Para Onde Devo Viajar: Recomendação de Cidades Baseada em Comunidades de Usuários

Ruhan Bidart[1], **Adriano C. M. Pereira**[1], **Jussara Almeida**[1], **Anisio Lacerda**[1]

[1]Departamento de Ciência da Computação (DCC)
Universidade Federal de Minas Gerais (UFMG)

{ruhanbidart, adrianoc, jussara, anisio}@dcc.ufmg.br

Abstract. *Recommendation systems play a key role in the decision making process of users in online systems. In this work, we propose a city recommendation system that exploits the user interests and the similarity between different users. The proposed method builds a social network among users where the edges are weighted by the similarity of interests between pairs of users. This network is then used as a component of a collaborative filtering strategy. We evaluate our method using a large dataset collected from TripAdvisor. Our experimental results show that our approach can double the precision achieved by baseline approaches, which exploit only the overall popularity of cities, reaching 65% for the most active users.*

Resumo. *Sistemas de recomendação têm desempenhado um papel vital nos processos de decisão feitos pelos usuários de sistemas online. Neste trabalho é proposto um sistema de recomendação de cidades que explora os interesses dos usuários e similaridade entre diferentes usuários. A solução proposta cria uma rede social virtual entre os usuários, na qual as conexões são ponderadas pelas similaridades de interesses entre pares de usuários. Esta rede é então utilizada em uma abordagem de filtragem colaborativa. O método proposto foi avaliado em uma grande base de dados coletados do sistema TripAdvisor. Os resultados mostram que a abordagem proposta dobra a precisão em relação a métodos alternativos, os quais são baseados na popularidade das cidades, atingindo 65% para usuários com um maior número de informações compartilhadas.*

1. Introdução

Existem muitos sistemas de recomendação disponíveis para as mais diversas áreas, no entanto, não foi encontrado nenhum sistema de recomendação de cidades. Mesmo o TripAdvisor[1], que é o maior *Web site* de apoio ao turismo do mundo, não possui um sistema de recomendação de cidades personalizado aos interesses dos usuários.

Dado este cenário, o trabalho propõe uma estratégia para recomendar cidades com base nos perfis de interesses dos usuários e na similaridade entre múltiplos usuários. A estratégia é dividida em três componentes principais:

1. Geração do Grafo: gera-se um grafo de relações de similaridade entre os usuários. Para inferir as similaridades, utilizou-se uma abordagem diferente da comum. Comumente são utilizadas avaliações feitas pelos usuários a itens semelhantes para

[1]http://www.tripadvisor.com.br

prever as relações de similaridade entre eles. Nos dados que foram utilizados, nenhum usuário determina explicitamente uma avaliação para as cidades, sendo assim o método proposto utiliza-se de informações implícitas para prever as relações entre os usuários;
2. Geração de Comunidades: são geradas comunidades de usuários utilizando a abordagem kNN. Essas comunidades são utilizadas pela terceira etapa, de ordenação, como recurso indispensável para filtrar cidades que seriam mais adequadas ao usuário alvo;
3. Ordenação: as comunidades geradas são utilizadas para selecionar cidades candidatas à recomendação para cada usuário e para atribuir um peso à cada uma delas. As cidades são então ordenadas em ordem decrescente de seu peso, de forma que aquelas com peso maior sejam recomendadas para o usuário.

Esta metodologia é avaliada em um estudo de caso com dados reais, coletados do site TripAdvisor, com 1.597.609 páginas e 266.392 usuários. É demonstrado como é possível inferir relações entre usuários com avaliações implícitas e que as relações inferidas são capazes de dobrar a precisão na recomendação de cidades para os usuários do TripAdvisor, quando comparadas aos métodos de referência.

Em suma, as principais contribuições deste artigo são:

1. Proposta de uma abordagem simples, genérica e escalável para inferência de relações entre usuários, a qual foi comprovada em um conjunto de dados reais como sendo útil para melhoramento da precisão das recomendações;
2. Uma metodologia para recomendação com avaliação implícita que, embora tenha sido avaliada em um conjunto de dados do TripAdvisor, pode ser facilmente generalizada;
3. Primeiro artigo de que se tomou conhecimento a focar em recomendação de cidades e avaliar estas recomendações em um conjunto de dados reais.

