

Desmistificando o uso de *Matrix Factorization* em domínios de recomendação

Nícollas Silva , Leonardo Rocha , Fernando Mourão

DCOMP/UFSJ - São João del Rei, Minas Gerais, Brasil

{nicollasilva, lcrocha, fhmourao}@ufsj.edu.br

Abstract. *Despite the popularity of Matrix factorization methods (MF) in Recommender Systems (RSs), it is not clear the practical distinction among them in real scenarios. In this context, this work organizes existing methods, proposing a taxonomic organization for the area, and characterizes their performance, correlating the application success with domain characteristics. Aiming at the practical use of this research, we present a characterization framework for MF methods applied to RSs, useful for contrasting the performance of different methods. Analyses of actual collections have shown that recommendations generated via SVD tend to focus on non-popular items, while recommendations issued by PCA focuses on popular items.*

Resumo. *Apesar da popularidade de métodos de Matrix Factorization (MF) em Sistemas de Recomendação (SsR), não é clara a distinção prática entre eles em domínios reais. Neste contexto, este trabalho organiza métodos existentes, propondo uma organização taxonômica para a área, e caracteriza seus desempenhos, correlacionando seus sucessos da aplicação com características de cada domínio. Visando o uso prático desta pesquisa, apresentamos um arcabouço para caracterização de métodos de MF aplicados a SsR, útil para se contrastar o desempenho de distintos métodos. Análises sobre coleções reais demonstraram que recomendações geradas via SVD tendem a focar em itens não populares, enquanto recomendações oriundas do PCA focam em itens populares.*

1. Introdução

Atualmente, Sistemas de Recomendação (SsR) constituem uma das mais importantes ferramentas utilizadas para auxiliar usuários, de maneira personalizada, na tomada de decisão em variados domínios reais [Koren et al. 2009]. Dentre as diversas propostas existentes na literatura, particular atenção é dada a SsR baseados em Fatoração de Matrizes (i.e., *matrix Factorization* - MF), dada a elegante formalização que MF proporciona ao problema de recomendação, bem como a alta eficiência frequentemente observada em sua aplicação neste tipo de problema. De fato, os principais recomendadores estado-da-arte atualmente utilizam em alguma etapa do processo de recomendação um método baseado em MF [Sharifi et al. 2013, Adomavicius and Tuzhilin 2005]. A maioria destes SsR correspondem a adaptações para os domínios de recomendação dos tradicionais métodos de Singular Value Decomposition (SVD) e o Principal Component Analysis (PCA).

Apesar da crescente popularidade destes métodos, os trabalhos atuais não se preocupam em entender e contrastar o real impacto de cada adaptação sobre a tarefa de

recomendação. Perguntas tais como: *Quais as diferenças e semelhanças entre as diversas propostas de adaptação? Diferenças conceituais determinam diferenças práticas quanto ao desempenho destes SsR? Quais domínios são mais apropriados para cada tipo de técnica?* apesar de pertinentes são negligenciadas na literatura. Responder tais perguntas pode nos permitir gerar recomendações com melhor qualidade. Assim, este trabalho propõe revisar, caracterizar, organizar e distinguir distintos métodos de recomendação baseadas em MF. Além disso, almejamos correlacionar características dos domínios de análise com a qualidade das recomendações geradas por cada método avaliado.

Neste sentido, primeiramente, consolidamos uma taxonomia de classificação para diversos métodos de recomendação baseados em MF, com base em características comumente exploradas por estes métodos. Tal taxonomia considera cinco dimensões, associadas a passos distintos do processo de recomendação, identificadas através de análises não-automáticas sobre diversos métodos existentes. Estas dimensões levam em conta a tarefa que está sendo realizada pela técnica, a forma de modelagem dos dados de entrada, o método de fatoração utilizado, a estratégia de recomendação e a fórmula de predição utilizada. Cabe ressaltar que a taxonomia proposta é não-fechada e extensível, uma vez que não engloba todas as características potencialmente relevantes, nem todos os métodos existentes. Em seguida, organizamos 16 SsR distintos baseados em MF de acordo com a taxonomia proposta, enfatizando diferenças e semelhanças entre estes. Dessa forma, observamos mais facilmente que métodos tradicionalmente apresentados como distintos, tais como *SimpleSVD* e *ConventionalSVD*, são na verdade idênticos. Ou ainda, constatamos que a diferença entre os métodos *NormalizedSVD* e *SVDP* restringe-se apenas à fórmula de modelagem dos dados utilizada. De maneira geral, por meio da taxonomia proposta, este trabalho auxilia pesquisadores da área a facilmente visualizarem conceitos e semelhanças entre técnicas, que são apresentadas de maneira confusa na literatura.

