PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

RECOMENDAÇÃO DE TAGS PARA MÍDIA SOCIAL COLABORATIVA: DA GENERALIZAÇÃO À PERSONALIZAÇÃO

ANGELINA DE CARVALHO A. ZIESEMER

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientador: Prof. João Batista Souza de Oliveira

Porto Alegre
2012

Z67r Ziesemer, Angelina de Carvalho A.
Recomendação de TAGS para mídia social colaborativa : da generalização à personalização / Angelina de Carvalho A. Ziesemer. – Porto Alegre, 2012.
106 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Prof. Dr. João Batista Souza de Oliveira.

1. Informática. 2. Sistemas de Recuperação da Informação. 3. Redes Sociais. I. Oliveira, João Batista Souza de. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "Recomendação de Tags para Mídia Social Colaborativa: da Generalização à Personalização", apresentada por Angelina de Carvalho Alvarez Ziesemer como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Sistemas Interativos e Visualização, aprovada em 20/03/2012 pela Comissão Examinadora:

Prof. Dr. João Batista Souza de Oliveira - Orientador — PPGCC/PUCRS

Prof. Dr. Ricardo Melo Bastos - PPGCC/PUCRS

Prof. Dr. Marcelo Blois Ribeiro - GE-Brasil

Prof. Dr. Leandro Krug Wives — UFRGS

Homologada em 22/05/2012, conforme Ata No. 011 pela Comissão Coordenadora.

Prof. Dr. Paulo Henrique Lemelle Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

"A menos que você tente fazer algo além do que você já domina,
você nunca crescerá."
Ralph Waldo Emerson

AGRADECIMENTOS

Agradeço primeiramente a Deus ao qual devo tudo que sou.

Ao meu orientador João Batista que sempre demonstrou disponibilidade, paciência, amizade e ao qual levo como inspiração e exemplo para o meu futuro.

À CAPES pelo fomento.

À professora Isabel e aos colegas de laboratório Piccoli, Nicole, Aline, Alexandre, Chamum, Francine, Felipe, Bruney, Jimmy e Anderson do projeto MCAL/HP com os quais convivi durante estes dois anos em momentos de trabalho e diversão.

Ao meu marido Adriel pelo exemplo que é para mim, por sempre me incentivar e apoiar em todas as minhas decisões, pela paciência e compreensão nos momentos em que estive ausente.

E por fim à minha Avó, que me ensinou com amor valores sem os quais não chegaria até aqui.

RECOMENDAÇÃO DE TAGS PARA MÍDIA SOCIAL COLABORATIVA: DA GENERALIZAÇÃO À PERSONALIZAÇÃO

RESUMO

Sistemas de mídia social como Flickr, Youtube e Picasa tornaram-se muito populares devido ao seu ambiente para compartilhamento de imagens, vídeos e suporte à atribuição de tags, avaliações e comentários. Sistemas colaborativos possuem grandes quantidades de conteúdo provido pelos usuários, os quais fornecem informações relevantes para *engines* de recomendação. O uso de *tags* também permite a *clusterização* e busca de conteúdo baseado em palavras-chaves. Neste trabalho foi investigado um mecanismo para recomendar tags, desenvolvendo medidas de co-ocorrência, popularidade e relevância de tags comumente usadas em itens similares e por usuários similares. Foi desenvolvido um sistema para recomendar possíveis tags relevantes baseadas na similaridade contextual de outras tags providas pelos usuários. Para o desenvolvimento do experimento, foi utilizado um *dataset* do *Flickr* para gerar recomendações e analisar o comportamento do algoritmo e as atribuições efetuadas pelos usuários participantes. Os resultados obtidos demonstraram padrões de atribuição e desempenho de acordo com o conteúdo/contexto da imagem. Utilizando a frequência de atribuição baseada no histórico de cada perfil é sugerido um novo modelo personalizado para recomendação de tags.

Palavras-chave: Folksonomia, Recomendação, Tags, Recuperação de Informação

TAG RECOMMENDATION FOR COLLABORATIVE SOCIAL MEDIA: FROM THE GENERALIZATION TO PERSONALIZATION

ABSTRACT

Social media systems such as Flickr, Youtube and Picasa have become very popular as they provide a collaborative environment to share photos and videos supporting tags, ratings and comments. This kind of interaction includes a lot of content provided by users, which may bring meaningful information to recommendation systems. The aggregation of tags is also a way to cluster items and provide tag-based search content. We investigate how to support tag recommendation by ranking the co-occurrence, popularity and relevance of commonly-used tags in similar items and by similar users. We developed a tag recommendation system to recommend of possibly relevant tags. We use Flickr's dataset to analyze our algorithm's behavior and present the results provide by the experiment. A new model using personalized recommendation was developed using the experiment results and the behavior of each user.

Keywords: Folksonomy, Recommendation, Tags, Information Retrieval.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

FC	*Filtragem Colaborativa*
BC	*Abordagem Baseada em Conteúdo*
TF-IDF	*Term Frequency/Inverse Document Frequency*
URL	*Uniform Resource Locator, Localizador-Padrão de Recursos*
i-users	*Usuários Influentes*
AJAX	*Asynchronous Javascript and XML*
PHP	*PHP: Hypertext Preprocessor, originalmente Personal Home Page*
HTML	*HyperText Markup Language*
CSS	*Cascading Style Sheets*
API	*Application Programming Interface*

SUMÁRIO

1. INTRODUÇÃO

Todos os dias e de forma natural as pessoas recebem e dão recomendações umas às outras através de conversas, e-mails, notícias, *reviews* de produtos, etc [SK09]. Com a popularização da internet e o advento da Web 2.0 o espaço de interação dos usuários aumentou, proporcionando uma diversidade de informações a serem exploradas por abordagens de recomendação e filtragem de conteúdo. Dentre estas, destaca-se a atribuição de tags em imagens, vídeos e textos, que servem para categorização e auxílio às *engines* de busca [LdGS+09, SGMB08].

A utilização de tags provê liberdade para usuários efetuarem a classificação de seus conteúdos e adicionarem significado aos documentos/fotos/vídeos/textos compartilhados. Sistemas colaborativos de atribuição de tags desenvolvidos para ambiente Web permitem a atribuição de palavras-chave para itens de forma arbitrária [JEHS09]. Geralmente a categorização de assuntos ou objetos é feita por especialistas, entretanto no ambiente Web isto também é feito por usuários comuns, e neste caso, chama-se *folksonomia* (do inglês *folks* (pessoas) + taxonomia). Dentre os ambientes mais populares da Web 2.0 que utilizam sistemas de atribuição de tags estão as redes de mídia social como o *Flickr*[1], *Picasa*[2] e *Youtube*[3]. No *Flickr* por exemplo, a estrutura de cada nova postagem é formada pelo item (foto/vídeo), a identificação do usuário e um conjunto de tags [LM10]. É possível que mais de uma tag seja atribuída para cada postagem, portanto o conjunto de tags acaba servindo de indicação para o significado do conteúdo ou para o contexto, como observa-se na Figura 1.1.

A atribuição de tags permite a organização e clusterização do conteúdo pelo significado das tags atribuídas, melhorando resultados de busca em ambientes de mídia social e permitindo uma livre categorização efetuada pelo usuário.

A atribuição de tags também se popularizou em ambientes de publicação de conteúdo textual. O Blogger[4], serviço para postagem de conteúdo, utiliza o histórico de atribuição das tags do usuário (Figura 1.2) para recomendação. Este tipo de atribuição, apesar de personalizada, não diversifica o vocabulário do usuário tornando as atribuições repetitivas e com pouco fundamento no conteúdo que está sendo publicado, pois utiliza somente as primeiras letras da palavra que está sendo digitada para recomendar outras já existentes no histórico de tags.

Já no Twitter[5], micro-blogging para publicação de mensagens curtas, as *hash tags* (Figura 1.3) são palavras-chave publicadas em conjunto com o texto e pré-fixadas por um sustenido (#). Esta prática permite a clusterização dos tópicos mais comentados no momento e também a busca pelo assunto baseada nesta atribuição.

---

[1] http://www.flickr.com
[2] https://picasaweb.google.com
[3] http://www.youtube.com
[4] http://blogger.com
[5] http://twitter.com

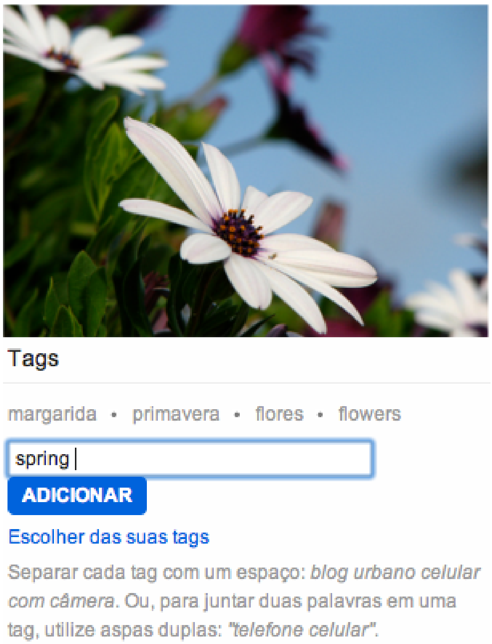

Figura 1.1: Sistema de *folksonomia* do *Flickr*, conjunto de tags sendo atribuídas para uma imagem.

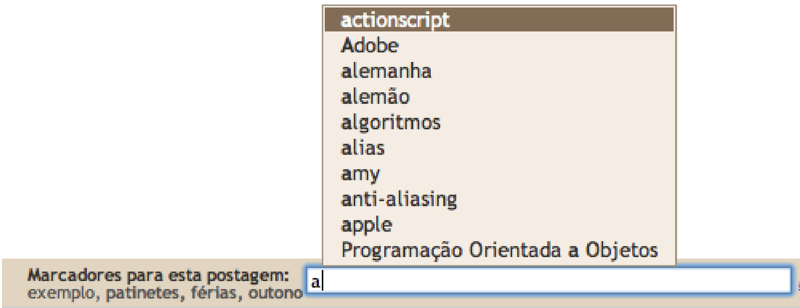

Figura 1.2: Recomendação baseada no histórico de atribuição de tags de um usuário em suas postagens.

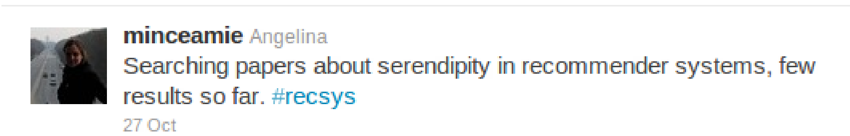

Figura 1.3: Exemplo de postagem do Twitter utilizando a hash tag #recsys.

Apesar das vantagens da *folksonomia*, a atribuição de tags é uma tarefa repetitiva, tediosa e são esses os motivos para os usuários não atribuirem tags ao seu conteúdo. Além disso, erros tipográficos que ocorrem durante o processo de atribuição podem prejudicar os resultados de busca.

1.1 Motivação

A atribuição repetitiva de tags afeta a quantidade e a qualidade das tags atribuídas, pois exige que o usuário dispense tempo e discernimento para escolher o conjunto adequado de palavras-chave para o objeto que está sendo publicado. A atribuição de tags em redes de mídia social é uma das principais abordagens para auxílio a pesquisas ou *queries* baseadas em palavras-chave. Prejudicada a qualidade da atribuição, informações ficarão ocultas das buscas ou resultados irrelevantes podem ser apresentados.

Em buscas efetuadas no *Flickr*, observou-se usuários que utilizam conjuntos de tags, frases, informações pessoais e repetitivas para suas imagens. Por exemplo, a Figura 1.4 mostra o resultado de uma busca no *Flickr* utilizando a palavra-chave "dogs". Ao observar as imagens deste usuário

Figura 1.4: Imagens resultantes de uma busca de imagens por tags no *Flickr* utilizando a palavra-chave "dogs".

do *Flickr* e suas tags, percebe-se que o mesmo costuma atribuir um conjunto definido/repetitivo de tags para grande parte de suas postagens. Este comportamento se assemelha ao relatado no trabalho de Strohmaier et. al [SKK10], onde é descrita a existência de motivações diferentes entre os usuários para o modo como utilizam as tags. Foram observados dois tipos de usuários, chamados categorizadores e descritores. Usuários categorizadores utilizam as tags para ajuda navegacional própria ou para outros usuários específicos, como integrantes de grupos, enquanto usuários descritores utilizam as tags para que seus conteúdos sejam apresentados nos resultados de buscas da comunidade em geral. Pelo ponto de vista da recuperação de informação [SM86], as tags atribuídas pelos usuários descritores são mais úteis já que estas descrevem o conteúdo do objeto e geralmente utilizam sinônimos para descrição do mesmo conteúdo.

Independente do perfil de atribuição do usuários a atribuição de tags é essencial para garantir a qualidade de resultados de busca por itens em uma rede de mídia social. Por ser uma tarefa tediosa para o usuário acaba sendo ignorada, deixando de fora resultados relevantes por conta da não categorização. Nestes casos a recomendação de tags pode promover a *folksonomia* já que conta com o histórico de atribuição de uma comunidade de usuários.

1.2 Objetivos

Sistemas para recomendação de tags analisam tags associadas pelos usuários aos itens para verificar sua relevância para um novo post/conteúdo. A combinação entre sistemas de *folksonomia* e abordagens de recomendação tem o objetivo de facilitar a tarefa de atribuição de tags a cada nova publicação de conteúdo, já que isto requer paciência e tempo. A *folksonomia* de imagens em especial exige mais trabalho do usuário pela quantidade de imagens que geralmente são postadas e que de alguma forma diferem uma da outra. Devido ao rápido crescimento da quantidade de conteúdo gerenciado em redes de mídia social, a classificação de conteúdo feita por especialistas se tornaria uma tarefa impossível pela quantidade de objetos publicados todos os dias, e para o usuário a atribuição de tags é uma tarefa trabalhosa, o que garante a ascensão das *engines* para recomendação de novas tags.

Motivados pela importância da *folksonomia* e pelas possibilidades que ela proporciona, o objetivo deste trabalho é o desenvolvimento de medidas para promover melhores resultados para recomendação de tags através de uma *engine* para redes de mídia social utilizando o *dataset* de tags do *Flickr* como estudo de caso. Para cada tag t digitada para uma imagem r serão recomendadas outras tags similares t_j utilizando filtragem colaborativa como abordagem para recomendação juntamente com as medidas desenvolvidas para recomendar tags mais relevantes baseadas em um ranking de tags co-ocorrentes.

A recomendação pode promover a serendipidade e um vocabulário comum nas redes de mídia social, melhorando os resultados das buscas no ambiente. Além disso, a recomendação de tags aumenta a qualidade das tags atribuídas através da homogenização do vocabulário fornecido pelos próprios integrantes de comunidades colaborativas *online*. Por exemplo, quando um usuário está atribuindo tags a uma imagem de um determinado local, ele pode obter recomendações de novas tags baseadas nas atribuições de outros usuários que também já fizeram atribuições similares aos seus itens.

1.2.1 Metodologia de estudo e pesquisa

Para alcançar os objetivos propostos neste trabalho foram executados alguns passos essenciais para entender e mesclar as áreas de *folksonomia* e recomendação de conteúdo. Primeiramente foi feito um levantamento bibliográfico da área de recomendação e *folksonomia* para identificar as pesquisas existentes. O estado da arte da área de recomendação proporcionou a capacidade de identificar quais as abordagens existentes para recomendação de tags e quais os desfechos encontrados em cada pesquisa.

Definida a abordagem de recomendação a ser utilizada, foi escolhida a utilização do *dataset* do *Flickr*, como fonte de tags para recomendação. Ao analisar a co-ocorrência das tags foram observadas necessidades específicas de desenvolvimento de medidas para refinamento da recomendação.

Após o desenvolvimento das medidas foi executado um experimento para a observação da aceitação da recomendação, as padronizações nas escolhas das tags, e a comparação entre o vocabulário resultante da recomendação em relação às tags digitadas. A análise dos dados possibilitou o desenvolvimento de um modelo de recomendação personalizado baseado na influência das medidas das tags atribuídas em relação a cada usuário.

Os resultados do experimento demonstraram que é possível identificar as tags irrelevantes e menos populares no *dataset* através das medidas desenvolvidas neste trabalho. A utilização das medidas permitiu um reposicionamento das tags candidatas proporcionando assim melhores resultados no vocabulário resultante e um maior número de tags aceitas pela recomendação do que tags digitadas pelos usuários. Este comportamento resultou em um vocabulário mais homogênio reduzindo a ocorrência de tags únicas no *dataset*. Através dos resultados do experimento também foi possível modelar uma nova abordagem para recomendação, utilizando medidas personalizadas para cada usuário.

1.3 Organização

Este trabalho está dividido de acordo com as etapas vivenciadas para o desenvolvimento, aplicação, experimento e análise da *engine* de recomendação que aqui será apresentada. Neste primeiro capítulo foram identificados os desafios e perspectivas que a *folksonomia* proporciona e a motivação para o uso de abordagens de recomendação nestes ambientes.

No próximo capítulo serão apresentados os conceitos e aplicabilidades das áreas que serão abordadas. O conteúdo apresentado irá delimitar a abrangência e utilização dos conceitos e onde/porquê eles foram aplicados.

No terceiro capítulo serão discutidos os trabalhos relacionados da área de *folksonomia* e recomendação e a comparação com a abordagem que será desenvolvida.

No capítulo quatro será apresentado o modelo e o desenvolvimento das medidas para recomendação de tags e logo em seguida no capítulo cinco o *dataset* utilizado para verificação no ambiente de experimento que está detalhado no capítulo seis.

Os resultados do experimento serão apresentados no capítulo sete e com base nestes um novo modelo descrito no capítulo oito.

Por fim, a conclusão e os trabalhos futuros serão abordados no último capítulo.

2. CONCEITOS INICIAIS

Neste capítulo será feita uma revisão da área de recomendação, *folksonomia* e as abordagens utilizadas para recuperação de informação e *feedback* dos usuários. As abordagens discutidas foram utilizadas para o desenvolvimento do modelo de recomendação de tags e da *engine* para interação.

2.1 Sistemas de *Folksonomia*

Sistemas de atribuição de tags tornaram-se populares principalmente em redes de mídia social onde os conteúdos são fotos/vídeos e a busca baseada em conteúdo é difícil de ser aplicada. Nestes ambientes também foram proporcionados recursos para interação e gerenciamento de conteúdo por parte dos usuários. Dentre estes encontramos informações textuais como *tags* e *reviews* [SS09, GSRM09, LXL+09, LdGS+09, SGMB08] explicitamente atribuídas para categorização e avaliação.

A atribuição de tags ao conteúdo publicado permite a busca baseada em palavras-chave dadas pelos usuários. Este tipo de categorização feita por usuários é conhecida como *folksonomia* e está relacionada à taxonomia efetuada por pessoas não especialistas.

Em sistemas como o *Flickr*, devido à grande quantidade de fotos enviadas diariamente, tornou-se conveniente a atribuição de significado ao conteúdo pelos usuários, já que a classificação de todas as imagens por especialistas seria uma tarefa impossível. Desta forma a colaboração dos usuários para a categorização e avaliação do que está sendo publicado tornou-se uma ferramenta importante na filtragem e entrega de resultados de buscas.

Em sistemas de *folksonomia*, a clusterização das tags pode representar aspectos visuais dos conteúdos mas também pode apresentar aspectos pessoais no conjunto de tags de cada usuário, como nomes próprios, datas etc. Este tipo de atribuição cria uma grande quantidade de tags únicas na base de dados que não são relevantes para os usuários da comunidade em geral. De acordo com Kennedy [KCK06], somente 50% das tags providas pelos usuários estão verdadeiramente de acordo com o conteúdo. Entretanto, sistemas de atribuição de tags são ferramentas poderosas para categorização de conteúdo e podem ser melhoradas quando desenvolvidas em conjunto com sistemas de recomendação para facilitar o seu uso e aproveitamento.

2.2 Sistemas de Recomendação

A quantidade de informação disponibilizada na Web tornou necessária a elaboração de filtragem de conteúdo de forma personalizada. Sistemas de Recomendação usam técnicas para entrega de conteúdo relevante de grandes *bases de dados* para os usuários *online* e recentemente tornaram-se muito populares [OLL08, LDP10, LSY03, LMWdO10, DK04]. No contexto da Web, a recomendação implica na entrega dinâmica de conteúdo como elementos textuais, *links*, anúncios, vídeos, imagens e recomendação de produtos, que são adaptados às necessidades ou interesses de um usuário ou um segmento de usuários [Mob07]. Existem diferentes abordagens sendo desenvolvidas para recomendação de conteúdo, em sua maioria baseadas em conhecimento coletivo, perfil de usuários, relacionamentos e regras de ambientes.

A estrutura básica de um sistema de recomendação consiste em identificar como os relacionamentos entre usuários e itens se assemelham e dão efeito à recomendação. Um conjunto de usuários $U = \{u_1, u_2, \ldots, u_m\}$ pode se relacionar com um conjunto de itens $I = \{i_1, i_2, \ldots, i_n\}$ e gerar relação entre as partes ($R \subseteq U \times I$) a cada ação efetuada, por exemplo, em um determinado item visualizado ou clicado pelos usuários.

Para obter a relevância de cada assunto, produto ou conteúdo, é necessário existir uma interface que permita analisar o comportamento do usuário ou obter *feedback* de suas preferências. Esta área está relacionada à forma como essas informações são obtidas, podendo ser extraídas de forma implícita [OLL08, Nic98, XJL08] ou explícita [JSK10, LXL$^+$09, DK04, APTO09, SLH09], descritas a seguir.

Abordagem implícita: é a técnica desenvolvida para monitoração automática das ações do usuário com o sistema. A principal vantagem desta técnica é que dispensa o trabalho do usuário de prover uma avaliação ou *feedback* em relação ao conteúdo acessado [KT03]. No site de compras *Amazon*[1], por exemplo, de acordo com o histórico de busca e acessos a uma ou mais categorias de produtos o sistema infere que se o usuário está procurando ou clicando em determinados itens, outros itens da mesma categoria também podem ser relevantes como sugestão de compra.

Abordagem explícita: pesquisas mostram que apesar dos bons resultados apresentados pela abordagem implícita, a técnica explícita é mais assertiva na criação do perfil para recomendação [JSK10, KT03, KSK97]. Isto ocorre porque nesta abordagem é o usuário quem informa a relevância dos assuntos ou conteúdos acessados. Neste tipo de abordagem geralmente o usuário possui um vínculo com o sistema de recomendação, ou seja, pode ser necessário um cadastro de perfil para acesso à área personalizada. No *Youtube*, site para compartilhar e assistir vídeos *online*, é possível ao usuário logado avaliar o nível de relevância do vídeo e também adicioná-lo à sua lista de favoritos. Esta abordagem explícita supõe que, quando a avaliação do vídeo é relevante, o usuário também demonstrará interesse por outros vídeos da mesma categoria ou conteúdo.

Definida a forma como será obtido o *feedback* do usuário, também devem ser escolhidas as estratégias para obter conteúdo relevante para a recomendação. Na seção a seguir serão apresentadas as subdivisões existentes em um sistema de recomendação.

2.2.1 Filtragem Colaborativa (*Collaborative Filtering*)

Sistemas de recomendação que utilizam Filtragem Colaborativa (FC) [SK09] supõem que se dois usuários avaliam itens de forma similar ou acessam itens de mesma categoria, também irão avaliar ou acessar outros itens similarmente. Esta associação tem como objetivo relacionar os interesses entre os usuários e sugerir aos grupos de usuários gerados pelas classificações outros conteúdos que possam

[1] http://www.amazon.com

ser interessantes. Sistemas FC são divididos em algoritmos baseados em memória (*memory-based*) que utilizam técnicas para identificar usuários com comportamentos similares e algoritmos baseados em modelos (*model-based*) que utilizam coleções de avaliações para aprender um modelo de perfil para receber recomendação de conteúdo. Neste caso, os algoritmos baseados em memória podem ser adaptados para a análise da aceitação das *tags* mais frequentemente utilizadas em conjunto pela comunidade de uma rede de mídia social, por exemplo. A FC para recomendação de tags analisa os metadados associados aos conteúdos pelos usuários para inferir a relevância das tags para conteúdos específicos [DFT10]. Neste trabalho o objetivo central é encontrar a similaridade entre tags para recomendar outros conjuntos de tags similares.

Algoritmos baseados em memória

Algoritmos baseadas em memória fazem recomendações através das similaridades entre usuários [SLH09] ou entre itens [LSY03, DK04], levando em consideração as avaliações passadas e relacionando usuários e itens para a criação de grupos classificados pelos seus interesses em comum.

O valor de avaliação $a_{u,i}$ é dado pela relação entre o usuário u e o item i, sendo agregado a outras avaliações feitas por outros usuários para o mesmo item. Nesse contexto um conjunto de usuários que possui os mesmos interesses passará a se relacionar como vizinhança próxima por possuir interesses semelhantes e efetuarem avaliações similares dos mesmos itens [AT05].

Algoritmos baseados em modelos

Algoritmos baseados em modelos são desenvolvidos utilizando cálculo de previsão de utilidade baseados em um modelo de dados com uso de técnicas de estatística e aprendizagem de máquina, isto é o que o difere da filtragem baseada em conteúdo, que será apresentada na próxima seção [AT05].

Sistemas colaborativos para recomendação que utilizam algoritmos baseados em modelo usam avaliações passadas salvas na base de dados do usuário para prever futuras atribuições de avaliações a conteúdos ainda não acessados [DDGR07]. Diferente do algoritmo baseado em memória, nesta abordagem o sistema precisa aprender as preferências de apenas um usuário por vez para poder definir o perfil de acordo com suas avaliações e assim recomendar conteúdo.

Breese [BHK98] desenvolveu dois modelos probabilísticos para filtros colaborativos baseados em modelos. O primeiro utiliza clusterização, onde os usuários são agrupados em classes e suas avaliações são independentes, e o segundo modelo utiliza redes de Bayes, onde o estado de cada nodo corresponde a possíveis valores de avaliação para cada item. Como modelos mais recentes, existem o processo de decisão de Markov [SHB05], *Latent Dirichlet Allocation* [Hof04] e PLSI (*probabilistic latent semantic indexing*) [Hof99].

A recomendação baseada em modelos é basicamente voltada para a personalização do conteúdo a ser entregue, pois antes da recomendação é feita uma análise do comportamento do usuário para aprender o modelo do seu perfil.

2.2.2 Filtragem Baseada em Conteúdo (*Content-based Filtering*)

Em geral, um sistema de recomendação baseado em conteúdo (BC) [PB07] analisa documentos avaliados individualmente por um usuário e utiliza o conteúdo destes documentos e a avaliação recebida para inferir um perfil que pode ser usado para recomendar itens relevantes [SLH09]. Sistemas baseados em filtragem de conteúdo foram desenvolvidos principalmente para recomendações baseadas em itens textuais, pois o conteúdo nesses sistemas é usualmente descrito com palavras-chave [AT05].

