Explorando a Dissimilaridade em Sistemas Colaborativos de Recomendação

Abstract. *The huge amount of options available in various commercial applications became Recommender Systems (RS) crucial tools to assist users in their choices. Despite recent advances in RS, there is still room for more effective techniques which are applicable to a larger number of domains. Most problems arise from the simplified model recurrently used. In this paper, we propose a richer user modeling which allows to extrapolate the usual similarity analysis. Furthermore, we propose a technique that, by exploiting an information type defined as dissimilarity, provides significant improvements over traditional techniques based on collaborative systems, as well as reduces the analysis cost required by such techniques.*

Resumo. *O grande volume de opções existentes em variadas aplicações comerciais tornaram Sistemas de Recomendação (SR) ferramentas cruciais para auxiliar os usuários em suas escolhas. Apesar dos avanços recentes em SR, há ainda uma necessidade por técnicas mais eficazes e aplicáveis a um número maior de domínios. Grande parte dos problemas existentes decorrem da modelagem simplificada utilizada. Neste trabalho, propomos uma modelagem de usuários mais rica que permite extrapolar a usual análise de similaridade. Além disso, propomos uma técnica que, explorando informações definidas como dissimilaridade, provê melhorias significativas sobre técnicas tradicionais de sistemas colaborativos, bem como reduz o custo de análise requerido por tais técnicas.*

1. Introdução

O grande volume de dados disponível na WEB gerou, nos últimos anos, um cenário desafiador para variadas aplicações. Usuários possuem mais opções que efetivamente podem manipular [Adomavicius and Tuzhilin 2005]. Neste cenário, cresce a necessidade por mecanismos especializados em filtrar conteúdo e auxiliar usuários em suas decisões. Sistemas de Recomendação (SRs) são ferramentas que, considerando o histórico dos usuários, realizam tal filtragem através de indicações de itens que pareçam interessantes ao indivíduo [Burke 2002]. Muitas aplicações WEB têm recorrido aos recomendadores a fim de auxiliar seus clientes na tomada de decisões e prover um serviço personalizado.

Uma das técnicas mais simples e eficazes de recomendação existentes são os "Sistemas Colaborativos de Recomendação" (SCR)[1]. SCR consistem, basicamente, em identificar grupos de K indivíduos com preferências ou hábitos de escolha similares, de forma que os itens a serem recomendados para um dado usuário sejam aqueles de maior interesse para o seu grupo. A correlação entre indivíduos similares é, usualmente, representada por meio de uma rede, em que usuários são vértices e as relações são definidas para vértices com similaridade acima de um limiar mínimo δ.

[1]Tal como feito em alguns trabalhos [Koren 2009], referenciamos SCR apenas como o conjunto de técnicas baseada em $k\text{-}nearest\text{-}neighghbor$.

Como argumentado em [Adomavicius and Tuzhilin 2005], embora haja inúmeras propostas para SCRs, os atuais sistemas ainda requerem melhorias para tornar a recomendação mais eficaz e aplicável a cenários mais diversificados. Algumas das limitações de SCRs decorrem do fato de tais métodos serem aplicados sobre uma modelagem simplificada de rede, que não agrega diversas informações potencialmente relevantes. Uma das informações negligenciadas pela maioria destes sistemas é a existência de níveis distintos de similaridade. Um indivíduo pode ser, ao mesmo tempo, similar a algumas pessoas e ter preferências completamente distintas das de outro conjunto de usuários. Tais níveis de similaridade são particularmente relevantes por endereçar um grande desafio existente em SRs: a escassez de informação sobre usuários ou itens.

Dada a complexidade de se avaliar todos os níveis de similaridade, considerando o balanceamento entre eficácia e eficiência de utilização de tais informações, este trabalho de iniciação científica limita-se a analisar a dissimilaridade em SCR. Definimos como dissimilaridade informações obtidas a partir do conjunto de usuários menos similares a cada usuário. Como principais contribuições destacamos: **(1)** a proposta de uma nova modelagem de rede para SCR, que permite considerar distintos níveis de similaridade; **(2)** a caracterização e verificação de duas hipóteses sobre os dados do Netflix[2]: *1- existem diversos níveis de similaridade no cenário de recomendação*; *2- a incorporação da dissimilaridade pode beneficiar a tarefa de recomendação*; **(3)** a proposta de uma nova técnica baseada em SCRs que, através do uso da dissimilaridade, permite endereçar a escassez de informação, bem como uma limitação de SCRs atuais identificada em nossas caracterizações, denominada o problema de *supervalorização*. Este problema refere-se ao fato dos SCRs superestimarem os valores de qualificações preditos, degenerando a qualidade das recomendações, sobretudo dos itens mais desaprovados. Avaliações sobre o Netflix demonstraram que nossa técnica é capaz não só de prover melhorias sobre técnicas tradicionais, atingindo 8,43% de ganhos, mas também de reduzir o custo de análise destas técnicas, exigindo um tamanho reduzido de vizinhança. Além disso, uma discussão sobre as análises realizadas esclarece o que diz respeito ao emprego apropriado da dissimilaridade em SCR. Cabe salientar ainda que não identificamos na literatura trabalhos que utilizem esse conceito de dissimilaridade em SCRs.