Além de organizar qualitativamente os diversos SsR baseados em MF usando suas premissas e abordagens, propomos uma metodologia de caracterização capaz de quantificar o comportamento de cada método, bem como correlacionar indicadores de qualidade de SsR com algumas características das coleções de entrada. Consequentemente, visamos entender melhor as características, pontos fortes e fracos de cada método, bem como o domínio mais adequado para a aplicação destes. Por fim, consolidamos um arcabouço para caracterização aplicável a distintos domínios (e.g., músicas, vídeos ou notícias). Para tal, construímos quatro módulos de análise distintos, baseado nas etapas da metodologia de caracterização. A partir deste arcabouço é possível, facilmente, quantificar as principais diferenças práticas entre os distintos SsR baseados em MF. De fato, avaliações em domínios reais sugerem que técnicas baseadas em SVD e PCA recomendam itens distintos e que combiná-las em uma técnica híbrida pode melhorar significativamente os resultados encontrados. Além disso, observamos que características do domínio e individuais de cada usuário podem afetar o desempenho do SVD e PCA em recomendação.

Enfatizamos que todas as implementações, execuções de experimentos e análises de resultados foram realizadas pelo aluno Nícollas Silva, sob a orientação dos professores Fernando Mourão e Leonardo Rocha. A construção da metodologia de caracterização foi conduzida em conjunto, aluno e professores.

2. Trabalhos Relacionados

Métodos de MF oriundos da Álgebra Linear (AL) e da Estatística (ES) têm se tornado cruciais para o sucesso da tarefa de recomendação em variados domínios [Yin and Peng 2012, Sarwar et al. 2000]. Estes métodos apresentam alta eficácia, especialmente, quando combinados a estratégias de Filtragem Colaborativa, uma vez que eles conseguem através da redução de dimensionalidade atenuar o problema da esparsidade comum a domínios de recomendação. Além da recorrente eficácia, a utilização de MF em recomendação permite uma elegante formalização do problema e requer apenas uma simples transformação dos dados para um formato numérico matricial.

Dessa forma, um número crescente de trabalhos vem utilizando métodos de MF baseados em SVD e PCA nas tarefas de recomendação. Por exemplo, em [Yin and Peng 2012], é proposta uma técnica de predição de *ratings* denominada *SVDP* (*SVD prediction algorithm*), que considera os fatores latentes resultantes da aplicação do SVD para relacionar itens a usuários. Em [Sharifi et al. 2013], é apresentado o *Simple SVD*, que implementa o SVD básico na tarefa de predição, o *Normalized SVD*, que normaliza os *ratings* de entrada antes de aplicar a predição, além de algumas variações da modelagem dos dados de entrada, tais como *SimpleSVD*, *MIDItemSVD*, *MIDUserSVD* e *MIDTotalSVD*. As técnicas *Conventional SVD* e *NSVD* são descritas em [Koren 2008]. Neste mesmo trabalho é apresentada uma das técnicas mais eficientes da literatura, o *Assymmetric SVD*, uma extensão da técnica do *SVD++*, proposta em [Bell and Koren 2007]. Ambas são técnicas de vizinhança inspiradas no SVD, que consideram os fatores latentes resultantes da aplicação do SVD relacionados aos usuários e aos itens. Outra técnica inspirada em SVD e de grande relevância na literatura é o *FunkSVD*, proposta em [Funk 2006] e descrita em [Paterrek 2007] como *RSVD* (*Regularized SVD*).

Em outro conjunto de abordagens, como em [Yin and Peng 2012], os autores focam em técnicas híbridas no intuito de se correlacionar as vantagens da aplicação de SVD e PCA com técnicas de Aprendizado de Máquina clássicas, tais como relacionados à tarefa de classificação. Entre elas, destacam-se as técnicas *PCA-KNN* (*Hybrid algorithm based on PCA and KNN*) e *SVD-KNN* (*Hybrid algorithm based on SVD and KNN*), que foram propostas na tentativa de se combinar as vantagens do PCA e SVD com o KNN (*k-Nearest Neighbour*). Outras técnicas híbridas eficientes na tarefa de recomendação que devem ser destacadas são a *Eigentaste*, proposta em [Goldberg et al. 2001], e a *Neighborhood Model*, proposta em [Koren 2010], também inspiradas em métodos de MF.