TF-IDF *(term frequency/inverse document frequency)* ou frequência de termos/frequência inversa de documentos é uma das abordagens mais conhecidas para recuperação de informação. TF é frequência normalizada de vezes que a palavra-chave i aparece em um documento d_j e IDF é a frequência inversa, medida de importância de i representada pelo logaritmo da divisão do número total de i pelo número de documentos que contêm i. Então, o peso p de TF-IDF para a palavra-chave em questão é dado por

$$p_{i,j} = TF_{i,j} \times IDF_i$$

Independente do tipo de abordagem, o que caracteriza um sistema de recomendação baseado no conteúdo é a relação estreita que um usuário possui com a categoria de itens que acessa.

2.2.3 Sistemas Híbridos

Sistemas híbridos para recomendação dizem respeito a diferentes técnicas empregadas em um mesmo sistema para a recomendação de conteúdo.

A iniciativa de utilizar sistemas de recomendação em conjunto partiu principalmente do esforço para tentar superar problemas conhecidos na área de recomendação de conteúdo. Os principais problemas estão relacionados aos "novos usuários" e "novos itens", pois novos itens não possuem avaliação e portanto não são recomendados para os usuários, por outro lado, novos usuários em um sistema baseado em conteúdo não possuem interação suficiente para a criação de perfil para conteúdo personalizado.

No trabalho de Burke [Bur07], os sistemas híbridos foram classificados da seguinte forma:

Média ponderada: sistemas de recomendação que mesclam resultados de diferentes técnicas. As avaliações de diferentes técnicas de recomendação são combinadas numericamente.

Switching: sistemas de recomendações diferentes são utilizados de acordo com critérios para recomendação;

Mixed: recomendações de diferentes sistemas são apresentadas juntas;

Cascata: os resultados de um sistema de recomendação são refinados por outro sistema de recomendação;

Combinação de resultados: características derivadas de diferentes fontes de conhecimentos são combinadas para serem executadas em um único algoritmo para recomendação;

Acréscimo de características: um sistema de recomendação é utilizado para computar um resultado ou conjunto de resultados e torná-los valores de entrada para um outro sistema de recomendação;

Fab [BS97] foi desenvolvido utilizando a combinação de técnicas FC e BC. Neste sistema o perfil encontrado pelo algoritmo baseado em conteúdo é utilizado para calcular a similaridade entre dois usuários através de algoritmos de filtragem colaborativa.

No trabalho de Spiegel [SKL09], avaliações e informações de conteúdo são utilizadas em um modelo unificado para a redução de parâmetros e recomendações mais efetivas. Em [LDP10] o sistema de recomendação de notícias utilizado para o *Google News*[2] que previamente utilizava FC passou a combinar juntamente o sistema BC para prever os interesses dos usuários utilizando um *framework* Bayesiano.

2.3 Recuperação de Informações

A recomendação de conteúdo está condicionada principalmente ao *feedback* sobre atividades do usuário. Sistemas de personalização geralmente diferem no tipo de dados e métodos utilizados para criar perfis e no tipo de abordagem usada para tomar as decisões [Mob07].

Inicialmente, um sistema de recomendação conseguia extrair conteúdo das atividades do usuário principalmente pelos *clicks* efetuados e pelas *queries* de pesquisa. Porém, com o advento da Web 2.0 o espaço de interação do usuário com o sistema aumentou, e outras fontes de obtenção das interações para a recomendação de conteúdos começaram a ser exploradas para gerar recomendações. Recentemente começou a ser explorado o uso das tags e das avaliações de conteúdo efetuadas explicitamente. Para extração destas informações diferentes estratégias podem ser aplicadas, dependendo da abordagem escolhida para efetuar a personalização do conteúdo que será recomendado.

Neste trabalho serão utilizadas três formas de obtenção do *feedback* do usuário. A primeira forma são as tags já atribuídas e que são encontradas no *dataset* e servirão para buscar conteúdo similar para atribuição. O outro método de obtenção de dados para recomendação é através das *queries*/tags digitadas pelos usuários e que servirão para recomendar diferentes tags para atribuição. Por fim, a outra forma de obtenção de informações referentes ao comportamento do usuário é através do armazenamento dos *clicks*/seleções das tags recomendadas, levando em consideração a posição da tag no ranking e as medidas que cada uma delas possui para utilização no algoritmo de personalização desenvolvido. Estas três formas de *feedback* serão descritas a seguir.

2.3.1 Tags

Tags são termos livremente escolhidos por usuários de sistemas *online* para classificar um assunto, imagem, vídeo, música ou qualquer conteúdo relacionado a uma categoria ou tema. É a simples atribuição de palavras-chave ao conteúdo que está sendo publicado.

[2] http://news.google.com.br/

Na Web 2.0 a utilização de informação textual explícita como tags e *reviews* está se tornando cada vez mais popular [SS09, GSRM09, LXL+09, LdGS+09, SGMB08]. Em sistemas de recomendação que utilizam FC as tags são usadas para definir os interesses dos usuários. Estes sistemas são baseados na ideia de que usuários com interesses similares compartilham tags semelhantes [DFT10]. Segundo Shepitsen et. al. [SGMB08] a clusterização hieraquizada de tags para recomendar navegação personalizada é eficaz para identificar os interesses dos usuários assim como determinar o tópico de um conteúdo.

A popularidade de sistemas que utilizam tags deve-se à liberdade do usuário em definir a categoria de acordo com suas preferências e a rápida proliferação de redes sociais que utilizam este recurso como o *Flickr*, *Last.fm*, *YouTube*, *Twitter* além de sistemas de *bookmarking* como o *del.icio.us*[3], *digg*[4] etc. Tags podem classificar o conteúdo que caracteriza o documento e atribuições como a relevância do assunto que está sendo publicado, utilizando palavras que dão ênfase à informação, o autor, o tipo de documento publicado, data, região e eventos.

Inicialmente as redes sociais utilizavam as tags para facilitar a navegação e como alternativa de busca, já que são essas informações que estarão relacionadas, por exemplo a uma imagem no *Flickr*. Porém, as tags apresentaram-se em um contexto mais amplo quando comparadas a *links* para pesquisa de conteúdo, através das interações e associações de similaridade entre usuários e itens tendo um recente crescimento dentro da área de recomendação.

2.3.2 *Query*

É comum o uso de *engines* de busca na Web que complementam a pesquisa de resultados com uma lista de buscas relacionadas [SMWH10]. O histórico de pesquisas de um usuário contém um dos *feedbacks* implícitos mais utilizados atualmente por que as palavras digitadas em uma pesquisa geralmente estão relacionadas com o conteúdo de interesse do usuário no momento do acesso.

No site da *Amazon*, o histórico de *queries* é utilizado para sugerir produtos desconhecidos pelo usuário ativo e já vistos por usuários similares. Esta sugestão infere que o usuário que está buscando um item i_1, pode também se interessar pelo produto i_2 já visualizado/avaliado/comprado por sua vizinhança. Da mesma forma, pode-se considerar a atribuição de tags como *queries* para indicar o assunto/contexto/descrição do objeto que esta sendo publicado e logo sugerir outras tags através das abordagens de recomendação.

2.3.3 *Clickstreams*

Clickstreams são os registros das ações de usuários em ambiente Web. Esta técnica de coleta de dados é utilizada para identificar um usuário e o item acessado durante a sua interação com a página ou documento, o que permite manter os registros de identificação e ações [EV03].

[3] http://www.delicious.com
[4] http://www.digg.com

A ideia de armazenar os *clicks* supõe que se o usuário clicar em um *link* de um documento, a informação contida neste *link* é considerada mais interessante do que outra que não foi clicada neste mesmo documento [XJL08]. Isto é importante como análise pois a comunicação com o usuário em um ambiente Web é mais restrita do que a comunicação de forma direta fora do ambiente do sistema. Portanto, toda ação efetuada passa a ser importante já que o acesso a um conteúdo, assunto ou anúncio é feito de forma espontânea, seja por necessidade, interesse ou curiosidade.

No trabalho de Das [DDGR07] foi utilizado o histórico de *clicks* dos usuários para o desenvolvimento de novas recomendações de notícias no *Google News*. Por exemplo, se um usuário costuma efetuar *clicks* em categorias como esporte, o sistema de recomendação irá identificar o grupo de usuários que também costuma acessar este tipo de notícia, e assim definir a similaridade entre o grupo e o usuário ativo para recomendar conteúdo ainda não visitado pelo mesmo. Esta recomendação também ocorre quando uma notícia entra na lista das *Top Stories*, as notícias mais visitadas pelos leitores *online*. Este sistema de recomendação se encaixa no contexto de sistemas FC, que utilizam a avaliação de conteúdo efetuada por grupos de usuários com mesmos interesses para gerar futuras recomendações.

Ao observar o comportamento do usuário pelos *clicks* efetuados, é possível monitorar diretamente os parâmetros passados entre as URL's. Por exemplo, caso uma URL traga em sua estrutura o usuário u e o item i significa que a recomendação do conteúdo se dará a partir dessas informações que desenvolverão o perfil P de u ou de i através da relação que ocorrerá entre eles.

Esta abordagem é importante pois não necessita da avaliação do usuário para identificar seus interesses já que infere que se determinada categoria de conteúdo está sendo mais clicada do que outra, então no momento este tipo de conteúdo é mais relevante.

Na *engine* desenvolvida para este trabalho foi utilizado o armazenamento das tags que foram recomendadas e aceitas pelo usuário para guardar atributos para personalização. Basicamente será guardada a informação relativa ao ranking de tags e se as tags posicionadas no topo do ranking estão sendo mais clicadas/selecionadas/atribuídas do que as que estão na base do ranking.

3. TRABALHOS RELACIONADOS

A inserção de tags permite a classificação de conteúdo por usuários em mídias sociais. Essa liberdade de classificação exige um significativo trabalho por parte do usuário quando existem muitos conteúdos a serem classificados. Para facilitar o processo de atribuição de tags, alguns sistemas recomendadores de tags foram desenvolvidos recentemente e serão descritos a seguir.

AutoTag [Mis06] foi um dos primeiros sistemas colaborativos para recomendação de tags. Foram utilizadas medidas de recuperação de informação para estimar a similaridade entre weblogs de postagens (Figura 3.1).

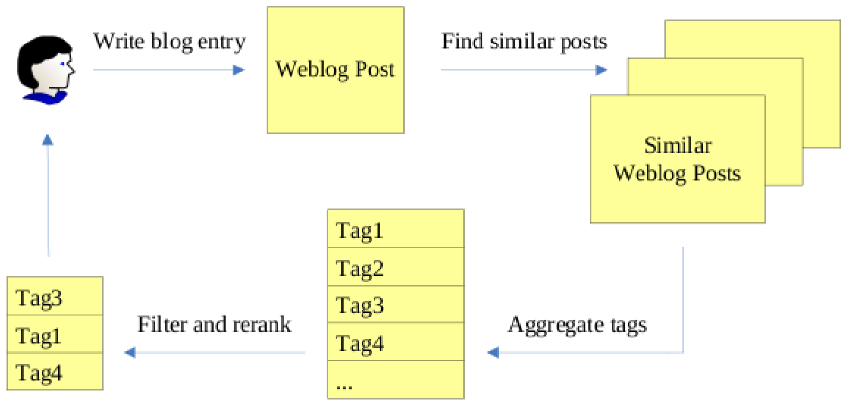

Figura 3.1: Fluxo de informação no ambiente Autotag [Mis06].

Cada tag recebe uma pontuação de acordo com sua frequência e logo passa por uma fase de reordenação onde é verificado se as tags já foram previamente utilizadas pelo usuário ativo, aumentando sua pontuação em caso positivo.

No trabalho de Brooks [BM06] o autor utilizou a abordagem $TF * IDF$ para recomendar tags para postagem em blogs. Os três termos com maiores resultados são sugeridos para atribuição. No trabalho de Zanardi e Capra [ZC08] foi desenvolvida uma abordagem para melhorar o resultado das pesquisas na Web 2.0: o *social ranking* explora a similaridade entre usuários e entre tags. Gemmel [GSRM09] propôs a adaptação do algoritmo de vizinhança *k-NN*, para usuários e tags e assim identificar o grupo ao qual eles pertencem pela similaridade obtida.

Sigurbjornsson et. al. [SvZ08] analizou o comportamento de atribuição de tags no *Flickr*. Foi concluído que a maioria dos usuários atribuem poucas tags a suas fotos e em geral estas são informações de onde/que/quem e quando a foto foi tirada. Sua pesquisa teve uma contribuição substancial para a compreensão do *long tail* [And06] de distribuição das tags. Este trabalho é o que mais se assemelha à abordagem que será apresentada, pois também usa a co-ocorrência para definir um ranking preliminar antes de refinar a recomendação. Os autores utilizaram duas estratégias para obter a co-ocorrência entre as tags: a primeira abordagem utiliza a similaridade baseada na fórmula de Jaccard

$$J(t_i,t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|}$$

onde t_i t_j são as tags para verificação da co-ocorrência.

Já a segunda forma de obtenção da co-ocorrência é chamada de medida assimétrica:

$$P(t_j|t_i) = \frac{|t_i \cap t_j|}{|t_i|}$$

Além disso foram desenvolvidos dois métodos de agregação, sendo um baseado em votos e outro em somatório. O método de agregação baseado em votos não leva em consideração a co-ocorrência das tags definidas pelo usuário enquanto a estratégia de somatório utiliza a co-ocorrência para a recomendação final de tags. Para entrega das tags mais relevantes, os autores desenvolveram uma função de promoção utilizando estratégias chamadas *stability*, *descriptive* e *rank*.

A estabilidade (*stability*) considera que as tags atribuídas pelos usuários com baixa frequência são menos desejáveis que tags com alta frequência. Foram promovidas as tags que são mais estáveis pela medida estatística de estabilidade:

$$stability(u) = \frac{k_s}{k_s + abs(k_s - log(|u|))}$$

A medida calcula o peso do impacto das tags candidatas para uma tag definida pelo usuário, onde $|u|$ é a frequencia da coleção da tag u pertencente a U referente ao conjunto de tags atribuídas pelo usúario à foto e k_s é um parâmetro determinado por treino.

A medida de descrição (*descriptive*) considera que tags com frequências muito altas provavelmente tendem a ser muito genéricas para fotos individuais. Para amortecer o impacto destas tags na recomendação eles desenvolveram a função:

$$descriptive(c) = \frac{k_d}{k_d + abs(k_d - log(|c|))}$$

Onde c pertence a C_u que é a lista das tags mais co-ocorrentes para a tag definida pelo usuário e k_d também é um parâmetro determinado por treino.

A última medida desenvolvida chamada de *rank* é a promoção do ranking que não utiliza a co-ocorrência mas sim a posição r da tag candidata $c \in C_u$

$$rank(u,c) = \frac{k_r}{k_r + (k_r - 1)}$$

Ao final estas três estratégias são computadas para o desenvolvimento da recomendação. A Figura 3.2 apresenta um resumo do processo do sistema de recomendação de tags.

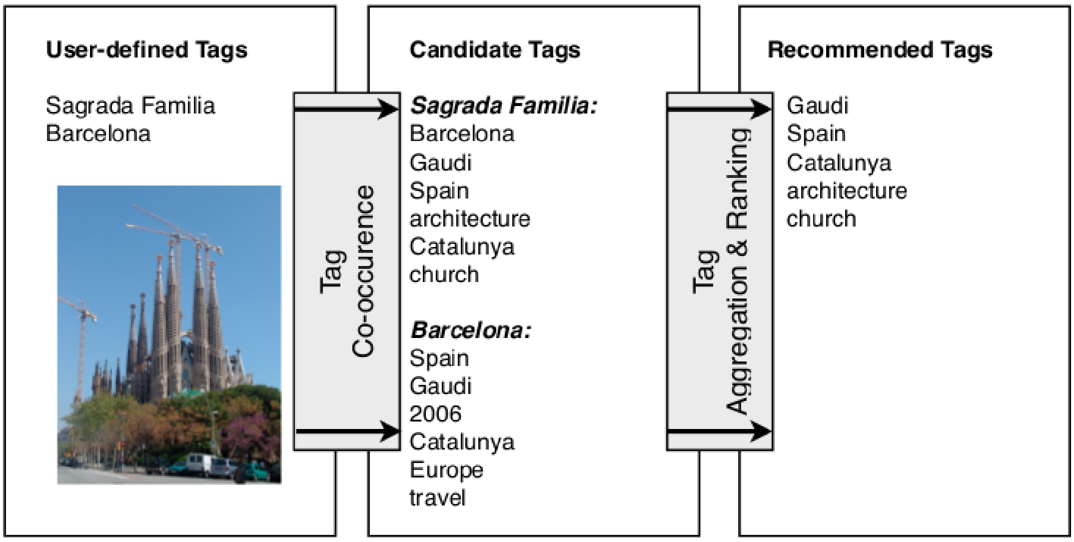

Figura 3.2: Exemplo de resultado e o processo para recomendação do ranking final de [SvZ08].

Já na *engine* desenvolvida para este trabalho foram criadas medidas para obter a popularidade e relevância de tags candidatas a recomendação combinadas com a co-ocorrência entre tags. Cada medida verifica a diversidade das tags no conjunto de utilizadores das tags e nos itens aos quais elas foram associadas. A abordagem desenvolvida efetua o cálculo das medidas para *queries* simples e compostas com o propósito de entregar tags relevantes através da combinação de tags. Com base nos resultados do experimento criado para recomendação generalizada, foi desenvolvido um modelo de recomendação utilizando as mesmas medidas porém de forma personalizada para refletir o perfil do usuário que está classificando seu conteúdo.

4. DESENVOLVIMENTO DO MODELO DE RECOMENDAÇÃO

Este capítulo apresenta o desenvolvimento das medidas do modelo e algoritmo desenvolvidos neste trabalho, além da abordagem para obtenção de tags co-ocorrêntes para criação de um ranking de tags para recomendação.

4.1 Modelo e Algoritmo

O processo de submissão de fotos de um usuário em uma rede de mídia social consiste principalmente em dois passos: seleção de itens e atribuição de tags para o conjunto de itens ou para cada um deles. Provavelmente em uma coleção de imagens existirá conteúdo referente aos lugares, pessoas e datas que poderá ser categorizado individualmente. O objetivo deste trabalho é recomendar conjuntos relevantes de tags durante o processo de categorização de itens para facilitar o uso de sistemas de *folksonomia*.

Geralmente um sistema de *folksonomia* pode ser modelado como um tripla $P_i =< u_i, r_i, T_i >$ para cada postagem, onde $T_i = \{t_1, t_2...t_n\}$ é um conjunto de tags atribuídas ao conteúdo r_i enviado pelo usuário u_i. Por exemplo, um usuário submete uma imagem de Londres ao *Flickr*, e adiciona à imagem tags como "inglaterra", "uk", "londres" e outras, o conjunto de tags será $T = \{inglaterra, uk, londres\}$.

A Figura 4.1 mostra alguns exemplos de fotos resultantes da busca no *Flickr* utilizando a tag *eiffel*. Os resultados mostram imagens diferentes categorizadas por usuários e com conjuntos diferentes de tags. Porém, observa-se que algumas das tags se repetem em diferentes fotos. Como no exemplo apresentado na Figura 4.1, neste trabalho serão utilizadas as associações entre tags-tags, usuário-tags, item-tags para sugerir outras tags similares para os usuários.

O modelo desenvolvido utiliza abordagens de FC, em primeiro momento baseado em memória *(memory-based)* tornando-se personalizado *(model-based)* logo após a análise do comportamento do usuário pelo histórico de atribuição.

Para o modelo de recomendação de tags estabeleceu-se que após o usuário fornecer uma tag t para um item, é definido o conjunto

$$P(t) = \{P_i | t \in T_i\}$$

de todas as triplas que contém t em seu conjunto de tags T_i.

No desenvolvimento da *engine*, será utilizada a co-ocorrência entre tags e serão criadas as medidas de relevância e popularidade. Para tanto, será obtido o ranking preliminar com as k tags mais co-ocorrentes baseadas na tag digitada pelo usuário em tempo real de atribuição. Assim que o usuário digitar uma tag para a imagem, ela será tratada como uma *query* que serve como sugestão para a busca de tags similares, ou seja, usadas em contexto semelhante. Será verificada similaridade

Figura 4.1: Exemplo de algumas fotos resultantes de uma busca no *Flickr* utilizando como tag principal t a palavra *eiffel*.

entre as tags baseada na co-ocorrência tentando indicar o contexto do objeto e não a semântica da palavra.

A seguir serão apresentados o ranking preliminar e as medidas desenvolvidas que foram baseadas nas observações do *dataset* do *Flickr*. Essas medidas retornarão valores que servirão para reorganizar o ranking preliminar de tags que será dado pela co-ocorrência e frequência, de forma que as tags mais relevantes fiquem no topo do ranking. Para a entrega de melhores resultados na recomendação foi gerado um ranking final através da combinação destas medidas.

4.1.1 Ranking Preliminar

De acordo com o modelo criado para fazer a recomendação de tags relevantes é necessário obter as k tags mais co-ocorrentes com t entre os objetos do conjunto $P(t)$ conforme a seguir:

$$exist(t,T) = \begin{cases} 1, t \in T \\ 0, t \notin T \end{cases}$$

A função $exist(t,T)$ irá sinalizar a existencia de t no conjunto T e ordenar a lista das tags co-ocorrêntes utilizando a função:

$$ranking(t,t_j) = \sum_{P_i \in P(t)} exist(t_j, T_i)$$

A função $ranking(t, t_j)$ conta quantas vezes a tag t_j co-ocorre com t. As k tags mais co-ocorrentes serão utilizadas para calcular as próximas medidas.

4.1.2 Medidas

A utilização da frequência de tags para recomendação não apresenta bons resultados para o usuário devido à presença de tags com significados pessoais como nomes próprios, datas e informações relevantes a um pequeno grupo, por exemplo.

Para evitar a recomendação de tags pessoais para a comunidade em geral, os valores obtidos pela co-ocorrência de cada tag foram normalizados e logo desenvolvidas duas medidas para a *engine* de recomendação com o objetivo de refinar o ranking preliminar e melhorar o ranking final.

Normalização da Co-ocorrência

Para a utilização das tags ordenadas pela co-ocorrência e obtenção de um valor normalizado, calcula-se o número de itens que contêm ambas as tags t e t_j e divide-se pelo número de itens que possuem apenas t.

$$coo(t, t_j) = \frac{ranking(t, t_j)}{|P(t)|}$$

O valor de $coo(t, t_j)$ para cada tag t_j irá resultar entre zero e um. Esta medida irá auxiliar no cálculo do ranking final, onde serão aplicadas as medidas que servirão para retirar do topo do ranking as tags menos relevantes, já que por enquanto a principal influência para o posicionamento das tags é a frequência de cada uma. Entretanto, a normalização da co-ocorrência não muda a ordem do ranking preliminar, pois o principal objetivo é a obtenção de um valor normalizado das tags mais frequentes para utilização em conjunto com as próximas medidas.

Relevância

Considerar a relevância da tag t_j tornou-se necessário pois o número de vezes que a mesma ocorre com t no *dataset* não indica que ela seja a melhor opção para atribuição. Por exemplo, existem muitos itens usando o mesmo conjunto de tags, mas atribuídos sempre pelo mesmo usuário. Apesar da elevada frequência e a relevância pessoal para o usuário, este tipo de tag não ilustra o comportamento geral da comunidade da rede de mídia social.

Durante o processo de desenvolvimento das medidas, foi possível observar no *dataset* perfis de usuários influentes (*i-users*) . Em geral, *i-users* adicionam o mesmo conjunto T_i de tags para várias das suas imagens, mesmo quando algumas palavras-chave não têm conexão com os itens.

Frequentemente existem itens que recebem como tag o nome dos usuários como pode-se observar na Figura 4.2. Neste exemplo, é apresentada uma recomendação de tags usando somente a co-ocorrência como medida para sugestão de tags. O usuário ativo com intenção de postar um item r juntamente com a tag principal $t =$"london" logo recebe outras tags t_j que ocorrem com t.

Entretanto, existe um *i-user* que usa a tag "aline" para fazer a categorização de todo o seu conteúdo. Isto não seria um problema quando outros usuários costumam atribuir esta tag também, mas neste caso não existem outros itens com a mesma tag atribuída, tornando "aline" uma tag personalizada.

Apesar do alto valor da co-ocorrência para a tag "aline", em geral tags personalizadas não ajudam a melhorar a recomendação para a comunidade em mídias sociais, e a co-ocorrência baseada nos itens de *i-users* não traz resultados desejáveis para recomendação colaborativa. Este comportamento provoca distorções na lista de tags, pois mostra tags irrelevantes para a comunidade.

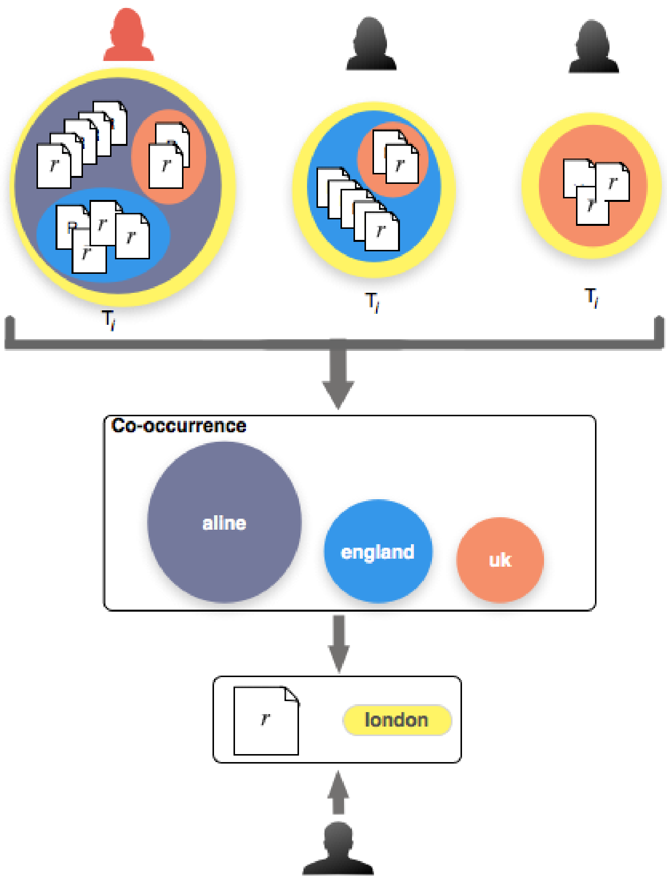

Figura 4.2: Recomendação colaborativa de tags pela medida de co-ocorrência sugerindo a tag "aline" como a tag mais frequente nas triplas que contém a tag principal "london" $P(t) = \{P_i | \text{london} \in T_i\}$

Para melhorar esta recomendação, foi utilizada a frequência baseada em itens e em usuários para calcular a relevância de t_j. A medida de relevância foi modelada pelo número de usuários u que possuem ao menos um item que recebeu as tags t e t_j e dividido pelo número total de itens que possuem t e t_j:

$$rel(t, t_j) = \frac{|users(t) \bigcap users(t_j)|}{ranking(t, t_j)}$$

O valor de $rel(t, t_j)$ apresentará baixos valores para tags menos relevantes e irá colocá-las na base do ranking mesmo se existirem muitos itens com a tag t_j, porém, atribuída por alguns poucos usuários verificados através da função $user(t)$.

