Aplicações de técnicas de predição de links de redes complexas ao domínio de filtragem colaborativa em sistemas de recomendação

Caio Túlio de Deus Andrade

27 de junho de 2025

Resumo

Este trabalho explora a aplicação de técnicas de predição de links em redes complexas para o domínio da filtragem colaborativa em sistemas de recomendação. A partir da matriz de avaliações do conjunto MovieLens 100k, construímos redes complexas de usuários baseadas em diferentes limiares de similaridade de cosseno entre linhas da matriz de avaliações. Comparamos métodos locais e globais de predição de links como medidas de similaridade entre usuários para realizar recomendações. Avaliamos o desempenho dos métodos utilizando Root Mean Squared Error (RMSE) e analisamos propriedades topológicas das redes geradas para diferentes limiares. Os resultados indicam que, embora a predição de links capture características estruturais interessantes, as abordagens tradicionais baseadas em similaridade de cosseno ainda apresentam melhor desempenho preditivo. Discutimos também as implicações dos limiares na topologia da rede e sugerimos como trabalhos futuros podem investigar a evolução dinâmica das preferências em sistemas de recomendação, aproveitando o potencial das redes complexas para modelar dinâmicas temporais.

1 Introdução

Sistemas de recomendação (SRs) tem se tornado cada vez mais prevalecentes no dia a dia de usuários de sistemas inteligentes. O principal objetivo de SRs é inferir as preferências gerais dos usuários de um sistema por meio de um conjunto de avaliações explícitas ou implícitas de usuários com itens (filmes, músicas, livros) de uma base histórica de interações(Aggarwal 2016).

As preferências dos usuários são geralmente representadas por uma matriz de avaliação $M \in \mathbf{R}^{|U|\times|I|}$ altamente esparsa, onde $M_{i,j}$ guarda a nota do usuário i para o item j, ou zero caso não tenha havido interação. Cada linha da matriz representa um usuário e cada coluna, um item.

[1]

[1] A implementação está disponível em `https://github.com/caiotda/link-prediction-for-recommender-systems`.

Uma abordagem comum de recomendação é a filtragem colaborativa, na qual as preferências de um usuário são aprendidas a partir do comportamento de outros usuários similares. A similaridade entre usuários costuma ser calculada com a distância de cosseno entre duas linhas da matriz de avaliação:

$$sim(u,v) = \frac{\sum_{i \in M_u \cap M_v} M_{ui} \times M_{vi}}{\sqrt{\sum_i (M_{ui})^2}} \tag{1}$$

onde M_u e M_v são vetores $\in \mathbf{R}^{|I|}$ que contêm as notas dos usuários u e v para os itens avaliados. Filtragem colaborativa pode ser definida, também, em termos de similaridade entre itens. Essa versão é análoga, porém, utilizando colunas da matriz M.

Em filtragem colaborativa baseada em usuários, a nota prevista que o usuário u daria ao item i pode ser estimada por:

$$\hat{M}_{ui} = \frac{\sum_{v \in N(u)} \text{sim}(u,v) \cdot M_{vi}}{\sum_{v \in N(u)} |\text{sim}(u,v)|} \tag{2}$$

onde $N(u)$ é o conjunto de vizinhos mais similares a u que avaliaram o item i.

A avaliação de um sistema de recomendação é feita pela comparação da nota prevista pelo recomendador com a nota real dada pelo usuário, ou analisando a lista de itens recomendados e comparando com um conjunto de teste contendo itens que o usuário interagiu.