2. Trabalhos Relacionados

Grande parte dos esforços presentes na literatura sobre recomendadores constituem os chamados "Sistemas Colaborativos de Recomendação" (SCR), que consideram somente informações históricas dos usuários ou itens. Em particular, a popularidade de técnicas deste tipo deve-se à sua simples modelagem de correlação entre usuários. Em SCR, tipicamente, para cada indivíduo, define-se um conjunto de outros usuários *"vizinhos"*, cujas avaliações anteriores sejam similares. Pontuações para cada item desconhecido pelo usuário são preditas a partir dos pesos atribuídos pelos vizinhos mais próximos do usuário [Breese et al. 1992]. Como SCRs baseiam-se, fundamentalmente, em fazer agrupamentos, sua efetividade depende dos grupos gerados expressarem boas correlações entre os usuários.

Apesar dos inúmeros estudos em SCRs, existem poucos questionamentos sobre a limitação de informações agregadas que são imposta pela modelagem quando se tra-

[2]A base de dados do Netflix contém qualificações explícitas de usuários sobre filmes.

balha com uma vizinhança restrita aos usuários mais similares. Em [Hu et al. 2008] é proposto um modelo que considera, além da vizinhança formada por usuários similares, informações implícitas, tais como preferências e níveis de confiança. Em [Kagie et al.] são discutidas métricas que podem ser adotadas para incorporação da dissimilaridade na construção do modelo. Nosso trabalho se diferencia dos demais da área por constituir uma proposta simples e, ao mesmo tempo, flexível e efetiva de uma nova modelagem em recomendação. Muitos trabalhos, apesar de apresentarem propostas sofisticadas de algoritmos para recomendação, negligenciam informações importantes ao adotarem um modelo restrito. Além disso, não encontramos trabalhos precedentes que investigaram o emprego da dissimilaridade em recomendação, por uma perspectiva similar à nossa.

3. Sistemas Colaborativos de Recomendação

De uma maneira geral, SCRs tradicionais seguem três passos: **(1)** cálculo da similaridade entre o usuário u para o qual se deseja recomendar e os demais usuários do sistema; **(2)** determinação do subconjunto de usuários N_u mais similares ao usuário u (denominados vizinhos de u); **(3)** ponderação das avaliações dos usuários em N_u para formação da avaliação predita para u. Valores de qualificação sobre um item[3], para um dado usuário, são derivados considerando uma soma ponderada dos votos de sua vizinhança. Nestes algoritmos, os votos dos vizinhos são, usualmente, tratados de maneira indistinta, atribuindo a mesma importância às avaliações de vizinhos com diferentes graus de similaridade. A equação a seguir descreve o algoritmo tradicional para estimar a avaliação geral $\hat{r}$ que o usuário u concederia ao item i. A variável $r_{v,i}$ representa o valor da qualificação que o usuário v atribuiu, anteriormente, ao item i.

$$\hat{r}_{u,i} = \frac{\sum_{v \in N_u} sim(u,v) r_{v,i}}{\sum_{v \in N_u} sim(u,v)}$$

A qualidade dessa abordagem de recomendação depende, fundamentalmente, da modelagem de rede utilizada. Dessa forma, modelagens simplórias sobre os dados podem limitar a qualidade do recomendador. Por exemplo, considerar apenas os comportamentos similares pode restringir a quantidade de informações disponíveis para os recomendadores. A seção seguinte detalha nossa proposta de uma nova modelagem dos dados, capaz de agregar informações potencialmente relevantes para a aplicação das técnicas de SCRs.

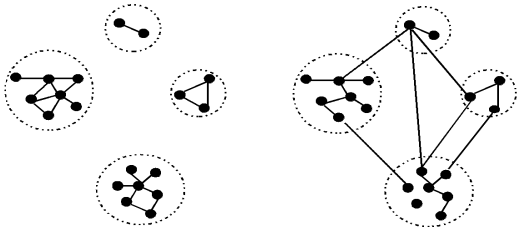
4. Modelo de Rede

Atualmente, a escassez de informações persiste entre os desafios mais estudados em SRs [Adomavicius and Tuzhilin 2005]. Neste trabalho, a fim de identificar outro tipo de informação potencialmente útil para SCRs em cenários de escassez, estamos interessados nas informações fornecidas pelo conjunto de usuários menos similares a cada usuário, informações estas definidas como **Dissimilaridade**. Entretanto, para que seja possível avaliá-las, faz se necessário o uso de um modelo mais rico e representativo de rede.