A grande diversidade de propostas para aplicação de MF em SsR, porém, faz surgir um desafio para a área. Em geral, os estudos não se preocupam em entender as diferenças teóricas e práticas entre as distintas propostas. Pesquisas recentes focam em propor novas estratégias, negligenciando o estudo de domínios e características que tornam as técnicas existentes diferentes. Dessa forma, questões como: *Qual a diferença prática de se aplicar SVD ou PCA em SsR?*, permanecem em aberto, dificultando a consolidação do conhecimento ao longo do tempo. De fato, SVD e PCA são confundidos inclusive em outros domínios além de recomendação, como mostramos no trabalho premiado no CTIC 2015 [Silva et al. 2015]. Este trabalho representa assim uma extensão do CTIC 2015, visando distinguir conceitos usualmente confusos quando aplicamos MF em SsR.

3. Taxonomia para SsR baseados em MF

Nesta seção, apresentamos uma taxonomia para SsR baseados em métodos de MF. Nosso intuito é criar uma classificação para as técnicas mais relevantes da literatura atual, com base em características comuns a estas técnicas. A partir de uma análise não-automática sobre os principais trabalhos da área, identificamos cinco dimensões de classificação comumente associadas a distintos passos do processo de recomendação. Enfatizamos que a taxonomia proposta é não-fechada, uma vez que não engloba todas as características potencialmente relevantes, nem todos os métodos existentes. Porém, tal taxonomia é facilmente extensível a qualquer outra característica ou método. Considerando o escopo deste trabalho, avaliamos cada método com base nas seguintes dimensões:

1. **Tipo de tarefa:** refere-se ao tipo de tarefa de recomendação que está sendo realizada, a qual pode ser a tarefa de **predição**, quando a técnica prediz os valores dos *ratings* para alguns itens, ou mesmo a tarefa **Top-N**, quando a técnica recomenda N itens aos usuários.
2. **Forma de imputação:** analisamos como cada uma das técnicas lida com os *ratings* desconhecidos no conjunto de dados, onde na sua maioria são preenchidos com valores: **zeros**, mediana dos *ratings* dados pelos usuários (**MID Users**), média dos *ratings* recebidos por cada item (**MID Items**), média total dos *ratings* do conjunto (**MID Total**) ou mesmo desconsiderando estes itens não classificados.
3. **Método de fatoração:** refere-se a cada um dos métodos, podendo ser, por exemplo, o SVD, PCA de Covariância ou Correlação. Embora existam outros métodos de fatoração aplicados a recomendação, nossas análises restringem-se apenas à comparação destes três métodos.
4. **Estratégia de recomendação:** compreende a estratégia de exploração do método de fatoração no processo de recomendação. As técnicas que se baseiam nas propriedades algébricas do SVD e PCA, usando-os diretamente para a tarefa de predição de *ratings*, são classificadas como *Conventional Strategy*. Por outro lado, as técnicas que tem como estratégia, adaptar o modelo de predição de *ratings* de SVD e PCA para a aplicação de outros métodos, são classificadas como *Adaptation Strategy*. Por sua vez, as técnicas que utilizam SVD e PCA apenas como modelos para refinar os dados de entrada, no intuito de se aplicar outros métodos de recomendação, são classificadas como *Model Refinement Strategy*.
5. **Fórmula de predição:** representa a função de geração de *score* ou *ratings*, baseada nas funções mais utilizadas na literatura.

Dada a relevância prática dos métodos de Fatoração de Matrizes em domínios de SsR, uma importante contribuição deste trabalho consiste na organização das técnicas mais relevantes da literatura. Basicamente, aplicamos a taxonomia proposta, nas técnicas citadas nos trabalhos relacionados. A tabela 1 sumariza todas nossas análises, classificando cada técnica quanto ao grupo em que a técnica se encaixa, o método que é utilizado, a forma com que o conjunto de dados é tratado, a regra de predição dos *ratings* e o tipo de dados que são considerados. Note que, as técnicas que se encaixam no grupo de *Model Refinement Strategy* não seguem as regras de predição específicas, comuns nos outros modelos, uma vez que SVD e PCA são aplicados apenas para refinar a modelagem dos dados e não para gerar as predições propriamente ditas.