Podemos observar certa similaridade de modelagem entre redes complexas e sistemas de recomendação. Enquanto que em filtragem colaborativa estamos interessados em conectar usuários a itens similares, em redes complexas modelamos as relações em ambientes dinâmicos capturando padrões de interação, formação de comunidades e evolução estrutural que podem influenciar, e ser influenciados por, os mecanismos de recomendação. Existem abordagens na literatura, por exemplo, que modelam o problema de filtragem colaborativa por meio de uma rede bipartite, onde usuários e itens constituem dois componentes da rede complexa (Xia et al. 2016)

O objetivo deste trabalho é iterar na semelhança entre os dois domínios. Iremos modelar filtragem colaborativa como um problema de predição de links, utilizando técnicas listadas em (Martínez, Berzal e Cubero 2016). Iremos detalhar na sessão 2 como geramos uma rede complexa a partir de um conjunto de dados de sistemas de recomendação, além das comparações realizadas. A sessão 3 traz resultados observados e uma discussão. Finalmente, concluímos o trabalho na sessão 4 onde apontamos trabalhos futuros.

2 Metodologia

Nessa sessão iremos detalhar o processo de geração de uma rede complexa a partir do conjunto de dados movielens-100k (Harper e Konstan 2015), composto por 100 mil interações entre usuários e itens na plataforma de avaliação de filmes movielens. Em seguida, iremos detalhar o processo de avaliação das técnicas de predição de links.

2.1 Construindo uma rede complexa

Um dos principais desafios deste trabalho foi compreender como modelar uma rede complexa a partir de uma matriz de avaliações típica de um sistema de recomendação. Em outras palavras, quais são os nós da rede e como criar links entre esses vértices? Embora fosse possível adotar a abordagem relativamente comum de representar os dados como uma rede bipartida, os métodos de previsão de conexões geralmente se baseiam em medidas de proximidade local, que não são bem aplicáveis a esse tipo de rede, pois não existe vizinhança compartilhada entre os nós que pertencem à mesma classe.

Portanto optamos por realizar a construção da rede por uma metodologia baseada em similaridade: sejam dois usuários $u, v \in U$ e $H(u), H(v) \in \mathbf{R^M}$ a linha da matriz de avaliações $\mathbf{R^{M \times N}}$ correspondentes ao histórico de ambos usuários, consideramos cada usuário um nó na rede complexa e geramos uma aresta entre os nós se:

$$cos(H(u), H(v)) \geq L \tag{3}$$

Onde L é um limiar de similaridade de cossenos entre os usuários na base e Cos() representa a similaridade de cossenos entre dois vetores, como detalhado na equação 1. Exploraremos na sessão 3 como o limiar foi escolhido e diferentes estatísticas de topologia das redes geradas para cada limiar. Modelamos uma rede não direcionada e sem pesos. Alternativamente, poderíamos utilizar a similaridade inicial como peso da aresta.

2.2 Métodos testados

Os métodos de predição de link podem ser classificados em duas categorias: os locais e globais. Métodos locais, frequentemente mais simples, tendem a serem baseados nas vizinhanças de um nó enquanto que métodos globais, mais complexos, utilizam medidas de distâncias e topologias da rede para inferir a similaridade entre dois nós. Por simplicidade, utilizamos apenas métodos locais e um método global: o método global foi selecionado por ser implementado nativamente na biblioteca networkx, nos possibilitando ter representantes das duas categorias de métodos de predição de link sem muita complexidade adicional.

Selecionamos os seguintes métodos de predição de link para serem comparados em relação a uma filtragem colaborativa baseada em usuários:

1. Vizinhos em comum: definido como o tamanho da intersecção de vizinhos. Um método simples e utilizado como base para todos os outros métodos de vizinhança

$$\mathrm{CN}(a, b) = |\Gamma(a) \cap \Gamma(b)| \tag{4}$$

2. Índice de Sørensen semelhante ao índice Jaccard, mas menos sensível a outliers. Utilizado para evitar influências de itens populares no histórico dos usuários.

$$\mathrm{Sørensen}(a, b) = \frac{2 \cdot |\Gamma(a) \cap \Gamma(b)|}{|\Gamma(a)| + |\Gamma(b)|} \tag{5}$$

3. Indice detrator de hubs: favorece a conexão entre nó de grau baixo e hubs, conectando usuários novos a usuários ativos.

$$\text{HPI}(a, b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{\min(|\Gamma(a)|, |\Gamma(b)|)} \tag{6}$$