O restante do artigo está organizado da seguinte forma. Os trabalhos relacionados são discutidos na Seção 2. Na Seção 3 o problema é definido. A seção 4 descreve a solução proposta. A Seção 5 apresenta a metodologia experimental desenvolvida, bem como o detalhamento de seus módulos e as técnicas utilizadas em cada um deles. Na Seção 6 os resultados obtidos são apresentados para duas filtragens de dados, uma com usuários mais ativos e outra com um grupo maior e mais heterogêneo de usuários. Por fim, na Seção 7, o artigo é concluído e há direções de trabalhos futuros.

2. Trabalhos Relacionados

A ideia de sistemas de recomendação consiste em recomendar automaticamente itens para um usuário, procurando prever seu interesse nesses itens [Yang et al. 2012]. Esses sistemas auxiliam o usuário a lidar com a sobrecarga de informação, provendo-lhes recomendações personalizadas de produtos e serviços [Adomavicius and Tuzhilin 2005].

Sistemas de recomendação são tipicamente classificados em três categorias: *baseados em conteúdo*, onde recomenda-se itens similares a outros que o usuário preferiu no passado [Linden et al. 2003]; *colaborativos*, onde são recomendados itens que pessoas com gosto similar preferiram no passado [Herlocker et al. 2002]; e abordagens *híbridas*, as quais combinam as duas anteriores para recomendação de itens [Balabanović and Shoham 1997].

As abordagens *colaborativas* são as mais populares e mais largamente implementadas [Ricci et al. 2011], tendo uma precisão melhor do que as abordagens *baseadas em conteúdo* em várias aplicações. Nos algoritmos que utilizam abordagens *colaborativas* enfrenta-se tipicamente o problema de se identificar grupos de usuários semelhantes. Existem várias técnicas para idenficação de grupos em grafos [Fortunato 2010], a mais simples e genérica delas é o kNN ($k - NearestNeighbor$), a qual é utilizada em vários trabalhos [Herlocker et al. 2004, Cremonesi et al. 2010]. Em todos os trabalhos estudados, a comparação de similaridade entre os usuários ocorre para dados que possuem avaliação explícita, ou seja, os itens avaliados possuem uma nota específica dada pelo usuário e a função do sistema de recomendação é prever as notas que o usuário ainda não preencheu.

Formalmente, a utilidade $u(c, s)$ de um item s para um usuário c é estimada com base nas utilidades $u(c_j, s)$ assimiladas para o item s, por aqueles usuários c_j que são "similares" ao usuário c. Embora esta tarefa seja bem definida, existem diversas maneiras de se implementar uma solução que resolva este problema. Em [Herlocker et al. 2002], é definido um arcabouço para a implementação de uma solução para este problema, a qual é utilizada neste trabalho. A implementação desse arcabouço é detalhada na Seção 5.

Este arcabouço pode ser dividido em três componentes principais: computação de similaridades, seleção de vizinhos e ordenação. Existem trabalhos explicitamente focados em desenvolver e testar algoritmos que resolvam cada um desses problemas:

1. **Computação de similaridades**: o trabalho [Fortunato 2010] é focado em descrever e relacionar técnicas de computação de similaridades;
2. **Seleção de vizinhos**: esse problema é tratado especialmente por [Herlocker et al. 2002, Jamali and Ester 2009] e de maneira secundária por [Yang et al. 2012], que aplica sistemas de colaboração utilizando uma rede social real, o *last.fm*;
3. **Ordenação**: a ordenação é especialmente desenvolvida em [Herlocker et al. 2004], onde diversas técnicas são analisadas empiricamente. [Cremonesi et al. 2010] compara métricas de avaliação para a tarefa de recomendação top-N, a qual consiste em indicar alguns itens específicos que seriam os mais adequados para o usuário.

Existem trabalhos de recomendação focados em diferentes tipos de aplicações. Em [Schafer et al. 2001] podem ser vistas algumas aplicações, em especial uma análise de serviços de comércio eletrônico. No caso do contexto de turismo, foram encontrados alguns poucos trabalhos. Destaca-se [Kurashima et al. 2010], que utiliza dados de localização existentes em fotos para inferir rotas de usuários e, então, recomendar rotas para outros usuários. Pode-se destacar também [Wang et al. 2013, Noulas et al. 2012], que são trabalhos direcionados a recomendação de atrações (no contexto de sites de turismo atrações são locais que podem ser visitados em uma cidade).