Ressaltamos que, por meio dessa taxonomia, podemos analisar a semelhança entre as técnicas já existentes. As técnicas intituladas *SimpleSVD* e *ConventionalSVD* são semelhantes, enquanto que, a técnica *Asymmetric SVD* é uma extensão da técnica *SVD++*. Da mesma forma, nota-se que o *Normalized SVD* consiste nos mesmos passos do *SVDP*, porém normalizando o conjunto de entrada através da técnica *z-score*. Entender

tais semelhanças é uma importante contribuição, auxiliando pesquisadores da área a compreender melhor conceitos e métodos apresentados de forma confusa na literatura.

Tabela 1. Classificação das técnicas mais relevantes da literatura atual seguindo a taxonomia proposta por este trabalho.

	Técnica	Fatoração	Imputação	Fórmula de predição
Conventional	PureSVD *	SVD	zeros	$r_{ui} = r_u \cdot Q \cdot q_i^T$
	SVDP	PCA-Corr	zeros	$r_{ui} = \overline{b_{ui}} + \sigma_i(Z_d) \cdot u_k$
	Simple SVD	SVD	zeros	$r_{ui} = \overline{b_{ui}} + r_u \cdot Q \cdot q_i^T$
	Conventional SVD	SVD	zeros	$r_{ui} = \overline{b_{ui}} + r_u \cdot Q \cdot q_i^T$
	Normalized SVD	PCA-Corr	zeros (z-score)	$r_{ui} = \overline{b_{ui}} + \sigma_i(Z_d) \cdot u_k$
	MIDItem SVD	SVD	MID Items	$r_{ui} = \overline{b_{ui}} + r_u \cdot Q \cdot q_i^T$
	MIDUser SVD	SVD	MID Users	$r_{ui} = r_u \cdot Q \cdot q_i^T$
	MIDTotal SVD	SVD	MID Total	$r_{ui} = r_u \cdot Q \cdot q_i^T$
Adaptation	NSVD	SVD	ratings existentes	$r_{ui} = \overline{b_{ui}} + q_i^T \left(\sum_{j \in R(u)} r_j \cdot R(u) ^{-\frac{1}{2}} \right)$
	SVD ++	SVD	valores binários	$r_{ui} = \overline{b_{ui}} + q_i^T \left(p_u + \sum_{j \in N(u)} y_j \cdot N(u) ^{-\frac{1}{2}} \right)$
	Asymmetric SVD	SVD	ratings existentes ou não	$r_{ui} = \overline{b_{ui}} + q_i^T \left(\sum_{j \in R(u)} (r_{uj} - b_{uj}) x_j \cdot R(u) ^{-\frac{1}{2}} + \sum_{j \in N(u)} y_j \cdot N(u) ^{-\frac{1}{2}} \right)$
Refinement	Funk SVD	SVD	ratings existentes	técnica de regressão linear
	SVD-KNN	SVD	zeros	-
	PCA-KNN	PCA-Corr	zeros (z-score)	-
	Eigentaste	PCA-Corr	zeros (z-score)	-
Neighbor Model	SVD	zeros ou somente os itens com ratings	-	

* representa técnicas que pertencem a tarefa Top-N, enquanto as demais pertencem a tarefa de predição.

4. Sobre as diferenças práticas entre SsR baseados em MF

Nesta seção, apresentamos uma metodologia e um arcabouço para caracterização do comportamento dos métodos de Fatoração de Matrizes aplicados a recomendação. Por meio deste arcabouço propomos formas de, facilmente, quantificarmos as principais diferenças entre as distintas técnicas baseadas em MF. Consequentemente, visamos entender melhor as características, pontos fortes e fraquezas de cada método, bem como o domínio mais adequado para a aplicação destes.

4.1. Metodologia de Caracterização

De forma a quantificar os impactos da aplicação de métodos de MF sobre a tarefa de recomendação, propomos uma metodologia de caracterização. Nossa metodologia estabelece métricas e procedimentos que nos permitem correlacionar características inerentes às coleções de entrada com indicadores de qualidade na aplicação destes métodos. Um uso prático dessa caracterização seria a identificação de padrões relevantes e recorrentes que nos permitam prever o desempenho dos métodos de MF de acordo com características de cada domínio. Nossa metodologia consiste nos quatro passos descritos abaixo. A tabela 2 apresenta uma sumarização dos indicadores de qualidade avaliados, características selecionadas, bem como as principais métricas de correlação utilizadas.

