4
Análise de recomendadores para a rede social Flickr

Neste capítulo apresentaremos os resultados obtidos com o emprego de recomendadores para itens da rede social *Flickr* [fli, 2008a] que é um sistema web para compartilhamento de fotos. Esta rede será melhor descrita na seção seguinte.

Como motivos para escolha da rede *Flickr* para essa análise mais detalhada, podemos enumerar:

- API pública: é disponibilizada[1] uma interface para integração de aplicativos clientes de terceiros com o *Flickr*. Tal interface é implementada usando o protocolo HTTP para comunicação, através de chamadas REST [Fielding, 2000]. Estas chamadas são implementadas por bibliotecas disponíveis em diversas linguagens de programação. Através dela é possível programaticamente listar, obter detalhes e executar ações sobre as principais entidades do sistema.
- maduro: o serviço *Flickr* começou a operar em 2001 e teve um grande crescimento ao longo dos seus primeiros 5 anos de existência. Hoje conta um pouco mais de 3 bilhões de fotos e possui uma comunidade grande e estável, oferecendo um volume enorme de inter-relacionamentos entre pessoas, grupos, fotos, tags etc.
- disponibilização do conteúdo dos itens: o principal tipo de item gerenciado pelo sistema são fotos[2] e o conteúdo de cada um desses itens (marcado pelos usuários como "públicos") está facilmente disponível via sua API. A rápida e simples obtenção do conteúdo de cada item analisado (imagens em sí) é fundamental para a realização de experimentos onde o conteúdo de cada item deve ser levado em conta.
- interessante: trata-se de um sistema público, cujo foco – compartilhamento de fotografias e comunidades orientadas à paixão e interesse por fotos – é amplamente compreendido pelo público em geral. Esse aspecto oferece maiores possibilidades para condução de testes online com o feedback de usuários acerca do desempenho dos recomendadores.

[1] em http://www.flickr.com/services/api/
[2] atualmente também é permitido o compartilhamento de vídeos curtos

A *framework* descrita no capítulo 3 será empregada: na modelagem dos relacionamentos presentes nessa rede, como auxílio na identificação das tarefas de recomendação possíveis e — através de extensões — na realização dos experimentos com recomendadores em sí.

Neste capítulo também descreveremos como foi realizada a caracterização da rede *Flickr* através de análises estatísticas realizadas em dados coletados do próprio serviço online. Tais análises são realizadas sobre os relacionamentos entre as principais entidades dessa rede social e maior ênfase é dada aos aspectos que influenciam a implementação de recomendadores.

Em seguida são demonstrados os resultados obtidos com os recomendadores implementados, contando com a análise de métricas de desempenho de recomendadores e sua aplicabilidade para os fins propostos.

Por fim, são apresentados também resultados do emprego de algoritmos recomendadores baseados em conteúdo, onde a métrica de semelhança entre itens adotada é a semelhança visual entre imagens.

4.1 Conceitos da rede Flickr

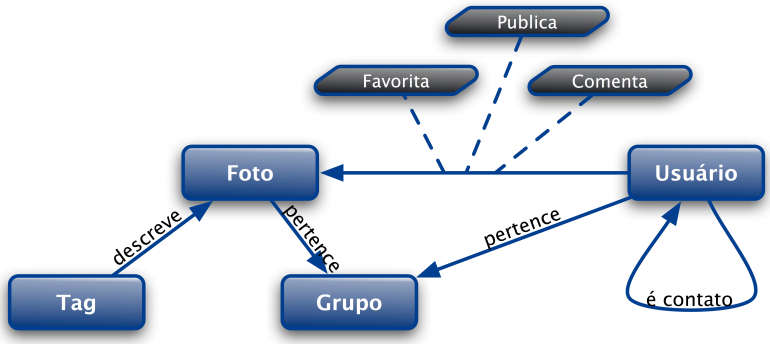


Figura 4.1: Principais entidades e relacionamentos do serviço *Flickr*

Os principais conceitos da rede *Flickr* e seus relacionamentos são apresentados na figura 4.1 e descritos a seguir[3].

Foto é a principal entidade. Sempre são contribuidas por usuários.

Tag são palavras-chave usadas livremente para a descrição de fotos. A associação de tags a fotos são realizadas primariamente pelos usuários responsáveis pela contribuição da foto à rede Flickr, no entanto demais contatos e usuários podem também contribuir com tags.

[3]um diagrama mais completo com todas as entidades e modos de interação do usuário no serviço *Flickr* são ilustrados na figura A.3.

Grupo representa tanto um conjunto de usuários com interesses em comum (usuários tomam a iniciativa de se associarem a um grupo). Todo grupo possui um "*pool*" de fotos e usuários membros do grupo tomam a iniciativa de submeter fotos a ele.

Usuário representa os usuários finais da rede *Flickr*, que são responsáveis por contribuir com todo o conteúdo existente.

Favorito usuários podem marcar fotos de seu interesse como favoritas, indicando alguma preferência pessoal por elas. A construção de sua coleção de favoritos pode ajudar o usuário na navegação e na recordação de imagens interessantes encontradas por ele no passado.

4.2 Coleta de dados

Foi realizada uma coleta de dados do *Flickr* considerando as seguintes entidades: `Foto`, `Grupo`, `Tag`, `Usuário`, `Fotos favoritas` e os seguintes relacionamentos entre eles: fotos em grupos, tags de fotos em grupos, grupos de usuários, fotos favoritas de usuários, tag de fotos favoritas de usuários, fotos de usuários e contatos de usuários.

Para tal foi utilizada a API pública [fli, 2008b] provida pelo *Flickr* e foram criados scripts em Python[4] para interagir com essa API e realizar a coleta de dados.

Como alternativa ao uso da API pública, existem protótipos [fli, 2008c] para realizar a exportação de contatos e grupos de um usuário *Flickr* em formato RDF[rdf, 2008]. Neste protótipo, as entidades representadas são mapeadas para classes das ontologias FOAF e SIOC. Um exemplo de saída deste protótipo pode ser vista na listagem 4.1.

Listagem 4.1: Exemplo de triplas RDF descrevendo dados *Flickr*

```
1 :myitempost rdf:type exif:IFD ;
2         dc:title ''Exemplo de item'';
3         sioc:has_creator :john ;
4         sioc:has_container :myflickrgallery .
5 :myflickrgallery rdf:type sioct:ImageGallery .
```

[4]Python [pyt, 2008] é uma linguagem de programação dinâmica, interpretada, orientada a objetos e de propósito geral, sendo utilizada para o desenvolvimento rápido de diversos tipos de softwares.

Como ponto de partida para obtenção desses dados, foram utilizadas as últimas imagens interessantes[5] do dia. Tal escolha acelerou a coleta de dados pois essas imagens fazem parte (na média) de um número alto de grupos e por serem mais populares, possuem um número maior de usuários, tags, etc conectados.

Os dados são coletados através de repetidas execuções do algoritmo 1.

Algoritmo 1: Obtenção de relacionamentos *Flickr* a partir de fotos interessantes aleatórias

```
Saída: fgu = lista de tuplas (grupo, usuario)
Saída: fgp = lista de tuplas (grupo, foto)
Saída: fup = lista de tuplas (usuario, foto)
Saída: fuf = lista de tuplas (usuario, fotofavorita)
Saída: fgt = lista de tuplas (grupo, tag, ocorrencias)
Saída: fuft = lista de tuplas (usuario, tag, ocorrencias)
para cada foto ∈ API.lista de fotos interessantes() faça
    ownerId = usuario que submeteu foto
    para cada ff ∈ API.lista de fotos favoritas(ownerId) faça
        fuf ← (ownerId, id de ff)
        para cada t ∈ API.tags da foto(ff) faça
            fuft ← (ownerId, id de t, ocorrencias de t)
    para cada g ∈ API.grupos do usuario(ownerId) faça
        fgu ← (ownerId, id de g)
        para cada f ∈ API.fotos do grupo(g) faça
            fgp ← (id de g, id de f)
            fup ← (ownerId, id de f)
            para cada t ∈ API.tags da foto(f) faça
                fgt ← (id de g, id de t, ocorrencias de t)
```

Em muitos experimentos a quantidade de dados utilizada para treinar e avaliar os recomendadores é utilizada como um parâmetro, para assim determinar como os resultados obtidos são influenciados pelo volume de dados empregado.