Popularidade

A popularidade de uma tag t_j está relacionada à sua frequência de uso pela comunidade em geral. Neste caso, o que importa é a relação com o conjunto de usuários que utilizaram a tag t. O resultado desta avaliação irá medir o quão popular é a tag t_j no conjunto de usuários que usam t em seus itens, ou seja, as tags trazidas pela co-ocorrência para o topo do ranking podem ser ou não populares quando avaliadas em relação à sua frequência de uso pela comunidade.

A popularidade utiliza a relação usuário-tag (Figura 4.3) para identificar os conjuntos de usuários que utilizam as tags t e/ou t_j e obter a interseção desta relação. Através desta observação é possível extrair a medida para a importância das tags para a comunidade em geral.

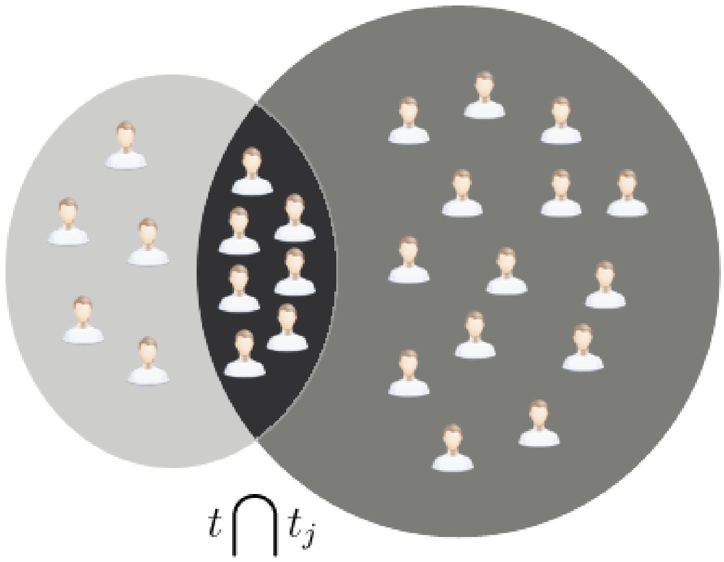

Figura 4.3: Conjuntos de usuários que usam a tag t e/ou a tag t_j para extração de resultados da popularidade.

Para o desenvolvimento desta medida foi utilizado o conceito da probabilidade condicional, no qual é verificada a probabilidade de um segundo evento acontecer depois da ocorrência de um primeiro evento, ou seja, a capacidade de identificar em um espaço amostral qual a probabilidade de A ocorrer dado que B já ocorreu.

Com base neste conceito, foi desenvolvida a medida de popularidade para a *query* t e suas tags candidatas t_j. A medida de popularidade de t_j com respeito a t é dada pelo número de usuários que usam ambas as tags t e t_j dividido pelo número de usuários que usam a tag t:

$$popRank(t, t_j) = \frac{|users(t) \bigcap users(t_j)|}{|users(t)|}$$

o resultado apresentará valores entre zero e um como nos resultados das outras duas medidas.

Para a análise da relação reversa entre as tags, foi calculado o quão popular é a tag t no universo de usuários utilizando t_j como tag principal. Esta ação permite observar a simetria entre as tags do ponto de vista dos usuários, ou seja, verificar se é frequente o uso deste conjunto de tags caso a co-ocorrência fosse obtida a partir de t_j como *query*. Por exemplo, para a tag "bear" existem tags co-ocorrentes como *polar, ice, urso*. A análise reversa irá mostrar a popularidade da tag "bear" para usuários da tag "polar". Ou seja, neste momento a tag "polar" é tratada como t e a tag "bear" como t_j. Se "bear" não é uma tag popular nos itens que usam a tag principal "polar", o valor computado para $popRank(\text{``}polar\text{''}, \text{``}bear\text{''})$ será baixo.

A popularidade combina os valores $popRank(t,t_j)$ e $popRank(t_j,t)$ pela média aritmética:

$$pop(t,t_j) = \frac{popRank(t,t_j) + popRank(t_j,t)}{2}$$

Ranking Final

A co-ocorrência, relevância e popularidade são as medidas usadas no ranking de recomendação de tags. É necessário combinar os resultados obtidos pelas medidas para posicionar as tags mais relevantes no topo do ranking. Foi definido o ranking final de tags pela média geométrica entre as três medidas:

$$mean(t,t_j) = \sqrt[3]{coo(t,t_j) * rel(t,t_j) * pop(t,t_j)}$$

O valor de $mean(t,t_j)$ irá definir a posição de t_j no ranking final de tags associadas a t para recomendação. Quanto mais no topo estiverem posicionadas as tags no ranking final, mais relevantes elas devem ser em relação à tag t.

4.2 Recomendação

Para cada tag t atribuída pelo usuário serão apresentadas outras dez tags na lista de tags recomendadas. Sabe-se que nem sempre todas estas tags serão relevantes para atribuição por isto é importante o reposicionamento das tags através das medidas desenvolvidas afim de tirar do início da recomendação tags menos relevantes.

4.2.1 *Query* simples

Uma *query* simples faz a recomendação baseada em uma única tag. Quando o usuário digitar uma tag t para o objeto r, ele receberá tags t_j associadas a t. A consideração de um caso de sucesso na recomendação acontecerá quando as tags aceitas pelo usuário estiverem posicionadas nas primeiras cinco posições do ranking final gerado por $mean(t,t_j)$.

A *engine* considera as tags digitadas pelo usuário como *queries* de busca para outras tags similares, portanto, é necessário verificar a digitação de tags para que não seja uma atribuição irrelevante para a pesquisa. Por exemplo, erros de digitação são ocorrências comuns em sistemas de *folksonomia* e acabam por aumentar o *long tail* de tags em um dataset. Além disso, a tag

t é a principal ligação entre o histórico de tags atribuídas e as tags que serão sugeridas para o usuário. Portanto, erros tipográficos levariam a recomendação ao fracasso. O uso da distância de Levenshtein [Gil09] é uma medida popular de similaridade que serve para comparar duas *strings*. Esta medida foi utilizada para sugerir palavras similares com o objetivo de "corrigir" a digitação, quando o usuário incluir tags que não existam no banco de dados e que sejam similares a outras tags já atribuídas.

Outra necessidade para recomendação é o uso de conjuntos de tags para o usuário definir o assunto/contexto das tags que ele deseja receber. Neste caso, uma *query* é um conjuntos de tags já atribuídas, esta abordagem para recomendação foi chamada de *query composta*.

4.2.2 *Query* composta

A combinação de tags direciona o recebimento de outras tags que possuam o contexto similar entre si. A *query* composta verifica tags já atribuídas para recomendar outras similares com base nas combinações e pelas medidas apresentadas neste capítulo, adaptadas para este caso. Nesta adaptação é feita a combinação de duas ou mais tags $S = \{t_1, t_2 ...\}$ para obtenção da co-ocorrência e aplicação das medidas de relevência e popularidade para o conjunto de tags T de cada item.

$$exist(S,T) = \begin{cases} 1, S \subseteq T \\ 0, S \nsubseteq T \end{cases}$$

Por exemplo, na Figura 4.4 observa-se os níveis da estrutura de recomendação da *engine* para a *query* composta.

No primeiro nível *(a)* a tag t inserida pelo usuário como uma *query* simples dá origem a uma lista *(b)* de outras tags t_j sugeridas pela *engine*. A combinação *(c)* das tags atribuídas pelo usuário, seja pela inserção manual ou pela aceitação da recomendação, será interpretada *(d)* para recomendar outros resultados *(e)* mais específicos do que o resultado de uma *query* simples.

A Figura 4.5 é um exemplo de aplicação real para a recomendação de tags baseadas em *queries* compostas. A *query* "ny" foi atribuída para a imagem e as tags t_j são as tags aceitas pela recomendação. A combinação destas tags formam a *query* composta para a apresentação de outras tags que não foram recomendadas pela *query* simples.

O refinamento do assunto efetuado pela combinação das tags "statueofliberty" e "newyork" em substituição a t gerou informações mais de acordo com o contexto da imagem.

No próximo capítulo serão apresentados os detalhes do *dataset* utilizado para aplicação do experimento com os usuários.

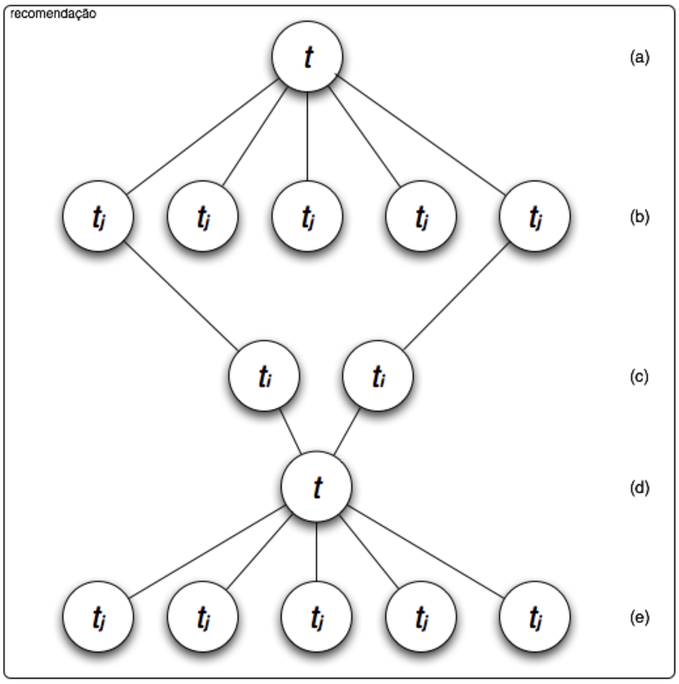

Figura 4.4: Combinação das tags atribuídas e aceitas pelo usuário para geração de novas recomendações a partir da co-ocorrência do conjunto de tags

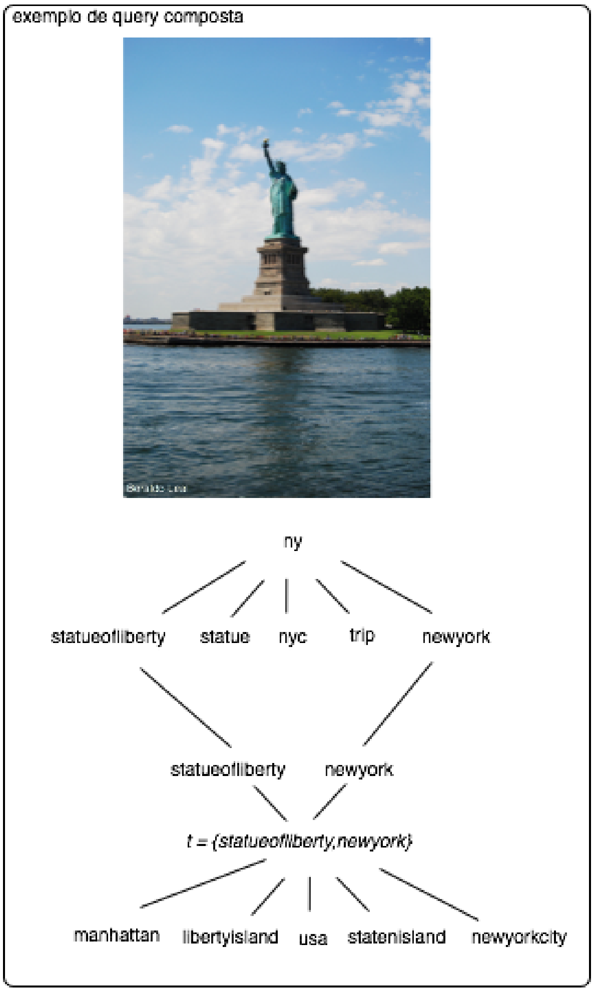

Figura 4.5: Exemplo da combinação de tags. *Query* composta para busca de outras similares resulta em tags mais específicas.

5. *DATASET* EXPERIMENTAL

Para desenvolver um experimento sobre a aceitação das tags sugeridas pelo algoritmo foram utilizadas as tags atribuídas pela comunidade do *Flickr*. Cada tag pertence a um usuário e a um item mesmo que existam tags iguais atribuídas a outros itens. Ou seja, a postagem de um objeto r_i contendo as tags $T_i = \{brazil, curitiba, sol\}$ pelo usuário u_i é individual e não leva em consideração outras postagens r_n com as tags $T_n = \{sol, curitiba, brazil\}$. No trabalho de Sigurbjörnsson [SvZ08] investigou-se como os usuários costumam atribuir tags no *Flickr*, e o resultado apontou que em 52 milhões de fotos o *long tail* consiste em mais de 15 milhões de fotos com somente uma tag atribuída e 17 milhões de fotos tendo 2 ou 3 tags. Fotos com uma tag representam 29.82% do total, fotos com 2-6 tags representam 56.72% do *dataset* e fotos com mais de 6 tags representam 13.46% do *dataset*.

Foram selecionados aleatóriamente 154 mil objetos/fotos do *dataset*[1] do *Flickr* restringindo apenas o número mínimo e máximo de tags por foto entre 2-6 tags como fonte para recomendação.

Após construir o banco de dados para experimento, foi efetuada uma limpeza na base, excluindo itens com tags que possuíssem caracteres especiais. O resultado foi uma base que servirá como fonte de dados para recomendação onde existem mais de 600 mil tags incluídas por mais de 36 mil usuários pertencentes a comunidade do *Flickr*.

Tabela 5.1: *Dataset* do *Flickr* com o total de tags, itens e usuários.

tags		usuários distintos	itens
distintas	total		
49.120	605.043	36.397	154.124

A distribuição de tags no *dataset* é um fator importante e deve ser considerado na utilização para recomendação. Informações como reusabilidade de tags e o histórico de tags atribuídas apenas uma vez influenciarão no impacto da recomendação e a aceitação dos usuários. Para tanto foi efetuada uma análise no conjunto de tags atribuídas para os mais de 154 mil itens selecionados, que será mostrada a seguir.

5.1 Análise do *Dataset*

A qualidade das tags atribuídas pelos usuários está relacionada à reusabilidade de cada palavra-chave. Foi analisado o *long tail* gerado pela frequência de tags no *dataset*. A curva gerada (Figura 5.1) representa a heterogeneidade do vocabulário gerado pela *folksonomia* no *Flickr*. Essa

[1]Systems Lab Amsterdam, University of Amsterdam.
http://staff.science.uva.nl/˜xirong/index.php?n=Main.Dataset

distribuição implica que uma abundância de tags é utilizada poucas vezes para a categorização de objetos, e uma minoria de tags é aplicada para a maioria das categorizações [PdS08]. Pela análise do *dataset* utilizado no experimento na Figura 5.1 foi observado que dentre as mais de 49 mil tags distintas, aproximadamente 24 mil foram utilizadas apenas uma vez pelos usuários. Isto representa

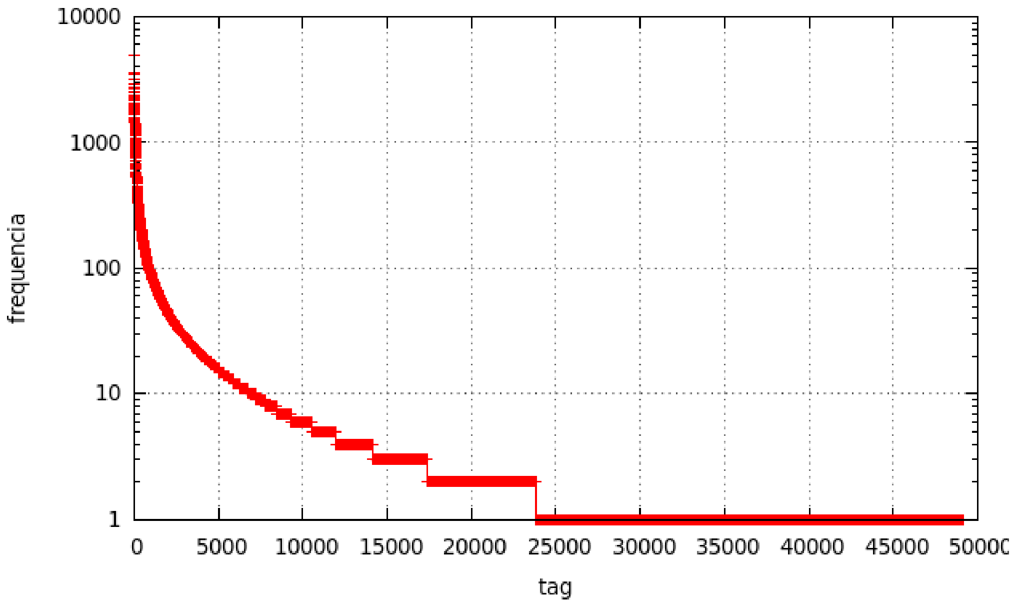

Figura 5.1: Representação em escala logarítmica do *long tail* gerado pela frequência de uso das tags no *dataset* que será utilizado para recomendação.

uma grande heterogeneidade no vocabulário das postagens e esta é umas das motivações para a utilização de sistemas de recomendação em sistemas de *folksonomia* pois um vocabulário mais homogêneo pode melhorar resultados de busca e classificação de imagens que possuem uma mesma descrição/contexto/significado.

Na Figura 5.2 estão representadas as quinze tags mais frequentes no *dataset*. A tag "2006" é a tag mais frequente no ambiente ocorrendo mais de 5 mil vezes. Logo em seguida as palavras-chave "dog", "china", "kitchen", "greatwall", confirmando que o comportamento de atribuição é basicamente relacionado ao tempo, lugar e a quem/que está na imagem.

Dentre as tags únicas encontradas, observa-se problemas relacionados a erros de tipografia, tags compostas, erros de inclusão e também à utilização de tags com motivo navegacional mencionadas em [SKK10]. Por exemplo, erros de digitação como "minneasota" (minnesota) e "manhatten" (manhattan), podem ser facilmente contornados com a implementação de uma verificação de palavras similares como já mencionado no capítulo anterior.

No próximo capítulo serão apresentadas a preparação do ambiente para experimento, as tarefas solicitadas para execução dos testes e o questionário para identificação do perfil do usuário relativo a sua experiência com o ambiente e com sistemas de *folksonomia*.

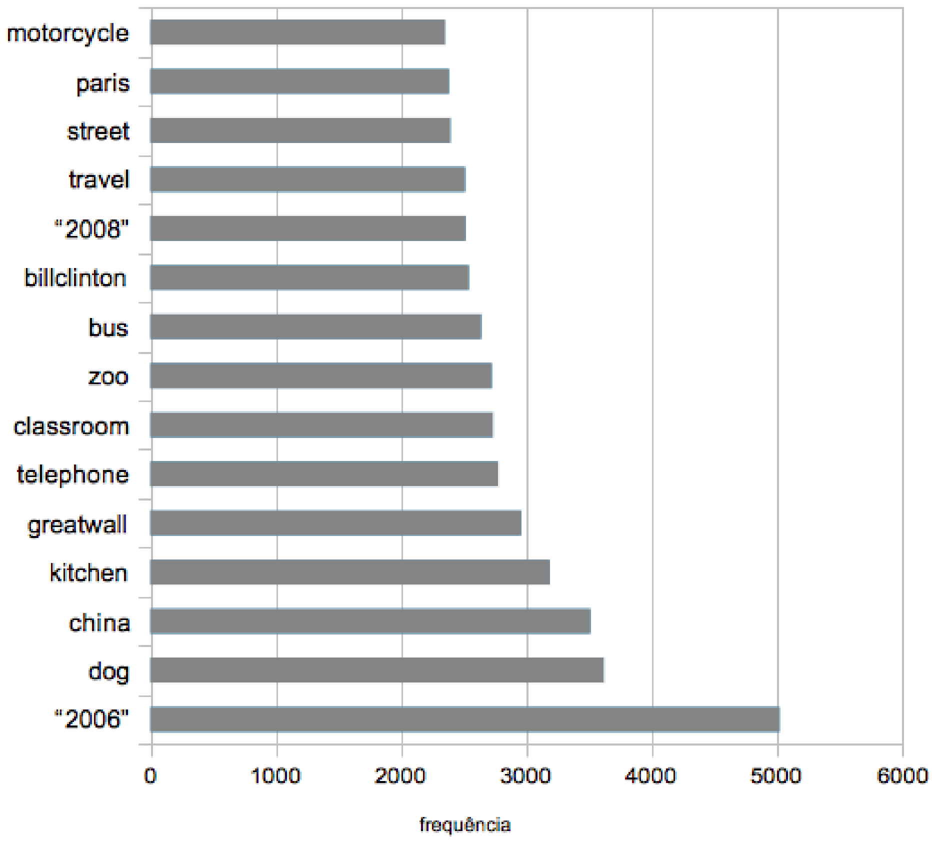

Figura 5.2: Tags mais frequentes no *dataset* do *Flickr*, mostrando palavras que geralmente estão relacionadas ao tempo, lugar e quem/que está na imagem.

6. AMBIENTE PARA EXPERIMENTO

Para a avaliação das recomendações geradas pela *engine*, foi desenvolvido um ambiente para verificar a aceitação das tags pelos usuários. Ao final do processo de atribuição foi aplicado um questionário para obter informações qualitativas da experiência de cada participante. A *engine* para recomendação foi implementada utilizando linguagens populares para ferramentas Web 2.0 como JavaScript, AJAX, PHP, HTML e CSS e para gerenciamento dos dados foi utilizado o gerenciador de banco de dados PostgreSQL.

6.1 Interação com a Engine

O ambiente para interação constitui-se de uma sequência de imagens para atribuição de tags. As imagens são apresentadas uma a cada vez e para cada imagem é solicitado ao usuário que inclua uma tag para receber recomendação de outras similares. Ao incluir uma tag/*query* para a imagem são apresentadas outras dez tags em uma lista ordenada pelo peso da média geométrica gerada pelo algoritmo desenvolvido. O usuário pode selecionar as tags que considerar mais relevantes para atribuir à imagem. Estas tags atribuídas aparecerão em uma sub lista onde é possível combinar mais de uma tag para solicitar outras, ou seja, neste cenário surgem as *queries* compostas, que são combinações entre as tags.

A representação do ambiente de interação pode ser observada pelo fluxograma apresentado na Figura 6.1. A sequência de ações para a recomendação começa com a primeira tag digitada para a imagem que está sendo postada. Quando a tag t é inserida pelo usuário o sistema interpreta-a como uma *query* para a busca de outras tags no *dataset*. Quando o sistema encontra outras tags na verificação da co-ocorrência é iniciada a fase de aplicação dos cálculos da normalização da co-ocorrência, das medidas de relevância e popularidade.

Processadas as medidas, ocorre a apresentação da lista de tags, ou seja, o ranking final, deixando por conta do usuário a opção de atribuí-las à sua foto ou não, tendo a opção de novas inserções. Porém, quando não existe co-ocorrência de t com as outras tags do *dataset*, são verificadas através da distância de Levenshtein outras tags atribuídas anteriormente pelos usuários da comunidade e que são semelhantes à tag t. Quando houverem tags similares o sistema apresenta a opção de atribuição em substituição à tag digitada. O usuário pode optar por aceitar a tag similar e logo é iniciado o processo de verificação de co-ocorrência até a apresentação do ranking, ou o usuário pode negar a sugestão para atribuição e continuar sua inserção atual. Após efetuadas todas as atribuições desejadas, o sistema finaliza o processo através da submissão da postagem.

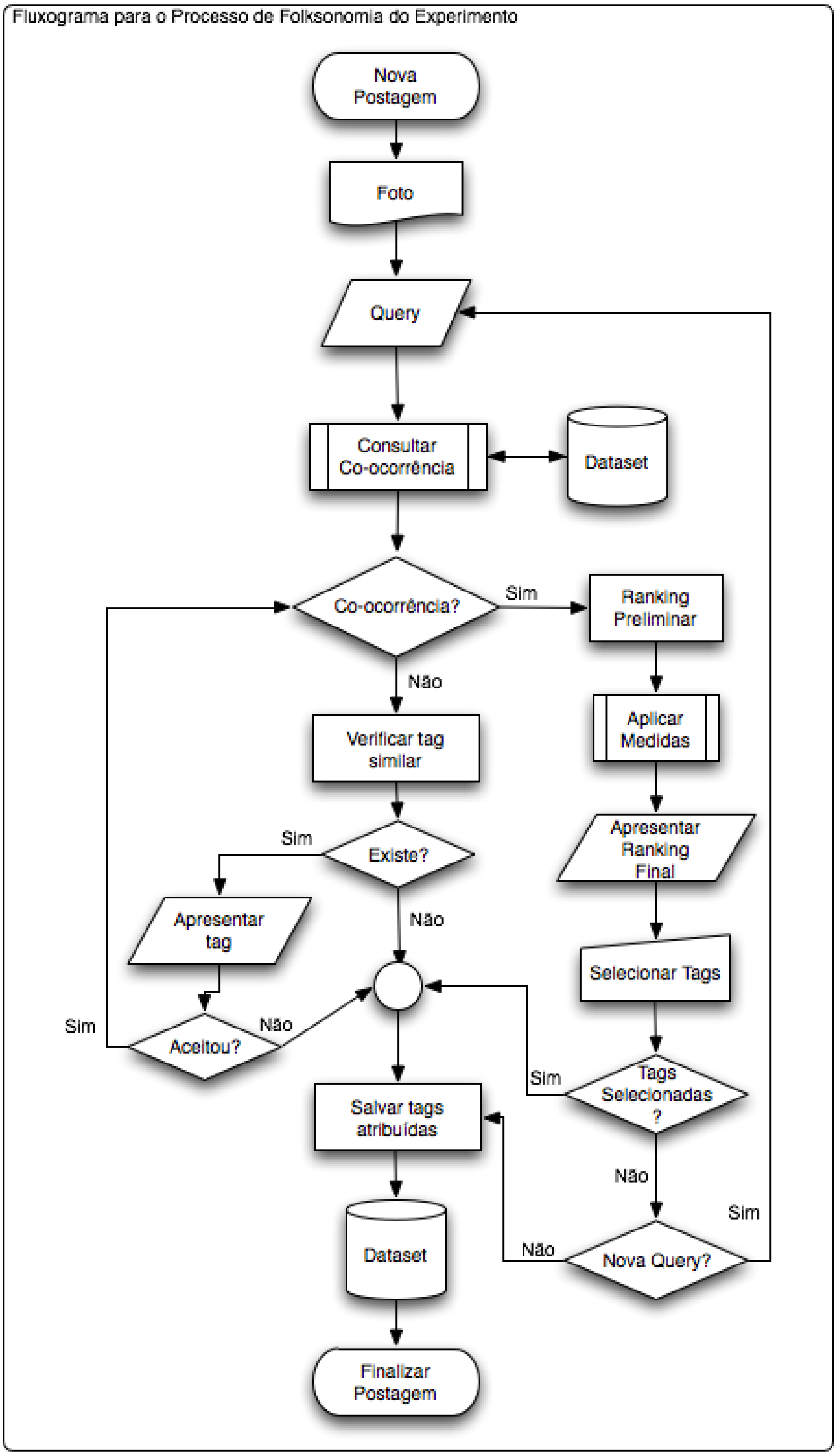

Figura 6.1: Fluxograma do ambiente de experimento. Estrutura da *engine* desenvolvida para teste com usuários.

6.2 Coleção de imagens

A escolha das imagens para o ambiente do experimento foi motivada pela compreensão geral do que está sendo apresentado, ou seja, o usuário não necessita ser especialista no conteúdo da imagem.