Não foram encontrados trabalhos com foco específico em recomendação de cidades, espaço em que identificamos um campo a ser investigado.

3. Definição do Problema

Este trabalho lida com três problemas principais no algoritmo de filtragem por colaboração. O primeiro problema trata-se de inferir o relacionamento entre os usuários,

uma vez que o TripAdvisor não possui uma rede social explícita entre seus usuários. O problema foi descrito como: seja $U = \{u_1, u_2,...,u_n\}$ o conjunto contendo todos os usuários e $\mathbf{T} \in \mathbb{R}^{nxn}$ a matriz representando as similaridades entre eles. O objetivo é preencher a matriz **T**, assim chegando a uma rede social virtual.

O segundo problema consiste em separar os usuários em grupos. Definiu-se o problema como: dado um usuário u (alvo da recomendação) e uma matriz de similaridades **T**, a tarefa é descobrir um grupo G_u com usuários relacionados a u. Note que esta tarefa é análoga à de detecção de comunidades.

O terceiro problema é uma tarefa de aprendizado de ordenação, cujo objetivo é ordenar (*ranking*) uma lista de cidades para cada usuário. O problema foi definido como: seja um usuário alvo $u \in U$ e que possui um grupo de usuários similares $G_u = \{u_1, u_2, ..., u_m\}$, se $C_g = \{c_1, c_2,..., c_k\}$ denota o conjunto de todas as cidades visitadas pelos usuários pertencentes a G_u e $C_u = \{c_1, c_2, ..., c_l\}$ denota o conjunto de cidades visitadas por u, a tarefa consiste em ordenar o conjunto de cidades C_g - C_u para cada um dos usuários contidos em U. Assim, somente as cidades que nunca foram visitadas pelo usuário alvo serão recomendadas.

4. Solução Proposta

A solução proposta para endereçar o problema de pesquisa deste trabalho foi segmentada em três módulos. Cada módulo funciona de maneira independente, utilizando apenas a saída da etapa anterior. A Figura 1 apresenta uma visão geral da solução proposta. Nas próximas seções cada parte será detalhada.

Figure 1. Visão geral da solução proposta. Na parte 1 o grafo é criado. A segunda parte detecta as comunidades no grafo e na parte 3 ocorre a ordenação das cidades para recomendação.

4.1. Geração do Grafo

O primeiro ponto no processo de geração do grafo diz respeito à questão: como podemos criar uma rede social relacionando os usuários? Esta questão se relaciona com o primeiro problema definido, de relacionar similaridades entre os usuários. Como não há avaliações explícitas dos usuários em relação às cidades, decidiu-se por utilizar o grupo de cidades C_u de cada usuário como um parâmetro para o índice de Jaccard, conforme Equação 1.

$$J(C_{u1}, C_{u2}) = \frac{|C_{u1} \cap C_{u2}|}{|C_{u1} \cup C_{u2}|} \tag{1}$$

Esta fórmula foi então utilizada para preencher a matriz **T**. É importante notar a simplicidade do índice de Jaccard, que proporciona um melhor desempenho para todo o modelo. Tentou-se utilizar outros índices de maior complexidade (como similaridade

de cossenos) no entanto esta mudança não trouxe melhoras estatisticamente significativas para precisão do modelo como um todo. Assim, decidiu-se pela utilização do índice de Jaccard.

Utilizando o índice de Jaccard, toda a matriz **T** será preenchida. Todavia, é preciso compreender qual é o valor mínimo para o índice que representa uma relação válida entre dois usuários. Para realizar esta verificação, foi traçado um gráfico relacionando os resultados do índice de Jaccard e o número de relações de usuários que possuem índice maior do que este valor, como pode ser visto na Figura 2.

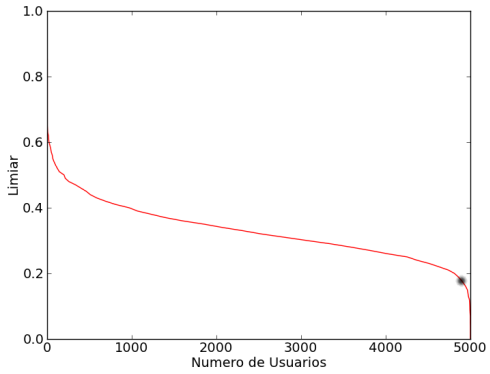


Figure 2. Número de usuários para cada valor de limiar. Note que o joelho da curva está aproximadamente em 0.2 (marcado com um círculo preto).