4. Índice promotor de hubs: favorece a conexão entre hubs, conectando usuários ativos.

$$\text{HDI}(a, b) = \frac{|\Gamma(a) \cap \Gamma(b)|}{\max(|\Gamma(a)|, |\Gamma(b)|)} \tag{7}$$

5. Indice de preferential attachment: inspirado no modelo Barabási-Albert (Barabási 2013), define similaridade entre nós com base na lei de potência

$$\text{PA}(a, b) = |\Gamma(a)| \cdot |\Gamma(b)| \tag{8}$$

6. Negative shortest path: O caminho mais curto entre dois nós é a menor quantidade de arestas ou soma de pesos que os conecta. No NetworkX, usamos a função 'nx.shortest_path', que aplica BFS para grafos não ponderados e Dijkstra para ponderados. Como medida de similaridade, utilizamos o valor negativo dessa distância, atribuindo pontuações maiores a pares de nós mais próximos.

$$\text{NSP}(a, b) = -d(a, b) \tag{9}$$

Compararemos o desempenho de cada um dos métodos para a tarefa de prever a nota que um usuário dará a um item. Sendo u, v dois usuários na base, utilizamos cada um dos métodos acima como uma medida de similaridade entre u, v. Consequentemente, fomos capazes de prever a avaliação dada por um usuário a um item i utilizando a equação 2. A vizinhança de usuários mais similares a u foi definida como os 'k=20' vizinhos mais próximos, seguindo as medidas de similaridade definidas acima.

Para compararmos a performance das diferentes variantes de filtragem colaborativa, medimos o erro quadrático médio:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_{u,i} - \hat{r}_{u,i})^2} \tag{10}$$

Neste trabalho, optamos por usar toda a base do MovieLens 100k tanto para calcular as similaridades quanto para avaliar as predições via RMSE. Sabemos que isso pode introduzir certo viés por *Data Leakage*, já que estamos avaliando com dados que também ajudaram a construir o modelo. No entanto, como todos os métodos comparados estão sujeitos exatamente ao mesmo processo, esse efeito acaba sendo neutralizado nas comparações. O foco aqui está menos no valor absoluto do erro e mais em como os diferentes métodos se comportam em relação uns aos outros.

3 Resultados e Discussão

3.1 Análise da topologia das redes

Primeiramente, iremos apresentar os resultados obtidos ao criar a rede complexa. Como definido na equação 2.1, criamos diversas redes complexas utilizando diferentes limiares mínimos de usuários para criarmos links entre nós. Naturalmente, quanto maior o limiar, mais esparsa se torna a rede. Para analisar esse efeito, testamos diferentes valores de limiar de similaridade: 0,5, 0,6, 0,7, 0,8, 0,9, 0,95 e 0,99, cobrindo desde cenários moderadamente restritivos até configurações extremamente seletivas. Para cada rede, medimos o grau médio, variância, segundo momento, agrupamento médio, clustering global, caminho mais curto médio, índice de assortatividade, número de componentes na rede e número de arestas (Costa et al. 2007). Realizamos as medidas na rede inteira, exceto em casos que a rede apresentou mais de uma componente. Nesse caso, a medida foi realizada na maior componente. Os resultados podem ser encontrados na tabela 1

Tabela 1: Métricas topológicas para diferentes valores de limiar de similaridade

L	$\langle k \rangle$	$\langle k^2 \rangle$	σ^2	$\langle C \rangle$	C	$\langle d \rangle$	r	n_c	$\lvert E \rvert$
0,50	469,52	229999,58	9548,40	0,564	0,570	1,501	0,113	1	221145
0,60	375,62	148883,65	7794,89	0,473	0,479	1,601	0,105	1	176916
0,70	281,71	86021,52	6659,10	0,386	0,384	1,701	0,011	1	132687
0,80	187,81	40875,10	5602,91	0,301	0,283	1,800	-0,124	1	88458
0,90	93,90	12045,36	3227,31	0,198	0,166	1,933	-0,249	1	44229
0,95	48,13	3756,03	1439,69	0,103	0,087	2,211	-0,301	1	22115
0,99	2,21	8,32	3,45	0,000	0,000	7,361	-0,271	48	191