- Seleção e implementação dos métodos de Fatoração de Matrizes:** baseados na nossa organização taxonômica, o primeiro passo consiste em selecionar e implementar técnicas oriundas de categorias distintas de forma a quantificar o desempenho de cada uma na tarefa de recomendação;
- Seleção das características de análise das coleções:** Selecionamos neste passo quatro características comumente relevantes em domínios de recomendação;

3. **Seleção dos indicadores de qualidade:** selecionamos três métricas bem difundidas no domínio de recomendação, capazes de avaliar a qualidade das recomendações geradas, considerando diferentes perspectivas de qualidade;
4. **Análise de correlação entre características e indicadores:** no intuito de se identificar quais características de cada domínio foram relevantes para o desempenho das técnicas implementadas, correlacionamos as características analisadas com os resultados obtidos pelos indicadores de qualidade.

Tabela 2. Características e indicadores de qualidade utilizados na metodologia.

Indicadores	<ol style="list-style-type: none"> 1. Precision: representa a probabilidade que um item selecionado seja relevante e é definida como a razão dos itens relevantes selecionados pelo número de itens recomendados. 2. Recall: definida como a razão entre o número de itens relevantes selecionados pelo número de itens relevantes existentes. <i>Recall</i> representa a probabilidade de um item relevante ser selecionado. 3. Taxa de Acertos: representam o número de acertos do recomendador. Simplesmente verificamos quantos itens da lista Top-N, gerada pelo recomendador, foram de fato <i>consumidos</i> pelos usuários de acordo com as informações retidas no conjunto teste.
Características da Coleção	<ol style="list-style-type: none"> 1. Popularidade dos itens: De maneira simples, para cada item i da coleção, calculamos quantos usuários distintos consumiram-no. Em seguida, ordenamos decrescentemente os valores obtidos construindo uma distribuição de popularidade. Essa distribuição, quando correlacionada às técnicas, permite analisar se os itens recomendados estão relacionados à distribuição <i>long tail</i> ou não. 2. Histórico de consumo dos usuários: Para cada usuário u da coleção, mensuramos a quantidade de itens distintos que este usuário consumiu. Novamente, ordenamos decrescentemente os valores obtidos, agora resultando em uma distribuição de histórico de consumo. Tal distribuição nos permite demonstrar se o número de acertos da técnica está ou não relacionado à assiduidade de cada usuário dentro do conjunto de dados. 3. Nota média de cada usuário: Para cada usuário u da coleção, calculamos a média dos <i>ratings</i> atribuídos por u, simplesmente somando os valores existentes e dividindo pela quantidade de itens avaliados por u. No fim, ordenamos decrescentemente os valores de média, obtendo uma distribuição de média que representa todos os usuários da coleção. Com isso, podemos verificar se o desempenho das técnicas, estão relacionadas às notas atribuídas pelos usuários. 4. Variância média das notas do usuário: Para cada item i da coleção, calculamos a variância dos valores de <i>ratings</i> atribuídos a i por todos os usuários que o consumiram. Posteriormente, ordenamos decrescentemente os valores obtidos, a fim de construir uma distribuição de variância dos <i>ratings</i> da coleção. Com essa distribuição, podemos verificar se o desempenho das técnicas, estão relacionados ou não com a variância das notas atribuídas pelos usuários.
Correlação	<ol style="list-style-type: none"> 1. Correlação de Spearman: teste não-paramétrico, que não exige nenhum pressuposto de distribuição normal, que aplica uma métrica de correlação de <i>rankings</i> ou "postos". 2. Correlação de Kendall-tau: métrica de correlação de <i>rankings</i> mais sensível ao número de observações existentes.