Esta análise em função do volume de dados foi fundamental na fase inicial para determinar a quantidade de dados adequada a ser coletada.

O volume de dados coletado é resumido na tabela 4.1 e os resultados das análises realizadas com esse volume de dados são descritos em 4.3.

[5] o *Flickr* utiliza uma métrica interna (não revelada para usuários) para determinar quão interessante uma foto pode ser para o público em geral. Tal métrica (chamada *interestingness*) considera por exemplo o numero de visualizações, número de pessoas que escolhem a foto como favorita, numero de comentários etc. Uma lista com as últimas 500 imagens *interessantes* pode ser obtida com a chamada à API *flickr.interestingness.getList()*.

Volume de dados	Tupla de identificadores em cada linha			
1,9G	grupos:	15,6 mil	tags:	12,3 milhões
81M	usuários:	110 mil	fotos:	1,15 milhões
129M	grupos:	15,6 mil	fotos:	1,59 milhões
42M	usuários:	141 mil	tags:	229 mil
98M	usuários:	127 mil	usuários:	893 mil
43M	usuários:	51,5 mil	fotos favoritas:	1,47 milhões

Tabela 4.1: Resumo do volume total de dados coletado do serviço *Flickr*

4.3 Caracterização dos dados coletados

Com o objetivo de melhor entender quantitativamente os relacionamentos entre as principais entidades dessa rede, nessa seção apresentamos as análises realizadas sobre os dados coletados usando a API pública do *Flickr*.

O conhecimento quantitativo desses relacionamentos é pré-requisito para a construção de recomendadores sob alguns aspectos: determinação do tamanho mínimo necessário da massa de dados para treinamento e validação, seleção de subconjuntos mais relevantes dos dados de treinamento — por exemplo, algumas tags usadas por usuários para descrever fotos têm baixo poder discriminatório pois aparecem com muita frequência (baixo IDF[6]) — ou não tem valor semântico, como tags-de-máquina: usadas para geo-localização ou como metadados para determinados mash-ups — por exemplo para contribuir com fotos de um determinado evento musical do site online *Last.fm*, basta o usuário associar à foto uma tag pré-determinada como *lastfm:event=148549*.

As análises foram realizadas com volumes de dados distintos para averiguar a necessidade de obter ainda mais dados, baseado na convergência das características observadas. Podemos afirmar então que os gráficos mostrados nas seções seguintes refletem a grosso modo o universo inteiro do serviço *Flickr*.

4.3.1 Trabalhos relacionados

Na literatura encontramos alguns trabalhos que descrevem o serviço *Flickr* em maiores detalhes, realizando também análises quantitativas dos dados lá disponíveis.

[Sigurbjornsson and van Zwol, 2008] traz uma análise de uma massa representativa de dados dessa rede, apresentando uma caracterização para as

[6]IDF — do inglês *inverse document frequency* ou frequencia inversa em documentos — pode ser obtida através do logaritmo da divisão do número total de fotos pelo número de fotos contendo a tag i, ou seja $idf_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$.

tags usadas pelos usuários para descrever fotografias e quais informações estão contidas nessas associações. Baseada nessa análise, também são apresentadas e avaliadas estratégias para a recomendação de tags para fotos. Essa mesma tarefa de recomendação é estudada pelo trabalho [Garg and Weber, 2008], que apresenta um sistema para sugestão de tags personalizado ao usuário que desempenha a tarefa de anotação das fotos. Neste trabalho o sistema recomenda em tempo real tags para a foto que o usuário está atualmente anotando baseado nas tags que ele ou outras pessoas usaram no passado, levando em conta também as tags atualmente associadas à foto em questão.

Já [Nicolas Pissard, 2007] estuda os aspectos sociais da rede *Flickr* com mais detalhes e descreve uma metodologia para análise em redes de usuários do sistema, capaz de produzir uma caracterização dos grupos do *Flickr* em termos da temática ou seus aspectos sociais.

[Prieur et al., 2008] traz as principais estatísticas observadas numa massa de dados considerável de usuários (5 milhões), fotos (150 milhões) e outras entidades obtidas do *Flickr*. A partir desses dados é feito um estudo de como contribuições individuais simples podem compor fontes de dados sólidas para diversos usos. Também são estudados os vários procedimentos empregados pelos usuários para selecionar itens de qualidade e como suas interações podem formar comunidades.

4.3.2 Contribuições de fotos por usuários

Na figura 4.2 temos um gráfico de barras representando — dentro do universo de dados coletados — o número de usuários que contribuíram determinada quantidade de fotos. Este gráfico contempla apenas os usuários que contribuíram pelo menos uma foto. A média de fotos públicas submetidas por um usuário é 10,45 com um alto desvio padrão de 19,61. Nota-se um elevado número de usuários com menos de 10 fotos e poucos usuários com mais de 300 fotos.

4.3.3 Relacionamentos entre grupos e fotos

Para a análise deste relacionamento, apresentamos na figura 4.3 histogramas de probabilidade para (a) o número de grupos onde determinada foto faz parte e (b) o número de fotos que fazem parte de determinado grupo. Estes gráficos contemplam apenas os grupos com pelo menos uma foto e as fotos que se encontram em pelo menos um grupo. Em (a) vemos que as fotos participam em média de 1,85 grupos, com um baixo desvio padrão de 2,21 enquanto que

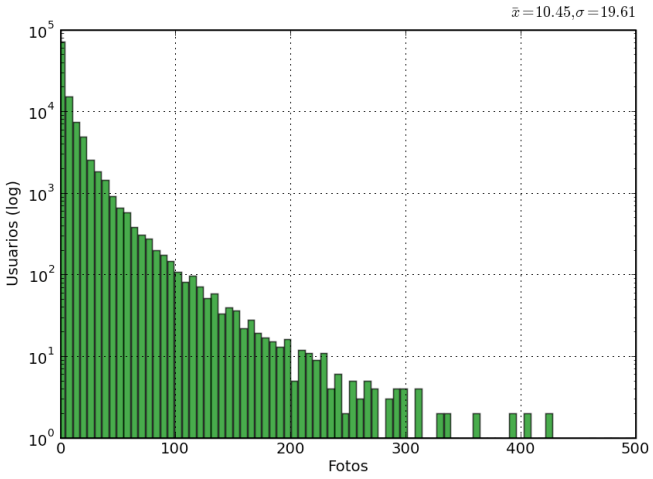
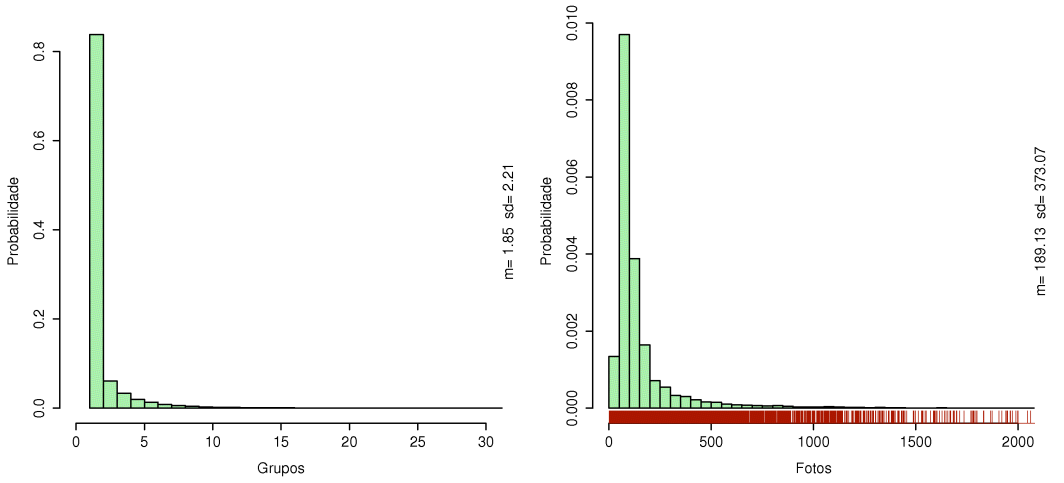


Figura 4.2: Freqüência de fotos contribuídas por usuário

em (b) temos em média 189 fotos associadas a um grupo e um alto desvio padrão de 373.