[3]A tarefa de quantificar o interesse de um usuário por um dado item é, frequentemente, chamada de predição. Outra modalidade de recomendação, chamada de "top-N", é derivada no formato de uma lista de itens ordenados pelo potencial de interesse para o usuário.

Em SCR o modelo de rede consiste em um grafo $G(V, A)$, em que o conjunto de vértices V representa os usuários e o conjunto de arestas A representa suas relações. Relacionamentos entre usuários existem somente se estes possuem similaridade acima de um limiar δ, dada uma função de similaridade tal como Cosseno ou Pearson [Adomavicius and Tuzhilin 2005]. Para fins de podagem, ainda, o número de vizinhos é reduzido para um parâmetro K, definindo os "k-vizinhos mais próximos". Apesar desta modelagem ser muito empregada, por ser simples e intuitiva, alguns problemas podem comprometer sua qualidade. Por exemplo, muitas informações são perdidas ao se considerar somente um grupo restrito de usuários.

Em nossa proposta de modelagem, a rede $G(V, A)$ é composta por relacionamentos entre todos os usuários que possuam uma interseção mínima entre seus históricos de transações. Por exemplo, considerando o domínio de vídeos, haverá uma relação entre dois usuários se existir uma quantidade I mínima (definida como a *intensidade* do relacionamento) de filmes que ambos avaliaram. Dessa forma, é possível definir relacionamentos entre usuários com preferências distintas sobre os mesmos itens. Com isso, a similaridade entre as preferências deixa de ser o fator determinante dos relacionamentos para ser um atributo S de cada relacionamento presente na rede. Outros dois atributos importantes que associamos a cada relacionamento (A_{u_i,u_j}) são denominados *representatividades* do relacionamento para u_i (R_{u_i}) e para u_j (R_{u_j}). (R_{u_i}) é definida como o tamanho da interseção entre os históricos de u_i e u_j sobre o tamanho do histórico de u_i. Essa medida expressa quanto do perfil de cada usuário um relacionamento representa. Assim, nosso modelo consiste de um grafo não direcionado, multi-valorado em que cada relacionamento possui atributos I, S, R_{u_i} e R_{u_j}.



(a) Modelo Tradicional (b) Modelo Proposto

Figure 1. Modelo de Rede

A figura 1 contrasta a construção da rede tradicional de usuários similares *(a)* com a rede segundo o modelo proposto *(b)*. Como podemos observar, deixar que usuários com preferências distintas se relacionem permite ao modelo agregar uma quantidade maior de informações sobre os usuários, beneficiando cenários típicos de SCRs, tais como presença de usuários com gostos peculiares ou com poucas informações em seus históricos. Entretanto, uma possível limitação desta modelagem é que, como o volume de informações agregadas pela rede aumenta, o custo computacional para processar e armazenar tal rede tende a aumentar. Este problema pode ser controlado assumindo um compromisso entre custo e quantidade das informações agregadas, que discutiremos no estudo de caso, uma vez que tal compromisso depende do domínio de análise.

5. Validação das Hipóteses

Nesta seção, verificamos, sobre um relevante domínio real, a existência das hipóteses: **(1)** *Existem diversos níveis de similaridade entre usuários em aplicações baseadas em SCR*;

(2) *O emprego da dissimilaridade pode beneficiar a tarefa de recomendação*. De forma a facilitar o entendimento da discussão que segue, descrevemos a base de dados utilizada e apresentamos uma breve caracterização sobre a mesma.

Caracterização do Netflix

O *Netflix* é um serviço *online* de aluguel de filmes que disponibilizou, para fins de pesquisa em recomendação, uma base de dados com informações de filmes, usuários e qualificações coletadas entre 01/10/1998 a 31/12/2005. Essa representa uma base interessante por conter características típicas de cenários reais onde as técnicas de recomendação são frequentemente aplicadas. Seus dados incluem mais de 100 milhões de qualificações atribuídas por 480.189 usuários distintos a 17.770 filmes distintos.

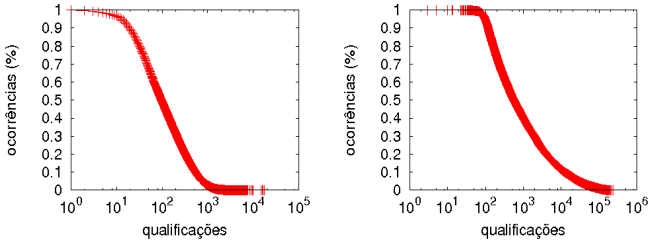
Realizamos uma caracterização da base, a fim de entender, dentre outras coisas, os hábitos dos usuários ao avaliar itens. Para tanto, foram extraídas informações estatísticas sobre as qualificações concedidas pelos usuários. No Netflix, o usuário qualifica um filme atribuindo um valor discreto entre 1 e 5. Investigando a distribuição destes valores de qualificação, verificamos que os mais freqüentes são 4 (33,60%), 3 (28,67%) e 5 (23,06%). A predominância dos valores mais elevados é um indício de que os usuários tendem a qualificar, preferencialmente, os filmes que lhes agradam mais.