4.2. Ferramenta de Caracterização

Com base na metodologia de caracterização apresentada, consolidamos um arcabouço capaz de realizar o estudo comparativo dos métodos de MF na tarefa de recomendação. Novamente, este trabalho foca especificamente nos métodos de MF denominados SVD e PCA. Motivados pela confusão dos conceitos de SVD e PCA no domínio de recomendação, vamos avaliar as técnicas que se encontram no grupo de *Conventional Strategy*, uma vez que estas técnicas são a base para os outros grupos. Além disso, limitamos nossas análises à tarefa Top-N, dada sua relevância prática em domínios reais. Para tal, selecionamos a técnica de *PureSVD* e implementamos uma adaptação deste algoritmo para a utilização do PCA, sendo chamado aqui de *PurePCA*. O arcabouço proposto é organizado em quatro módulos de análise distintos:

1. **Módulo de Recomendação:** consiste na aplicação das técnicas de MF selecionadas, sobre o conjunto treino gerado com 70% dos usuários de cada coleção, no intuito de gerar uma lista de recomendação para cada usuário;
2. **Módulo de Extração de Características:** extrai as características levantadas na metodologia sobre o conjunto de treino, gerando as distribuições de análise;
3. **Módulo de Avaliação:** responsável por avaliar a qualidade da recomendação, com base no conjunto teste existente, através dos indicadores de qualidade selecionados na metodologia. Para este trabalho, definimos como um item relevante ao usuário, um item que possui nota média superior a nota média dos *ratings* atribuídos por este;
4. **Módulo de Correlacionamento:** responsável por correlacionar as características da coleção com os indicadores de qualidade, através das métricas de correlação selecionadas.

5. Estudo de Caso

Nesta seção, apresentamos um estudo de caso sobre as técnicas de recomendação baseadas em SVD e PCA quando aplicadas a coleções relevantes da literatura. Basicamente, visamos responder as perguntas pertinentes levantadas por este trabalho por meio da aplicação do arcabouço proposto em coleções reais.

5.1. Coleções de Dados

Utilizamos em nossas análises os conjuntos de dados *MovieLens 100k* e *MovieLens 1M*¹ que foram consolidados e tornados públicos pelo grupo de pesquisa *GroupLens*. A tabela 3 sumariza as principais informações sobre cada coleção. Ambos são conjuntos de dados explícitos sobre recomendação de filmes, em que cada usuário assinala uma avaliação (*ratings*) entre 1 e 5 a cada filme.

Tabela 3. Informações sobre as bases de dados utilizadas.

Base	Usuários	Itens	Esparsidade
<i>MovieLens 100k</i>	943	1676	93,67%
<i>MovieLens 1M</i>	6040	3952	95,82%

5.2. Análise de Resultados

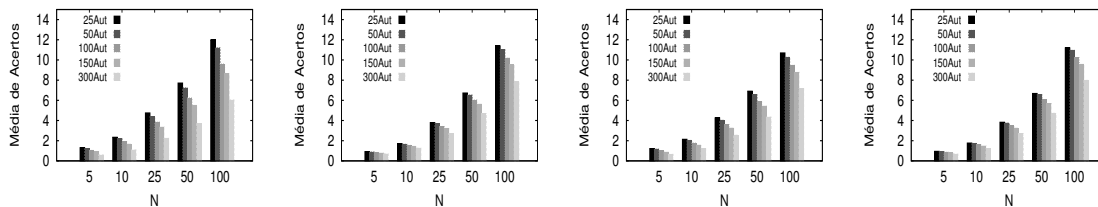
Motivados a caracterizar as distinções entre SVD e PCA quando aplicados em técnicas de recomendação, consolidamos um projeto de análise experimental que está relacionado aos Módulos de Avaliação e Correlacionamento, propostos pelo nosso arcabouço. Este projeto experimental é dividido em quatro etapas: (1) análise da extração de características das coleções de dados, onde avaliamos as distribuições de ranking geradas; (2) análise individual do comportamento de cada um dos métodos, onde nos preocupamos em calcular a taxa de acertos e os valores de precision e recall na aplicação destes; (3) análise pareada dos métodos, onde contrastamos os itens recomendados por cada um dos métodos; e (4) aplicação de um projeto fatorial 2^k , que nos permite verificar os fatores mais importantes que podem afetar o desempenho de métodos baseados em SVD e PCA.