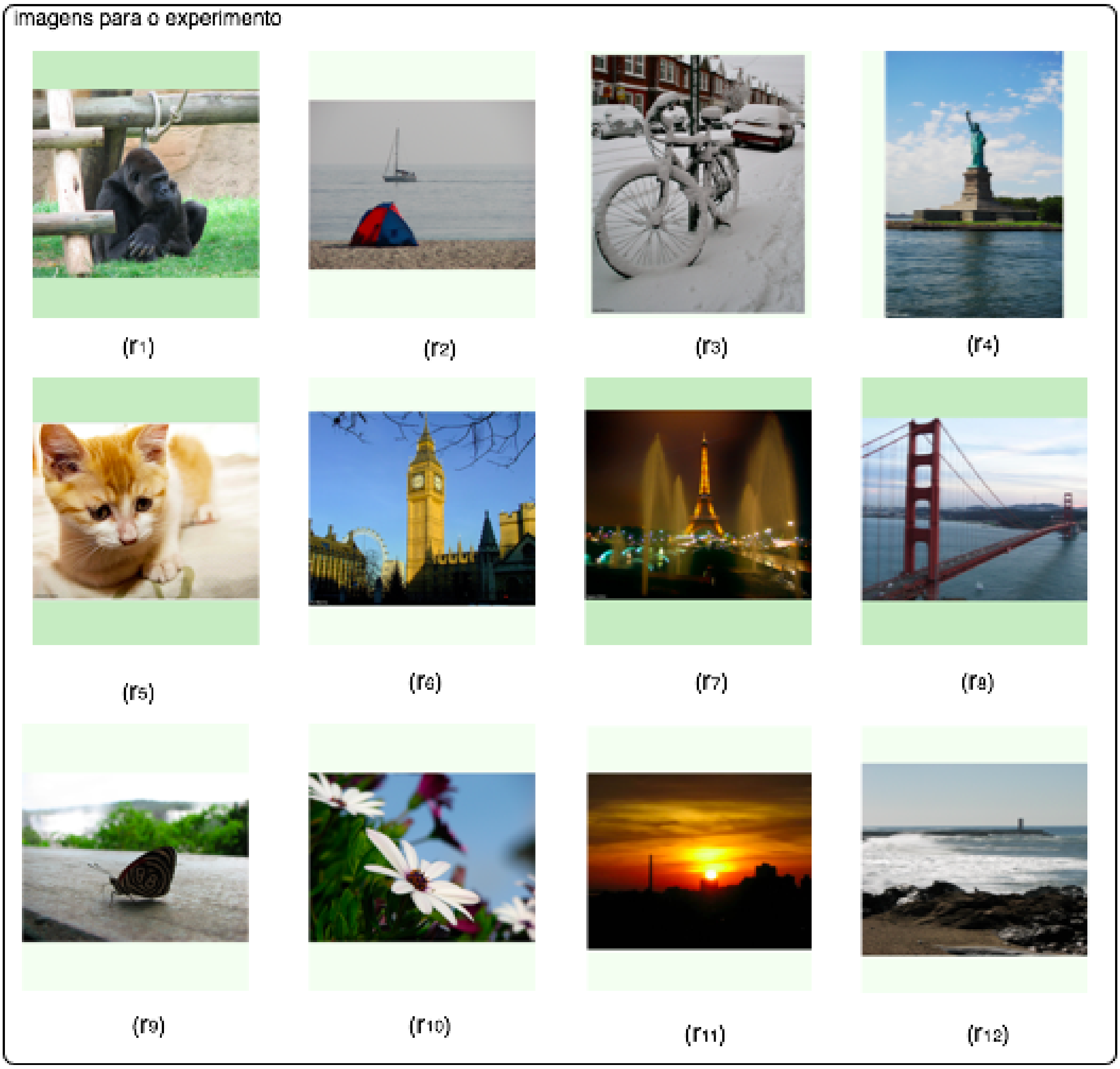

Figura 6.2: Imagens usadas para o experimento.

Durante o experimento as imagens foram apresentadas na sequência da Figura 6.2, uma a cada vez, ou seja, cada foto é tratada como r_n e cada participante u_n do experimento que fizer uma atribuição de tags receberá um identificador único para *seu* objeto para representação do cenário real de um sistema de *folksonomia*. Para cada tag atribuída à imagem, foram armazenadas informações das medidas de cada tag quando aceitas pela recomendação e a posição da tag no ranking oferecido. Os detalhes da estrutura do novo banco de dados para o experimento serão apresentados nas subseções 6.4.1 e 6.4.3.

6.3 Interface

Foram definidas as seguintes tarefas para o usuário conforme Figura 6.3 que apresenta a interface desenvolvida: dada a imagem (1) apresentada no ambiente, digite uma tag e adicione (2) para confirmar a atribuição da mesma. Ao executar esta ação dois processos ocorrem em paralelo, o primeiro (3) é a inserção da tag em uma lista/conjunto T de tags atribuídas para a imagem. Neste mesmo momento são processadas as medidas que geram o ranking apresentado como uma lista de dez tags (4) disponível para seleção. O usuário pode selecionar quantas tags achar pertinente de acordo com o contexto da foto. Quando selecionadas, é feita a confirmação das escolhas adicionando-as (5) à lista de tags (3). As tags são agrupadas (6) e é disponibilizada a opção de combinar as tags, ou seja, executar uma *query* composta através da seleção de um conjunto de tags para obter outras relevantes ou para refinar a busca clicando em (7) para obter estes resultados (8) no mesmo lugar (4) que foi apresentada a lista baseada na *query* simples.

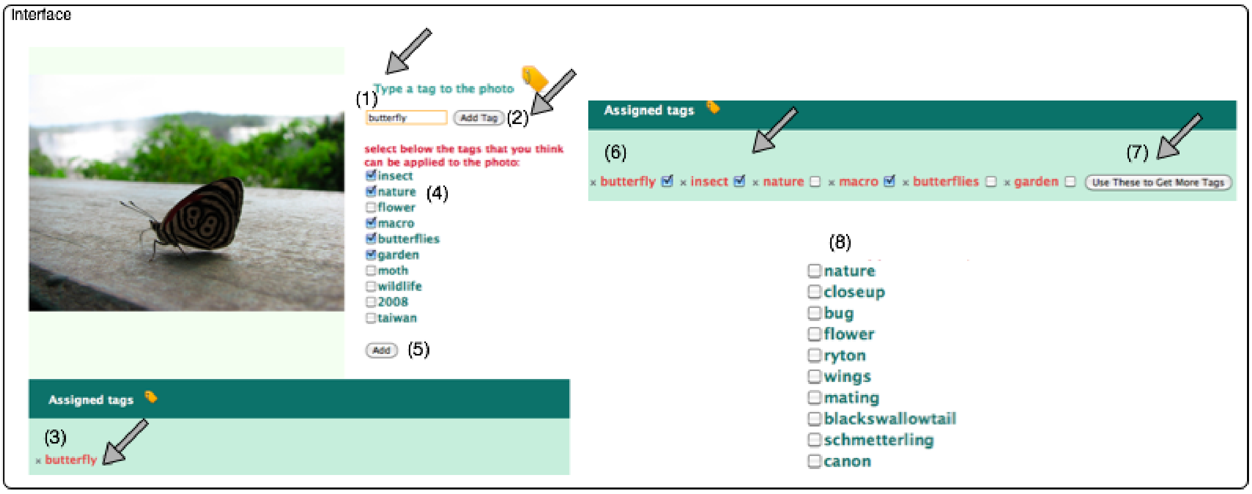

Figura 6.3: Interface do sistema Web para interação do usuário e obtenção do comportamento de atribuição de tags.

Estas tarefas são solicitadas para as doze fotos apresentadas para atribuição de tags. Após a finalização das tarefas foi solicitado o preenchimento de um questionário (Apêndice 10) para verificar a familiaridade do usuário com sistemas de *folksonomia* e também um relato de sua experiência utilizando este ambiente de recomendação.

6.4 Coleta de Dados

Para obter as informações referentes ao desempenho da *engine* o ambiente foi controlado pelas seleções/*clicks*/aceitações de tags as informações de posicionamento das tags aceitas no ranking, atribuição de recomendações e *queries* e o peso das medidas serão armazenadas a cada postagem que o usuário executar.

6.4.1 Armazenamento das medidas de cada tag

Cada tag recomendada pela *engine* possui atributos relativos às suas medidas gerados pelo ranking final. A *query* digitada pelo usuário gera uma lista de tags para recomendação e cada uma destas possui informações dos seus pesos para co-ocorrência, relevância e popularidade. Como exemplo, a Figura 6.4 apresenta os valores de co-ocorrência, popularidade e relevância em cada tag recomendada com base na tag t "desert". As tags sugeridas serão salvas no banco também contendo informações relativas às suas medidas e posicionamento no ranking. Neste caso, se o usuário selecionar a tag "utah" que foi recomendada, serão salvas as informações $t_j =< 3, 0.08, 0.18, 00.7 >$.

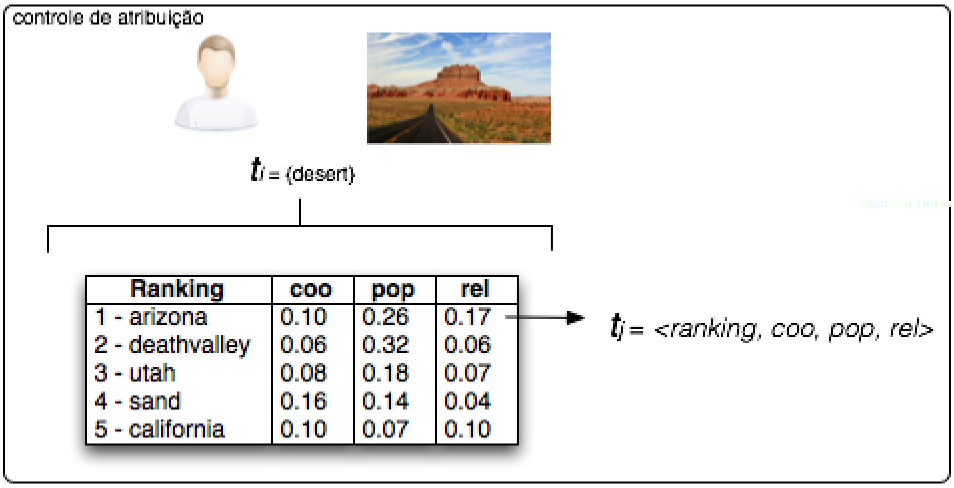

Figura 6.4: Exemplo de ranking gerado pela tag "desert" e o peso de cada medida para cada tag recomendada.

Este tipo de informação servirá para a verificação do perfil do usuário, ou seja, se ele geralmente costuma atribuir tags com maior relevância, popularidade ou co-ocorrência. Além disso, o posicionamento das tags no ranking indicará de maneira geral se as tags recomendadas no topo são as tags mais aceitas pelos participantes do experimento.

6.4.2 Queries vs. Tags Recomendadas

O tipo de tag atribuída pelo usuário também foi uma informação controlada neste experimento. A observação da atribuição está basicamente relacionada à forma como as tags são atribuídas ao objeto, ou seja, se estas foram recomendadas ou não. Esta informação permitirá a observação do *long tail* nesta relação, além de contribuir para o controle da homogeneidade do vocabulário em cada abordagem.

6.4.3 Modelo Relacional do Banco de Dados

O armazenamento das informações de atribuição é constituído por três entidades para controle de usuários, itens e tags. Em um sistema de *folksonomia* cada tag atribuída é tratada como única no banco de dados, ou seja, não é por que existe uma tag "amor" no banco de dados que o sistema vinculará esta tag ao *id* de outra idêntica atribuída. Este é um dos motivos que dá tanta liberdade para a classificação e organização dos dados, pois não existe um vínculo entre tags idênticas.

O modelo relacional das tabelas do banco de dados utilizado para o experimento está representado na Figura 6.5. As informações contidas em cada entidade permitirão controlar as ações

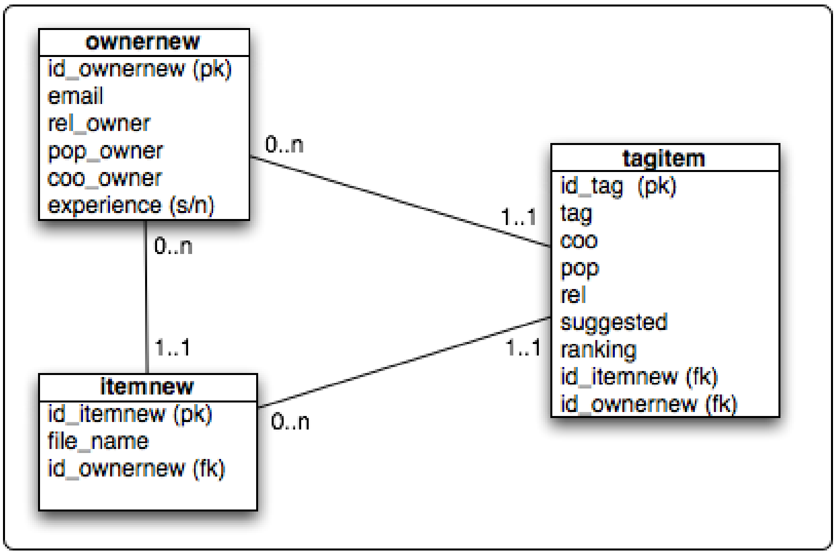

Figura 6.5: Modelo de relacionamento entre as entidades do banco de dados para o controle do experimento.

e comportamento geral de atribuição dos participantes através da análise do seu histórico. Por exemplo, o controle de tags atribuídas pela recomendação é feito através do atributo *suggested* pertencente à tabela com o nome *tagitem*. A informação binária contida neste campo (Recomendada? 1-sim, 0-não) permitirá a clusterização de tags de acordo com sua categoria. A clusterização de cada tipo de tag permite identificar os dois cenários da *folksonomia*, ou seja, o ambiente de atribuição puro, onde o usuário é o autor de cada tag, e o ambiente colaborativo gerado pela recomendação baseada tags de outros integrantes da comunidade. Os campos *coo, rel e pop* na mesma tabela, guardam informações referentes aos valores das medidas de cada tag atribuída para o objeto postado. Caso esta tag seja uma *query*, não haverá valores nos campos. Estes valores servirão para o algoritmo de recomendação personalizado, desenvolvido após os resultados do experimento (capítulo 8).

Outro campo na tabela *tagitem* é o *ranking*, que guarda informações sobre a posição em que a tag foi sugerida no momento da recomendação. Estes valores têm o objetivo de medir o desempenho da recomendação, ou seja, se as tags aceitas pelos usuários estão sendo posicionadas pelo algoritmo nas primeiras posições do ranking final.

Já na entidade *ownernew* são armazenadas as informações dos usuários participantes do experimento. Os campos *rel_owner, pop_owner e coo_owner*, armazenam a soma das vezes que o usuário teve em suas tags valores com maior índice para estas medidas como apresentado na Figura 6.4.

Utilizando o exemplo da tag "sand" que possui os valores $coo = 0.16$, $pop = 0.14$ e $rel = 0.04$, ao serem incluídas as informações na entidade *ownernew* será somado um ponto na tupla *coo_owner* pois é esta que tem o maior dentre os valores de cada medida. Estas informações definem o perfil de atribuição dos usuários, ou seja, quais medidas são mas influentes durante a *folksonomia*.

Mais detalhes sobre as entidades e suas tuplas estão apresentadas no dicionário de dados deste modelo relacional na Apêndice 11.

6.4.4 Escalabilidade

O tempo de resposta e a consequente aplicação prática de um sistema de recomendação baseado nas ações dos usuários dependem da escalabilidade do sistema. Para verificarmos a influência da quantidade de dados utilizados para efetuar a recomendação e o tempo de resposta do sistema, foram analisadas tags com alta frequência em relação a tags com baixa frequência no *dataset*. Outro ponto observado foi a influência da quantidade de dados disponível para consulta e qual a diferença que esta exercerá no tempo de resposta. Para a análise foram utilizadas as *queries* "london" e "love", sendo respectivamente suas frequências alta e baixa. A Figura 6.6 apresenta o tempo para a entrega de 10 e 20 tags para recomendação, utilizando como *query* a tag "love" que foi considerada com baixa frequência (baixa@10 e baixa@20) em relação à tag "london" de alta frequência (alta@10 e alta@20) no *dataset*.

O tempo de resposta foi observado em 6 *datasets* originários da mesma fonte, mas limitando o número máximo de tags consultadas, começando com 100 mil até 1 milhão de tags para consulta.

Como já esperado observou-se maior demora no tempo de resposta à medida que o número de tags aumenta. Além disso, o cálculo das medidas para 10 e 20 tags também influenciou no tempo de resposta, principalmente quando executado para a tag com alta frequência. Para o dataset com 1 milhão de tags, nota-se que houve um aumento considerável no tempo de resposta quando executado para as 20 tags mais frequentes resultantes do processo de co-ocorrência e aplicação das medidas. O tempo de resposta para as tags com alta frequência em @10 foi de 2.556 segundos enquanto para o cálculo executado em @20 foi de 0.672 milissegundos a mais, ou seja, 3.228 segundos.

O experimento realizado com usuários para análise dos dados da recomendação foi executado em um *dataset* com 600 mil tags, limitando o tempo de resposta para recomendação a menos de 2 segundos mesmo para tags com alta frequência.

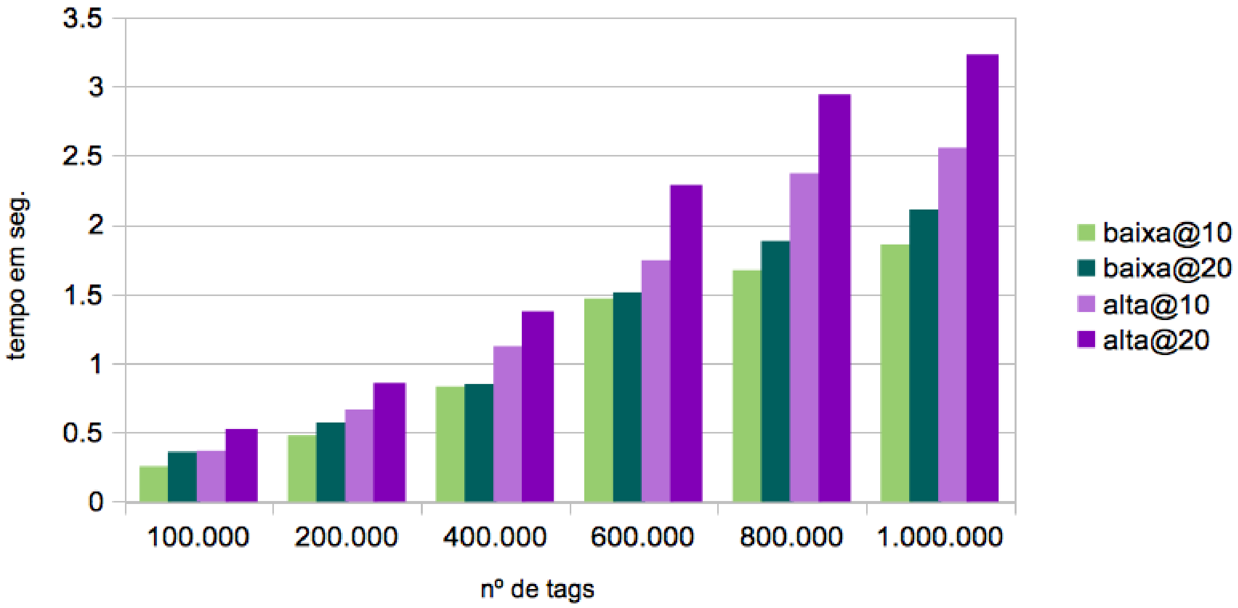

Figura 6.6: Resultado da verificação do comportamento do sistema em relação ao tempo de resposta para recomendação. Observações foram executadas para a entrega de 10 e 20 tags, utilizando tags com alta e baixa frequência para análise do tempo de resposta em relação à quantidade de dados no *dataset* e influência do cálculo das medidas.

7. RESULTADOS

A *engine* desenvolvida para o experimento e avaliação dos resultados de recomendação foi disponibilizada *online* durante duas semanas. O único requisito exigido dos participantes foi a utilização da língua inglesa para inserção de tags pois a base de dados possui em sua maioria palavras-chave em inglês e também para aumentar o número de possíveis participantes.

No total foram 50 participantes que completaram todas as etapas do experimento até o questionário aplicado. Dentre os participantes, 60% deles responderam que não costumavam fazer atribuição de tags em seus documentos digitais. Para o experimento, estes usuários foram considerados como inexperientes.

Para representação dos resultados, as tags serão apresentadas conforme o seu tipo de atribuição, ou seja, se foram digitadas pelo usuário (*queries*) ou se foram aceitas através *engine* de recomendação.

Os dados de cada tipo de tag apresentados na Tabela 7.1 demonstram que existe uma quantidade maior de tags aceitas pela recomendação em relação ao número de *queries* digitadas. Ainda assim o vocabulário das tags recomendadas foi mais homogêneo, pois apresentou uma quantidade menor de tags distintas no banco de dados enquanto as *queries* digitadas pelos usuários geraram mais de 20% de tags distintas. A relação de homogeneidade e heterogeneidade do vocabulário pode ser melhor analisada pelo *long tail* do *dataset*, que será apresentado na próxima seção.

Tabela 7.1: Informações resultantes do *dataset* gerado pelo experimento aplicado à 50 participantes.

queries		recomendadas	
distintas	total	distintas	total
182	891	145	1.235

No gráfico da Figura 7.1 estão apresentadas as quinze *queries* mais frequentemente atribuídas no *dataset* resultante do experimento. A palavra "cat" é a que aparece com mais frequência e seu uso está relacionado essencialmente à imagem r_5. A verificação do conjunto de *queries* atribuídas ao item r_5 (Figura 7.2) permite a observação da predominância do reuso desta tag em relação às outras tags também utilizadas como *query*.

O reuso representa a reincidência do uso das tags e é um fator importante em *folksonomia*, pois torna mais homogêneo o vocabulário do ambiente além de auxiliar na análise da compreensão da imagem por parte do usuário.

Já no gráfico da Figura 7.3 estão representadas as tags recomendadas com maior frequência de aceitação. A palavra "nature" é a tag recomendada com maior aceitação no *dataset* obtido pelo experimento e seu reuso acontece para os itens r_9 e r_{10} (Figura 7.4).

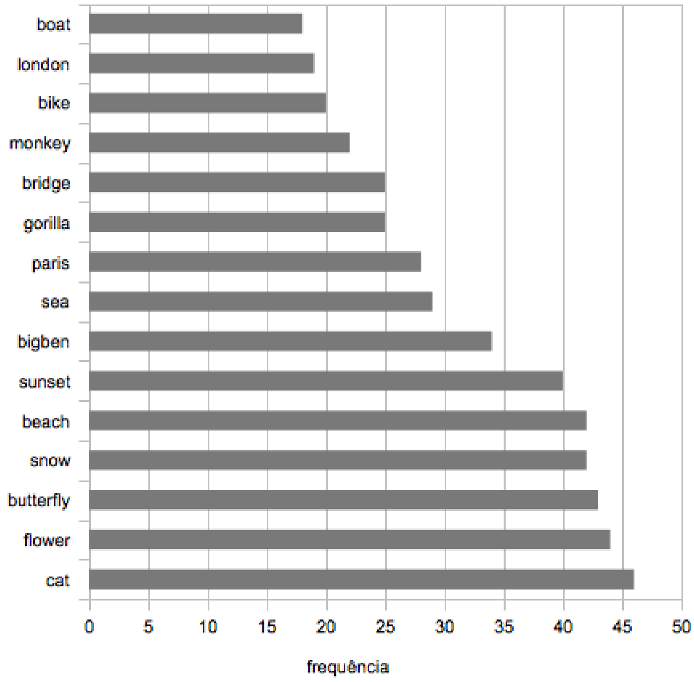

Figura 7.1: Número de casos de tags digitadas pelo usuário como atribuição e *query* de busca para outros conjuntos de tags similares.

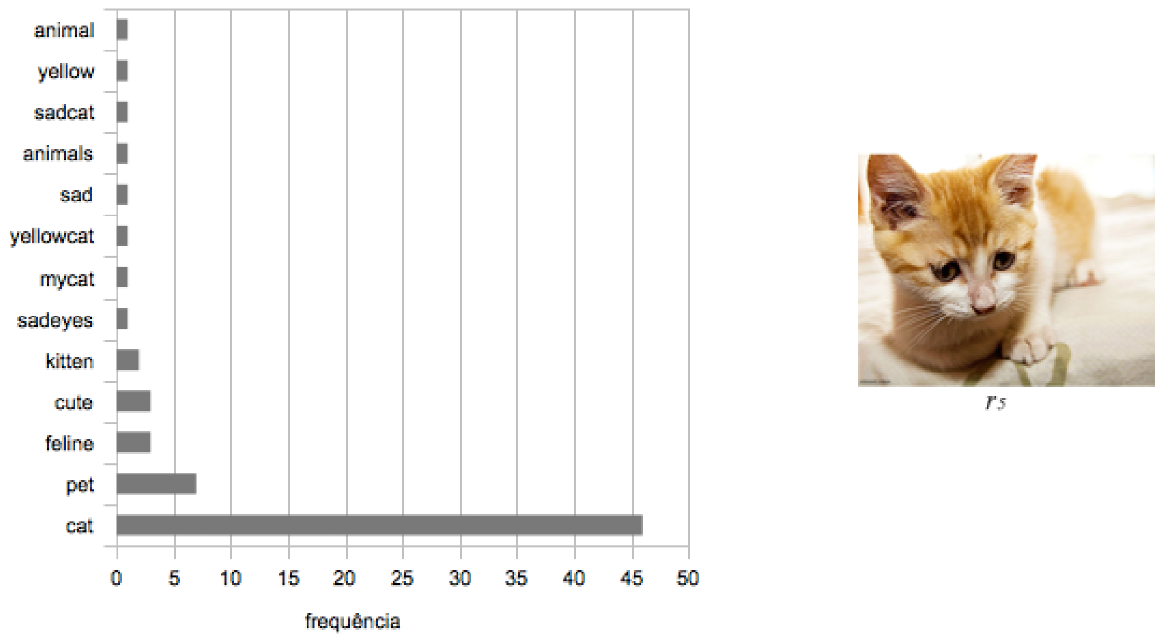

Figura 7.2: Frequência das *queries* digitadas para o item r_5.