Avaliamos também o número de filmes que os usuários qualificam e a popularidade dos filmes, através da distribuição acumulada complementar do número de qualificações fornecidas por usuário e recebidas por filme, respectivamente, tal como mostra a figura 2. Os gráficos indicam que o número de qualificações 100 representa uma divisão no comportamento de ambas as curvas. Existem muitos usuários/filmes que concederam/receberam pelo menos 100 qualificações. Mas este número reduz drasticamente para quantidades maiores que 100. Embora este valor seja específico da base de dados analisada, o comportamento observado é comum em cenários que envolvem a interação entre múltiplos usuários e itens [Park and Tuzhilin 2008].

Por fim, investigamos o número de relacionamentos entre usuários quando variamos o número mínimo I de interseção de histórico (i.e., intensidade mínima do relacionamento). Tal análise nos mostra, novamente, que o número 100 surge como divisor do comportamento da curva. Nesse caso, ele indica que o número de pares de usuários que qualificaram pelo menos I filmes em comum reduz drasticamente quando I ultrapassa 100. Nota-se que o valor de I não pode ser muito superior à 100, caso em que a condição seria muito restritiva. Um valor muito abaixo de 100, por outro lado, pode tornar a restrição pouco efetiva, gerando redes gigantescas, porém, pouco informativas. Os gráficos da figura 3 contrastam a distribuição de graus das redes considerando valores da intensidade mínima para um caso pouco restritivo (I igual a 20) com um caso muito restritivo (I igual a 120).

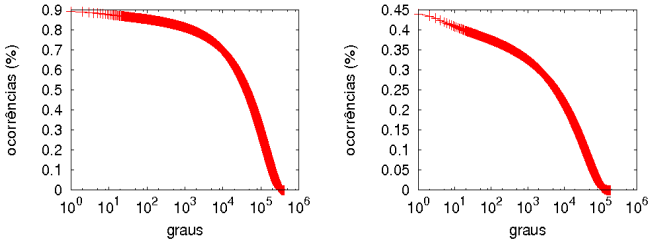
Verificação da Primeira Hipótese

A primeira hipótese proposta defende que, em contextos de recomendação, existem diversos níveis de similaridade. Apesar de seu caráter intuitivo, é necessário verificá-la para quantificar os níveis de similaridade em cada domínio de forma a melhor incorporá-los nos recomendadores. O experimento realizado para validar essa hipótese procurou mostrar que os usuários se relacionam de maneiras diferenciadas, podendo existir graus diversos de semelhanças entre eles. Para isso, todos os usuários foram combina-



(a) Atribuída por Usuário (b) Recebida por Filme

Figure 2. Distribuição acumulada do número de qualificações

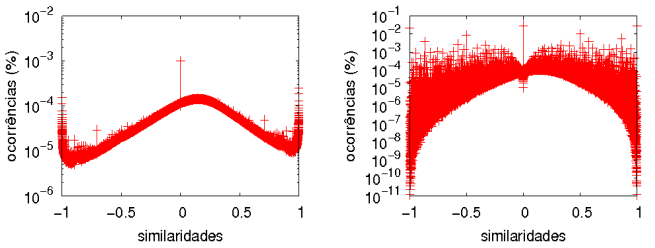


(a) Restrição $i = 20$ (b) Restrição $i = 120$

Figure 3. Distribuição de graus para redes formadas com diferentes restrições

dos em pares e a similaridade de cada par foi computada, considerando as duas métricas mais populares (i.e., *Cosseno* e *Pearson*). Os gráficos da figura 4 exibem distribuições de similaridade para ambas métricas.

As duas distribuições apresentaram muita heterogeneidade em seus valores, o que sugere que, entre os vários relacionamentos de usuários da rede, existe, de fato, uma diversidade de níveis de similaridade. Além disso, é importante observar a existência de pontos próximos aos valores extremos de cada métrica, mostrando que há tanto usuários quase idênticos (i.e., acima de um liminar mínimo δ), quanto quase antagônicos (i.e., abaixo de um liminar máxima ω).



