Sistemas de recomendação de notícias na Internet baseados em filtragem colaborativa

Allan Panossian Kajimoto - 5122986
Renato Shirakashi de Sousa - 5123027
Sidney Eduardo Serra Zanetti - 5122781
Victor Miranda Cirone - 5123006

Universidade de São Paulo
Instituto de Matemática e Estatística
MAC499 – Trabalho de Formatura Supervisionado
Orientador: João Eduardo Ferreira

Dezembro 2007

Sumário

1. Introdução

A quantidade de informações hoje disponíveis nos diversos meios de comunicação é enorme. O problema da falta de acesso à informação foi substituído pela necessidade de filtrá-la e apresentá-la de acordo com as necessidades de visualização de conteúdo dos interessados, que, por muitas vezes, se perdem em meio a tanta informação pessoalmente irrelevante. Diante desse contexto, encontrar formas de filtrar conteúdo, oferecendo relevância ao usuário, tornou-se um imenso diferencial competitivo entre as organizações.

A Internet, grande responsável por essa inversão, tem sido também o principal objeto de estudo para encontrar formas de encontrar, classificar e filtrar conteúdo.

Entretanto, a recente implementação de sistemas colaborativos através da Web 2.0 em serviços como Digg[1], Myspace[2], Via6[3], Twitter[4] e Wikipédia[5] começa a criar uma enorme base de dados de descrições, intenções e classificações de conteúdo, levantando a questão de como utilizar a participação dos usuários sobre o conteúdo para recomendação.

Por exemplo, um usuário pode ser apresentado a um conteúdo não apenas baseando-se em sua idade, sexo ou características pessoais, mas utilizando dados fornecidos por outros para definir o melhor conteúdo a ser oferecido. Dessa maneira, podemos, devido à crescente participação dos usuários sobre o conteúdo, utilizar-se dessas informações para encontrar métodos para filtrar conteúdo relevante.

O propósito desse estudo é, portanto, através de uma abordagem prática e adaptativa, estudar métodos para recomendar conteúdo através de classificação de usuários. O resultado desse estudo é a ferramenta colaborativa Rec6[6], na qual a ação de usuários, combinada a algoritmos, provém notícias de forma relevante. Embora tal ferramenta também inclua outras funcionalidades, esse estudo restringe-se apenas ao estudo do compartilhamento e recomendação de notícias como forma de classificação.

O objetivo desse estudo também não é introduzir o leitor à área de classificação, recomendação e recuperação da informação. Tais estudos, principalmente na área de recuperação da informação, podem ser encontrados em diversas outras publicações. Esse estudo restringe-se apenas a apresentar métodos específicos de ordenar conteúdo a partir de recomendação de usuários (filtragem colaborativa), detectar problemas e propor soluções.

[1] http://www.digg.com
[2] http://www.myspace.com
[3] http://www.via6.com
[4] http://www.twitter.com
[5] http://www.wikipedia.com
[6] http://www.rec6.com.br

1.1 A Web 2.0

Tim O'Reilly[7] em 2005 denominou uma série de fatores que começavam a mudar o paradigma da interação com a internet como Web 2.0. Esses fatores, principalmente ligados à maior participação do usuário na internet, propiciaram a criação de ambientes colaborativos, nos quais os próprios usuários geravam, compartilhavam e consumiam o conteúdo na rede. De certa maneira, o antigo modelo baseado em um produtor de conteúdo institucional estava sendo substituído por esse novo modelo, que têm se mostrado mais relevante para os usuários, já que não estão ligados diretamente aos interesses de um pequeno grupo de pessoas ou instituições.

Entre os fatores que descrevem a Web 2.0, descritos por O'Reilly, estão:

1. Conteúdo gerado pelo usuário
2. Web como plataforma
3. Informação sobre tecnologia
4. Software multi-plataforma

Essencialmente Web 2.0 pode ser descrita como colaboração. Muitas aplicações, como o Wikipédia, Digg, e redes sociais são frutos desse novo paradigma. Com ele, a criação de conteúdo ganhou proporções globais, aumentando a quantidade de informações disponíveis, e também informações qualitativas e quantitativas sobre as mesmas. Com essa quantidade enorme de informações e classificações sobre elas, através de sistemas de votação, folksonomia[8] e taxonomia, encontrar sistemas que possam utilizar essas informações para recomendar conteúdo para o usuário é um enorme diferencial.

[7] http://tim.oreilly.com/
[8] a folksonomia permite a cada usuário da informação a classificar com uma ou mais palavras-chaves, conhecidas como *tags* (em português, marcadores). [Wikipédia]

2. Sistema de recomendação

2.1 Introdução

Como resultado da enorme quantidade de informações geradas e disponíveis na Internet, através de ferramentas colaborativas ou produção da mídia tradicional, um dos maiores problemas de veículos de comunicação deixou de ser deter a informação, mas apresentá-la de forma relevante ao usuário final. Sistemas de recomendação tentam determinar qual conteúdo deve ser apresentado, de forma mais relevante.

Recomendar não é uma tarefa fácil. Normalmente, em nosso dia-a-dia, utilizamos recomendações para diversas tarefas, como realizar uma compra, agendar uma visita ao médico ou mesmo buscar uma referência para alguma informação. Segundo Resnick e Varian[9], sistemas de recomendação auxiliam no aumento da capacidade e eficácia deste processo de indicação já bastante conhecida na relação social entre seres humanos.