4.3(a): Número de grupos aos quais uma foto pertence

4.3(b): Número de fotos que são parte do *pool* de um grupo

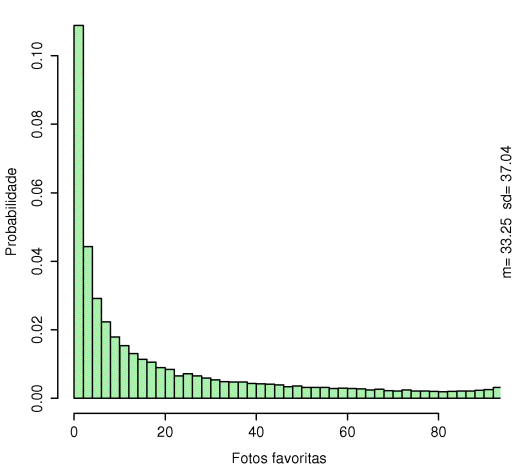
Figura 4.3: Relacionamentos entre grupos e fotos

4.3.4 Freqüência de imagens favoritadas por usuário

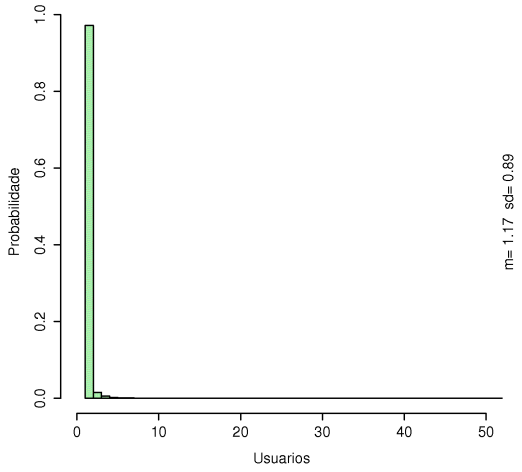
Na figura 4.4 podemos analisar o hábito dos usuários ao usar a funcionalidade de favoritos. Note que esses histogramas contemplam apenas as fotos que um ou mais usuários marcaram como favoritas e os usuários que favoritaram pelo menos uma foto.

Percebemos na figura 4.4(a) que os usuários favoritam em média 33 fotos, com um desvio padrão considerável. Vemos também que mais de 10% dos usuários favoritaram apenas uma foto.

Já a figura 4.4(b) representa o número de usuários que marcaram determinada foto como favorita. Dela podemos inferir que das fotos favoritadas por pelo menos um usuário, a imensa maioria foi favoritada por apenas um usuário. Como consequência para a construção de recomendadores, podemos concluir que a informação de co-ocorrência de usuários que favoritam uma determinada foto não seria um dado interessante pois há uma quantidade relativamente pequena de fotos que possuem mais de um usuário que declararam interesse por elas.



4.4(a): Número de fotos escolhidas como favoritas por um usuário



4.4(b): Número de usuários que escolheram determinada foto como favorita

Figura 4.4: Relacionamentos entre usuários e fotos favoritas

4.4 Avaliação das possíveis tarefas de recomendação

Fazendo uso da notação gráfica proposta na seção 3.1.3 podemos realizar a modelagem das principais tarefas de recomendação para a rede *Flickr* conforme representado na figura 4.5.

Nesta modelagem representamos **grupos** no Flickr usando a classe `SIOC:Usergroup` e os **usuários** como `SIOC:User`. Fotos contribuidas por usuários são representadas como `SIOC:Item` e o **pool** de fotos de um grupo pode ser modelado como uma extensão à classe `SIOC:Container`. Alternativamente, este conceito poderia ser modelado usando a classe `SIOC:ImageGallery` da ontologia *SIOC-Types*.

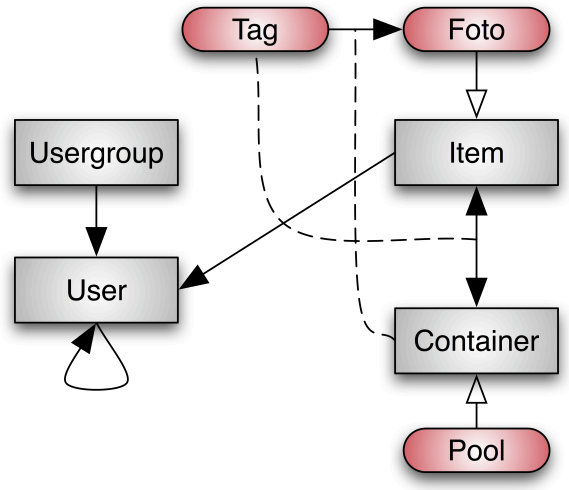


Figura 4.5: Tarefas de recomendação na rede social *Flickr*

Demais conceitos da rede *Flickr* como álbuns pessoais de fotos, comentários etc não são modelados já que não participam das principais tarefas de recomendação que destacamos a seguir:

Grupo para foto e vice versa seria útil para gestores de comunidades, ajudando na busca de novo conteúdo para o pool de fotos de um grupo, ou para usuários interessados em aumentar a exposição de suas fotos contribuídas. Neste caso, a recomendação de grupos relevantes para a foto alvo proporcionaria elevado número de visualizações da foto e por consequência maior número de tags submetidas para a foto, ocorrências de favoritos etc;

Grupo para usuário ajudaria usuários a encontrar novos grupos de seu interesse;

Usuário para usuário baseado no gosto dos usuários, poderia ajudar usuários a encontrarem outros usuários de gostos semelhantes, ajudando no estabelecimento de mais um link social ou na tarefa de encontrar mais conteúdo interessante para o usuário;

Tag para foto o sistema poderia sugerir tags a serem associadas a determinada imagem, baseando-se em tags comumente usadas pelo usuário para descrever suas fotos contribuídas, ou então baseada em conjuntos de tags frequentemente usadas para descrever fotos nos grupos onde a imagem se encontra. Ambas abordagens poderiam favorecer tags de alto teor discriminatório, uma vez que as tags associadas a imagens são usadas por demais usuários durante a busca por imagens;

Foto para usuário seria uma das recomendações mais úteis para usuários. Com esta ajuda, usuários teriam acesso a mais fotos de seu interesse, contribuindo para um melhor uso do serviço *Flickr*. Esta recomendação pode ser realizada através de filtragem colaborativa baseada em usuários, onde a proximidade de usuários é definida por contatos da rede social do usuário alvo e os itens associados aos usuários "vizinhos" são suas imagens favoritadas.

Nas seções seguintes serão apresentados os resultados obtidos com os recomendadores implementados para algumas das tarefas de recomendação acima. A figura 4.6 apresenta estes experimentos (setas em azul).

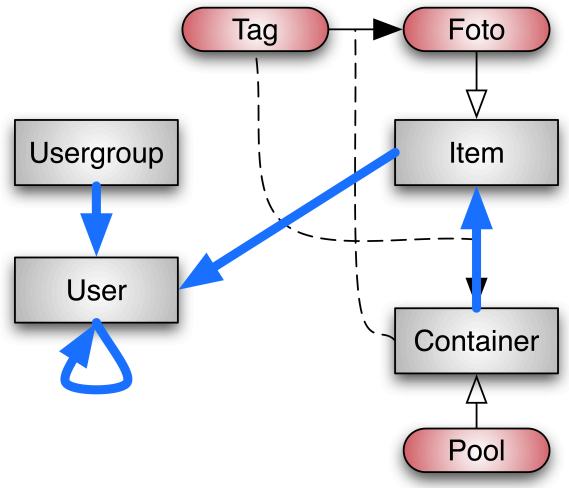


Figura 4.6: Experimentos de recomendação realizados na rede *Flickr* (setas em azul)

4.5 Recomendação de grupos para fotos

Nesta seção descrevemos a avaliação realizada acerca do desempenho de uma implementação de um algoritmo recomendador para a tarefa de sugerir grupos para fotos.

Trata-se então da implementação [Karypis, 1997][7] do algoritmo *Top-N* [Deshpande and Karypis, 2004, Sarwar et al., 2001] para filtragem colaborativa baseado na co-ocorrência de itens.

A avaliação do desempenho do recomendador nessa tarefa foi realizada segundo o algoritmo 2, onde as tuplas de entrada para treinamento $(usuário, item)$ são nesse caso tuplas $(fotos, grupos)$, que foram obtidos na

[7]Sua API foi exposta para programas Python e disponibilizada sob licença GPL em [Cabral, 2008].

seção 4.2. Este algoritmo baseia-se na técnica de *repeated hold-out* para composição dos subconjuntos de dados de treinamento e validação.