Durante o experimento também houveram tags sem reuso e sua utilização está ligada ao perfil de atribuição do usuário e a erros de digitação, como exemplos "white_flower", "libertystate", "bigbang", "bigbeng", "snafrancisco", "buterfly" e outras. O perfil categorizador geralmente faz atribuições de palavras compostas interpretadas como *links* no exemplo da tag "white_flower". Estas

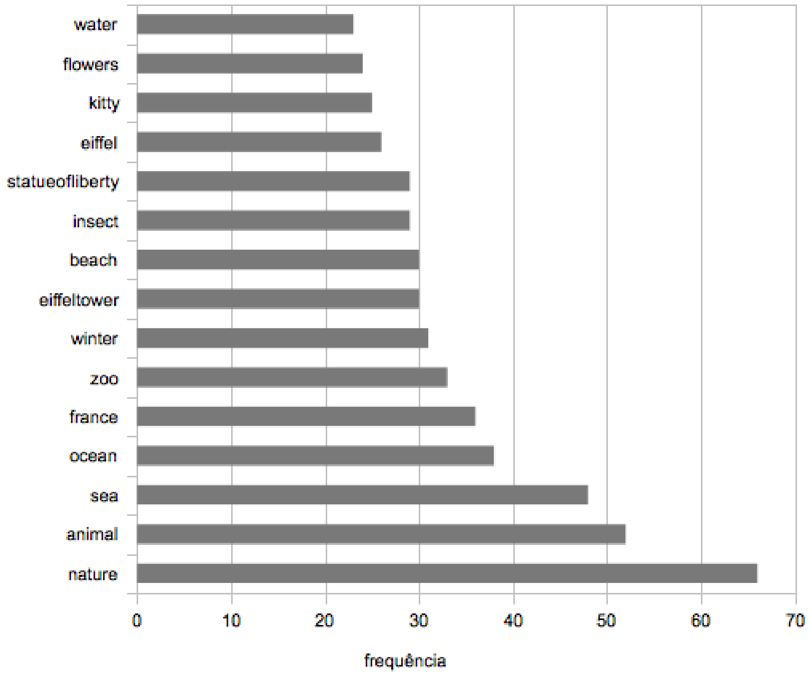

Figura 7.3: Tags mais aceitas pela recomendação.

Figura 7.4: Imagens que tiveram mais ocorrências da tag "nature", palavra-chave mais frequente no *dataset*.

são algumas das razões para a maior ocorrência de tags distintas no grupo de *queries* sem reuso, além da ocorrência de tags como exemplo "buterfly", que expressam erros de digitação.

Com base nestes resultados na próxima seção será apresentada a representação do *long tail* do *dataset* e as tags únicas.

7.1 Long Tail

Sistemas de *folksonomia* são caracterizados pela diversidade de vocabulário, idioma e escrita. As tags podem expressar informações relativas à descrição do objeto, contexto, emoções, conceitos, informações pessoais e eventos. Porém, esta diversidade de informação gera uma margem para a existência de tags únicas no banco de dados.

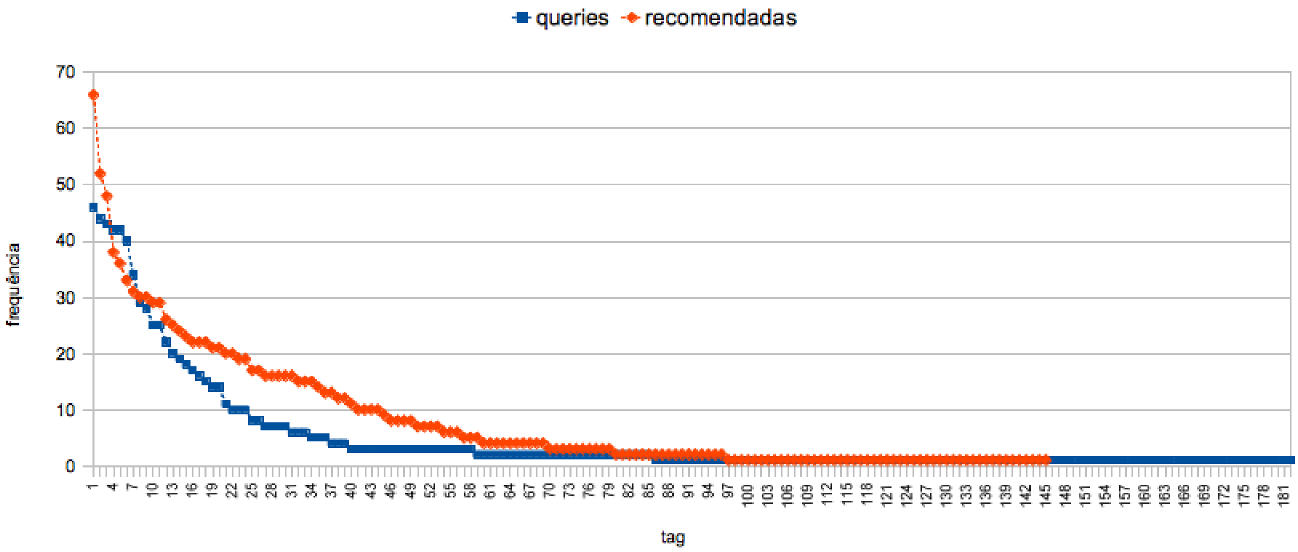

Figura 7.5: Representação do *long tail*

No gráfico da Figura 7.5 estão representados os dois cenários de atribuição: as *queries* e as tags recomendadas pelo sistema. O eixo x representa cada tag distinta atribuída durante o experimento e o eixo y representa a frequência de repetição de cada uma destas tags. Observa-se que o reuso das tags recomendadas é maior, pois a quantidade de tags recomendadas e atribuídas é maior que a quantidade de tags/*queries* digitadas pelos usuários. Além disso, a distribuição do vocabulário das recomendações é mais concentrado na repetição/aceitação da recomendação apresentando um *long tail* mais curto, ou seja, existe maior reuso do que para tags únicas.

Já na curva para as *queries* observa-se o *long tail* maior em relação às tags recomendadas pois a repetição das tags é menor comparada à curva de atribuição das tags recomendadas.

Para um total de 891 *queries* dentre as 182 distintas, 53% destas foram atribuídas apenas uma vez. Estes valores apontam que mais da metade do vocabulário formado pelas *queries* não foi reutilizado, gerando mais heterogeneidade na base. Por outro lado, dentre as 1.235 tags recomendadas com 145 distintas, o reuso foi de 66% e somente 34% das tags foram utilizadas apenas uma vez.

Percebe-se que a recomendação proporcionou um vocabulário mais homogêneo e com maior número de tags atribuídas do que digitadas. Na próxima seção será apresentada a análise do posicionamento das tags no ranking da recomendação.

7.2 Aceitação da Recomendação

Durante o processo de experimento foram armazenadas as posições das tags recomendadas permitindo assim obter informações em relação ao posicionamento de cada tag no ranking no momento em que foi sugerida.

O gráfico da Figura 7.6 representa a frequência das tags recomendadas e aceitas pelos usuários relativa ao posicionamento das mesmas no ranking de recomendação.

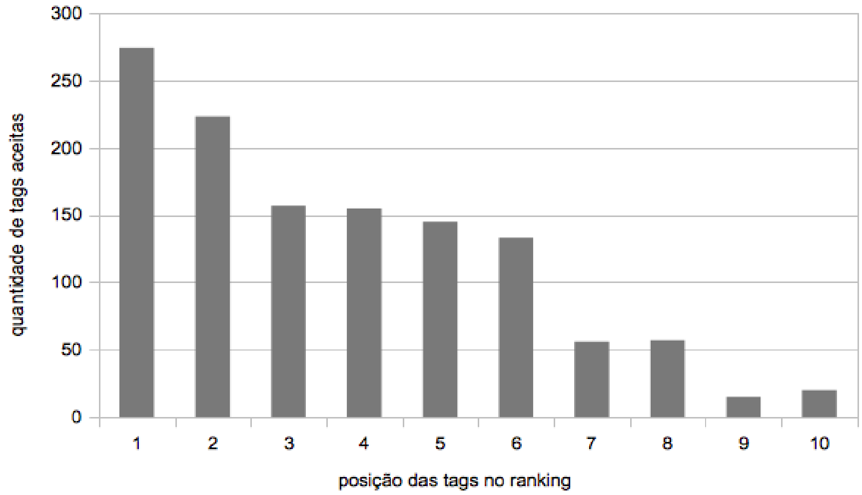

Figura 7.6: Relação entre a frequência e o posicionamento no ranking das tags recomendadas e atribuídas.

Observando o resultado do gráfico (Figura 7.6) nota-se a frequência maior de aceitação das tags posicionadas nas seis primeiras posições do ranking. Este resultado representa o bom desempenho da *engine* de recomendação, conseguindo posicionar tags relevantes no topo do ranking.

Como exemplo de reposicionamento das tags mais relevantes, observa-se na Tabela 7.2 os resultados gerados pelas medidas aplicadas e o novo ranking de recomendação. Neste exemplo a palavra "lion" foi utilizada como *query* e as dez tags apresentadas para recomendação foram reordenadas pelos cálculos das medidas e logo sugeridas no ranking final. As tags "08" e "stages" que apresentavam-se no início do ranking preliminar, foram reposicionadas no final do ranking. Por outro lado tags mais relevantes como "safari" e "animal" que encontravam-se na base do ranking foram para o topo do ranking para recomendação.

Para a observação do comportamento de atribuição no experimento foram desenvolvidos os gráficos do ranking individual de cada item r para a identificação das imagens que obtiveram melhor desempenho de recomendação em relação ao posicionamento das tags.

Como principais exemplos de bom desempenho na recomendação destacam-se os itens em que é possível identificar a sua localização, ou seja, as imagens onde a principal tag está relacionada ao lugar que ela representa.

Na Figura 7.7 observa-se este comportamento relativo a aceitação de tags para o item r_4 onde há informações da localização/lugar que a imagem representa. A concentração das tags aceitas encontra-se no topo do ranking confirmando o bom posicionamento das tags devido ao reuso daquelas que encontram-se nas primeiras posições.

A recomendação para itens que não possuem a localização específica no contexto da imagem mostrou maior heterogeneidade na distribuição das tags em relação ao posicionamento, como pode-se observar no gráfico da Figura 7.8 que representa o item r_9.

Durante o experimento foi comentado por alguns participantes que informações quanto ao lugar/localização no contexto da imagem são um importante fator para a inclusão de boas *queries*

Tabela 7.2: Resultados do reposicionamento no ranking gerado pelas medidas desenvolvidas para recomendação das tags utilizando a *query* "lion".

Ranking Preliminar	coo	pop	rel	$mean$	Ranking Final
zoo	0.155	0.176	0.462	0.233	zoo
08	0.123	0.054	0.008	0.142	safari
stages	0.123	0.500	0.004	0.132	animal
safari	0.059	0.211	0.230	0.130	tanzania
park	0.042	0.051	0.225	0.125	africa
sea	0.039	0.068	0.347	0.117	animals
africa	0.038	0.130	0.397	0.097	sea
tanzania	0.037	0.301	0.197	0.078	park
animal	0.035	0.091	0.706	0.064	stages
animals	0.035	0.083	0.552	0.038	08

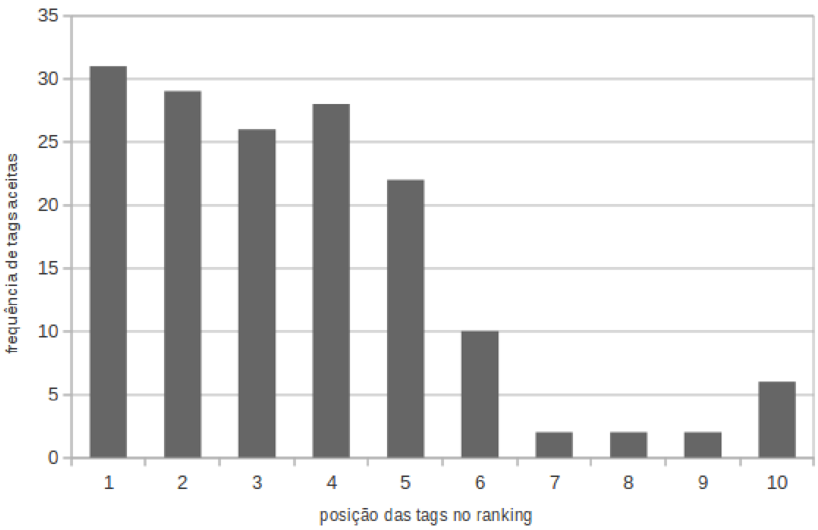

Figura 7.7: Posição das tags aceitas no ranking para o item r_4.

e aceitação da recomendação. Nestas observações percebe-se a estreita relação entre o perfil do usuário, o item, as tags atribuídas e o contexto representativo para o perfil.

O resultado do ranking de aceitação das tags por item está localizado no Apêndice 12 onde é possível observar a aceitação das tags em relação ao posicionamento de cada uma.

Na próxima seção será utilizada a medida de precisão para avaliar a recomendação em relação a cada item e usuário.

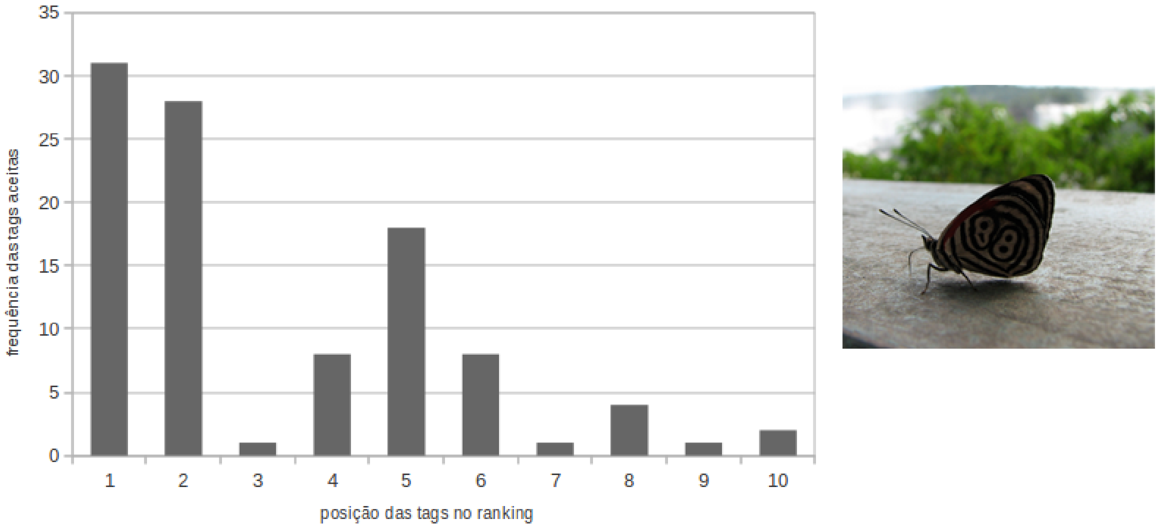

Figura 7.8: Posição das tags aceitas no ranking para o item r_9.

7.3 Precisão

Uma outra forma de medir o desempenho da recomendação é através da precisão (*precision*). Esta medida é popular na área de recuperação de informação e serve para analisar a proporção de documentos relevantes retornados em relação ao assunto de uma busca.

A fórmula da precisão leva em conta o número n de documentos retornados por uma consulta, indicando se todos os documentos recuperados estão de acordo com a solicitação. Para verificação dos resultados em um ranking utiliza-se variações para n, podendo obter-se a qualidade da recuperação da informação por seções do ranking apresentado.

Para este trabalho foi utilizada uma variação [JMH$^+$08] da fórmula de precisão adaptada para verificação de duas observações no ranking, sendo uma para as cinco primeiras tags recomendadas e a outra para as dez tags recomendadas. A verificação tem como objetivo identificar se as tags recomendadas apresentam resultados relevantes para cada item r através das atribuições de cada usuário u. A cada postagem P_i de um item, são atribuídas as informações sobre o tipo de tag (*query*, tag) e a quantidade de cada uma para cada item. Em cada postagem $P_i = <u_i, r_i, T_i>$ são comparadas as tags $T(t)$ que foram recomendadas para cada *query* t em relação às tags $T(u,r)$ recomendadas e aceitas para o objeto postado. Esses valores são utilizados para a fórmula de precisão

$$precision(u,r) = \frac{|T(t) \bigcap T(u,r)|}{|T(t)|}$$

onde o conjunto de tags recomendadas que foram consideradas corretas é dividido pelo número de tags recomendadas. O resultado apresentará valores entre zero e um, apontando melhor desempenho à medida que se aproxima de um.

Para verificar a precisão da recomendação, a fórmula foi aplicada a cada item r_i e por usuário como apresentado no gráfico da Figura 7.9. No gráfico está representada a precisão para as cinco primeiras recomendações do ranking ($P@5$) em relação ao desempenho das dez recomendações do ranking ($P@10$) para o item r_1.

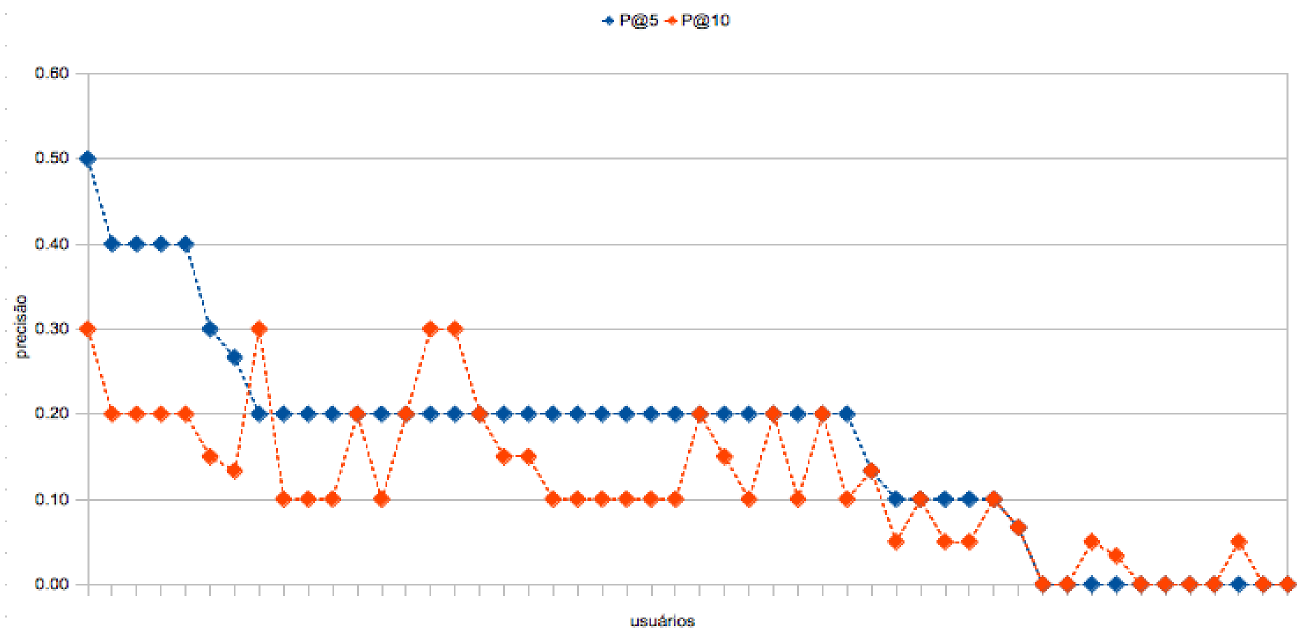

Figura 7.9: Resultados da precisão do objeto r_1 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

Observa-se melhor desempenho para as tags recomendadas no topo do ranking (5 primeiras posições) do que para as tags recomendadas na base do ranking. Este resultado se repete para outros dez objetos utilizados para o experimento, sendo que apenas um item (r_{10}) mostrou uma precisão maior para $P@10$.

Dentre os itens que obtiveram melhor resultado para $P@5$, observou-se como já esperado o melhor desempenho em imagens relacionadas a lugares/cidades facilmente reconhecidas, ou seja, objetos que tiveram tags como paris, france, ny, newyork etc. Estes resultados podem ser observados nos gráficos das Figuras 7.10 e 7.11.

Observou-se também durante o experimento que alguns usuários consideraram as recomendações bem mais válidas devido ao alto valor do índice de precisão em suas atribuições, enquanto para usuários apresentados à direita dos gráficos os resultados retornados não foram relevantes. Por exemplo, dentre as *queries* utilizadas para a classificação da imagem r_4 a mais frequente é a palavra "statueofliberty". O retorno desta tag recomendou nas cinco primeiras posições do ranking as palavras "newyork", "nyc", "newyorkcity", "manhattan" e "ny", que foram aceitas por grande parte dos usuários. Porém, alguns participantes que utilizaram esta mesma *query* neste mesmo item não aceitaram as recomendações.

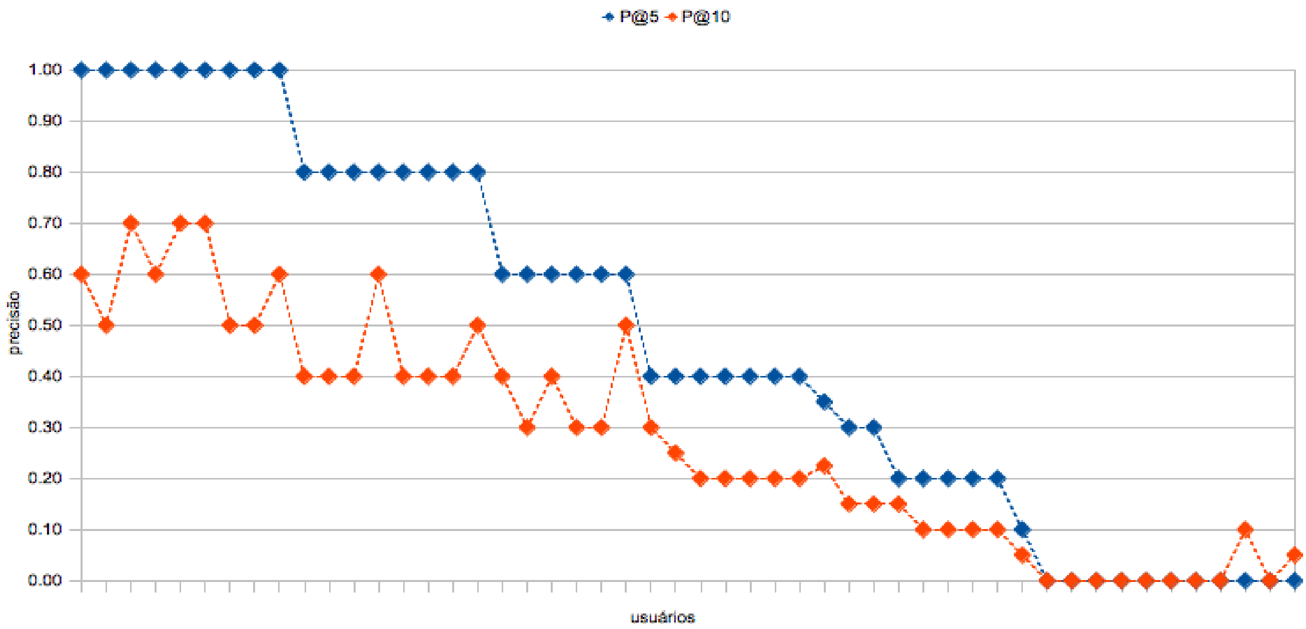

Figura 7.10: Resultados da precisão do objeto r_4 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

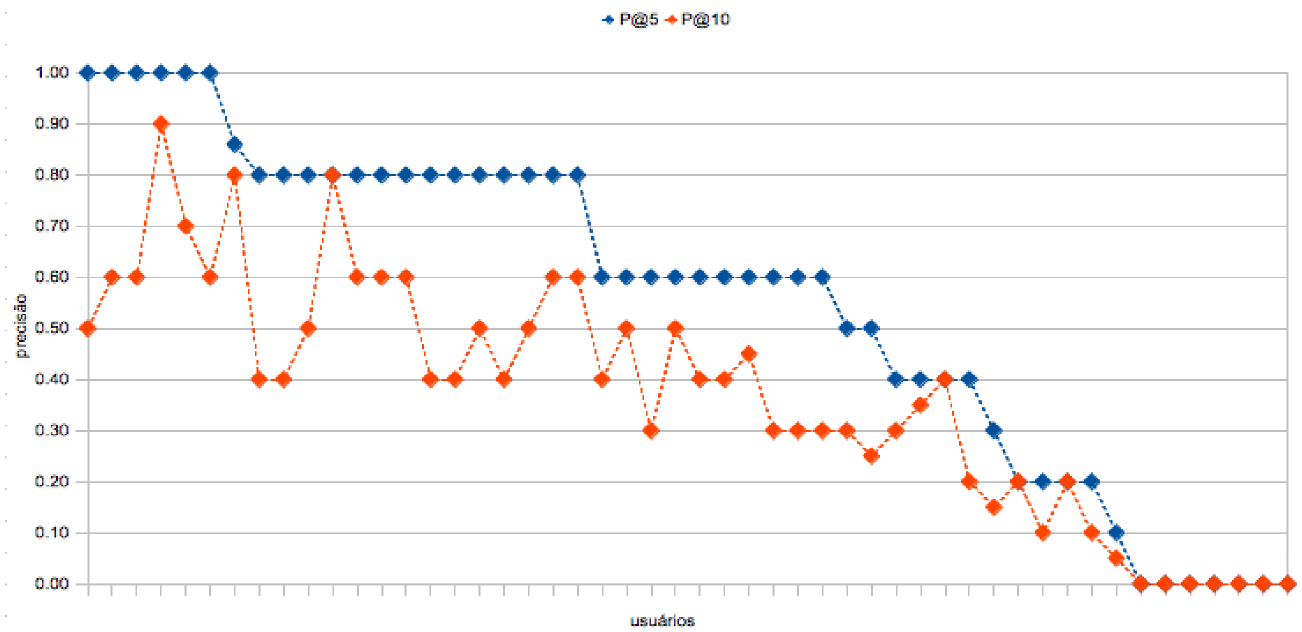

Figura 7.11: Resultados da precisão do objeto r_7 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

Sabe-se que a motivação do usuário para a utilização das tags também é um fator relevante para os resultados devido as respostas obtidas pelo questionário que será apresentado na próxima seção.

7.4 Resultados do Questionário Aplicado

A aplicação do questionário ao final do experimento teve como principal objetivo entender as motivações de atribuição dos usuários e avaliar a experiência durante a execução do experimento.

A primeira pergunta efetuada aos participantes é relativa à frequência do uso de tags em conteúdos próprios. Os resultados apontaram que apenas 13% dos participantes sempre utilizam a atribuição de tags ao seus conteúdos. Em seguida foram questionadas as razões para que as pessoas não atribuíssem tags aos seus itens. Esta questão permitia mais de uma resposta e apresentou principal concentração na dificuldade de atribuição, ou seja, a atribuição de tags é um trabalho exaustivo. Porém, todas as outras opções que citam o desconhecimento do tipo de tag que deve ser atribuída e ao motivo do uso das tags, também foram apontadas como obstáculos para a *folksonomia*.

Em relação à ajuda na atribuição de tags pelos sistemas de recomendação, 54% dos participantes concordaram que a recomendação ajuda no processo de atribuição de tags, 46% responderam que às vezes a recomendação ajuda, e apenas 2% acharam que a recomendação não ajuda na atribuição.

Estes resultados confirmam o esforço percebido para atribuição de tags e o desconhecimento do conjunto de tags apropriadas para a *folksonomia*. A recomendação semi-automática de tags permite facilitar o processo de escolha das tags apropriadas e a melhoria da qualidade e quantidade do conjunto de tags atribuídas aos objetos.

Na próxima seção poderá ser observado o comportamento de aceitação das tags em relação às medidas desenvolvidas para recomendação.