Na primeira etapa de nosso projeto experimental, após extrair as características de cada coleção e plotar as distribuições de ranking, verificamos que o consumo dos usuários está relacionado a popularidade dos filmes em ambas coleções. Notamos também que os

¹Disponíveis em: <http://grouplens.org/datasets/>

filmes mais populares são também aqueles que possuem muitas opiniões distintas, sugerindo que a aplicação de PCA tende a obter melhor resultado nestes cenários, visto que os autovetores do PCA priorizam as dimensões com maiores variâncias nos dados. Em contrapartida, notamos que os filmes com notas médias maiores são os mais relevantes em ambas coleções, o que nos permite dizer, que a aplicação de SVD é mais recomendada, visto que os autovetores de SVD tendem a apontar para dimensões de maior média.

Na segunda etapa, aplicamos os indicadores de qualidade levantados em cada coleção e para cada método, a fim de se caracterizar o desempenho do SVD e PCA de maneira individual. Nesta análise, variamos o número de autovetores relativos ao processo de decomposição de ambos considerando os valores [25, 50, 100, 150, 300]. De maneira análoga, variamos a quantidade de itens recomendadas na tarefa Top-N em [5, 10, 25, 50, 100]. Com base na taxa de acertos e nos valores de *precision* e *recall* calculados para estas configurações, conforme mostra a figura 1 observamos que utilizar apenas 25 autovetores eleva a taxa de acertos para ambas as técnicas, e portanto, será o parâmetro que utilizaremos em nossas análises seguintes.



(a) PureSVD - ML-100k (b) PureSVD - ML-1M (c) PurePCA - ML-100k (d) PurePCA - ML-1M

Figura 1. Influência dos parâmetros no comportamento de cada método.

Na etapa de análise pareada dos métodos, ao considerar os indicadores de qualidade resultantes do processo de recomendação Top-N, identificamos uma similaridade entre o desempenho do *PureSVD* e *PurePCA*. Sabe-se que existe uma diferença prática na aplicação dos dois métodos, porém é necessário mensurar as diferenças entre ambos métodos na tarefa de recomendação. Dessa forma, analisando as distribuições de similaridade entre as listas de recomendação geradas por cada método para todos os usuários, observamos que de 30% a 60% dessas recomendações são idênticas para a maioria dos usuários. Com isso, ainda temos uma porcentagem significativa (i.e., $\geq 40\%$) de recomendações que são geradas por apenas um dos métodos. De maneira complementar, verificamos se a taxa média de acertos relativa aos métodos de *PureSVD* e *PurePCA* estão relacionadas aos itens recomendados por ambos, ou se estão relacionadas aos itens diferentes recomendados por cada técnica. Neste intuito, dividimos cada lista de recomendação gerada em três grupos e comparamos a média de acertos de cada um deles, com base na média de acertos existente sobre toda a lista de recomendação.

- **Grupo 1:** grupo de itens idênticos em ambas as listas de recomendações;
- **Grupo 2:** grupo de itens existentes apenas na lista gerada por *PureSVD*;
- **Grupo 3:** grupo de itens existentes apenas na lista gerada por *PurePCA*.

A tabela 4 demonstra os resultados encontrados para cada um dos grupos, evidenciando que a média de acertos não está relacionada apenas aos itens recomendados por ambos, mas também aos itens recomendados apenas por uma das técnicas. Observe que estes itens recomendados apenas por um método (grupos 2 e 3) apresentam uma

média de acerto significativa, quando comparada à média de acertos total de cada uma das técnicas. Em outras palavras, não só existe diferenças ao aplicar SVD e PCA como base para técnicas de recomendação, mas também, combiná-los em uma técnica híbrida, pode trazer resultados promissores para este contexto de análise.

Tabela 4. Média de acertos para cada um dos grupos de análise.

Conjunto	Grupo 1	Grupo 2	Grupo 3	PureSVD	PurePCA
<i>MovieLens 100k</i>	1,0858	1,2535	1,0520	2,3393	2,1378
<i>MovieLens 1M</i>	0,8079	0,8919	0,9551	1,6998	1,7630