Algoritmo 2: Avaliação de algoritmos de recomendação

Entrada: $inData$ = lista de tuplas $(usuario, item)$ `/* usuario é a entidade que deseja recomendações e item a entidade recomendada */`
Entrada: $nRec$ = número de recomendações feitas por tentativa
Saída: $resultados$ = lista de tuplas $(taxaAcerto, parametro1, parametro2)$

```
1  repita
2      selData = subconjunto de tamanho aleatório de inData
3      para cada usuario ∈ selData faça
4          escolhe uma tupla (usuario, item) de selData e esconde
5      parametro1 = valor aleatório dentre os possíveis
6      parametro2 = valor aleatório dentre os possíveis
7      inicializa modelo de recomendação com
       selData, parâmetro1, parâmetro2
8      acertos = 0
9      tentativas = 0
10     para cada usuario ∈ selData faça
11         calcula nRec sugestões de itens para usuario
12         se alguma sugestão concorda com itens escondidos de usuario
           então
13             acertos++
14         tentativas++
15     adiciona tupla (acertos/tentativas, parametro1, parametro2) a
       resultados
16 até obter quantidade adequada de resultados
```

A figura 4.7 contextualiza tal algoritmo, representado-o como o processo `Testador`. Nesta mesma figura, a implementação do algoritmo *Top-N* utilizada seria responsável pelos processos `Aprender modelo` e `Recomendador`.

Nesta mesma figura, vemos que os dados coletados são particionados em dois: relacionamentos de treinamento e de teste. O processo de particionamento, treinamento do recomendador e avaliação do desempenho é realizado segundo a técnica de *repeated hold-out* [Blum et al., 1999]. Nesta técnica, instâncias de dados são selecionadas aleatoriamente para compor uma partição ou outra e todo o processo de partição, treinamento e avaliação é repetido diversas vezes para que os resultados obtidos convirjam para o valor correto.

Em alguns dos experimentos realizados, o tamanho da partição de teste em relação à de treinamento é também variado, com o objetivo de determinar a proporção ideal (a partir da qual o desempenho observado fica estável).

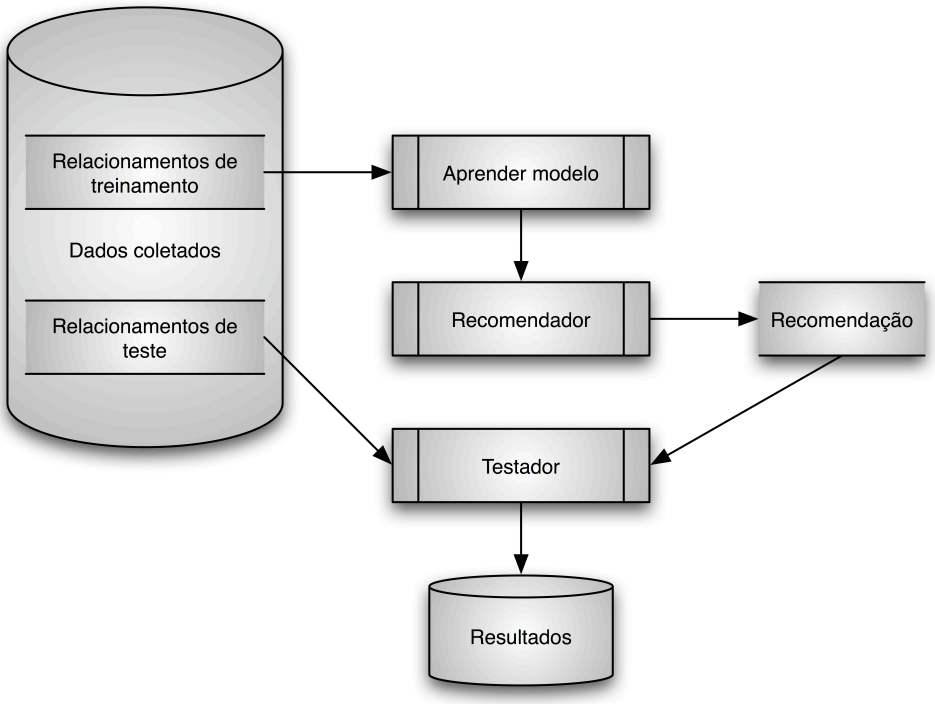


Figura 4.7: Fluxo de dados e entidades envolvidas no processo de avaliação de recomendadores

O algoritmo 2 proposto pode ser reutilizado para a avaliação de recomendadores quando há dois graus de liberdade e a tarefa a ser avaliada é, por exemplo, a de encontrar itens bons (ver seção 2.3). De uma forma, este algoritmo realiza um cálculo da precisão, que pode ser definida como a taxa de itens bons dentre os recomendados, ou seja, os grupos recomendados cujas fotos alvo de recomendação estavam de fato associadas a eles no conjunto de dados de validação.

Para avaliar o desempenho do algoritmo de recomendação *Top-N* para esta tarefa, foram empregados como parâmetros do algoritmo 2 o volume de dados utilizado para a construção do modelo interno (aprendizagem) e o tamanho (em número de itens) da vizinhança avaliada. Como entrada do algoritmo temos também o parâmetro fixo *nRec*, usado para determinar o *rank* da precisão medida.

Esta vizinhança, melhor definida em [Deshpande and Karypis, 2004], significa o número de itens semelhantes considerados na hora de construir recomendações.

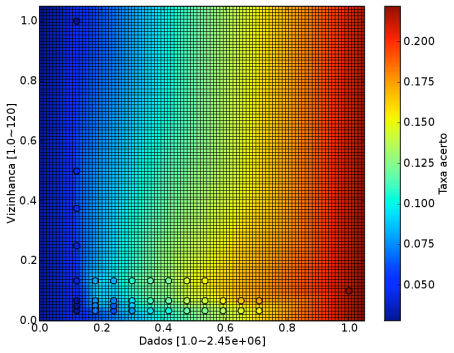
As tuplas contendo os resultados obtidos foram plotados num gráfico de espalhamento, contendo os parâmetros avaliados como eixos `X` e `Y` e o desempenho como a cor dos pontos.

Tais gráficos de espalhamento usam a técnica "interpolação natural de

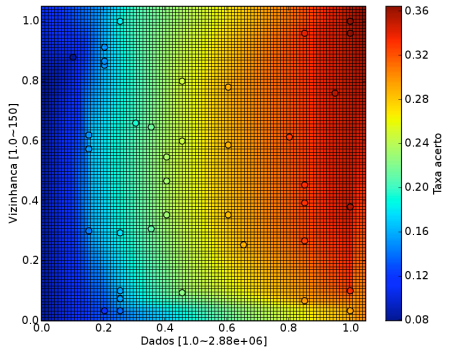
vizinhança"[8] para calcular o gradiente de cor utilizado no fundo, que indica o desempenho das recomendações.

Para esta tarefa são mostrados os resultados nas figuras 4.8 e 4.9. Na primeira o parâmetro de entrada *nRec* do algoritmo é 6 e na segunda é 12. O objetivo de realizar o experimento variando tal parâmetro é verificar como seria o desempenho em função do número de sugestões (rank) apresentadas na interface com o usuário final.[9]

Dos resultados mostrados na figura 4.8 podemos concluir que, como esperado, há uma forte correlação positiva entre a proporção dos dados utilizados para treinamento e o desempenho do recomendador. Observa-se que há pouca ou nenhuma correlação entre o tamanho da vizinhança (parâmetro do algoritmo *Top-N*) e o desempenho. Por fim, observa-se que há ganhos na precisão (ou desempenho) do algoritmo quando são apresentadas o dobro de recomendações (12) ao usuário, mas nota-se que a melhor precisão obtida com 12 recomendações não chega a ser o dobro do que é obtido com apenas 6 recomendações, o que indica que o número ótimo de recomendações — não levando em conta aspectos de usabilidade ao apresentar os grupos recomendados para o usuário: paginação, formato da listagem etc — deve estar dentro da faixa 6–12.