(a) Similaridade Cosseno (b) Similaridade Pearson

Figure 4. Análise da Hipótese 1

Verificação da Segunda Hipótese

A segunda hipótese afirma que a dissimilaridade pode ser utilizada em favor dos SCRs. Verificamos tal hipótese através de uma avaliação comparativa do desempenho de algoritmos de recomendação tradicionais com algoritmos que incorporam a dissimilaridade. Para tanto, é necessário, primeiramente, definir como empregar a dissimilaridade e como medir a qualidade das recomendações. A dissimilaridade foi utilizada adaptando

o algoritmo tradicional, discutido na seção 3, para considerar além dos vizinhos similares os vizinhos dissimilares. Ou seja, todos os indivíduos da rede com valores de similaridade com cada usuário u_i acima de δ são considerados vizinhos similares de u_i, enquanto indivíduos com similaridade abaixo de ω, são tidos como vizinhos dissimilares de u_i. Posteriormente, selecionamos apenas K vizinhos de cada usuário u_i para geração das predições para u_i. Em nossos experimentos, avaliamos versões distintas desta estratégia em que variamos percentualmente a quantidade de usuários dissimilares entre os K selecionados. Para avaliar a qualidade da recomendação, utilizamos a métrica RMSE [Koren 2009], que considera a distância entre o valor predito e a qualificação real como o erro da predição.

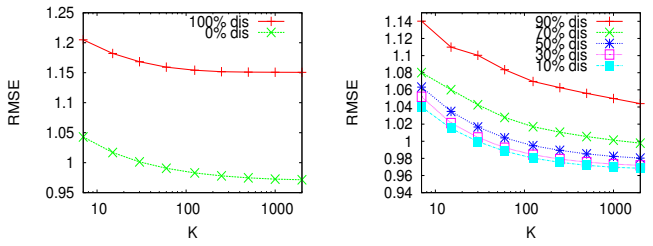
Nossas análises do uso da dissimilaridade para diferentes tamanhos K de vizinhança são sumarizadas nos gráficos da figura 5. Por questões de espaço, apresentamos apenas os resultados relacionados à utilização da correlação de Pearson como medida de similaridade mas, os resultados para testes com a distância cosseno foram similares. Além disso, por apresentar melhores resultados, utilizamos como medida de similaridade entre cada par de usuários o produto da correlação Pearson e os valores de representatividade entre os históricos de consumo de ambos, tal como definido na nossa rede. As predições foram realizadas para lista de pares <*usuário,item*> presentes em um conjunto de teste [4], utilizando informações contidas em um conjunto de treinamento. O valor real da qualificação atribuída pelo usuário ao item, presente no conjunto de teste, foi utilizado para o cálculo de RMSE.

A figura 5 *(a)* contrasta o desempenho do uso restrito da similaridade (i.e., 0% de vizinhos dissimilares) com o uso restrito da dissimilaridade (i.e., 100% de vizinhos dissimilares). Nota-se, claramente, que a similaridade é mais precisa para caracterizar relacionamentos do que a dissimilaridade. A figura 5 *(b)*, por sua vez, apresenta resultados para estratégias que combinam vizinhos similares e dissimilares na análise. Neste caso, temos dois pontos importantes a mencionar. Primeiramente, o melhor resultado foi obtido para combinações de 90% de similaridade e 10% de dissimilaridade. A qualidade alcançada com esta proporção superou, ligeiramente, o resultado obtido considerando apenas a similaridade. Segundo, o RMSE estabiliza-se em valores mais baixos com tamanhos muito menores de vizinhança. Enquanto que com o uso restrito da similaridade o RMSE estabiliza-se com aproximadamente $K = 1000$ vizinhos, com a melhor combinação de similaridade e dissimilaridade o RMSE estabiliza-se com aproximadamente $K = 200$ vizinhos. Isso representa um grande impacto sobre o custo computacional de SCRs, mostrando que com muito menos informações alcançamos resultados quase tão bons que os encontrados considerando-se grandes redes de vizinhos similares. Dessa forma, os resultados mostram indícios de que a dissimilaridade pode beneficiar SCRs.

Discussão

Um entendimento mais geral sobre o impacto da dissimilaridade em SCRs requer analisar mais a fundo as derivações decorrentes da definição de indivíduos dissimilares. Dizer que dois indivíduos são dissimilares quanto ao gosto, consiste em dizer que tais indivíduos, de maneira geral, não gostam do mesmo tipo de itens. Conseqüentemente,

[4] O conjunto de teste utilizado foi concedido, juntamente com a base de dados, pelo Netflix.