7.5 Atribuição e medidas de recomendação

A recomendação de tags utiliza as *queries* do usuário para inferir recomendações relevantes utilizando a similaridade baseada na sua co-ocorrência e nas medidas desenvolvidas neste trabalho. Estas medidas têm objetivo de identificar as tags que apenas estão no ranking pela sua frequência de ocorrência e logo retirá-las do topo da recomendação. As medidas de popularidade e relevância possuem objetivos distintos, mas que trabalham em conjunto para identificação de tags menos relevantes, logo espera-se que sua combinação na média geométrica resulte em tags mais relevantes no topo do ranking final.

Observando o comportamento de aceitação das tags em relação às suas medidas é possível definir perfis de usuários baseados nas medidas que mais enfatizam suas preferências ao utilizar uma tag.

A Figura 7.12 apresenta o comportamento de atribuição de cada usuário e a sua relação com as medidas de co-ocorrência, relevancia e popularidade. Foi verificado quais medidas eram predominantes para cada tag de cada perfil.

Percebe-se que houve uma predominância de tags aceitas que possuem a medida de relevância maior em relação as outras medidas. Ou seja, das tags aceitas o valor obtido para a medida de relevância é maior que o valor de popularidade e co-ocorrência.

Geralmente a primeira interpretação é considerar a relevância como medida mais importante para recomendação, excluindo a importância da popularidade e co-ocorrência do ranking, porém,

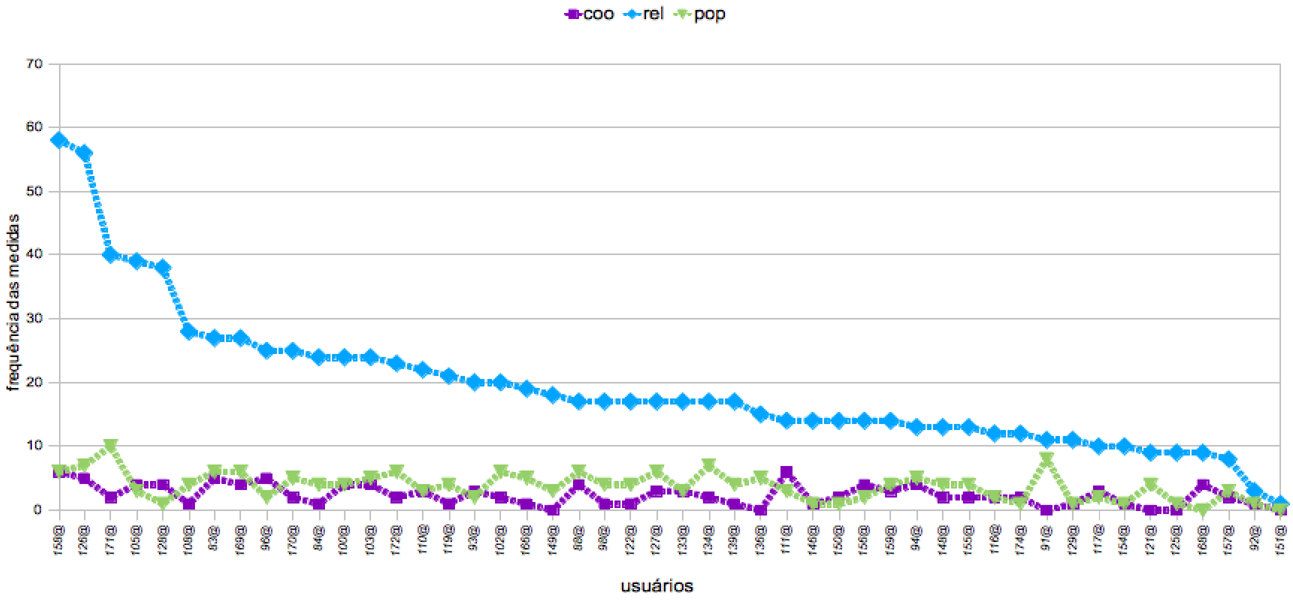

Figura 7.12: Comportamento geral dos usuários em relação a aceitação de tags. Há destaque para a relevância como medida mais influente na escolha das tags.

as três medidas possuem papéis distintos para a recomendação. A co-ocorrência traz as tags que ocorrem juntas com mais frequência de acordo com cada *query*, porém esta medida sozinha não é capaz de identificar os *i-users* identificados pela medida de relevância. Já a popularidade obtém as informações sobre a aceitação da tag pela comunidade em geral através da simetria entre a tag recomendada e a *query*. Esta informação não pode ser obtida pelo cálculo da relevância já que sua tarefa é excluir do ranking tags que foram incluídas repetidas vezes pelos mesmos usuários. Desta forma, cada medida tem seu papel para a melhoria da recomendação.

A observação do comportamento serviu para o desenvolvimento da recomendação personalizada que será apresentada no próximo capítulo.

8. UM MODELO DE RECOMENDAÇÃO PERSONALIZADA

Através da observação do comportamento de atribuição de cada usuário a recomendação de tags pode ser melhorada com uso de abordagens personalizadas.

O modelo de recomendação inicial baseado no conhecimento coletivo e nas *queries* utiliza a memória de atribuição da comunidade e é chamado de filtragem colaborativa baseada em memória. Em um novo modelo, o perfil de cada usuário pode ser definido por seu comportamento de atribuição passado. Este perfil receberá tags de outros usuários assim como na recomendação generalizada, porém as medidas utilizadas para a obtenção do ranking de tags utilizarão informações relativas ao seu histórico de tags, favorecendo as medidas mais importantes para o perfil. Esta abordagem de recomendação é conhecida como filtragem colaborativa baseada em modelo e tem por objetivo traçar o perfil do usuário para efetuar recomendações personalizadas e colaborativas.

Na fórmula apresentada para recomendação generalizada, o ranking final é dado por $mean(t,t_j) = \sqrt[3]{coo(t,t_j)^1 * rel(t,t_j)^1 * pop(t,t_j)^1}$, onde o peso de cada medida é seu expoente que neste caso é igual a um. O novo modelo considera as preferências relativas às medidas de popularidade, co-ocorrência e relevância. Para tanto, o controle desta informação será feito através de cada tag aceita pela recomendação que traz consigo uma tripla $t =< coo, rel, pop >$ com as medidas calculadas. Esta abordagem será utilizada para obter o modelo do perfil do usuário baseado em suas atribuições passadas. Para a personalização, o perfil de atribuição de cada usuário é definido como uma tripla $u =< w_c, w_r, w_p >$ onde cada w equivale à quantidade de ocorrências do maior valor entre os medidas de cada tag, onde o limite inferior é igual a um e o limite superior k pode ser definido conforme as preferências do sistema.

A maior medida da tripla t será verificada pela função $max(coo, rel, pop)$. No modelo personalizado o expoente para cada medida é alterado de acordo com cada seleção e a partir destes resultados é feito o cálculo do novo ranking. Seguindo esta abordagem para o novo modelo a tripla de cada usuário inicialmente também possuirá peso igual a um ($u =< 1, 1, 1 >$), pois não há interações suficientes para gerar recomendações personalizadas. Como exemplo, na Tabela 8.1 observa-se os valores das medidas de cada tag recomendada e aceita por um usuário. A cada vez que uma tag é atribuída através da recomendação, é verificado dentre as três medidas qual possui o maior valor. Por exemplo, a tag "kitchen" teve como maior medida a relevância, somando um ponto ao peso relacionado à relevância (w_r). Esta mesma abordagem é aplicada a todas as outras tags que fazem parte do histórico de atribuição neste exemplo.

Dentre as nove tags do exemplo o usuário atribuiu quatro que possuem a relevância como maior medida, então a relevância será o maior peso para o perfil e os expoentes substituídos para as medidas correspondentes para as próximas tags t_j:

$$meanProfile(t,t_j) = \sqrt[12]{coo(t,t_j)^3 * rel(t,t_j)^5 * pop(t,t_j)^4}$$

Tabela 8.1: Exemplo de atribuição de tags de um perfil de usuário para as medidas de co-ocorrência, popularidade e relevância.

tag	coo	rel	pop	$< w_c, w_r, w_p >$
kitchen	0.323	0.649	0.303	$< 1,2,1 >$
house	0.060	0.513	0.085	$< 1,3,1 >$
sandiego	0.185	0.125	0.159	$< 2,3,1 >$
holiday	0.054	0.029	0.215	$< 2,3,2 >$
utah	0.087	0.075	0.204	$< 2,3,3 >$
camping	0.083	0.048	0.109	$< 2,3,4 >$
sand	0.164	0.049	0.143	$< 3,3,4 >$
sanfrancisco	0.081	0.379	0.154	$< 3,4,4 >$
harbor	0.079	0.152	0.042	$< 3,5,4 >$

Assim como no ranking generalizado o resultado deste novo ranking apresentará uma lista de tags para recomendação. Entretanto, as ações passadas do usuário definirão qual o peso utilizado em cada expoente de acordo com o perfil obtido. Para o uso dos pesos na média geométrica ponderada foram sugeridas duas abordagens: substituição direta e definição fixa.

Na substituição direta, os pesos são as próprias quantidades obtidas pelo algoritmo, ou seja, os valores obtidos pela tripla de cada usuário. Porém, ao passo que cada tag recomendada é aceita por um perfil, os valores crescem, criando expoentes com números altos. Para solucionar este problema pode-se fazer uma redução dos pesos a partir do menor peso da tripla.

Utilizando os pesos do exemplo apresentado (Tabela 8.1), a tripla contendo os valores $< 3,5,4 >$ tem como menor peso a co-ocorrência (w_c). Para reduzir estes números é necessário que o menor peso não fique com valores negativos ou zero. Portanto, na redução geral dos pesos será subtraído k de todos os pesos, de modo que o limite inferior do menor peso volte a ser igual a um.

Já na abordagem de definição fixa, os valores atribuídos aos expoentes de cada medida são definidos pelo ranking entre pesos. Por exemplo, para a tripla $u =< 3,5,4 >$ os pesos serão constantes atribuídas conforme a ordem de importância que cada medida tem para o perfil. Os expoentes definidos para utilização desta abordagem foram os valores 1, 2 e 3. A importância das medidas é dada pela sua frequência de utilização, assim medidas maiores terão pesos maiores. Para o exemplo da tripla $u =< 3,5,4 >$ os pesos serão substituídos na média geometrica ponderada por $< 1,3,2 >$ para serem atribuídos aos expoentes, pois neste exemplo a primeira medida mais importante da tripla é a relevância, seguida da popularidade e logo a co-ocorrência.

8.1 Resultados Preliminares

O modelo de recomendação personalizada foi criado a partir das observações do tipo de tag atribuída e do histórico das ações dos usuários.

Na Figura 8.1 podem ser observados os resultados completos de um dos perfis resultantes das atribuições durante o experimento. Neste exemplo o perfil aceitou 52 tags recomendadas que dentre elas 40 possuem a relevância como medida predominante. Portanto, os pesos do perfil para as três medidas serão $< 2, 40, 10 >$, conforme seu histórico através da recomendação generalizada. Este perfil foi utilizado para gerar os resultados da abordagem personalizada usando os expoentes de forma fixa e direta.

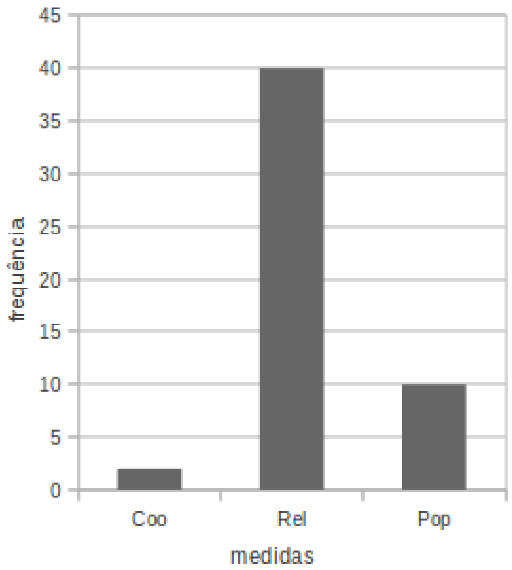

Figura 8.1: Comportamento de atribuição de um dos participantes do experimento em relação às medidas de recomendação.

Para apresentar as duas estratégias de utilização dos expoentes foi feito o uso da palavra-chave "nature" como *query*. A ilustração da diferença entre os resultados do ranking generalizado e o personalizado está representada na Tabela 14.4. Os resultados da utilização da média geométrica ponderada e as abordagens para o uso dos expoentes resultaram em recomendações distintas da abordagem generalizada.

Tabela 8.2: Recomendação generalizada em relação à recomendação personalizada utilizando a *query* "nature" e apresentando os cinco primeiros resultados

Abordagem Generalizada	Abordagem Personalizada $< 2, 40, 10 >$	Abordagem Personalizada $< 1, 3, 2 >$
butterfly	bird	bird
bird	flower	butterfly
wildlife	water	macro

Para esta abordagem foram buscadas as vinte tags mais frequentes no *dataset*, e a partir destas foram aplicadas as medidas e os respectivos pesos para o usuário extraído no exemplo. O posicionamento diferente das tags apresentado pelos resultados permite a entrega de tags distintas para cada abordagem utilizada. Esta abordagem não fez parte do experimento, pois foi desenvolvida a partir dos resultados da recomendação generalizada, porém a aplicação para outras *queries* está apresentada na Apêndice 14. A observação dos resultados demonstra um comportamento promissor e que merece atenção em trabalhos futuros.

9. CONCLUSÃO

A análise e organização de dados a partir de comunidades *online* como redes de mídia social, é um tema recente e que está em ascensão pela variedade de informações disponibilizadas e pelas diversas possibilidades de exploração que fornecem. As tags são uma das formas mais recentes para categorização/organização de conteúdo utilizadas livremente em sistemas de compartilhamento de imagens e vídeos como *Flickr*, *Picasa* e *Youtube*.

Neste trabalho foi desenvolvida uma *engine* de recomendação de tags com o objetivo de sugerir tags relevantes para aumentar o reuso das palavras-chave e proporcionar um vocabulário mais homogêneo para o processo de construção de uma *folksonomia*. Para tanto, foi utilizada a medida de co-ocorrência e desenvolvidas as medidas de relevância e popularidade com o objetivo de recomendar tags mais relevantes a partir das *queries* dos usuários.

A utilização da relevância foi necessária pela observação do comportamento dos resultados baseados somente na co-ocorrência relacionados aos *i-users*. Esta medida permitiu que resultados estimados através de poucos usuários influentes sofressem rebaixamento no posicionamento no ranking, deixando espaço para recomendação de tags mais relevantes. Da mesma forma a medida de popularidade trouxe resultados relevantes da base do ranking e retirou tags menos populares do topo do ranking.

Através do experimento de recomendação generalizada, foi possível verificar o comportamento de aceitação dos usuários e a influência da recomendação em relação ao *long tail* de comparação entre *queries* e tags recomendadas. Houve uma visível diminuição do *long tail* de distribuição das tags recomendadas demonstrando a melhoria do reuso das palavras recomendadas. Além disso, durante a análise dos dados percebeu-se que a recomendação proporcionou maior quantidade de atribuições de tags recomendadas em relação às *queries*, além de um vocabulário mais homogêneo no *dataset* devido ao reuso das recomendações.

Outra importante contribuição deste trabalho foram os resultados que as medidas desenvolvidas proporcionaram para o posicionamento das tags no ranking final. A atribuição da recomendação apresentou melhor desempenho para as tags sugeridas no topo do ranking conforme resultado para precisão de $P@5$ em relação a $P@10$. Estes resultados confirmam a importância da utilização de medidas de exclusão de tags do topo do ranking para a entrega de melhores resultados na recomendação e quando as tags possuirem comportamento referente a *i-users* e baixa popularidade.

Além da contribuição para a recomendação generalizada, foi também importante a observação dos resultados de acordo com as medidas preferidas pelos usuários conforme as tags escolhidas. De acordo com o histórico do perfil, foi possível modelar a recomendação personalizada de tags utilizando os perfis gerados pelos históricos de atribuição de tags recomendadas. Os resultados preliminares demonstraram bons resultados principalmente quando comparados à abordagem generalizada, entregando resultados distintos de acordo com o perfil de atribuição obtido.

Vale a pena citar que a recomendação teve especial aceitação principalmente quando foram apresentados itens referentes a lugares ou quando foi possível identificar a localização geográfica do conteúdo. Com base nestas observações, outras possibilidades de aplicação e recomendação foram levantadas para utilização em trabalhos futuros.

9.1 Trabalhos Futuros

O processo de análise dos dados do experimento e os resultados obtidos proporcionaram, além das contribuições já relatadas, a possibilidade de trabalhos futuros e melhoria da recomendação por abordagens mais específicas.

A popularização de dispositivos móveis como *smartphones*[1] e *tablets*[2], tornou possível obter informações relacionadas à localização geográfica através da latitute e longitude obtidas por estes aparelhos. Informações referentes à geolocalização dos usuários são valiosas fontes de dados para recomendação. A abordagem para recomendação basedada em localização é importante principalmente quando não existe um histórico de interações suficientes para a geração de recomendação.

Através da API[3] do *Flickr* é possível obter conjuntos de tags em um raio definido em relação às coordenadas de longitude e latitude contidas nas imagens com formato *jpeg*. O processo de utilização de tags que possuem latitute e longitude é conhecido como *geotagging*. Esta abordagem utilizada juntamente com as medidas desenvolvidas neste trabalho pode proporcionar aos usuários de *smartphones* e *tablets* receber recomendação de tags no momento em que tirarem fotos com seus aparelhos através da verificação de sua localização como fonte de informação para gerar sugestões de tags de acordo com o raio de localização do histórico de outros usuários.

Além de novas abordagens para a utilização das tags, também poderia ser efetuado um experimento com as medidas de recomendação personalizada. Este experimento necessitará de usuários fixos para recomendar tags e logo personalizar seus resultados pelo histórico de atribuição obtido. Outra possibilidade para recomendação de tags é utilizar a co-ocorrência de tags do perfil em conjunto com a co-ocorrência das tags da comunidade para obter um vocabulário personalizado e colaborativo ao mesmo tempo.

---

[1] Telefone celular com funcionalidades avançadas pela utilização de um sistema operacional além de GPS (*Global Positioning System* - sistema de posicionamento global), acesso a internet, câmera fotográfica, etc.

[2] Computador móvel com sistema operacional, acesso a internet, etc. Maior que um smartphone mas com funcionalidades similares.

[3] *Application Programming Interface* -(interface para programação de aplicações).

REFERÊNCIAS BIBLIOGRÁFICAS

[And06] Anderson, C. "The Long Tail: Why the Future of Business Is Selling Less of More". Hyperion, New York, 2006, 256p.

[APTO09] Amatriain, J; Pujol, J. M.; Tintarev, N; Oliver, N. "Rate it again: increasing recommendation accuracy by user re-rating". In: 3rd ACM Conference on Recommender systems, RecSys'09, 2009, pp. 173–180.

[AT05] Adomavicius, G.; Tuzhilin, A. "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.*, vol. 17-6, Jun 2005, pp. 734–749.

[BHK98] Breese, J. S.; Heckerman, D.; Kadie, C. "Empirical analysis of predictive algorithms for collaborative filtering". In: 14th Conference on Uncertainty in Artificial Intelligence, 1998, pp. 43–52.

[BM06] Brooks, C. H.; Montanez, N. "Improved annotation of the blogosphere via autotagging and hierarchical clustering". In: 15th international conference on World Wide Web, 2006, pp. 625–632,

[BS97] Balabanovic, M.; Shoham, Y. "Fab: Content-based, collaborative recommendation". *Communications of the ACM*, vol. 40-3, Mar 1997, pp. 66–72.

[Bur07] Burke, R. "The adaptive web". Berlin, Heidelberg: Springer-Verlag, 2007, 763p.

[DDGR07] Das, A. S.; Datar, M.; Garg, A.; Rajaram, S. "Google news personalization: scalable online collaborative filtering". In: 16th international conference on World Wide Web, 2007, pp. 271–280.

[DFT10] Dattolo, A.; Ferrara, F.; Tasso, C. "The role of tags for recommendation: a survey". In: 3rd International Conference on Human System Interaction, 2010, pp. 548–555.

[DK04] Deshpande, M.; Karypis, G. "Item-based top-n recommendation algorithms". *ACM Transactions on Information Systems*, vol. 22-1, Jan 2004, 143–177.

[EV03] Eirinaki, M.; Vazirgiannis, M. "Web mining for web personalization". *ACM Transactions on Internet Technology*, vol.3-1, Fev 2003, 3:1–27.

[Gil09] Gilleland, M. "Levenshtein distance, in three flavors", Capturado em: http://www.merriampark.com/ld.htm, Outubro 2012.

[GSRM09] Gemmell, J.; Schimoler, T.; Ramezani, M.; Mobasher, B. “Adapting K-nearest neighbor for tag recommendation in folksonomies”. In: 7th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems, 2009.

[Hof99] Hofmann, T. “Probabilistic latent semantic analysis”. In 22nd annual international ACM SIGIR Conference on Research and Development in Information retrieval, 1999, pp. 289–296.

[Hof04] Hofmann, T. “Latent semantic models for collaborative filtering”. *ACM Transactions on Information Systems*, vol. 22-1, Jan 2004, pp. 89–115.

[JEHS09] Jäschke, R.; Eisterlehner, F.; Hotho, A.; Stumme, G. “Testing and evaluating tag recommenders in a live system”. In: 3rd ACM conference on Recommender systems, 2009, pp. 369–372.

[JMH+08] Jäschke, R.; Marinho, L.; Hotho, A.; Lars, S.; Gerd, S. “Tag recommendations in social bookmarking systems”. *AI Communications*, vol. 21, Dez 2008, pp. 231–247.

[JSK10] Jawaheer, G.; Szomszor, M.; Kostkova, P. “Characterisation of explicit feedback in an online music’ recommendation service”. In 4th ACM conference on Recommender systems, 2010, pp. 317–320.

[KCK06] Kennedy, L. S.; Chang, S. F., Kozintsev, I. V. “To search or to label?: predicting the performance of search-based automatic image classifiers”. In 8th ACM international workshop on Multimedia information retrieval, 2006, pp. 249–258.

[KSK97] Kamba, T.; Sakagami, H.; Koseki, Y. “Anatagonomy: a personalized newspaper on the world wide web”. *International Journal of Human-Computer Studies*, vol. 46–6, Jun 1997, pp. 789–803.

[KT03] Kelly, D.; Teevan, J. “Implicit feedback for inferring user preference: a bibliography”. *ACM SIGIR*, vol. 37-2, Sep 2003, pp. 18–28.

[LdGS+09] Lops, P.; Gemmis, M.; Semeraro, G.; Gissi, P.; Musto,C.; Narducci, F. “Content-based filtering with tags: The first system”. In 9th International Conference on Intelligent Systems Design and Applications, 2009, pp. 255–260

[LDP10] Liu, J.; Dolan, P.; Pedersen, E. R. “Personalized news recommendation based on click behavior”. In 14th international conference on Intelligent user interfaces, 2010, pp. 31–40.

[LM10] Lipczak, M; Milios, E. E. “Learning in efficient tag recommendation”. In 4th ACM Conference on Recommender Systems, 2010, pp. 167–174.

[LMWdO10] Lopes, G. R.; Moro, M. M.; Wives, L. K.; Oliveira, J. P. M. "Collaboration recommendation on academic social networks". In 9th International Conference on Advances in Conceptual Modeling, 2010, pp. 190–199.

[LSY03] Linden, G.; Smith, B.; York, J. "Amazon.com recommendations: Item-to-item collaborative filtering". *IEEE Internet Computing*, vol. 7-1, Jan 2003, pp. 76–80.

[LXL+09] Liang, H.; Xu, Y.; Li, Y.; Nayak, R.; Weng, L. T. "Personalized recommender systems integrating social tags and item taxonomy". In IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, 2009, pp. 540–547.

[Mis06] Mishne, G. "Autotag: a collaborative approach to automated tag assignment for weblog posts". In 15th International Conference on World Wide Web, 2006, pp. 953–954.

[Mob07] Mobasher, B. "Data mining for web personalization". The Adaptive Web, Berlin: Spring-Verlag, 2007, pp. 90–135.

[Nic98] Nichols, D. M. "Implicit rating and filtering". In 5th DELOS Workshop on Filtering and Collaborative Filtering, 1998, pp. 31–36.

[OLL08] Oh, J.; Lee, S.; Lee, E. "A user modeling using implicit feedback for effective recommender system". In International Conference on Convergence and Hybrid Information Technology, 2008, pp. 155–158.

[PB07] Pazzani, M. J.; Billsus, D. "Content-based recommendation systems". The Adaptive Web, Berlin: Spring-Verlag, 2007, pp. 325–341.

[PdS08] Pereira, R.; Silva, S. R. P. "Folksonomias: uma análise crítica focada na interação e na natureza da técnica". In VIII Brazilian Symposium on Human Factors in Computing Systems, 2008, pp. 126–135.

[SGMB08] Shepitsen, A.; Gemmell, J.; Mobasher, B.; Burke, R. "Personalized recommendation in social tagging systems using hierarchical clustering". In ACM conference on Recommender systems, 2008, pp. 259–266.

[SHB05] Shani, G.; Heckerman, D.; Brafman, R. I. "An mdp-based recommender system". *Journal of Machine Learning Research*, vol. 6, Dez 2005, pp. 1265–1295.

[SK09] Su, X.; Khoshgoftaar, T. M. "A survey of collaborative filtering techniques". *Advances in Artificial Intelligence*, vol. 2009, Jan 2009, pp. 2–4.

[SKK10] Strohmaier, M.; Körner, C.; Kern, R. "Why do users tag? detecting users' motivation for tagging in social tagging systems". In 4th International AAAI Conference on Weblogs and Social Media, 2010, pp. 339–342.

[SKL09] Spiegel, S.; Kunegis, J.; Li, F. "Hydra: a hybrid recommender system cross-linked rating and content information". In 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management, 2009, pp. 75–80.

[SLH09] Shi, Y.; Larson, M.; Hanjalic, A. "Exploiting user similarity based on rated-item pools for improved user-based collaborative filtering". In 3rd ACM conference on Recommender systems, 2009, pp. 125–132.

[SM86] Salton, G.; McGill, M. J. "Introduction to Modern Information Retrieval". New York:McGraw-Hill, 1986, 400p.

[SMWH10] Sadikov, E.; Madhavan, J.; Wang, L.; Halevy, A. "Clustering query refinements by user intent". In 19th international conference on World wide web, 2010, pp. 841–850.

[SS09] Siersdorfer, S.; Sizov, S. "Social recommender systems for web 2.0 folksonomies". In 20th ACM conference on Hypertext and hypermedia, 2009, pp. 261–270.

[SvZ08] Sigurbjörnsson, B.; Zwol, R. "Flickr tag recommendation based on collective knowledge". In 17th international conference on World Wide Web, 2008, pp. 327–336.

[XJL08] Xu, S.; Jiang, H.; Lau, F. C. M. "Personalized online document, image and video recommendation via commodity eye-tracking". In ACM conference on Recommender systems, 2008, pp. 83–90.

[ZC08] V. Zanardi and L. Capra. "Social ranking: Finding relevant content in web 2.0". In Intl. Workshop on Recommender Systems, 2008, pp. 2–6.

10. QUESTIONÁRIO APLICADO

10.1 Recomendação de Tags - Survey

Por favor responda as seguintes questões para ajudar-nos a entender sua experiência em nosso sistema de recomendação de tags.

a) Você costuma categorizar suas fotos através da utilização de tags?