4.8(a): 6 recomendações

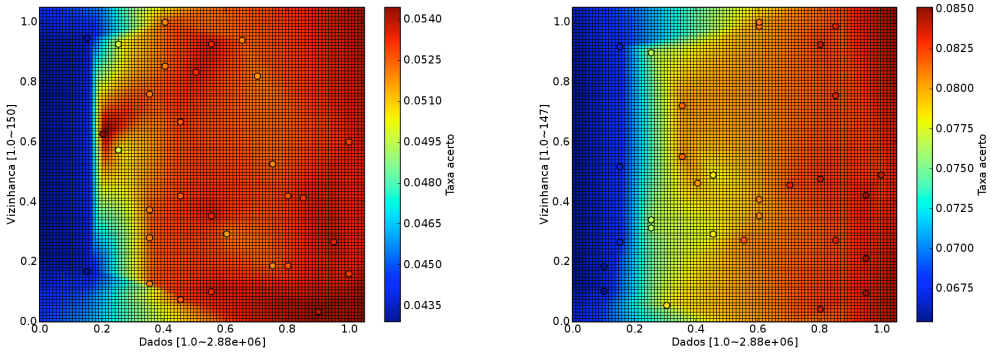


4.8(b): 12 recomendações

Figura 4.8: Desempenho da recomendação de grupos para fotos usando algoritmo *Top–N* (filtragem colaborativa baseada na co-ocorrência de fotos em grupos)

[8]ver Sibson, R., *A Brief Description of Natural Neighbor Interpolation*, em *Interpreting Multivariate Data*, ed. por V. Barnett, John Wiley & Sons, New York, 1981, pp. 21-36, conforme implementado na biblioteca NCAR natgrid (`http://code.google.com/p/griddata-python/`)

[9]Teoricamente, quanto mais sugestões, maior a chance de recomendar um item de interesse e maior o desempenho (segundo definição usada no algoritmo 2), no entanto na prática há um limite de usabilidade para o número de resultados apresentado aos usuários.



4.9(a): 12 recomendações

4.9(b): 24 recomendações

Figura 4.9: Desempenho da recomendação de grupos para fotos (comparativo naïve: grupos com maior número de fotos)

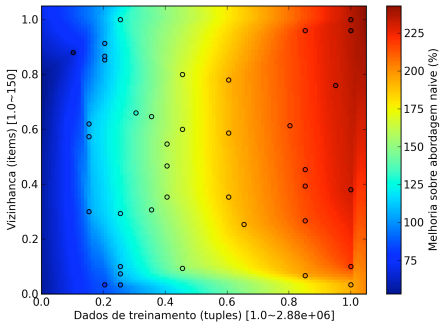


Figura 4.10: Comparativo entre o algoritmo *Top–N* para recomendação de grupos para fotos e a abordagem naïve (rank de 12 recomendações)

Como comparativo para o desempenho obtido nas recomendações, o algoritmo 2 foi executado usando um recomendador "míope" ou "naïve", que apenas recomenda os itens de maior freqüência nos dados de entrada. Isso equivaleria a sempre recomendar ao usuário os itens mais populares, independentemente das preferências dele e de sua vizinhança.

Para a precisão ao recomendar 12 itens para o usuário (rank), o desempenho desse recomendador "naïve" é apresentado na figura 4.9 e um comparativo com o algoritmo *Top–N* na figura 4.10.

Desta última figura podemos observar que com uma vizinhança de aproximadamente 150 e praticamente todos os dados coletados disponibilizados para o treinamento, a implementação *Top–N* alcançou uma precisão mais de 2,25 vezes melhor que a abordagem naïve.

4.6 Recomendação de grupos para fotos utilizando conteúdo visual

Como alternativa à abordagem anterior (seção 4.5) para a construção de um recomendador que ajude na tarefa de encontrar grupos relevantes para uma foto alvo, apresentamos nessa seção experimentos realizados durante a construção de recomendadores de grupos para fotos que levem em conta o conteúdo visual de imagens dos grupos recomendados.

4.6.1 Métrica para semelhança visual de fotos

Inicialmente foi realizada uma análise do conteúdo visual das imagens de alguns grupos escolhidos aleatoriamente. A métrica utilizada para determinar a distância entre duas imagens — e assim calcular a distância visual média entre todas as fotos de um grupo — consiste na comparação de duas assinaturas calculadas utilizando a transformada de *Haar*, segundo [Jacobs et al., 1995].

As transformadas *wavelet* [Stollnitz et al., 1995], em especial a de *Haar*, são indicadas para a determinação da semelhança visual de imagens e para a construção de assinaturas para uma base de dados de imagens por permitir uma boa caracterização aproximada da imagem com poucos coeficientes (ver exemplo na figura 4.11).

4.11(a): 20 coefs.

4.11(b): 100 coefs.

4.11(c): 400 coefs.

4.11(d): 16k coefs. (original)

Figura 4.11: *Wavelets* da base *Haar* permitem uma boa caracterização aproximada da imagem com poucos coeficientes

Dentre as vantagens do uso de transformadas *wavelet* para caracterizar imagens, podemos destacar[10]:

- Independentes da resolução original da imagem.
- Desacoplamento entre resolução da imagem de consulta e alvo.
- Fáceis de implementar.
- Extraem e codificam bem informação de bordas: mostrando-se útil para recuperação de imagens baseado no conteúdo e geometria da imagem alvo da busca, em especial para consultas realizadas a partir de desenhos realizados por usuários, onde há predomínio de formas e áreas (há pouca informação de textura).
- Rápidas: tempo linear em relação às dimensões das imagens em tempo de indexação e do tamanho da base já indexada em tempo de busca.
- Caracterização econômica (em termos de memória usada) de imagens: Considera apenas os coeficientes de módulo elevado.

Todas as fotos de um subconjunto aleatório dos grupos coletados na secão 4.2 foram analisadas segundo as transformadas *wavelet* descritas acima e calculou-se para cada grupo a distância visual média entre as fotos desse grupo (Δ_G) segundo

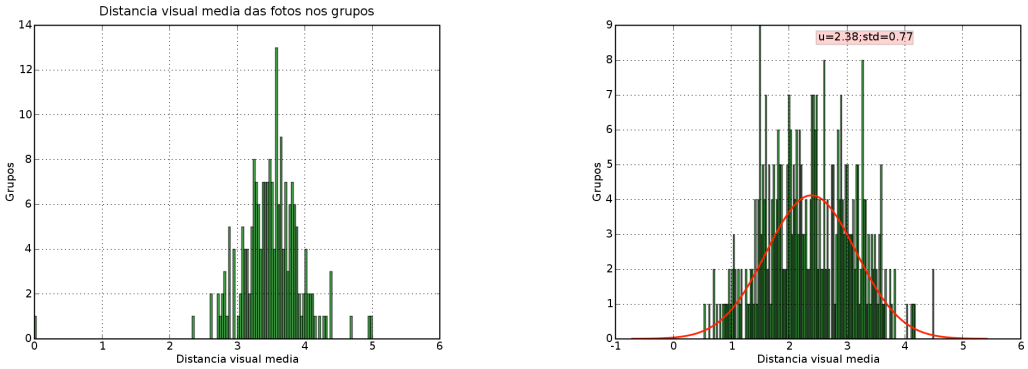
$$\Delta_G = \frac{\sum_{(a,b)\in\bar{G}} \delta(a,b)}{|G|}, \tag{4-1}$$

onde $\bar{G}$ é o conjunto de todas as combinações entre pares de fotos distintas (a,b) do grupo G, $\delta(a,b)$ o grau de semelhança entre duas fotos e $|G|$ o número de fotos de G.

Maiores valores da média Δ_G (chamaremos essa métrica de **homogeneidade visual** do grupo) indicam maior grau de semelhança entre as fotos constituintes do grupo, ou seja, que as fotos desse grupo têm baixa variedade de cor e forma. Os valores máximos e mínimos para Δ_G devem ser obtidos empiricamente, uma vez que eles dependem de detalhes da implementação dos algoritmos para processamento de imagens utilizada. Estima-se no entanto que grupos onde todas as imagens são idênticas possuam Δ_G=6 e um grupo teórico onde a variação visual entre as imagens é máxima possua Δ_G=0.

Na figura 4.12 temos o gráfico de barras da distribuição da homogeneidade visual dos grupos quando são consideradas amostras de 210 e 1248 grupos.

[10]Tais propriedades de *wavelets* também já foram exploradas para compressão de imagens com perdas, como por exemplo no formato JPEG2000 [Taubman et al., 2002].