Note que a distribuição forma um joelho claro: quando o limiar está em 0,2, somente 3,78% dos usuários são removidos. Além disso, cerca de 95% das arestas são removidas. O limiar que se procura deve levar a um número mínimo de usuários removidos ao mesmo tempo que a um número grande de arestas removidas, as quais representam conexões fracas entre os usuários. Assim, a escolha de limiar como 0,2 parece ser boa.

4.2. Detecção de Comunidades

Como primeira abordagem decidiu-se por utilizar o algoritmo kNN, o qual é um ponto de partida por ser simples e escalar para o volume de dados tratado neste trabalho. Fixou-se o k = 10, o que significa que o kNN irá selecionar os 10 usuários mais similares relacionados ao usuário alvo da recomendação. Estes 10 usuários formam o grupo do usuário avaliado. Pode-se perceber que cada usuário terá um grupo possivelmente diferente dos outros, algo natural para uma abordagem kNN.

As comunidades geradas terão uma importância crucial para as próximas etapas, visto que é o ponto em que o método proposto consegue remover da etapa de ordenação grande parte das cidades que possivelmente não teriam importância para o usuário alvo. Isso torna a próxima etapa mais simples de ser realizada, por ter um espaço de busca menor do que o espaço que consiste de todas as cidades.

4.3. Ordenação

O algoritmo de ordenação (*ranking*) de de cidades proposto se baseia na detecção de comunidades de usuários semelhantes para efetuar a recomendação. A hipótese é que usuários que visitaram mais cidades semelhantes ao usuário alvo u possuem uma maior probabilidade de visitar outras cidades que o usuário u também gostaria.

A partir dessa hipótese e do fato de não existir avaliações explícitas das cidades, foi proposto o algoritmo *ReCWEE* (Recommendation using Communities and Without Explicit Evaluation). O *ReCWEE* utiliza um tipo de técnica de recomendação Top-N, considerando somente o grupo de cidades $C_g - Cu$ (veja Seção 3). Pode-se verificar seu pseudocódigo no Algoritmo 1. Nas linhas identificadas por $1 - 11$ pode-se ver a função $GerarGrafo$, a qual gera o grafo completo entre os usuários, utilizando o índice de Jaccard entre as cidades já visitadas por cada um deles.

A função $Ordenacao$ (linhas $13 - 20$) recebe o usuário alvo da recomendação e a comunidade de usuários semelhantes a ele e retorna a ordem de cidades a ser recomendada para este usuário. Detalhadamente, a linha 14 recupera apenas as cidades que estão na comunidade selecionada para o usuário, porém removendo aquelas que ele já visitou, com o objetivo de sempre se recomendar cidades ainda não visitadas. Na linha 16 a cidade recebe um peso que depende do número de vezes que ela apareceu entre os usuários componentes daquela comunidade e da média das avaliações das atrações contidas na cidade. A linha 19 retorna a lista de cidades em ordem reversa porque se deseja que cidades com um índice maior sejam recomendadas primeiro para o usuário.

Nas linhas $22 - 31$ está a função $ReCWEE$, método principal do algoritmo proposto. Sua função principal é coordenar a execução de cada uma das partes da técnica proposta, retornando as ordenações de cidades para todos os usuários. Na linha 23 o grafo é gerado. A comunidade do usuário u é gerada na linha 26, comunidade esta que é utilizada na linha 27 para se gerar a lista de recomendações para o usuário. O processo é então repetido para cada usuário pelo laço contido na linha 25 para depois serem retornadas as ordenações de cidades para todos os usuários na linha 30.

Algorithm 1 ReCWEE

```
1: function GERARGRAFO(U)
2:     Grafo ← ∅
3:     for u1 in U do
4:         for u2 in U do
5:             if u1 ≠ u2 then
6:                 Grafo[u1][u2] ← Jaccard(Cidades(u1), Cidades(u2))
7:             end if
8:         end for
9:     end for
10:    return Grafo
11: end function
12:
13: function ORDENACAO(u, G_u)
14:    CidadesParaOrdenar ← Cidades(u) − Cidades(G_u)
15:    for c in CidadesParaOrdenar do
16:        v ← Votos(G_u, c) * MediaAvaliacoesAtracao(c)
17:        OrdemDasCidades[c] ← v
18:    end for
19:    return OrdemReversa(OrdemDasCidades)
20: end function
21:
22: function RECWEE(U)
23:    Grafo ← GerarGrafo(U)
24:    Ordenacoes ← ∅
25:    for u in U do
26:        G ← GerarComunidadesKNN(u, Grafo)
27:        RecomendacaoUsuario ← Ordenacao(u, G)
28:        Ordenacoes[u] ← RecomendacaoUsuario
29:    end for
30:    return Ordenacoes
31: end function
```

