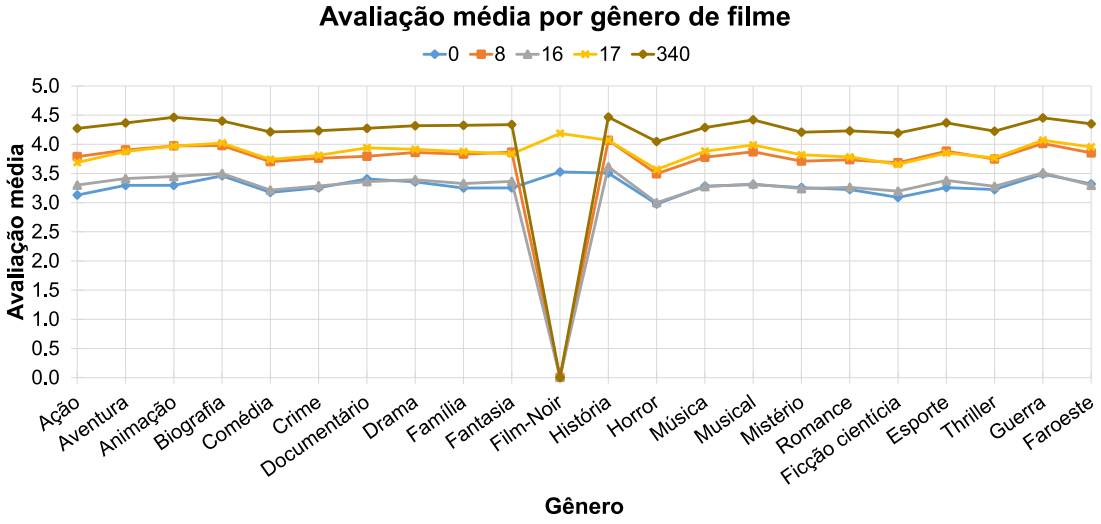

Figura 4. Avaliação média por gênero de filme por comunidade para as cinco maiores comunidades encontradas na rede de 50 mil usuários.

4.2. Rede de usuários via filtragem colaborativa

A rede construída com filtragem colaborativa apresentou características bem diferentes da anterior. Na Figura 5 podemos observar que a formação de comunidades foi bem diferente. Embora também existam vários usuários sem uma comunidade definida (com 551 comunidades no total), as quatro maiores comunidades estão todas misturadas. Embora as comunidades 45 e 391 estejam mais bem definidas, as comunidades 258 e 533 estão bem misturadas entre as duas anteriores.

De forma análoga à análise anterior, a Tabela 2 agrega informações sobre as quatro maiores comunidades da rede. Ao contrário do que é esperado após a análise da primeira, notamos que nesta rede as comunidades refletem mais a característica global da rede nas comunidades em termos de avaliações por usuário, que são bem mais baixas do que as encontradas na rede anterior. Vale notar que, entre os 5 mil usuários, também temos um número médio parecido com o da rede anterior e, a seleção aleatória feita anteriormente obteve uma distribuição de graus parecida com a original. Isto nos leva a crer que não foi uma diferença causada por uma seleção de usuários com características muito diferentes, mas sim pela própria filtragem colaborativa.

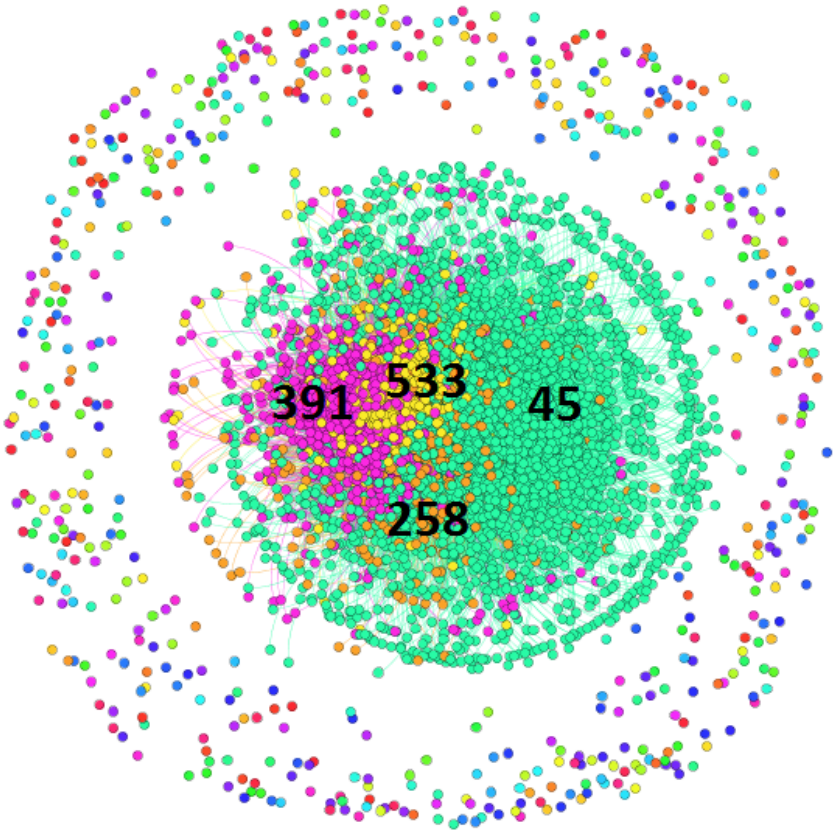

Figura 5. Visualização da rede de 5 mil usuários, com posicionamento dos nós computado pelo algoritmo Yifan Hu. As quatro maiores comunidades estão com seu respectivo número indicado.