Notas sobre a notação:

- L: limiar de similaridade aplicado à matriz de adjacência.
- $\langle k \rangle$: grau médio dos nós da rede.
- $\langle k^2 \rangle$: segundo momento da distribuição de grau.
- σ^2: variância da distribuição de grau.
- $\langle C \rangle$: coeficiente de agrupamento médio (clustering local médio).
- C: coeficiente de agrupamento global (proporção de triângulos na rede).
- $\langle d \rangle$: comprimento médio do caminho mais curto entre pares de nós conectados.
- r: coeficiente de assortatividade (correlação entre graus dos nós conectados).
- n_c: número de componentes conexos na rede.
- $|E|$: número total de arestas.

À medida que aumentamos o valor do limiar de similaridade, a rede resultante torna-se progressivamente mais esparsa. Isso é evidenciado pela queda acentuada no grau médio dos nós ($\langle k \rangle$), que diminui de aproximadamente 470 para apenas 2,2 quando o limiar passa de 0,50 para 0,99. Esse comportamento reflete a eliminação de conexões consideradas fracas, restringindo os vínculos na rede apenas àqueles de alta similaridade.

Como consequência direta dessa filtragem, observa-se uma redução tanto no segundo momento do grau ($\langle k^2 \rangle$) quanto na variância (σ^2), indicando menor heterogeneidade na distribuição de conexões. O coeficiente de agrupamento médio ($\langle C \rangle$) e global (C) também caem substancialmente, revelando que a rede perde densidade local e estrutura de comunidade à medida que conexões mais fracas são descartadas.

Outro efeito relevante é o aumento do comprimento médio dos caminhos mais curtos ($\langle d \rangle$), que cresce de 1,5 para mais de 7, à medida que a rede se torna mais desconectada e menos eficiente em termos de conectividade. Esse efeito é acentuado no limiar 0,99, em que surgem múltiplas componentes desconexas ($n_c = 48$), refletindo uma fragmentação significativa da estrutura global.

Por fim, nota-se uma mudança no padrão de assortatividade (r), que parte de um valor levemente positivo e se torna negativo à medida que o limiar aumenta. Isso sugere que, em redes mais filtradas, os nós de alto grau tendem a se conectar preferencialmente com nós de baixo grau, um comportamento característico de redes mais hierárquicas ou estruturadas. Provavelmente nesses casos, usuários ativos que consomem muito conteúdo se tornam hubs para outros usuários.

Um outro resultado interessante (relacionado com as discussões acima) se observa ao visualizarmos a distribuição de grau para cada limiar na figura 1. A curva azul representa a distribuição real e a curva laranja uma suavização utilizando uma janela de tamanho 5.

Observamos que, ao aumentar o limiar de similaridade entre usuários, a distribuição de grau da rede se aproxima de uma lei de potência em vez de uma curva aproximadamente normal (observada para limiares menores que 0,9). A distribuição de grau observada segue uma lei de potência, o que é uma característica típica de redes sem escala como as geradas pelo modelo de Barabási–Albert. No entanto, essa semelhança não garante que a rede tenha sido formada por um processo de *preferential attachment*. Como proposta para trabalhos futuros, podemos investigar se a rede construída a partir da matriz de similaridade realmente exibe as propriedades do modelo de Barabási–Albert. Caso isso se confirme, seria possível explorar o mecanismo de *preferential attachment* como estratégia para incluir novos usuários na rede, oferecendo uma alternativa mitigar o problema de cold-start de preferências em sistemas de recomendação (Lika, Kolomvatsos e Hadjiefthymiades 2014).

3.2 Comparativo de performance entre os diferentes métodos

Como as medidas de vizinhança funcionam melhor para uma única componente (e, no caso do caminho mais curto, só estão definidas em uma única component), escolhemos utilizar a rede com limiar de similaridade L = 0.95. Isso nos permite utilizar uma rede com um único componente, mas com um grau médio menor, potencialmente obtendo menos ruído ao determinar os vizinhos mais similares a um usuário.