5. Metodologia Experimental

A metodologia experimental é dividida em três partes: inicia-se com a aquisição de dados, em seguida são explicados os métodos utilizados como referência comparativa e por fim as técnicas de avaliação.

5.1. Aquisição de Dados

Todos os dados que foram utilizados neste trabalho foram coletados do *Web site* TripAdvisor [2], de 23/10/2013 a 23/03/2014. Um resumo dos dados coletados pode ser visto na Tabela 1.

Entidade	Quantidade
Páginas	1.597.609
Cidades	85.505
Atrações	162.168
Revisões	599.629
Usuários	266.392

Table 1. Lista de entidades e quantidade de dados coletados do TripAdvisor

Embora tenham sido coletadas várias entidades, neste trabalho será dado enfoque a três entidades em particular: *Cidades*, *Atrações* e *Usuários*, as outras estão enquadradas para serem utilizadas em trabalhos futuros. É importante ressaltar que todos os usuários coletados tiveram seu nome suprimido para preservação de suas identidades. Os principais atributos coletados para cada uma das entidades são apresentados na Tabela 2.

Entidades	Atributos
Atração	Nome, Cidade, Nota, Número de Revisões, Endereço.
Usuário	Identificador, Seenlist de cidades, Seenlist de atrações, Localização.

Table 2. TripAdvisor - Descrição dos dados coletados

Este estudo está direcionado ao comportamento de turistas brasileiros, com o objetivo de descobrir como recomendar cidades para eles. Assim, torna-se necessário compreender o comportamento que eles possuem ao viajar pelo Brasil e pelo mundo. Foram utilizados dois grupos de sementes para o coletor: um contendo as trinta maiores cidades turísticas do Brasil e outro contendo as quinze maiores cidades turísticas do mundo, onde essa grandeza diz respeito à popularidade desses locais no *TripAdvisor*.

O coletor seguiu a seguinte ordem de coleta: acessa primeiramente uma página semente, em seguida vai para a página de cada atração encontrada ali. A partir da página da atração ele coleta as revisões relacionadas ao estabelecimento. Logo após coletar as revisões, ele obtém os dados de usuário de cada revisão e, a partir destes últimos dados, continua coletando a lista de cidades visitadas pelo usuário analisado. Pode-se conferir este comportamento na Figura 3.

[2]http://www.tripadvisor.com.br

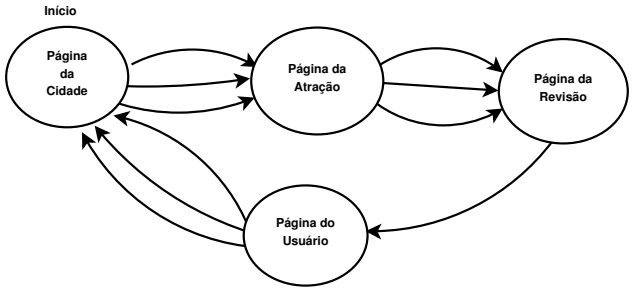


Figure 3. Ordem de páginas visitadas pelo coletor. Relações com uma seta indicam que apenas uma entidade de destino é encontrada na página de origem, enquanto as que possuem mais setas indicam que várias entidades podem ser encontradas.