(a) Similaridade vs. Dissimilaridade (b) Similaridade + Dissimilaridade

Figure 5. Análise da Hipótese 2

estes indivíduos tendem a não consumirem itens do mesmo tipo frequentemente. Com isso, menos informações temos para definir a dissimilaridade, repercutindo em menos "confiança" sobre tal métrica. Essa é uma diferença essencial entre similaridade e dissimilaridade. Enquanto a similaridade possui um conjunto maior de indícios para se apoiar (i.e., indivíduos similares consomem os mesmos tipos de itens), o mesmo não ocorre com a dissimilaridade (embora o caráter multi-gosto dos usuários garanta sua existência). Intuitivamente, isto explica o comportamento pouco eficaz obtido para a dissimilaridade em nossos resultados. Além disso, temos uma certa redundância nas informações contidas na dissimilaridade. Por exemplo, um item que não se adeque ao gosto de um dado usuário possivelmente não se adequaria aos gostos dos vizinhos deste usuário e, conseqüentemente, já não seria recomendado a ele. Tais distinções entre similaridade e dissimilaridade sugerem que a visão local propiciada pela análise da vizinhança de cada indivíduo em uma rede de relacionamentos privilegia a similaridade.

Uma solução para estes problemas seria mudar nosso foco de análise para uma visão global. Identificar tendências globais do domínio pode trazer um conjunto de informações acerca da dissimilaridade muito mais úteis. Nossa proposta reside justamente em realizar tal análise global através do uso de regras de associação que expressam padrões recorrentes entre os usuários do domínio.

6. Sistema Colaborativo com Regras Semânticas

A exploração de informações expressas como regras semânticas, tais como *usuários que gostam do item X, não gostam do item Y*, corresponde à principal idéia do nosso algoritmo. A motivação para isso vem de três observações distintas. Primeiramente, explorar o domínio como um todo provê uma massa muito maior de informações, endereçando o problema de escassez de informação para a dissimilaridade. Segunda, regras similares à exemplificada evidenciam correlações freqüentes e consistentes presentes no domínio. Terceiro, este tipo de análise é complementar à análise local definida por estratégias de recomendação baseada em sistemas colaborativos. Combinar aspectos de similaridade locais com fatores de dissimilaridades globais pode ser uma estratégia eficaz, principalmente por prover mais informações em casos onde a rede de vizinhança de um usuário não possui informações sobre determinados itens.

Dessa forma, apresentamos um novo algoritmo de recomendação que combina a técnica de filtragem colaborativa com regras de associação extraídas por técnicas tradicionais de mineração de dados. Descreveremos em detalhes tal proposta, bem como discutiremos seu impacto em termos de custo computacional. Nossa abordagem consiste,

basicamente, na definição de dois passos fundamentais: a geração de regras semânticas e a incorporação das regras no algoritmo de predição.

Geração de Regras Semânticas

De forma a gerar as regras semânticas utilizadas no algoritmo, é necessário primeiramente definir o que são transações neste contexto. Uma transação é o conjunto de todos os itens qualificados por um usuário. Dessa forma, teremos tantas transações quanto usuários distintos, bem como um tamanho variado de transação por usuário. Além disso, itens bem qualificados foram diferenciados de itens qualificados negativamente. Ou seja, para cada usuário, geramos uma transação em que, os itens cujo valor de qualificação seja maior que a média de qualificações do usuário mais X desvios padrões são classificados como "positivos", e itens qualificados com valores abaixo dessa média menos X desvios padrões são tidos como "negativos". Este valor X deve ser obtido empiricamente, uma vez que depende essencialmente do domínio considerado. O algoritmo *FP-Growth* [Han et al. 2000] foi utilizado para gerar *itemsets* freqüentes considerando o conjunto de transações e um parâmetro ρ que define o suporte mínimo. A partir dos *itemsets* gerados, e de um parâmetro τ que define a confiança mínima, obtemos as regras de associação.

Incorporação de Regras em SCRs

O passo seguinte à geração das regras semânticas consiste em incorporá-las ao sistema de recomendação. O algoritmo proposto consiste, basicamente, em uma adaptação do método tradicional de predição, discutido na seção 3. O processo de incorporação das regras possui, como primeira etapa, uma filtragem das regras de interesse. Uma vez mineradas no domínio todas as regras de associação com suporte mínimo maior que ρ e confiança mínima superior a τ, selecionamos, por questão de simplicidade, apenas regras, de qualquer tamanho, com um único item como conseqüente. Selecionadas as regras de interesse para o domínio, nosso próximo passo consiste em derivar as qualificações a partir do conjunto de regras resultantes.