Para definir a performance de cada variante, comparamos a avaliação prevista utilizando cada variante de função de similaridade com a nota real dada pelo usuário. Os resultados

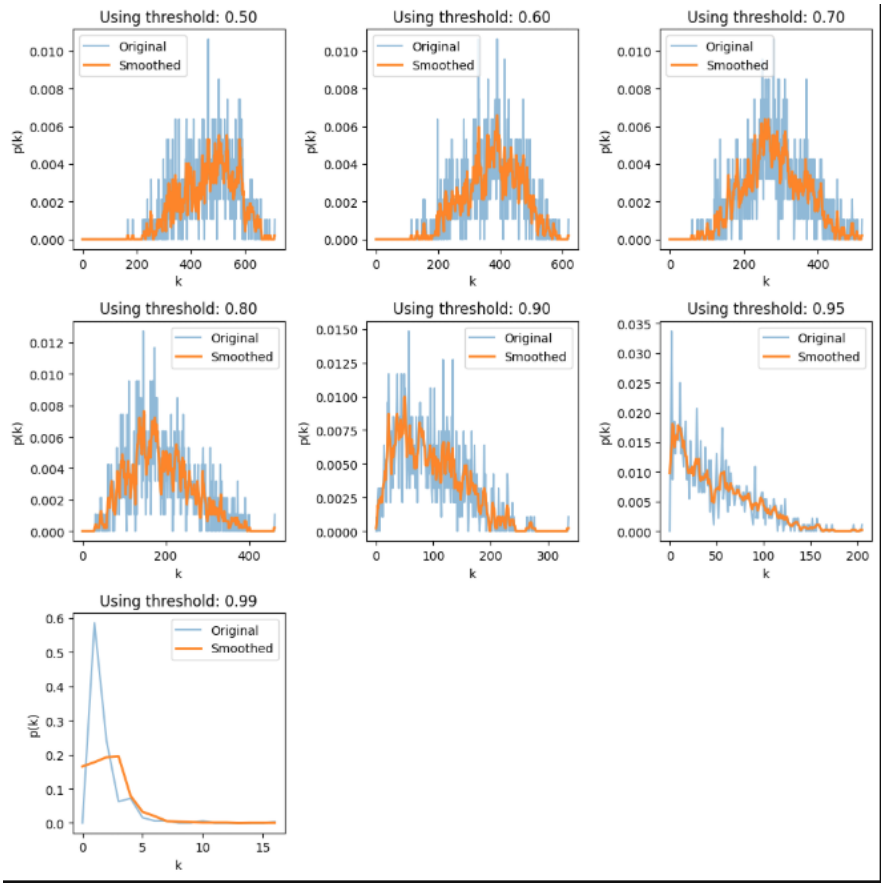

Figura 1: Distribuição de grau para diferentes limiares de similaridade.

reportados na tabela 2 foram medidos em termos de RMSE. Os experimentos foram realizados considerando 20 vizinhos para calcular a predição de avaliação.

Pelos nossos experimentos, as variantes baseadas em predição de link não superaram o baseline para a rede com L=0.95 e considerando k=20 vizinhos. É interessante notar que a variante baseada em topologia global (Caminho mais curto negativo) obteve o pior desempenho, mostrando que métricas baseadas em vizinhança são mais apropriadas à tarefa. Analisamos também a distribuição de erro médio na figura 2

4 Conclusão

Neste trabalho, testamos medidas de predição de link como uma medida de similaridade entre usuários para a tarefa de filtragem colaborativa baseada em usuários. Notamos que as variáveis não foram capazes de superar o baseline baseado em similaridade de cossenos. Em trabalhos futuros, seria interessante verificar se esse comportamento se mantém utilizando diferentes limiares de similaridade para geração de redes e utilizando um número diferente de vizinhos para realizar a recomendação.