1. Sempre
2. As vezes
3. Nunca
4. Outro

b) Baseado na sua própria experiência: Quais são as razões que você acha para as pessoas não atribuírem tags em suas fotos?

1. Atribuição de tags é um trabalho exaustívo.
2. As pessoas não sabem por que usá-las.
3. As pessoas não sabem quais conjuntos de tags são boas para atribuição.
4. Outro

c) Você acha que recomendar tags ajuda na atribuição de tags?

1. Sim.
2. As vezes.
3. Não ajuda.
4. Outro

d) Questão opcional: Deixe seu comentários, sugestões sobre os resultados de nossa recomendação.

11. DICIONÁRIO DE DADOS

Tabela 11.1: Dicionário de dados para a tabela *ownernew*

Ownernew: relação que armazena informações dos usuários donos das tags atribuídas			
Atributo	Descrição	Tipo	Restrições
id_ownernew	Atributo de identificação do usuário	Inteiro - autoinc.	Chave-primária
email	endereço eletrônico do usuário	Varchar	Não-nulo
rel_owner	Atributo relativo à medida de relevância para personalização.	Inteiro	Sem restrições
pop_owner	Atributo relativo à medida de popularidade para personalização.	Inteiro	Sem restrições
coo_owner	Atributo relativo à medida de co-ocorrência para personalização.	Inteiro	Sem restrições
experience	Atributo relativo à experiência de atribuição de tags do usuário.	Char	Não-nulo

Tabela 11.2: Dicionário de dados para a tabela *itemnew_new*

Itemnew: relação que armazena informações dos itens/objetos postados			
Atributo	Descrição	Tipo	Restrições
id_itemnew	Atributo de identificação do item	Inteiro - autoinc.	Chave-primária
file_name	Atributo relativo ao nome do arquivo.	Varchar	Não-nulo
id_ownernew	Atributo relativo a chave primária da tabela ownernew.	Inteiro	Chave estrangeira, não-nulo.

Tabela 11.3: Dicionário de dados para a tabela *tagitem*

Tagitem: relação que armazena informações das tags atribuídas aos objetos			
Atributo	Descrição	Tipo	Restrições
id_tag	Atributo de identificação da tag	Inteiro - autoinc.	Chave-primária
coo	Atributo relativo a medida de co-ocorrência de cada tag atribuída.	Real	Default 0
pop	Atributo relativo a medida de popularidade de cada tag atribuída.	Real	Default 0
rel	Atributo relativo a medida de relevância de cada tag atribuída.	Real	Default 0
suggested	Atributo que controla se a tag foi recomendada.	Char	Default 0
ranking	Atributo que guarda a posição da tag no ranking quando recomendada.	Inteiro	Default 0
id_itemnew	Atributo relativo a tabela itemnew.	Inteiro	Chave-estrangeira
id_ownernew	Atributo relativo a tabela ownernew.	Inteiro	Chave-estrangeira

12. RECOMENDAÇÃO

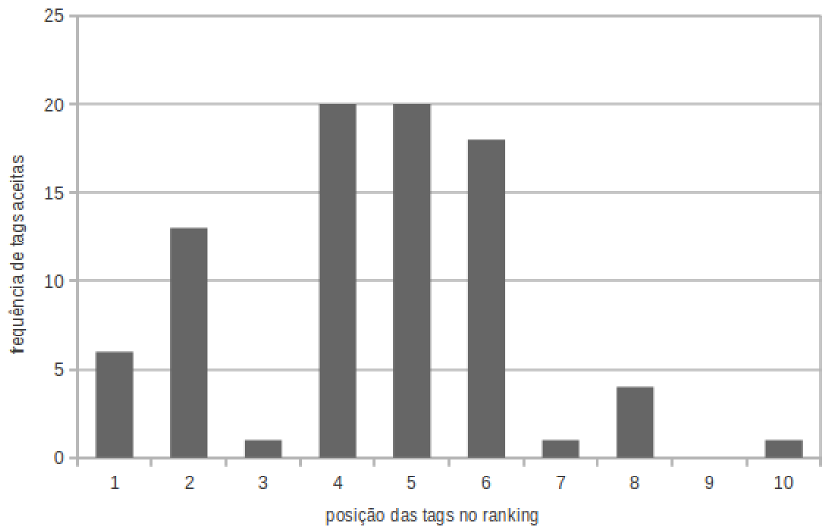

Figura 12.1: Posição das tags aceitas no ranking para o item $r1$.

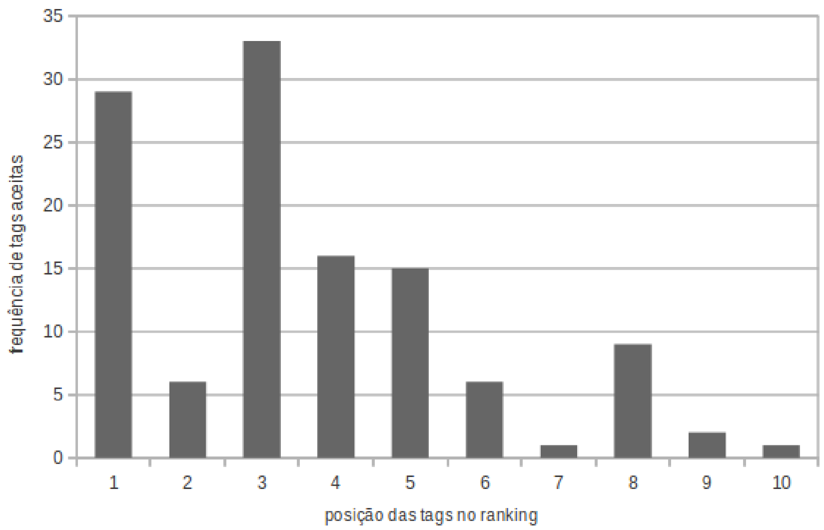

Figura 12.2: Posição das tags aceitas no ranking para o item $r2$.

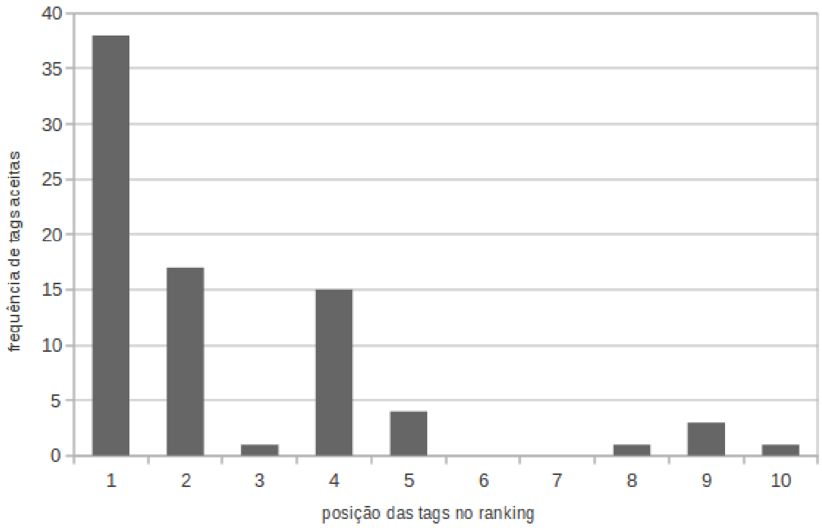

Figura 12.3: Posição das tags aceitas no ranking para o item $r3$.

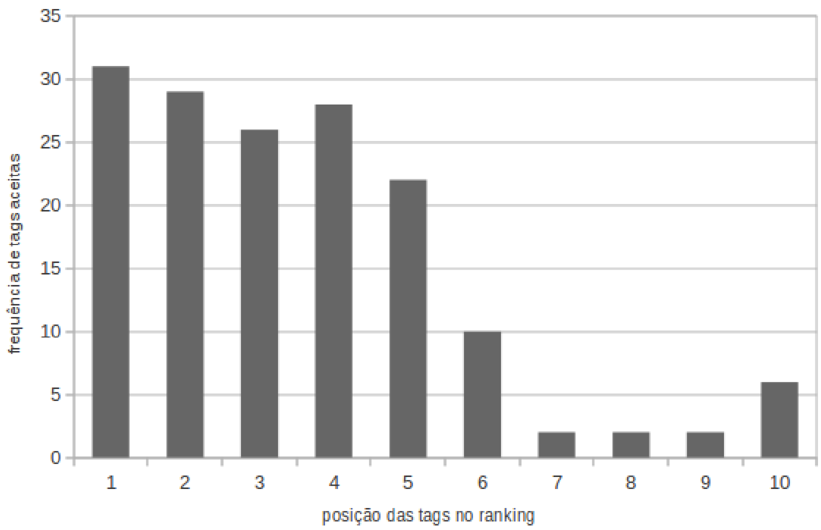

Figura 12.4: Posição das tags aceitas no ranking para o item $r4$.

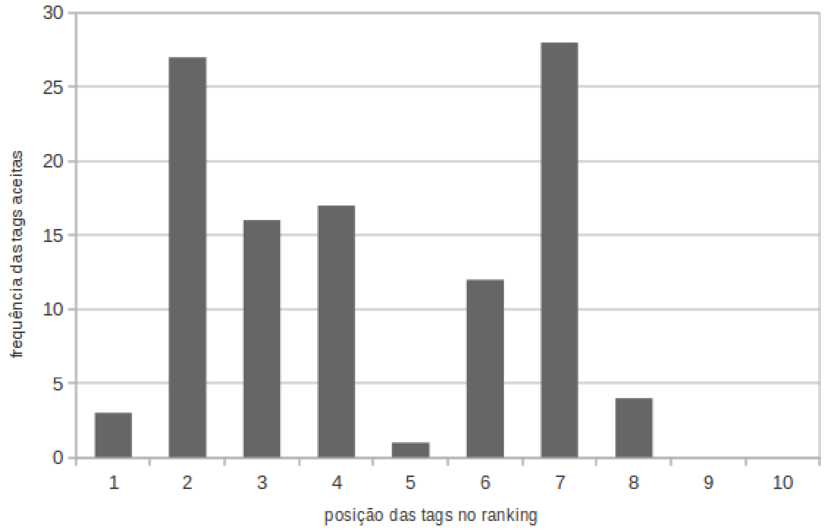

Figura 12.5: Posição das tags aceitas no ranking para o item $r5$.

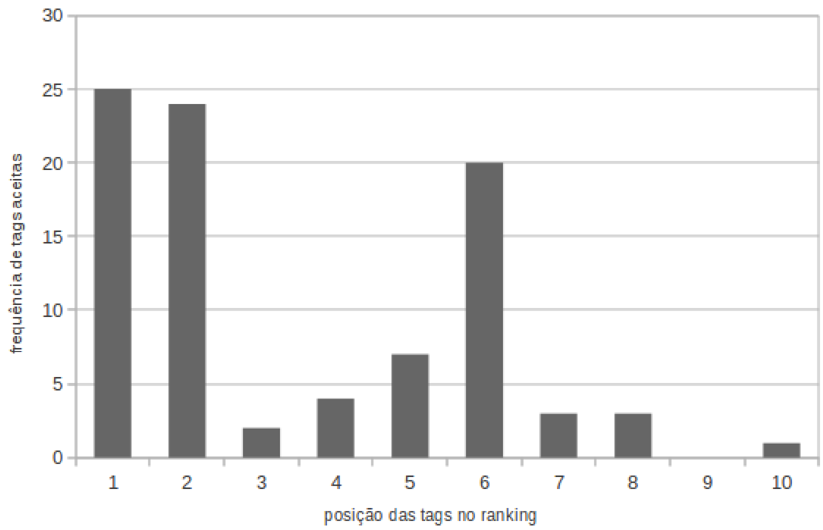

Figura 12.6: Posição das tags aceitas no ranking para o item $r6$.

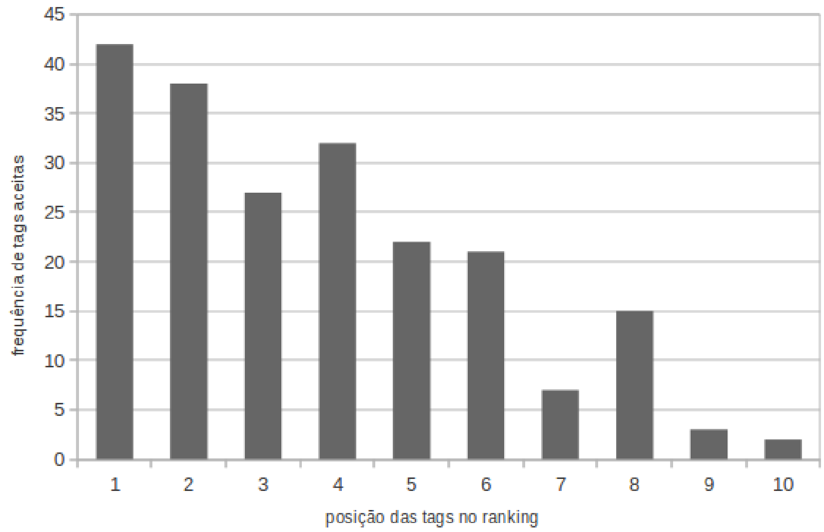

Figura 12.7: Posição das tags aceitas no ranking para o item $r7$.

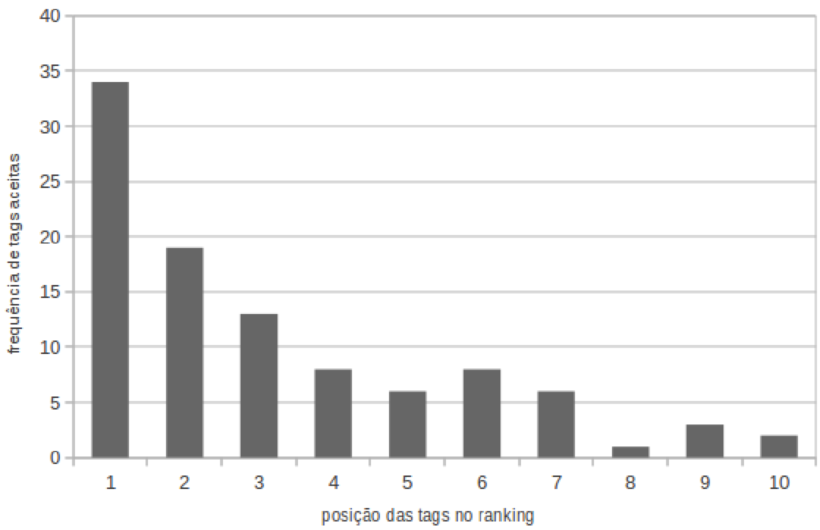

Figura 12.8: Posição das tags aceitas no ranking para o item $r8$.

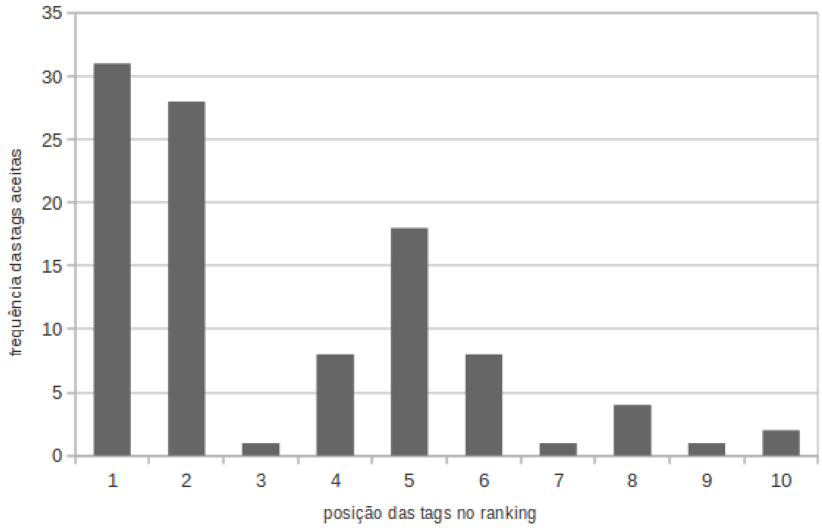

Figura 12.9: Posição das tags aceitas no ranking para o item $r9$.

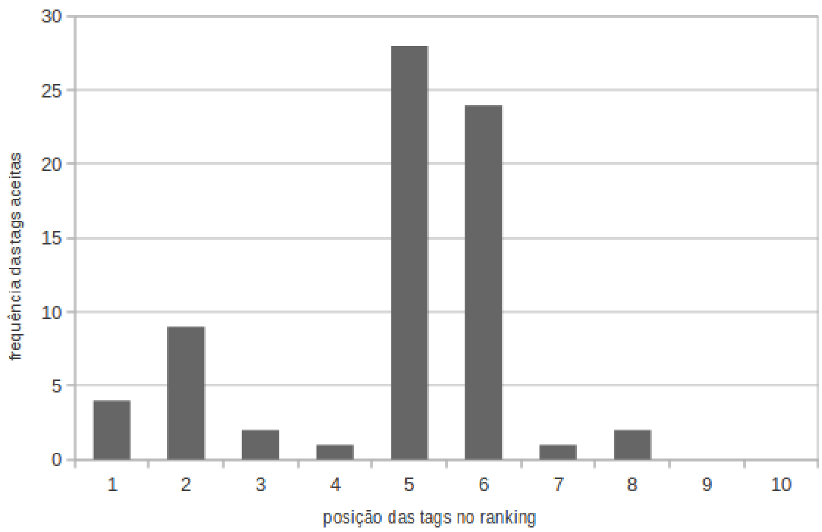

Figura 12.10: Posição das tags aceitas no ranking para o item $r10$.

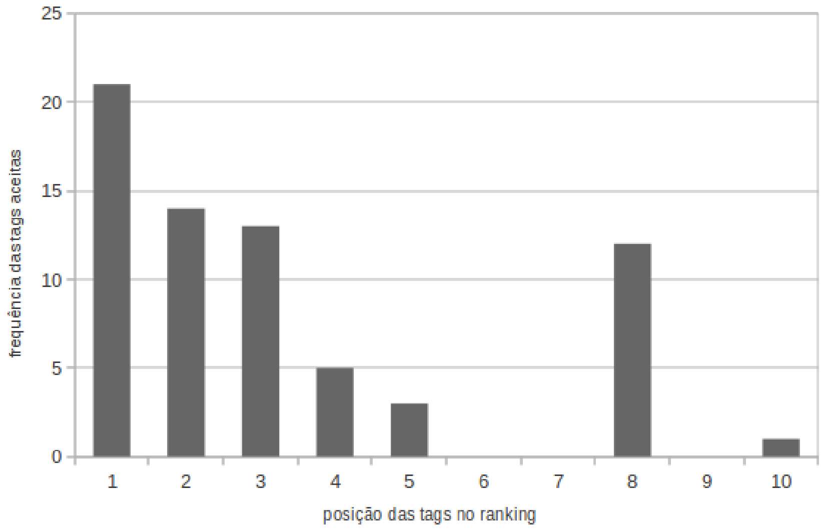

Figura 12.11: Posição das tags aceitas no ranking para o item $r11$.

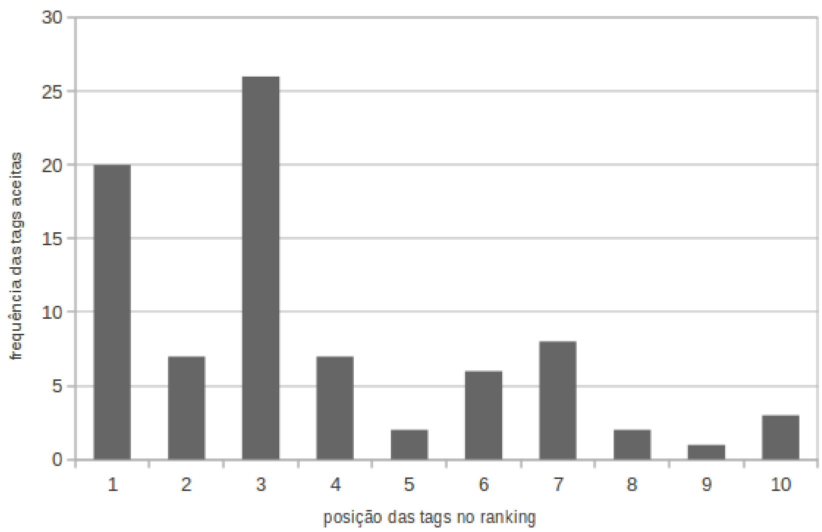

Figura 12.12: Posição das tags aceitas no ranking para o item $r12$.

13. GRÁFICOS DE PRECISÃO

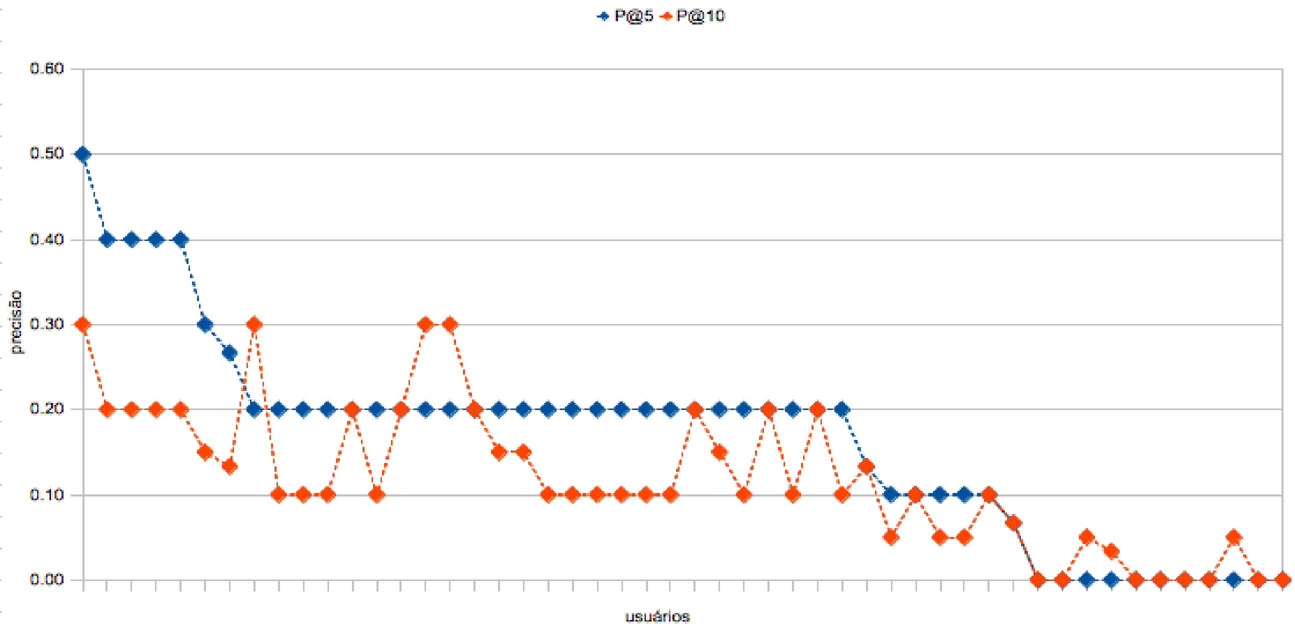

Figura 13.1: Resultados da precisão do objeto r_1 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

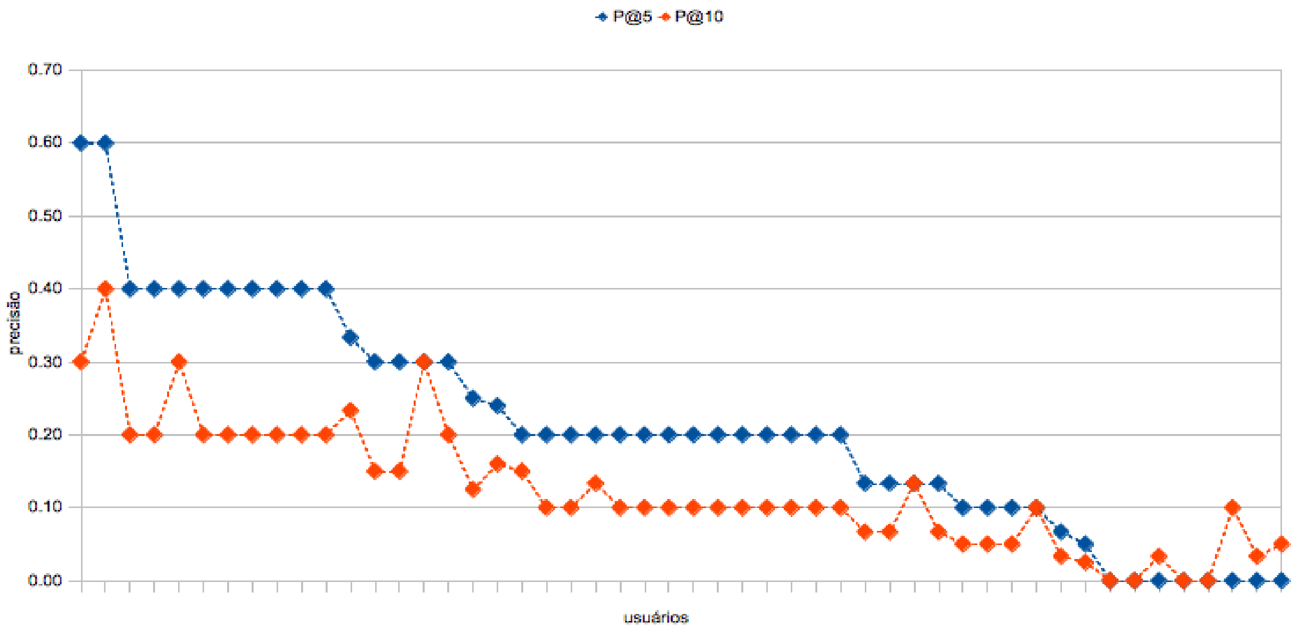

Figura 13.2: Resultados da precisão do objeto r_2 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