(a) 210 grupos (b) 1248 grupos

Figura 4.12: Distribuição da homogeneidade visual dos grupos

Embora na figura 4.12(b) a curva normal não seja uma boa aproximação, os parâmetros (média, variância, etc) de curvas que se ajustem a esses dados podem ser utilizados, por exemplo, para classificar grupos entre os que são ou não "visualmente consistentes" ou "visualmente homogêneos" (daí o nome "homogeneidade visual dos grupos" proposto para essa métrica). Essa métrica de distância visual média das fotos nos grupos pode ser usada também como peso ao usar técnicas híbridas de recomendação (quando o resultado de múltiplos recomendadores são combinados para formar uma recomendação final).

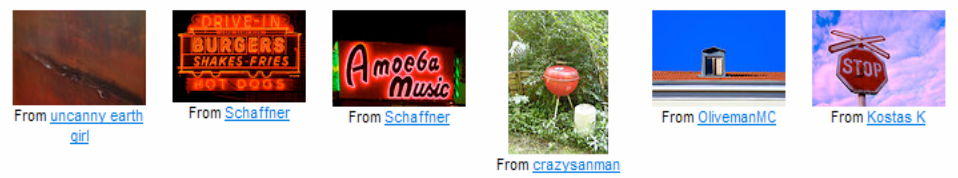


Figura 4.13: Exemplos de imagens do grupo *Absolut Red*

Figura 4.14: Exemplos de imagens do grupo *BURMA FREE - Birmania Libera*

Para uma análise qualitativa são listados na tabela 4.2 exemplos de grupos com as maiores e as menores médias para distância visual entre suas fotos. Da figura 4.13 à figura 4.16 são mostrados exemplos de imagens dos dois primeiros e dois últimos grupos desta tabela.

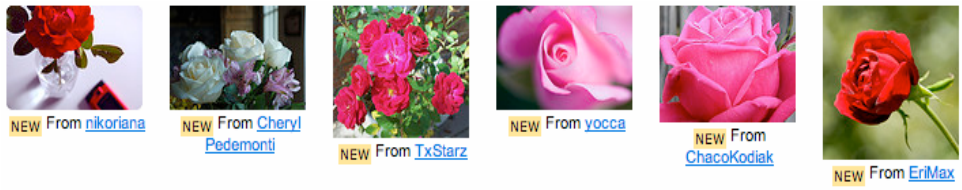


Figura 4.15: Exemplos de imagens do grupo *Roses*

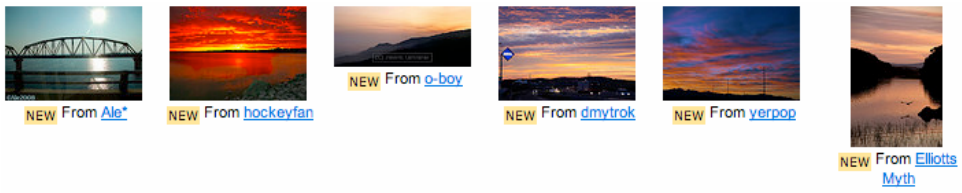


Figura 4.16: Exemplos de imagens do grupo *Sunrise, Sunset – Anything Sun!*

Semelhança média	Grupo
4.94594995937	Roses
4.87667388368	Sunrise, Sunset – Anything Sun!
4.86137230043	Sunsets and sunrises around the world
4.46156954561	i love food group
4.40303794329	Babies
4.31499678572	Flower Macro FIVE photos per day
4.29776448016	flower power
...	
2.78482034757	DeviantArt
2.72887774904	Art Now
2.71548516829	Moleskinerie
2.70778573965	no limits HDR
2.68105224647	It's magical - A photo manipulation group
2.6782242351	BURMA FREE - Birmania Libera
2.59634364472	Absolut Red

Tabela 4.2: Grupos com maiores e menores homogeneidades visuais. Maiores valores indicam maior grau de semelhança entre as fotos do grupo (valores entre 0 e ~ 6).

4.6.2
Algoritmo para recomendação baseada em conteúdo visual

Para avaliar a contribuição que heurísticas baseadas em conteúdo visual ofereceriam, foi realizado um experimento para a tarefa de recomendar grupos para fotos. Neste, foi usada como informação apenas aspectos visuais das fotos, ou seja, não foi considerado para o algoritmo de recomendação o contexto da foto alvo dentro da rede social.

O algoritmo 3 baseia-se na intuição de que fotos visualmente semelhantes devem pertencer ao mesmo grupo. Assim, bons grupos para uma determinada foto seriam aqueles mais recorrentes dentre os grupos onde imagens semelhantes à imagem alvo da recomendação estão.

4.6.3
Resultados

Este algoritmo foi então executado dezenas de vezes, gerando o gráfico de espalhamento na figura 4.17. Para esse experimento foi utilizado um universo de 900 imagens distintas e a métrica de semelhança visual de imagens descrita no início desta seção.

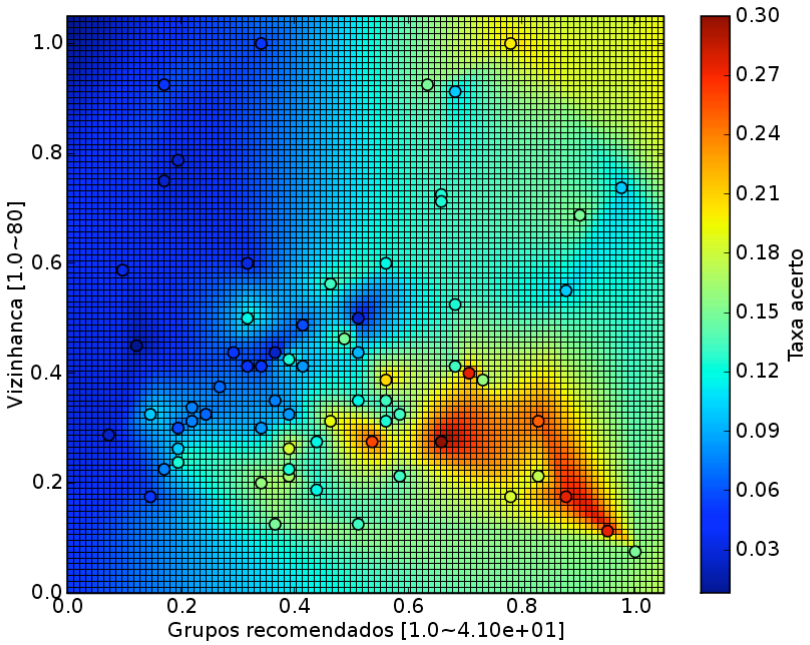


Figura 4.17: Desempenho de recomendador de grupos para fotos baseado apenas no conteúdo visual dos grupos

Nesta figura, podemos observar a maior precisão obtida como sendo cerca de 30%, quando foram recomendados aproximadamente 26 grupos e com a

vizinhança de 22 fotos. Pode-se inferir visualmente da interpolação realizada no gráfico que valores menores da vizinhança implicaram em melhores precisões.

Algoritmo 3: Avaliação de algoritmos de recomendação de grupos para foto baseado em conteúdo visual

Entrada: $inData$ = lista de tuplas $(foto, grupo)$
Entrada: $sim(imagem, n)$ = comparador de imagens baseado em semelhança. Retorna as n imagens mais semelhantes à imagem passada
Saída: tupla de resultado $(taxaAcerto, nRec, simImages)$

```
1  nRec = número de recomendações feitas (escolhida aleatoriamente)
2  simImagesNum = tamanho da vizinhança (imagens semelhantes)
   considerada (escolhida aleatoriamente)
3  acertos = 0
4  tentativas = 0
5  repita
6      imagem = imagem de inData (escolhida aleatoriamente)
7      imagensSemelhantes = sim(imagem, simImageNum)
8      gruposRecomendados = os nRec grupos mais comuns dentre os
       grupos que as imagens imagensSemelhantes pertençam
9      se imagem pertence de fato a algum dos grupos em
       gruposRecomendados então
10         acertos++
11     tentativas++
12 até obter quantidade adequada de resultados
13 retorna tupla (acertos/tentativas, nRec, simImages)
```

Como melhoria possível para os resultados obtidos, podemos por exemplo restringir (ou dar maior peso) o universo de grupos a serem recomendados e treinados apenas aos mais visualmente homogêneos (i.e. com maior semelhança visual média entre as fotos do grupo). A intuição por trás dessa melhoria é que encontramos no serviço *Flickr* diversos temas associados a grupos: alguns deles são completamente ortogonais a aspectos visuais das imagens, como temas abstratos (*amor, viagem, nação, Alemanha ...*) enquanto outros são mais relacionados ao conteúdo visual, como *arquitetura, paisagens, retratos, carros, mar, vermelho* etc. Faria mais sentido recomendar grupos para fotos (quando baseia-se no conteúdo visual da foto alvo) somente se os grupos recomendados pertencem a esse último conjunto de temas. Nota-se também que as imagens encontradas nesse último conjunto de temas possuem determinadas características visuais recorrentes entre elas, por conseqüência aumentando a semelhança visual média do grupo.