5.2. Métodos de Referência

Foram utilizados dois métodos de referência para a avaliação comparativa dos algoritmos propostos. Estes métodos estão listados e descritos a seguir:

1. **Top-N**: é um algoritmo básico e intuitivo para recomendação. Como estão sendo recomendadas cidades, cria-se uma grande lista contendo todas as cidades visitadas pelos usuários e esta lista é ordenada em ordem reversa de contagem de visitas, ou seja, o primeiro elemento é a cidade que aparece em um número maior de *seenlists* (neste contexto *seenlists* são as cidades já visitadas pelo usuário) de usuários, o segundo elemento é composto pela segunda cidade em ordem reversa de visitação pelos usuários e assim sucessivamente. Desta forma, a tarefa de recomendação torna-se simples, basta que seja construída esta lista ordenada para todos os usuários existentes na base de dados;
2. **Weighted Top-N**: este algoritmo é similar ao Top-N, no entanto utiliza avaliações de *Atrações* para ponderar os elementos da lista ordenada, modificando sua ordem anterior. No $TripAdvisor$, cada atração pode ser marcada pelos usuários com uma nota avaliativa que varia entre 1 a 5, onde 1 é ruim e 5 é bom. Foi calculada a média das avaliações de todas as atrações de uma cidade para determinar o peso que seria multiplicado pelo número de vezes que a cidade ocorre nas *seenlists* dos usuários.

5.3. Avaliação

Uma vez que modelamos o problema de recomendação como um problema de aprendizagem de máquina, adotamos a metodologia de avaliação proposta em [Murphy 2012]. Utilizou-se validação cruzada com 5 *folds*, o que significa que os usuários têm sua *seenlist* separada em 5 partes de 20% cada, utilizando 4 partes para treino (U_{train}) e uma para teste (U_{test}), conforme ilustrado na Figura 4. Cada teste feito será desmembrado em 5 testes, utilizando grupos diferentes de dados treino/teste. É importante citar que na etapa 1 do modelo proposto, geração do grafo, somente os dados de U_{train} são utilizados para gerar o grafo.

Como métricas principais de avaliação foram escolhidas **Precisão** e **Revocação**. Essas métricas são calculadas pelas fórmulas descritas nas Equações 2 e 3, respectivamente.

$$Precision_{@k} = \frac{|\{U_{test}\} \cap \{Ranking_{@k}\}|}{|\{Ranking_{@k}\}|} \tag{2}$$

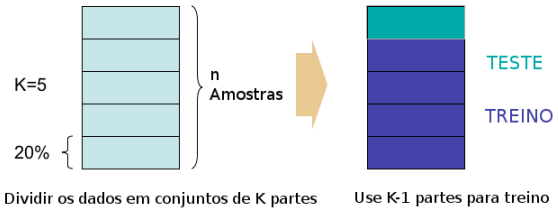


Figure 4. Validação cruzada com 5 folds. Os dados são divididos em 5 partes sendo uma separada para teste. O processo é executado mais 4 vezes separando uma parte para teste e as outras para treino.

$$Recall_{@k} = \frac{|\{U_{test}\} \cap \{Ranking_{@k}\}|}{|\{U_{test}\}|} \tag{3}$$

Nesta equação k corresponde à posição do *ranking* que está sendo analisada. Note que a diferença entre a avaliação de cada algoritmo citado na seção 4.3 é o parâmetro $Ranking_{@k}$, o qual é gerado diferentemente para cada um dos algoritmos. Todos os testes foram executados 5 vezes e é reportado o intervalo de confiança destes resultados, os quais serão apresentados na próxima seção.

6. Resultados

Esta seção apresenta os resultados dos experimentos, aplicando a técnica proposta ao conjunto de dados reais do TripAdvisor. Um ponto importante acerca dos experimentos é o número de cidades nas *seenlists* de usuários. Como esperado, esta distribuição possui uma cauda longa, o que significa que a maioria dos usuários possuem poucos elementos em sua *seenlist* 5. o que pode ser observado na Figura 5. A Seção 6.1 apresenta os experimentos para um grupo maior de usuários, em seguida é feita uma análise com grupo mais seleto de usuários, na Seção 6.2.

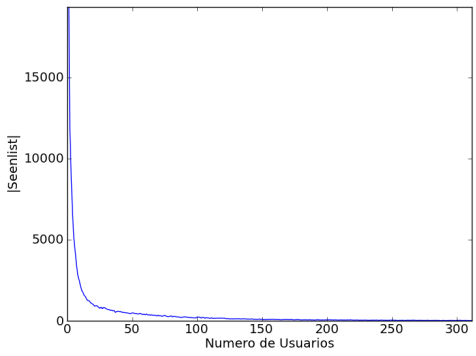


Figure 5. Relação entre número de usuários e tamanho de sua *seenlist*.