Por fim, na tentativa de se identificar quais fatores influenciam o sucesso de uso do SVD e PCA em recomendação, avaliamos a correlação existente entre os indicadores de qualidade e as características das coleções utilizando um projeto fatorial 2^k . Utilizamos as métricas de correlação propostas, capazes de enfatizar melhor a relação existente entre os fatores de análise. Especificamente, correlacionamos os indicadores de *precision* e *recall* com cada característica extraída, conforme proposto pelo Módulo de Correlação do nosso arcabouço. Esta análise nos permitiu dizer que a taxa de acertos está relacionada ao consumo dos usuários, devido ao aprendizado supervisionado dos métodos. Em uma análise mais detalhada, onde analisamos a popularidade de cada item da lista de recomendação gerada por *PureSVD* e *PurePCA*, notamos que apesar de pequena diferença de correlação entre média de popularidade e os valores de *recall*, para *PureSVD* e *PurePCA* (i.e., como enfatizado pelo tom de cinza mais claro), quando analisado qual dos métodos recomenda itens mais populares, notamos uma diferença entre eles. Nestes cenários, a técnica de *PurePCA* tende a recomendar itens mais populares que *PureSVD*, uma vez que itens mais populares apresentam uma maior variância das notas, o que influencia na aplicação do PCA.

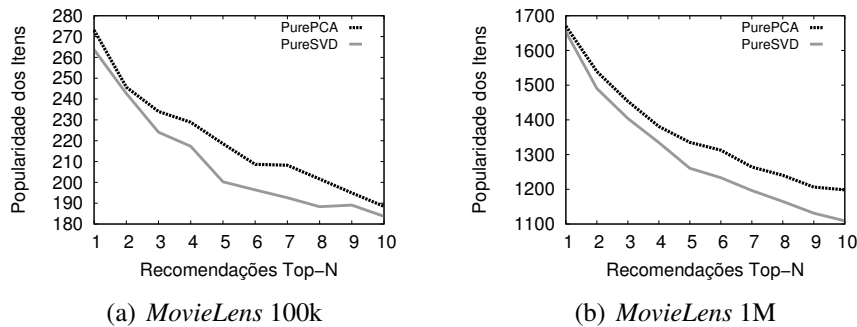


Figura 2. Popularidade média por posição no Top-N.

6. Conclusões e trabalhos futuros

Neste trabalho, revisamos, caracterizamos, organizamos e comparamos distintos métodos de recomendação baseadas em MF. Para tanto, primeiramente, propomos uma taxonomia para métodos de MF baseada em cinco dimensões associadas aos passos do processo de recomendação. Em seguida, classificamos 16 técnicas mais relevantes da literatura através dessa taxonomia, no intuito de auxiliar pesquisadores da área a entender conceitos e semelhanças entre elas. Com esta classificação fomos capazes de analisar as semelhanças entre as técnicas já existentes, bem como, entender melhor as modelagens existentes na aplicação de métodos de fatoração de matrizes. Posteriormente, apresentamos uma metodologia de caracterização e a consolidamos em um arcabouço capaz de

quantificar o comportamento e as diferenças de SVD e PCA quando aplicados à tarefa de recomendação. Avaliações em domínios reais, a partir do arcabouço proposto, sugerem que técnicas baseadas em SVD e PCA recomendam itens distintos e que combiná-las em uma técnica híbrida pode melhorar significativamente os resultados encontrados. Além disso, observamos que características do domínio e individuais de cada usuário podem afetar o desempenho do SVD e PCA em recomendação. Dessa forma, a definição de um projeto experimental apropriado, que nos permitirá validar cada padrão encontrado, é uma direção de pesquisa futura. A avaliação de outros métodos de MF, por meio de nosso arcabouço, representa outra importante direção de pesquisa.

Referências

- Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749.
- Bell, R. M. and Koren, Y. (2007). Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79.
- Funk, S. (2006). Netflix update: Try this at home (december 2006). *URL <http://sifter.org/~simon/journal/20061211.html>*.
- Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151.
- Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM.
- Koren, Y. (2010). Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(1):1.
- Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, (8):30–37.
- Paterek, A. (2007). Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD cup and workshop*, volume 2007, pages 5–8.
- Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2000). Application of dimensionality reduction in recommender system-a case study. Technical report, DTIC Document.
- Sharifi, Z., Rezghi, M., and Nasiri, M. (2013). New algorithm for recommender systems based on singular value decomposition method. In *Computer and Knowledge Engineering (ICCKE), 2013 3th International eConference on*, pages 86–91. IEEE.
- Silva, N., Neves, A., Rocha, R. C., and Mourão, F. (2015). Sobre as diferenças de aplicação entre svd e pca: Um estudo pragmático. In *Anais do XXXV Congresso da Sociedade Brasileira de Computação*.
- Yin, C.-X. and Peng, Q.-K. (2012). A careful assessment of recommendation algorithms related to dimension reduction techniques. *Knowledge-Based Systems*, 27:407–423.