A Figura 6 ilustra a média de avaliações por gênero de filme para cada uma das quatro comunidades da rede de 5 mil usuários. Nesta, notamos um fato que contradiz o estudo citado na seção anterior. Mesmo o número de avaliações por usuário sendo bem menor nas comunidades, notamos que as avaliações em si são mais duras em média. Entretanto, sabendo que a filtragem colaborativa possui a característica de capturar tanto informações globais quanto informações locais e olhando para os gêneros populares entre estas comunidades, podemos concluir que esta rede representa comunidades que tenham interesses que refletem filmes mais "comerciais".

5. Conclusões

Neste trabalho, investigamos e avaliamos a criação de redes de usuários que refletem os interesses em filmes, utilizando a base de dados do "Netflix Prize". Partindo da hipótese de que as avaliações dos usuários refletem o gosto por filmes, propusemos uma forma de construir a rede de interesses baseada nas avaliações feitas pelos usuários, agrupadas pelas médias de cada gênero de filme. Para comparação, construímos também uma rede de usuários com base na filtragem colaborativa, uma técnica que define a similaridade entre usuários que está entre as mais utilizadas em sistemas de recomendação.

Os resultados apresentados mostram que a abordagem proposta captura melhor as características de gosto por filmes de usuários mais experientes, revelando que estes

Tabela 2. Informações sobre uas quatro maiores comunidades encontradas na rede de 5 mil usuários. Em destaque, as mesmas informações considerando todos os usuários da rede.

Comunidade	Avaliações por usuário	Usuários
45	509,09	2548
391	343,18	973
533	648,12	648
258	702,09	283
Total	**463,05**	**5000**

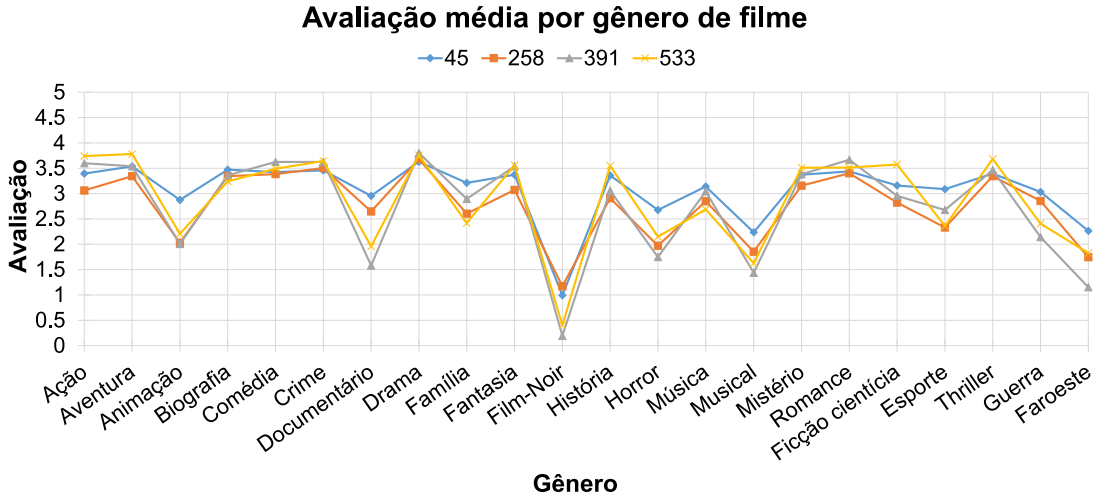

Figura 6. Avaliação média por gênero de filme por comunidade para as quatro maiores comunidades encontradas na rede de 5 mil usuários.

possuem um perfil mais bem definido com relação ao gênero de filme. Por outro lado, a filtragem colaborativa, por sua característica de capturar informações tanto globais quanto locais de interesses, revelou estruturas de comunidades que refletem usuários menos experientes do que no caso anterior, com avaliações mais duras mas mais variadas, refletindo um gosto por gêneros conhecidamente mais "comerciais".

Referências

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*.

Bennett, J. and Lanning, S. (2007). The netflix prize. In *Proceedings of KDD Cup and Workshop*.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424:175–308.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53.

Moon, S., Bergey, P., and Iacobucci, D. (2010). Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *Journal of Marketing*, 74:108–121.

Zhou, T., Ren, J., Medo, M., and Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *The American Physical Society*.