4.7 Recomendação de fotos para usuários

A tarefa para recomendação automática de fotos para usuários pode ocorrer dentro de alguns contextos (fonte de imagens candidatas para a recomendação): dentre fotos de todos usuários, dentro de fotos de um grupo, dentro de fotos dos seus contatos etc.

Como principais abordagens para a construção deste recomendador podemos destacar: filtragem colaborativa por imagens favoritadas em comum, por semelhança visual (via construção de perfil visual de usuários), imagens favoritadas por vizinho (no grafo social) que ainda não foram favoritadas pelo usuário etc.

Uma abordagem seria a filtragem colaborativa por co-ocorrência de tags das fotos contribuidas com as de outros usuários. Para tal, deve-se considerar que tags contém "impurezas": geo-localização, convites para submeter fotos para grupos etc, pois estas contém informação com pouco teor semântico e discriminatório.

Para esta tarefa realizamos três experimentos diferentes, todos eles com rank de valor 24:

1. comparativo naïve, que recomenda ao usuário as fotos mais favoritadas por seus contatos: Figura 4.18(a);

2. comparativo naïve que recomenda (independente do usuário alvo da recomendação) as fotos mais favoritadas globalmente: Figura 4.18(b);

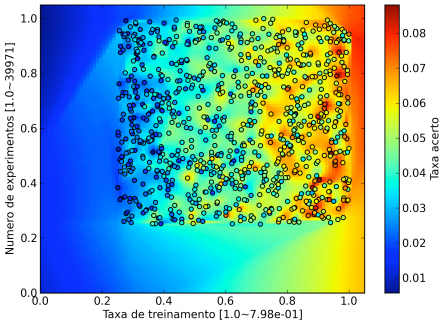
3. usando filtragem colaborativa *Top-N*, tendo como base a co-ocorrência de fotos favoritas em comum entre usuários: Figura 4.19.

Conforme resumido na tabela 4.3, a primeira abordagem naïve alcançou precisão máxima de 8% e a segunda de 5,4%. Já o algoritmo de filtragem colaborativa alcançou precisão de 13%.

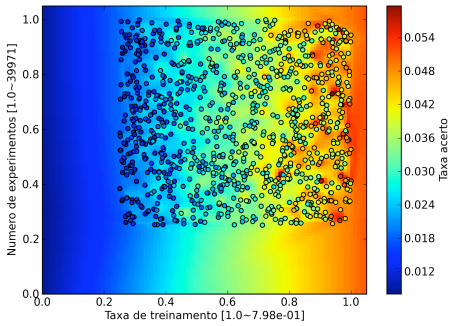
4.8 Recomendação de usuários para usuários

Como experimento para a tarefa de recomendação automática de usuários com gostos semelhantes, foi implementado um protótipo[11] que realiza recomendações de usuários para usuários (usuários semelhantes). Para tal faz uso de uma métrica baseada na distância cosseno entre os usuários, calculada a

[11]hospedado no ambiente TecWeb [tec, 2008] disponível em http://server2.tecweb.inf.puc-rio.br:8080/fs26/Users



(a) comparativo naïve (fotos mais favoritadas entre contatos do usuário alvo)



(b) comparativo naïve (fotos mais favoritadas globalmente)

Figura 4.18: Desempenho da recomendação de fotos para usuários usando duas abordagens naïve: "fotos mais favoritadas pelos contatos do usuário" vs. "fotos mais favoritadas por todos os usuários")

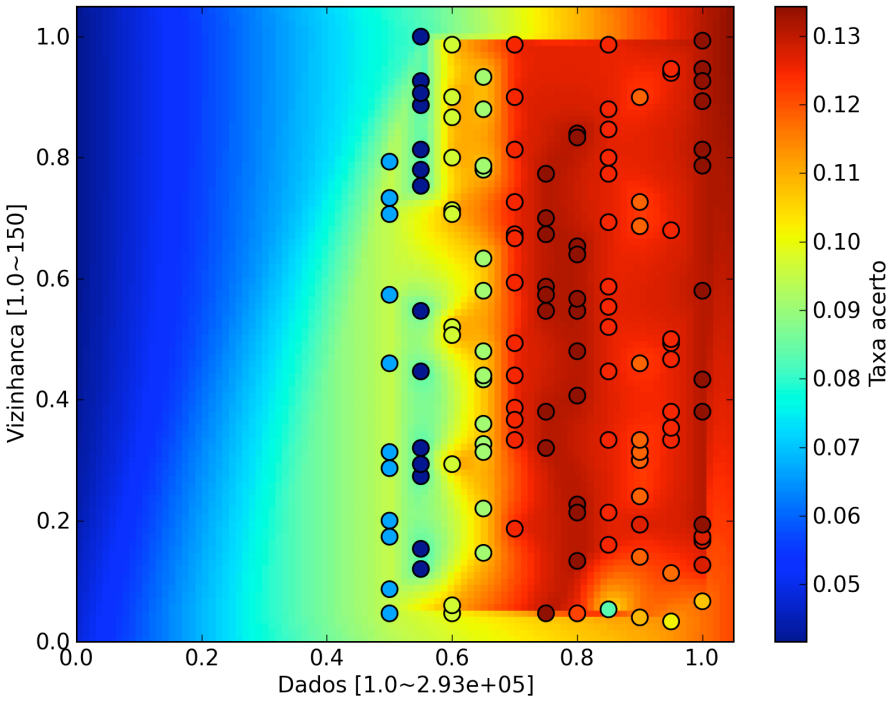


Figura 4.19: Desempenho da recomendação de fotos para usuários usando filtragem colaborativa *Top-N*, tendo como base co-ocorrência de fotos favoritas em comum entre usuários

partir da frequência de ocorrência de tags nas fotos contribuídas por cada usuário.

Trata-se então de um *mash-up*[12] *Flickr*, capaz de obter dados sobre a rede social (contatos) do usuário experimentando o sistema em tempo real, apresentando a ele os resultados da recomendação.

Na interface com o usuário, para realizar uma recomendação basta entrar um nome/id de usuário *Flickr* ou e-mail e clicar no botão "Search". Ao carregar no lado do servidor todas as fotos e tags to usuário alvo, o sistema constrói o perfil de gosto do usuário alvo da recomendação. Para alimentar sua base de usuários candidatos para futuras recomendações, o sistema faz também a obtenção via API *Flickr* das fotos e tags associadas para quatro contatos aleatorios de um usuário que fez uso do sistema.

Atualmente o protótipo conta com cerca de 180 mil tuplas *(id usuário, id tag)* e uma recomendação demora cerca de seis segundos para ser realizada (excluindo tempo de rede e reposta da API *Flickr* para obter os dados necessários).

4.8.1 Algoritmo de recomendação

O algoritmo utilizado é baseado nos conceitos de filtragem colaborativa e leva em conta a frequência de tags das fotos contribuidas de um usuário. Neste algoritmo o usuário é representado (equação 4-2) como um vetor normalizado cujas dimensões são as tags possíveis e o valor absoluto em cada dimensão é o número de ocorrência dessa tag em imagens do usuário. Faz-se uso então (equação 4-3) do produto vetorial com todos os outros usuários para definir uma distância cosseno como métrica para semelhança de usuários.

$$\vec{u} = (t_1, t_2, ..., t_n), n = |T| \tag{4-2}$$

$$\delta(u_1, u_2) = cos(\vec{u_1}, \vec{u_2}) = \frac{\vec{u_1} \cdot \vec{u_2}}{\|\vec{u_1}\| \cdot \|\vec{u_2}\|} \tag{4-3}$$

Como variação no algoritmo para calculo do grau de semelhança entre os usuários candidatos e o usuário alvo da recomendação, é usado o *log* do resultado desse mesmo produto vetorial, e um dos objetivos do experimento é avaliar a vantagem dessa variação.