O algoritmo tradicional adota um sistema de votação baseado em informações dos vizinhos de cada usuário u para prever a qualificação que este concederia para um determinado item i. Assim, apenas os vizinhos de u que qualificaram o item i no passado são considerados para a predição. Caso nenhum vizinho tenha qualificado o item i, é retornado um valor derivado a partir das médias de todas qualificações atribuídas pelo usuário u e recebidas pelo item i no sistema. Em nosso algoritmo as regras semânticas são utilizadas como complemento para a predição apenas na ausência de uma qualificação atribuída por um vizinho. Ou seja, para cada vizinho v_j de u que não tenha qualificado o item i, derivamos uma qualificação ${}'r_{ji}$ de v_j para i a partir das regras semânticas previamente mineradas. Para derivar uma qualificação ${}'r_{ji}$, é necessário buscar regras que possuam i como conseqüente. Também é necessário que o vizinho v_j atenda aos antecedentes das regras a serem utilizadas (i.e., tenha qualificado da mesma forma, positivamente ou negativamente, cada um dos itens que formam o antecedente da regra). Atendendo a estes critérios, somente a regra resultante com maior confiança ($\mathcal{R}$) é escolhida para derivarmos ${}'r_{ji}$. Tal derivação pode ser realizada de variadas formas. A estratégia adotada, visando eficiência computacional, consiste em definir ${}'r_{ji}$ como a qualificação média que todos os usuários do domínio que atendem a regra $\mathcal{R}$ selecionada atribuíram ao item i. Caso nenhuma regra seja seleciona, tal como no algoritmo tradicional, derivamos a qualificação a

partir das médias de qualificações dadas pelo usuário v_j e recebidas pelo item i.

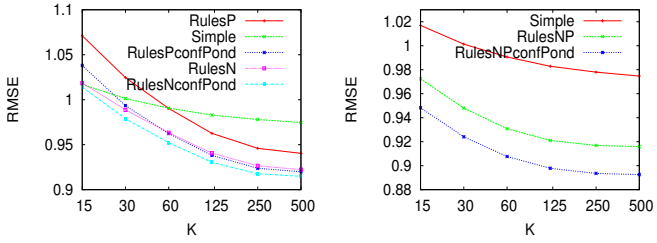
Análise experimental

De forma a avaliar a qualidade do algoritmo proposto, foram definidos dois cenários de análise. Em um primeiro conjunto de experimentos, denominado cenário de regras homogêneas, avaliamos o uso de regras semânticas constituídas somente por itens qualificados positivamente ou negativamente pelos usuários. No segundo conjunto experimental, denominado cenário de regras heterogêneas, avaliamos regras constituídas por itens positivos e negativos simultaneamente. Em ambos cenários de análise, avaliamos além das derivações de qualificações $'r_{ui}$ que cada usuário u atribuiria ao item i, tal como definido acima, duas variações do uso das regras semânticas. Um delas pondera cada qualificação $'r_{ui}$ predita pelo valor de confiança da regra $\mathcal{R}$ selecionada para derivar a qualificação. Na segunda variação, ponderamos $'r_{ui}$ por um fator γ caso o valor predito seja menor que média de qualificações atribuídas pelo usuário u, ou seja, caso $'r_{ui}$ represente uma predição negativa para u. Como os usuários no Netflix tendem a qualificar mais itens que lhes agradam, regras com itens positivos tornam-se mais freqüentes globalmente. Dessa forma, a fim de valorizar ocorrências das qualificações negativas, e tornar a análise mais balanceada utilizamos o fator γ. Tal fator deve ser determinado empiricamente em cada domínio. Para o Netflix o melhor valor encontrado foi 1,4.

A figura 6 apresenta os valores de RMSE para ambos cenário, considerando distintos tamanhos de vizinhança. Nestes gráficos, a curva denominada *'Simple'*, refere-se ao SCR tradicional, tal como definido na seção 3. *'RulesP'* refere-se ao nosso algoritmo de regras semânticas considerando apenas itens classificados como positivo nas regras e *'RulesN'* considera apenas itens negativos nas regras. A curva *'RulesPN'* refere-se ao cenário heterogêneo de análise. Além disso, as curvas contendo rótulos *'conf'* e *'Pond'* referem-se às versões do nosso algoritmo que consideram a confiança da regra e a valorização das qualificações negativas, respectivamente. De maneira geral, todos as variações de algoritmo avaliadas se mostram superiores ao algoritmo de SCR tradicional. Considerando o cenário de regras homogêneas, os melhores resultados foram alcançados considerando itens negativos, ponderação de confiança e valorização das regras negativas, atingindo 5,62% de ganho sobre o algoritmo tradicional. Para o cenário heterogêneo, os ganhos foram ainda maiores, atingindo 8,43% para o algoritmo que utiliza a ponderação de confiança e valorização das regras negativas. Outro aspecto pertinente a ressaltar é o tamanho da rede de vizinhança necessário para atingir tais valores de qualidade. Utilizando apenas 1/4 do tamanho de vizinhança máximo avaliado nos experimentos dos gráficos 5, atingimos valores significativamente melhores. Isso representa um grande impacto sobre o custo computacional para a recomendação. Como a obtenção do conjunto de regras de associação é realizada uma única vez, e o custo de incorporação de tais regras é linear com o número de itens a serem recomendados, temos uma grande redução neste custo computacional por analisar uma vizinhança significativamente menor para cada usuário.