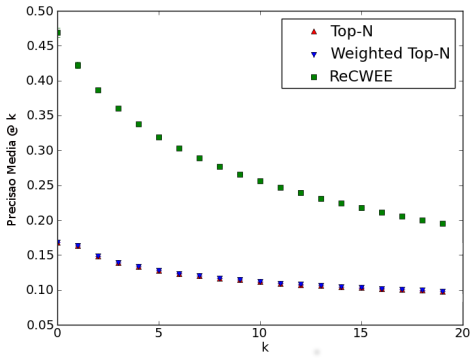
6.1. Resultados para grupo pouco seleto de usuários

Como o foco deste trabalho não é o problema de *cold start*, preferiu-se filtrar os usuários que possuem *seenlists* menores do que 10 cidades, o que resultou em 63.822 usuários. Não obstante, esses usuarios incluem usuários de todas as nacionalidades e deseja-se

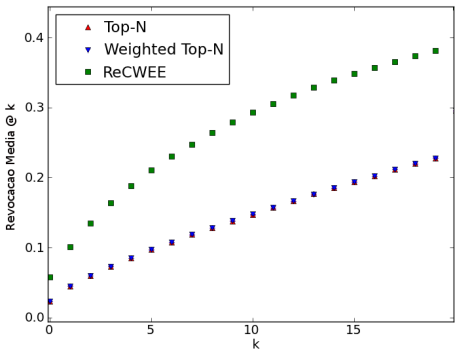
somente os que são brasileiros. Foi feito um filtro neste sentido, mas por conta de os dados serem não estruturados e pela falta de dados de vários usuários acerca do lugar em que vivem, este filtro por usuários brasileiros resultou em 26.549 usuários.

Foram geradas ordenações a partir das três abordagens algorítmicas em cinco grupos de 5.000 usuários, escolhidos aleatoriamente destes 26.549, com reposição. Para cada um destes 5 grupos foram executados testes com validação cruzada de 5 *folds*. Os resultados para esses experimentos são reportados na Figura 6(a), com intervalo de confiança de 95%. É importante perceber na Figura 6(a) que o algoritmo $ReCWEE$ é significantemente melhor que os outros dois em termos de precisão e é importante que se note também que os algoritmos $Top - N$ e $WeightedTop - N$ não apresentam resultados estatisticamente diferentes entre si, rejeitando a hipótese de que a ponderação pelas avaliações de atrações poderia melhorar a recomendação do $WeightedTop - N$.

Também são apresentados os resultados para revocação na Figura 6(b). Pode-se perceber que o algoritmo $ReCWEE$ possui uma revocação significativamente melhor que os outros dois e pode-se perceber também que os algoritmos $Top - N$ e $WeightedTop - N$ não apresentam resultados de revocação (*recall*) diferentes entre si.



(a) Resultados de precisão para os três algoritmos propostos com intervalo de confiança de 95%. Note que o algoritmo $ReCWEE$ é significantemente melhor que os outros dois.



(b) Resultados de revocação para os três algoritmos propostos com intervalo de confiança de 95%. Note que $ReCWEE$ possui revocação significantemente melhor que os outros dois.

Figure 6. Precisão e Revocação para os três algoritmos propostos

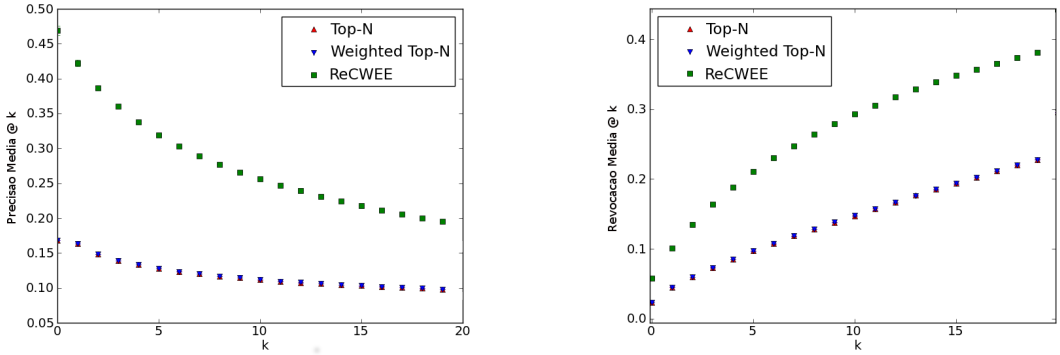
6.2. Resultados para grupo mais seleto de usuários

Foram executados os mesmos testes feitos anteriormente para grupos de usuários que possuem $|seenlist| \geq 200$, com o objetivo de verificar se o número de elementos na *seenlist* dos usuários poderia melhorar a precisão da recomendação. Os resultados destes experimentos são demonstrados na Figura 7(a).