[12]composição ad-hoc de aplicativos web que fazem uso de conteúdo oriundo de distintas fontes externas de dados, disponibilizando assim ao usuário final um novo serviço.

4.8.2
Avaliação de desempenho do protótipo online

Como medida do desempenho desse algoritmo de recomendação de usuários para usuários, foram empregadas avaliações subjetivas submetidas por usuários do protótipo.

De forma invisível ao usuário final é realizado um teste A/B[13] para medir a eficácia dos experimentos. A cada recomendação é escolhido aleatoriamente uma de duas métricas (um produto vetorial normal ou uma implementação com o produto vetorial em base log) e o feedback do usuário é armazenado em conjunto com qual variação foi utilizada para o cálculo das sugestões apresentadas.

A figura 4.20 mostra os resultados de uma recomendação conforme apresentados pelo protótipo. Note na parte inferior da interface uma escala de 1 a 5 onde os usuários podem opinar sobre a relevância das recomendações, sendo 5 a nota representando satisfação máxima.

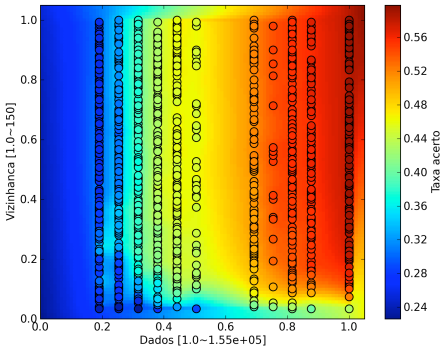
Figura 4.20: Exemplo de tela do protótipo implementado para recomendação de novos contatos para usuários

[13]método de teste online onde uma amostra base de controle é comparada com uma variedade de outras amostras onde há a variação apenas uma variável. Mais detalhes em [Montgomery, 2004]

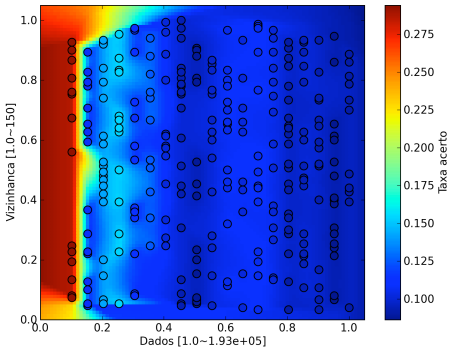
O protótipo encontra-se em operação há 10 meses, tendo recebido nesse período cerca de 320 ratings para recomendações realizadas. As recomendações calculadas usando o produto vetorial normal possuem uma rating média de 3,1 (escala de 1 a 5) enquanto as com o produto vetorial em base log possuem média de 3,5.

4.8.3 Experimento com filtragem colaborativa

Os experimentos realizados para a construção desse recomendador utilizaram os mesmos algoritmos (para recomendação e avaliação) descritos na seção 4.5 (*Top-N*), substituindo apenas fotos e grupos por usuários. Os resultados são mostrados na figura 4.21(a) e os resultados de desempenho para uma abordagem comparativa naïve encontram-se na figura 4.21(b) onde o recomendador naïve retorna sempre os n usuários com maior número de contatos (entre todos os usuários conhecidos), independente do usuário alvo da recomendação.



(a) *Top-N* (filtragem colaborativa baseada na co-ocorrência de usuários como contatos de outros usuários)



(b) comparativo naïve (usuários com mais contatos)

Figura 4.21: Desempenho da recomendação de usuários para usuários usando algoritmo de filtragem colaborativa *Top–N* e abordagem *naïve* comparativa

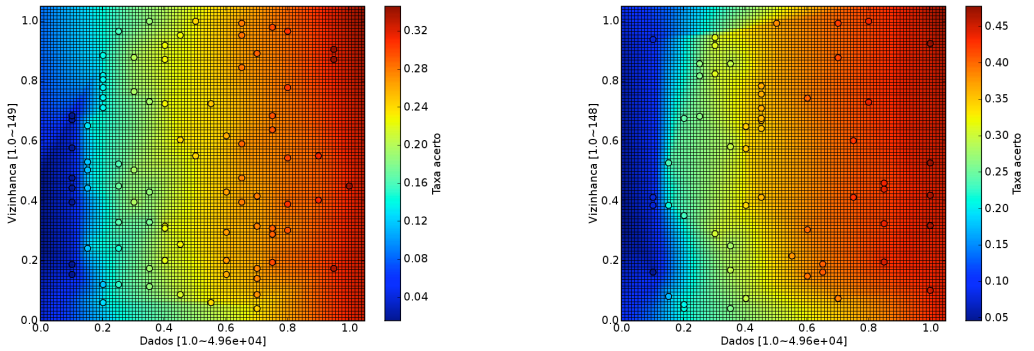
Nestes experimentos podemos observar que o desempenho do recomendador genérico *Top-N* pode ser quase duas vezes superior à abordagem *naïve*: 56% (*Top-N*) versus 27,5% (*naïve*).

4.9 Recomendação de grupos para usuários

Os experimentos realizados para a construção desse recomendador utilizaram os mesmos algoritmos (para recomendação e avaliação) descritos na seção 4.5 (*Top-N*), substituindo apenas fotos por usuários e os resultados são mostrados na figura 4.22. Os resultados de desempenho para uma aborda-

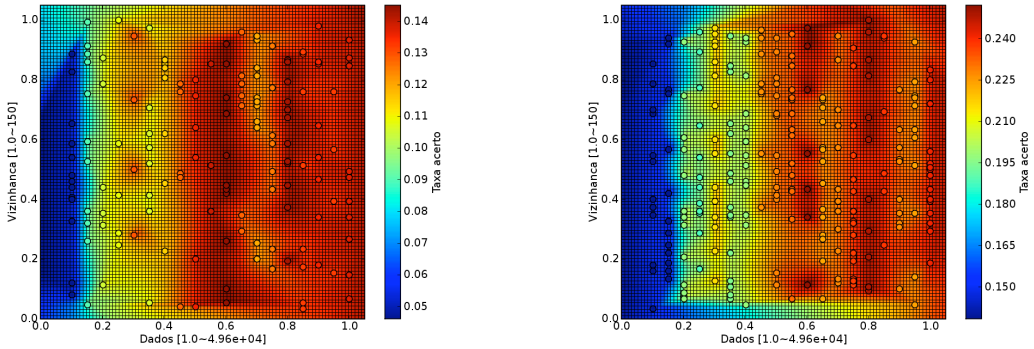
gem base comparativa encontram-se na figura 4.23 onde o recomendador base retorna os n grupos mais populares (entre todos os usuários conhecidos), independente do usuário alvo da recomendação.

Nestes experimentos podemos observar que o desempenho do recomendador genérico *Top-N* pode ser quase duas vezes superior à abordagem *naïve* em ambos os ranks de precisão (@12 e @24): 32% e 45% (*Top-N*) versus 14% e 24% (*naïve*).



(a) 12 recomendações (b) 24 recomendações

Figura 4.22: Desempenho da recomendação de grupos para usuários usando algoritmo *Top–N* (filtragem colaborativa baseada na co-ocorrência de usuários em grupos)



(a) 12 recomendações (b) 24 recomendações

Figura 4.23: Desempenho da recomendação de grupos para usuários (comparativo: grupos com maior número de usuários)

4.10 Resumo dos experimentos realizados

A tabela 4.3 abaixo apresenta um resumo dos experimentos com recomendadores para a rede *Flickr* descritos nesse capítulo, destacando em negrito os resultados mais interessantes.

Tarefa	Algoritmo	Rank	Precisão
Grupo para foto	*Top–N*	6	0,2
		12	**0,2**
	Grupos populares	**12**	**0,054**
		24	0,085
	Conteúdo visual	**12**	**0,14**
		40	0,30
Grupo para usuário	*Top–N*	12	0,32
		24	**0,45**
	Grupos populares	12	0,14
		24	**0,24**
Usuário para usuário	*Top–N*	24	0,56
	usuários populares	24	0,275
Foto para usuário	*Top–N*	24	0,13
	Populares entre contatos	24	0,08
	Fotos populares	24	0,054

Tabela 4.3: Resumo dos experimentos realizados com recomendadores para a rede *Flickr*