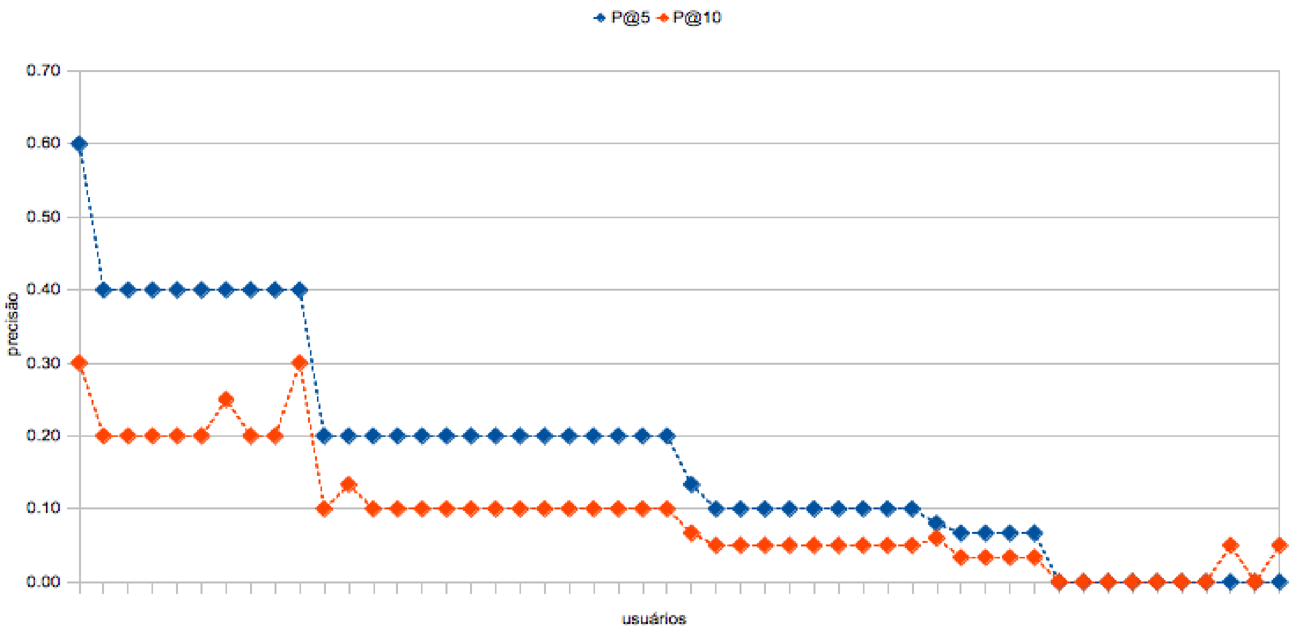

Figura 13.3: Resultados da precisão do objeto r_3 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

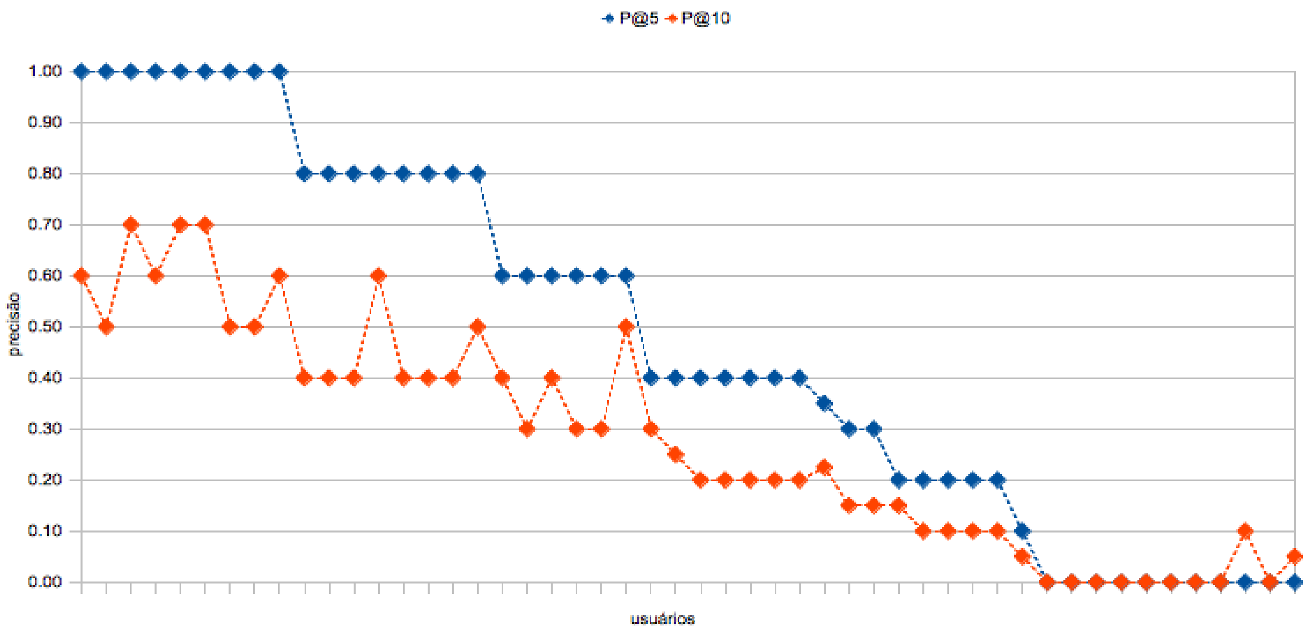

Figura 13.4: Resultados da precisão do objeto r_4 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

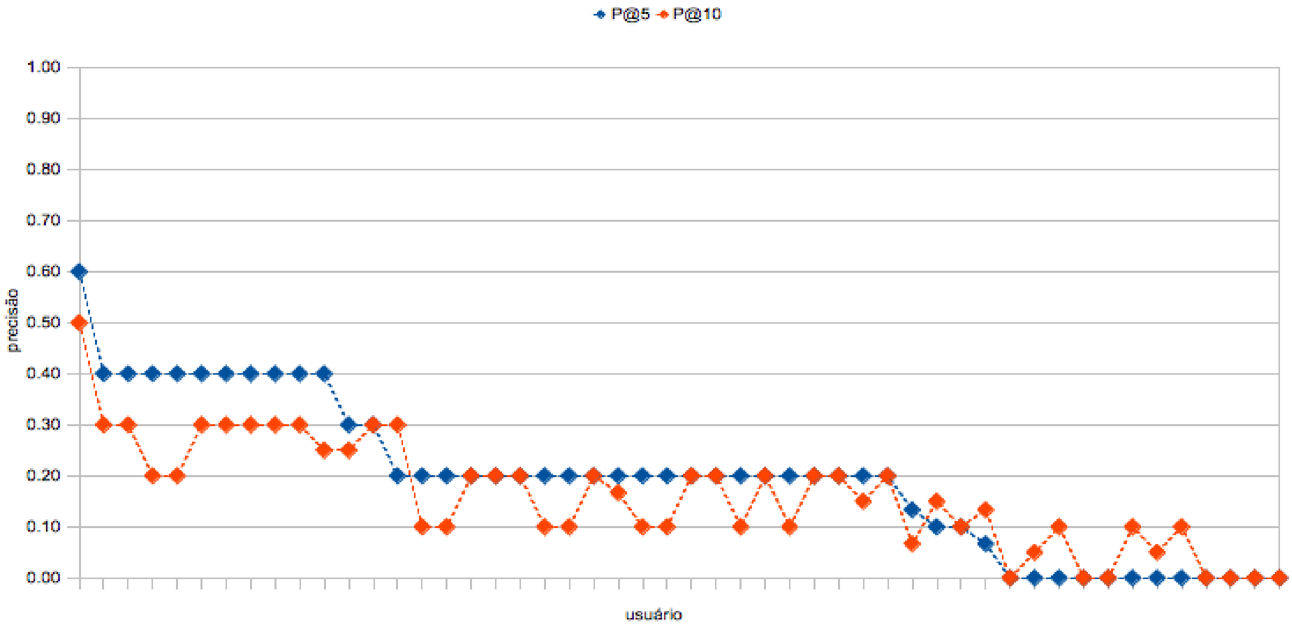

Figura 13.5: Resultados da precisão do objeto r_5 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

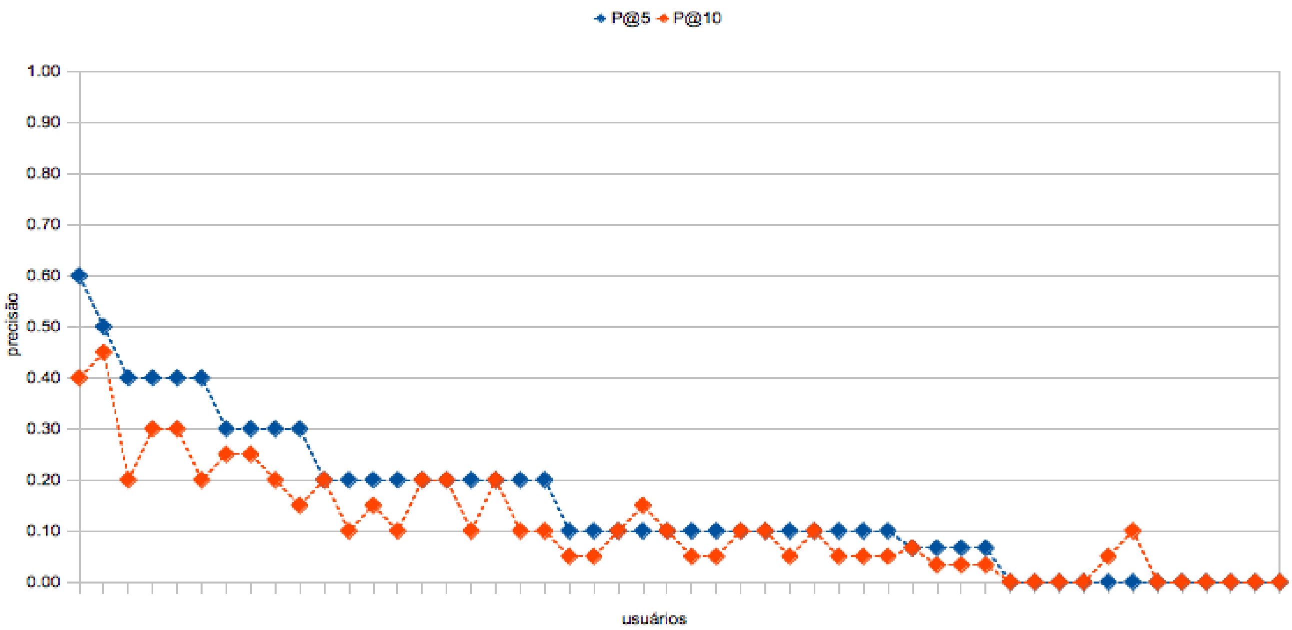

Figura 13.6: Resultados da precisão do objeto r_6 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

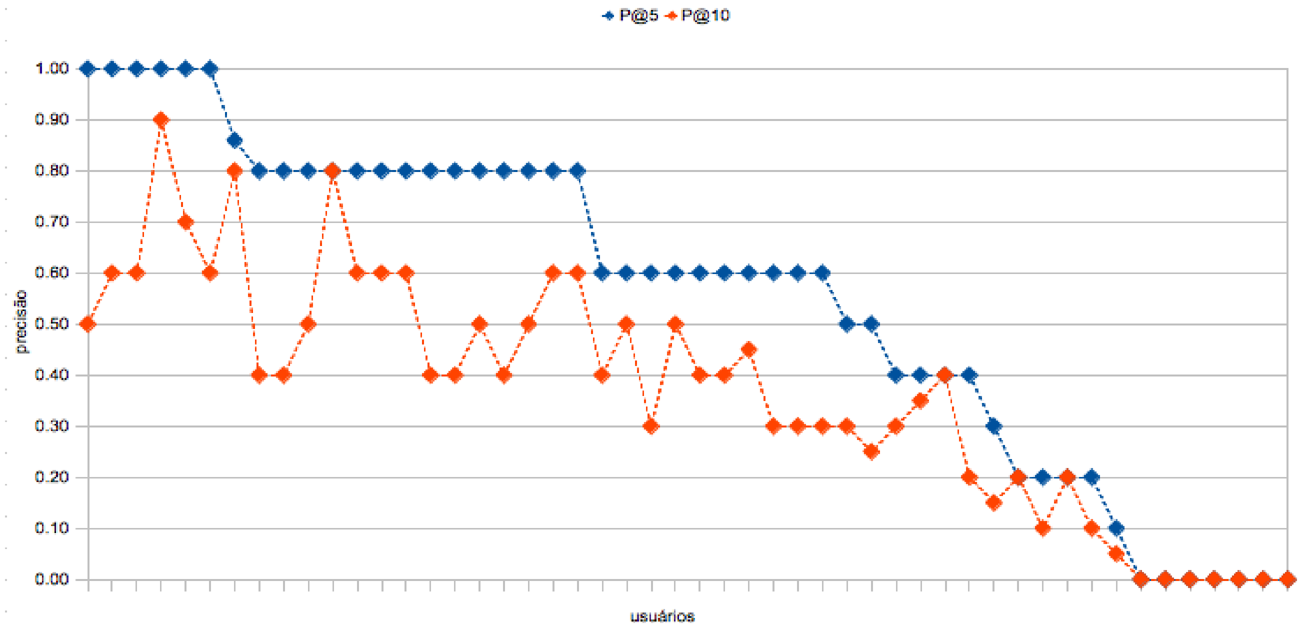

Figura 13.7: Resultados da precisão do objeto r_7 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

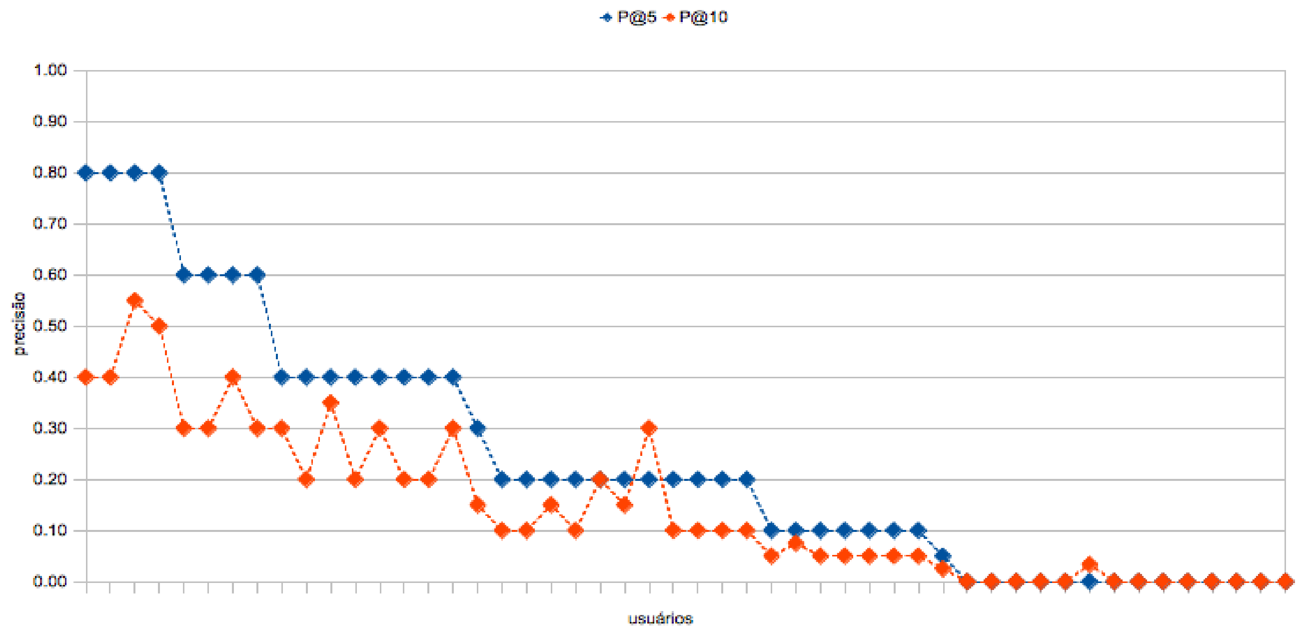

Figura 13.8: Resultados da precisão do objeto r_8 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

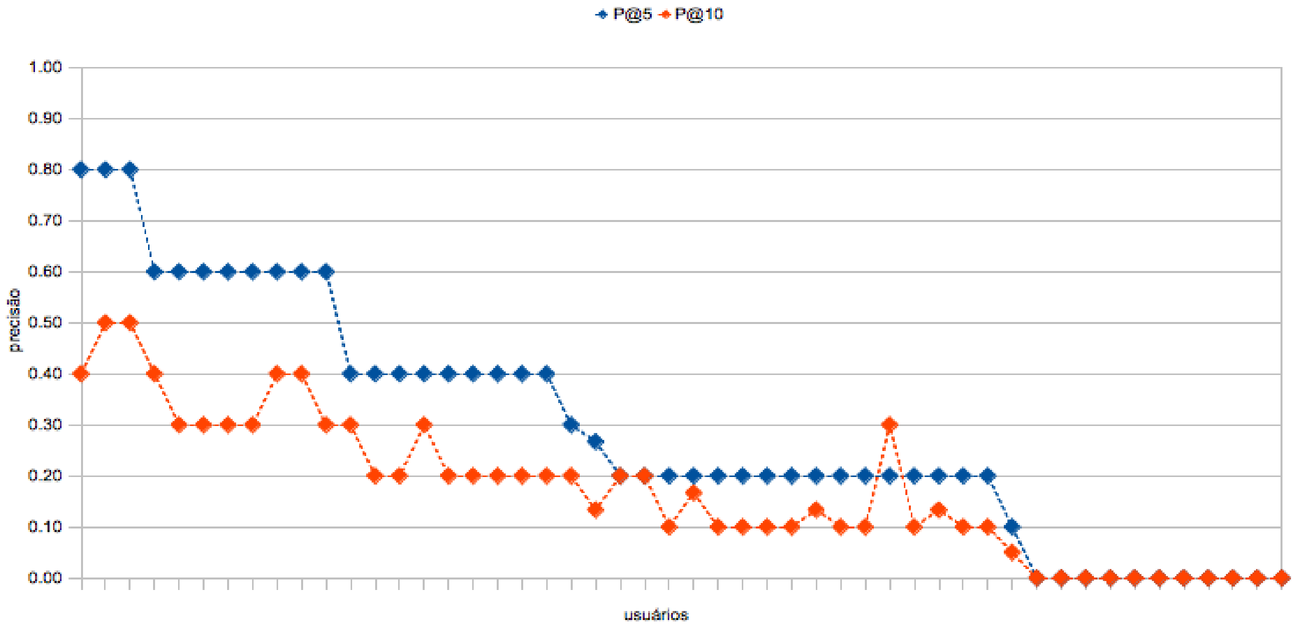

Figura 13.9: Resultados da precisão do objeto r_9 para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

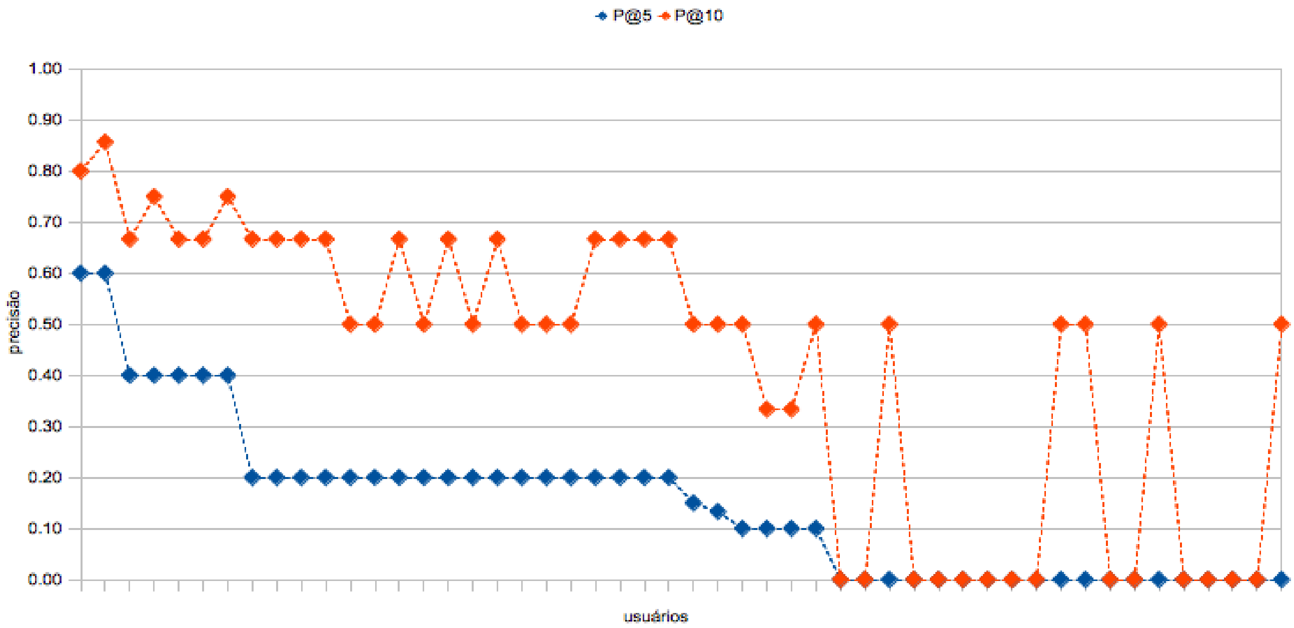

Figura 13.10: Resultados da precisão do objeto r_{10} para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

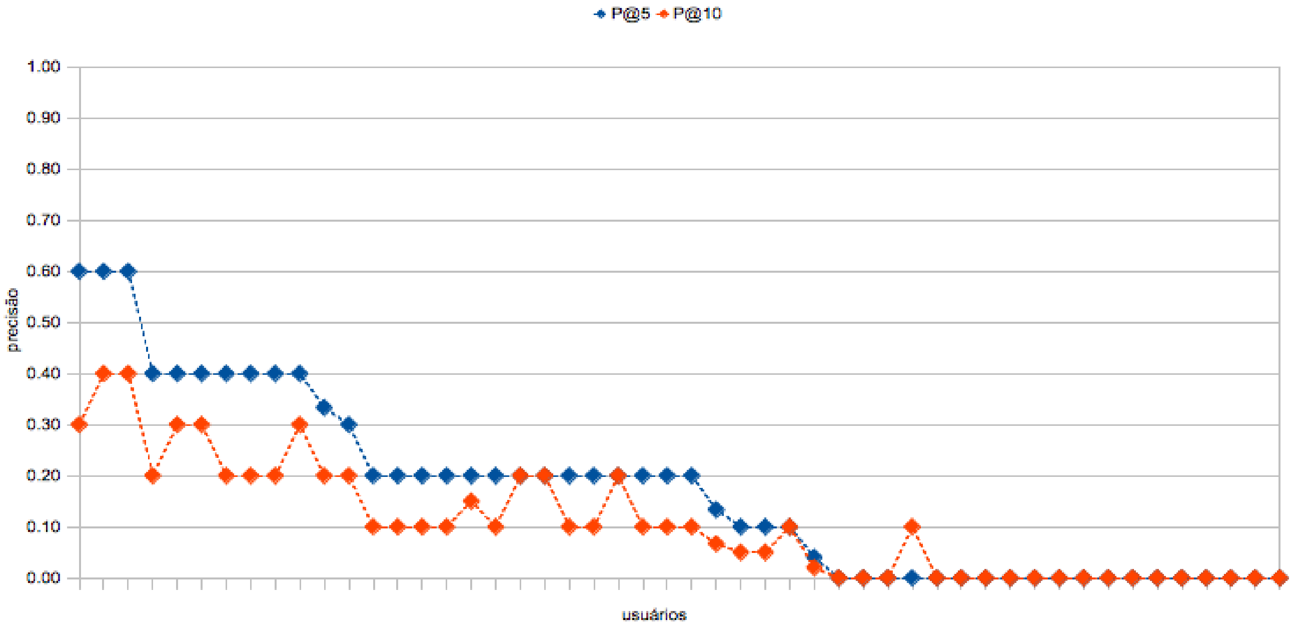

Figura 13.11: Resultados da precisão do objeto r_{11} para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

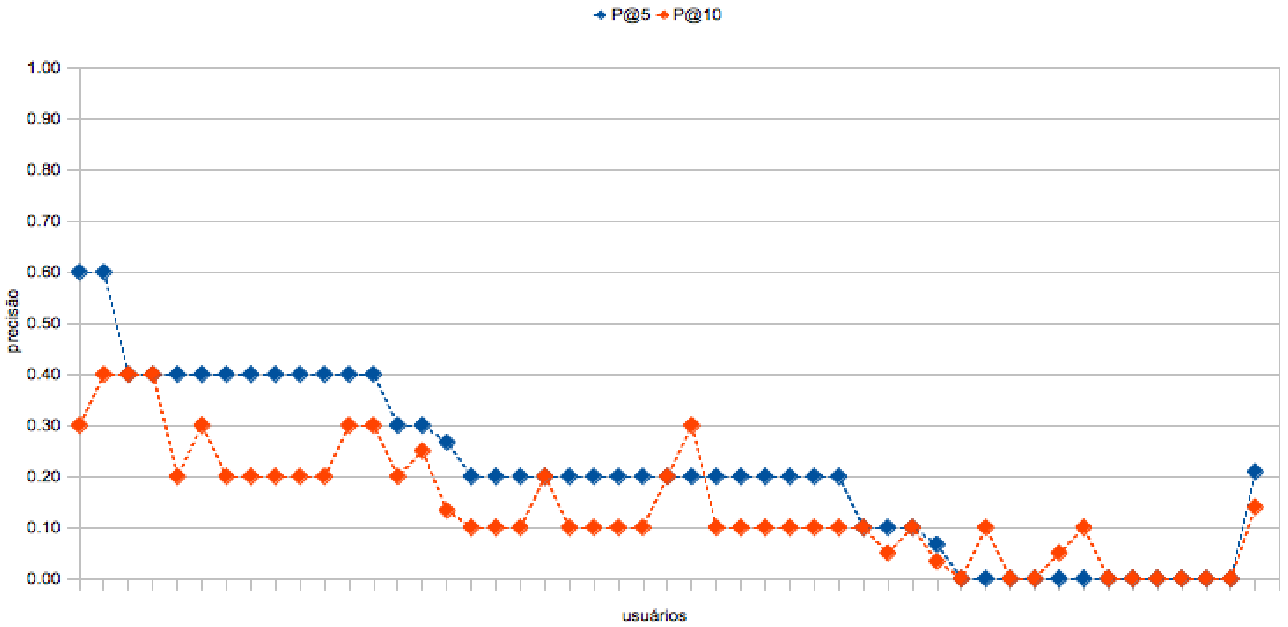

Figura 13.12: Resultados da precisão do objeto r_{12} para as tags recomendadas nas cinco primeiras posições do ranking ($P@5$) em relação as dez primeiras posições ($P@10$).

14. NOVO MODELO DE RECOMENDAÇÃO PERSONALIZADA

Tabela 14.1: Recomendação generalizada em relação à recomendação personalizada utilizando a *query* “ny” e apresentando os cinco primeiros resultados

Abordagem Generalizada	Abordagem Personalizada $< 2, 40, 10 >$	Abordagem Personalizada $< 1, 3, 2 >$
statueofliberty	usa	usa
newyork	policecar	statueofliberty
nyc	harbor	newyorkcity

Tabela 14.2: Recomendação generalizada em relação à recomendação personalizada utilizando a *query* “beach” e apresentando os cinco primeiros resultados

Abordagem Generalizada	Abordagem Personalizada $< 2, 40, 10 >$	Abordagem Personalizada $< 1, 3, 2 >$
sand	sand	sand
ocean	boat	ocean
dog	ship	boat

Tabela 14.3: Recomendação generalizada em relação à recomendação personalizada utilizando a *query* “venice” e apresentando os cinco primeiros resultados

Abordagem Generalizada	Abordagem Personalizada $< 2, 40, 10 >$	Abordagem Personalizada $< 1, 3, 2 >$
italy	gondola	italy
gondola	water	gondola
bridge	street	water

Tabela 14.4: Recomendação generalizada em relação à recomendação personalizada utilizando a *query* "zoo" e apresentando os cinco primeiros resultados

Abordagem Generalizada	Abordagem Personalizada $< 2, 40, 10 >$	Abordagem Personalizada $< 1, 3, 2 >$
polarbear	rhino	rhino
rhino	animal	polarbear
penguin	lion	animal