Estes testes demonstraram que é possível que sejam criados grupos mais seletos de usuários a fim de realizar uma tarefa de recomendação hierárquica, onde grupos de usuários com número de *seenlist* semelhante podem ser separados para se obter uma melhor taxa de acertos entre si.

7. Conclusões

Recomendação é uma forma de auxiliar o usuário na tarefa cada vez mais difícil de tomar decisões on-line, seja qualquer que for a natureza desta decisão. Neste trabalho enfrentou-



(a) (b)

Figure 7. Precisão e Revocação para grupos de usuários com mais elementos na *seenlist*.

se pela primeira vez a tarefa de recomendar cidades para usuários. Esta tarefa se demonstra como sendo importante pelo simples fato de existirem sistemas como o TripAdvisor, onde as pessoas compartilham suas opiniões acerca dos mais diferentes destinos na intenção de auxiliar umas as outras em sua tomada de decisões de viagem.

Primeiramente foram formalizados os problemas a serem endereçados, em seguida apresentou-se a metodologia utilizada para sua resolução. A metodologia proposta, embora aplicada em um estudo de caso, não é específica e pode ser utilizada para diferentes cenários, constituindo-se na principal contribuição deste trabalho.

Os resultados demonstraram que a solução de inferência de uma rede social no TripAdvisor, em conjunto com sua segmentação em comunidades, consiste em uma boa técnica para melhoramento da precisão na recomendação de cidades ainda não visitadas pelos usuários, dobrando a precisão dos métodos de referência e ainda triplicando em casos para usuários mais seletos. Os resultados também demonstram que técnicas de segmentação e hierarquização de usuários podem ser utilizadas para melhorar ainda mais a precisão encontrada, uma vez que ao separarmos usuários com um maior número de informações e recomendar itens somente para eles foi observada uma precisão maior.

Como trabalhos futuros pretende-se melhorar a técnica de ordenação atualmente utilizada, levando em consideração a diversidade das recomendações. Também se pretende utilizar dados das *Revisões* feitas para cada atração a fim de encontrar características mais descritivas para as cidades, podendo agrupá-las por semelhança, o que possibilitaria a recomendação de cidades por grupos e não somente individualmente.

Adicionalmente, pretende-se relacionar os usuários por uma rede social real, utilizando dados da rede social $Facebook$, que estão disponíveis para um conjunto representativo de usuários do conjunto de dados coletados do TripAdvisor.

8. Agradecimentos

Esta pesquisa é parcialmente apoiada pelo Instituto Nacional de Ciência e Tecnologia para a Web (INWEB - CNPq no. 573871/2008-6), CAPES, CNPq, Finep e Fapemig.

References

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 Trans. on Knowl. and Data Eng.*, 17(6):734–749.

Balabanović, M. and Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Commun. ACM*, 40(3):66–72.

Cremonesi, P., Koren, Y., and Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 39–46.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174.

Herlocker, J., Konstan, J. A., and Riedl, J. (2002). An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Inf. Retr.*, 5(4):287–310.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., John, and Riedl, T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22:5–53.

Jamali, M. and Ester, M. (2009). Using a trust network to improve top-n recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 181–188.

Kurashima, T., Iwata, T., Irie, G., and Fujimura, K. (2010). Travel route recommendation using geotags in photo sharing sites. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 579–588.

Linden, G., Smith, B., and York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Noulas, A., Scellato, S., Lathia, N., and Mascolo, C. (2012). A random walk around the city: New venue recommendation in location-based social networks. In *SocialCom/PASSAT*, pages 144–153. IEEE.

Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors (2011). *Recommender Systems Handbook*. Springer.

Schafer, J. B., Konstan, J. A., and Riedl, J. (2001). E-commerce recommendation applications. *Data Min. Knowl. Discov.*, 5(1-2):115–153.

Wang, H., Terrovitis, M., and Mamoulis, N. (2013). Location recommendation in location-based social networks using user check-in data. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 374–383.

Yang, X., Steck, H., Guo, Y., and Liu, Y. (2012). On top-k recommendation using social networks. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 67–74.