Um experimento que não foi realizado neste trabalho, por limitação de tempo, envolve a avaliação de recomendações em um cenário dinâmico. Em vez de apenas uma rodada de

Tabela 2: Performance de diferentes funções de similaridade para filtragem colaborativa em RMSE

Função de Similaridade	**RMSE**
Common Neighbors	0.968
Sorensen Index	0.964
Hub Promoter Index	0.969
Preferential Attachment	0.990
Caminho Mais Curto (negativo)	1.278
User-kNN (cosseno)	0.926

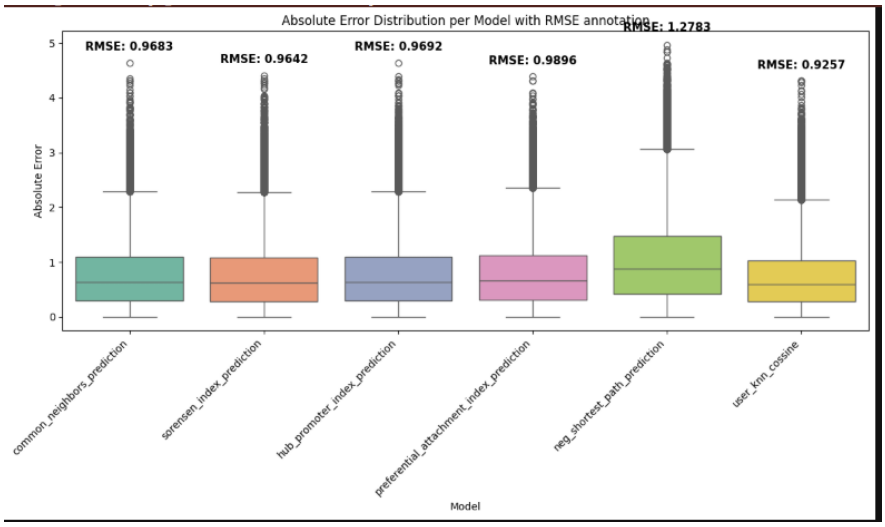

Figura 2: Boxplot do erro absoluto e RMSE de cada variante

recomendação seguida de avaliação, seria interessante testar como as variantes se comportam ao longo de múltiplas iterações, em que as recomendações são geradas sequencialmente e comparadas com dados históricos dos usuários. Em outras palavras, trata-se de investigar qual abordagem lida melhor com a evolução temporal das preferências, simulando um ambiente mais próximo de uso real. um cenário onde redes complexas se mostraram modelos poderosos para capturar dinâmicas e estruturas temporais em sistemas variados.

Referências

Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. 1ª ed. Springer Publishing Company.

Barabási, A.-L. (2013). "Network science". Em: *Philosophical Transactions of the Royal Society A* 371, p. 20120375. URL: http://doi.org/10.1098/rsta.2012.0375.

Costa, L da F et al. (2007). "Characterization of complex networks: A survey of measurements". Em: *Advances in physics* 56.1, pp. 167–242.

Harper, F. M. e J. A. Konstan (2015). "The MovieLens datasets: History and context". Em: *ACM Transactions on Interactive Intelligent Systems (TIIS)* 5.4, pp. 1–19.

Lika, Blerina, Kostas Kolomvatsos e Stathes Hadjiefthymiades (2014). "Facing the cold start problem in recommender systems". Em: *Expert Systems with Applications* 41.4, Part 2, pp. 2065–2073. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2013.09.005. URL: https://www.sciencedirect.com/science/article/pii/S0957417413007240.

Martínez, V., F. Berzal e J.-C. Cubero (2016). "A survey of link prediction in complex networks". Em: *ACM Computing Surveys (CSUR)* 49.4, pp. 1–33.

Xia, Jianxun et al. (2016). "Modeling recommender systems via weighted bipartite network". Em: *Concurrency and Computation: Practice and Experience*. INBIw algorithm proposed; validação em MovieLens. DOI: 10.1002/cpe.3895.