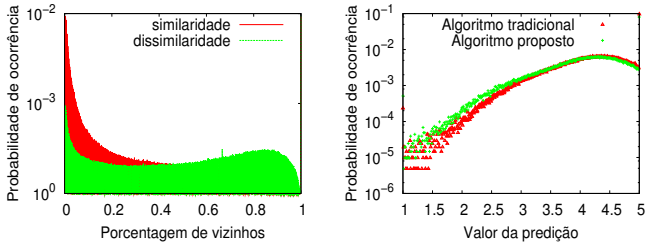
Discussão

De forma a entender os resultados alcançados, caracterizamos dois aspectos acerca das predições realizadas. O primeiro aspecto refere-se à porcentagem de vizinhos de cada usuário u que qualificaram no passado os itens a serem recomendados a u. O gráfico



(a) Regras Homogêneas (b) Regras Heterogêneas

Figure 6. Qualidade das Recomendações Usando Regras Semânticas



(a) Ocorrência de Vizinhos (b) Supervalorização das Qualificações

Figure 7. Caracterização dos Resultados

da figura 7 (a) apresenta as probabilidades de ocorrência de cada porcentagem de vizinhos similares e dissimilares. Claramente percebemos que a dissimilaridade prioriza porcentagens mais elevadas de vizinhança. A maior parte das vizinhanças dissimilares já qualificaram vários itens a serem recomendados, enquanto entre os vizinhos similares uma porcentagem muito menor de vizinhos já qualificaram estes itens. Ou seja, a dissimilaridade é capaz de endereçar o problema de escassez de informação acerca da vizinhança analisada. Um dos motivos das melhorias propiciadas pelo uso das regras decorre desse fato. Uma evidência disso é que os maiores ganhos em qualidade do algoritmo proposto ocorrem para tamanhos menores de vizinhança, situações em que o problema de escassez de informação é mais acentuado.

O segundo aspecto avaliado é a distribuição dos valores de predição. Distribuições tanto para o nosso algoritmo quanto para o algoritmo tradicional evidenciaram o que denominamos problema de supervalorização das qualificações. Ou seja, as predições tendem a gerar valores mais elevados que os valores reais, tal como mostrado no gráfico da figura 7 (b). Entretanto, o uso de regras, sobretudo com valorização das qualificações negativas, atenua este problema. De forma a mostrar isso, geramos o *Third Standardized Moment Skewness* para ambas curvas. A distribuição para o algoritmo tradicional apresentou um *skewness* de $-0,80$, enquanto para nosso algoritmo este valor foi de $-0,82$. Isso mostra que nossa distribuição é mais desbalanceada para valores menores, evidenciando o fato de gerarmos predições ligeiramente menos positivas. Dessa forma, uma segunda explicação para as melhorias alcançadas vem da atenuação do problema de supervalorização. É importante ressaltar que os resultados obtidos mostram o potencial da consideração dos diversos níveis de similaridade associados com uma visão global no processo de recomendação. A relevância de nossa proposta reside não só em sua aplicação para prover melhorias sobre técnicas tradicionais, mas também para reduzir o custo de

análise destas técnicas, exigindo um tamanho muito menor de vizinhança.

7. Conclusão e Trabalhos Futuros

Neste trabalho, apresentamos uma nova proposta para modelagem de dados em Sistemas Colaborativos de Recomendação (SCRs), bem como uma avaliação do uso de dissimilaridade no cenário de recomendação de vídeo, além de uma nova técnica que explora a dissimilaridade em SCRs. A partir da modelagem proposta, foi possível verificar a existência de diversos níveis de similaridade em nossa base de dados, além de executar predições considerando, para um dado usuário, vizinhos similares e regras semânticas que definem aspectos globais de dissimilaridade no domínio. Uma análise comparativa de qualidade das predições geradas mostrou que a técnica proposta é particularmente relevante não só por prover melhorias significativas sobre técnicas tradicionais de SCRs, mas por reduzir o custo de análise destas técnicas. Como trabalho futuro, destacamos o emprego da técnica proposta em outros cenários, com escassez de informação mais acentuada.

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. *IEEE TKDE*, 17(6):734–749.

Breese, J., Heckerman, D., and Kadie, C. (1992). Empirical analysis of predictive algorithms for collaborative filtering. *Learning*, 9:309–347.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD Record*, 29(2):1–12.

Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *8th IEEE ICDM*, pages 263–272.

Kagie, M., Van Wezel, M., and Groenen, P. An Empirical Comparison of Dissimilarity Measures for Recommender Systems.

Koren, Y. (2009). Collaborative filtering with temporal dynamics. In *Proc. of the 15th SIGKDD*, pages 447–456. ACM New York, NY, USA.

Park, Y. and Tuzhilin, A. (2008). The long tail of recommender systems and how to leverage it. In *Proc. of the 2008 ACM RecSys*, pages 11–18. ACM.