Sistemas de recomendação já têm sido largamente utilizados, principalmente em sites de compra, como Amazon[10] e Submarino[11], na intenção de sugerir produtos que possam ser de interesse do usuário. Também são utilizados em espaços publicitários, para melhor adequar o teor do anúncio ao seu alvo. Por exemplo, uma mulher pode ser apresentada a um anúncio de uma loja de cosméticos a partir de uma escolha de um sistema de recomendação, que assim julgou relevante tal anúncio; ou de forma invertida, na qual o anunciante define qual o público alvo de sua campanha.

Dois principais tipos de entidade podem ser identificados em sistemas de recomendação: usuários e conteúdo (itens). A relação entre ambos é responsável pela recomendação. O problema típico de sistema de recomendação pode ser descrito da seguinte maneira:

Para um usuário u, um conjunto C de conteúdo (itens), $c_i \in C$, recomendamos c tal que $F(u, c) = MAX\ F(u, c_i)$, onde F é uma função que determina a relevância de c a u.

2.2 Tipos de Sistemas

2.2.1 Filtragem Colaborativa

Nesse tipo de filtragem, as recomendações são feitas através da previsão das preferências do usuário baseadas em interações de outros usuários. Em geral, esse tipo de filtragem oferece um maior grau de surpresa ao usuário com boas recomendações e, em alguns casos, pode oferecer conteúdo de forma totalmente irrelevante.

Uma primeira abordagem para esse tipo de filtragem [Resnick] determina recomendações baseadas em conteúdo consumido por usuários com mesmo padrão de

[9] Paul Resnick and Hal R. Varian. Recommender systems. Commun. ACM, Março 1997.
[10] http://www.amazon.com
[11] http://www.submarino.com

consumo do usuário atual (figura 1). É utilizada principalmente em sistemas de comércio eletrônico, como Amazon e Submarino (figura 3). Para facilitar o entendimento desse tipo de abordagem, considere um usuário x e xi os i usuários com padrão de compra mais parecido com x (compraram alguns dos mesmos produtos que x comprou). Considere agora os vetores (p, n), onde p é um produto e n é o número de vezes em que esse produto foi adquirido por algum xi. Ordenando-se o conjunto dos vetores (p,n) decrescentemente em n, temos uma ordem de recomendação de produtos para x (tabela 1). Uma variação pode trabalhar com pesos entre usuários xi, baseados em seu nível de relação com x.

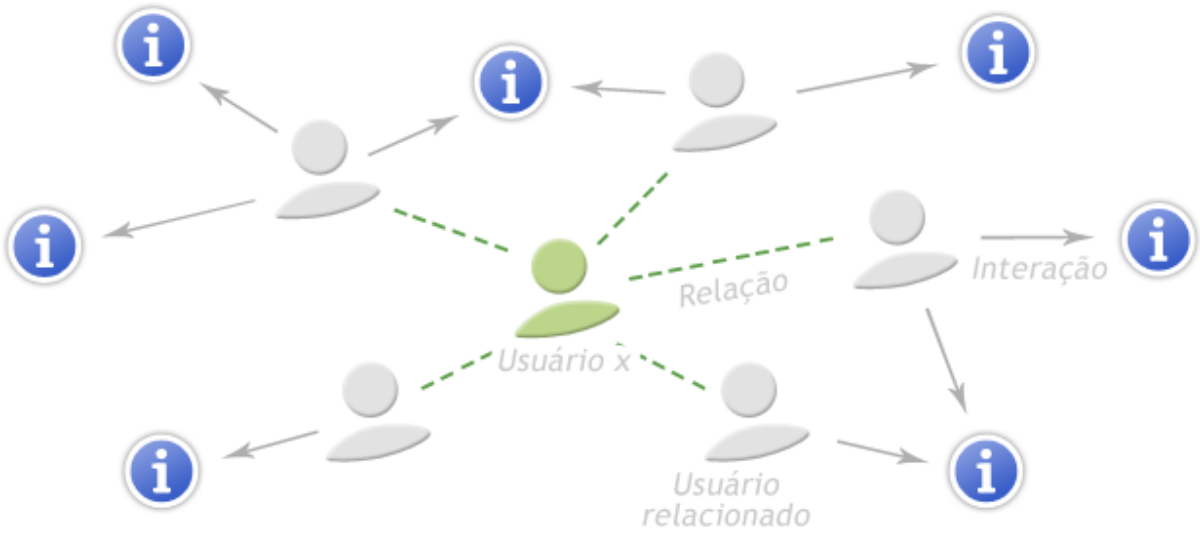

Figura 1: Filtragem colaborativa

Usuário (xi)	Produto
1	Livro1
2	Cd1
3	Cd2
4	Livro1
5	DVD1
6	Livro1
7	Cd2

Tabela 1: Filtragem colaborativa

V = {(Livro1,3), (Cd2, 2), (Cd1, 1), (DVD1, 1)}

Uma segunda abordagem, utilizada nesse estudo (figura 2), determina recomendações baseadas nas classificações realizadas por outros usuários dentro de

um grupo restrito de conteúdo, ordenadas pela soma da relevância de tais classificações (sistema de votação com pesos). É utilizada principalmente em sistemas de notícias, como o Digg. Em geral, esse tipo de abordagem oferece recomendações menos pessoais e mais dirigidas a um grupo determinado de usuários, restritos a um tema. Em contrapartida, sua implementação enfrenta menos problemas de escalabilidade e em geral é mais viável.

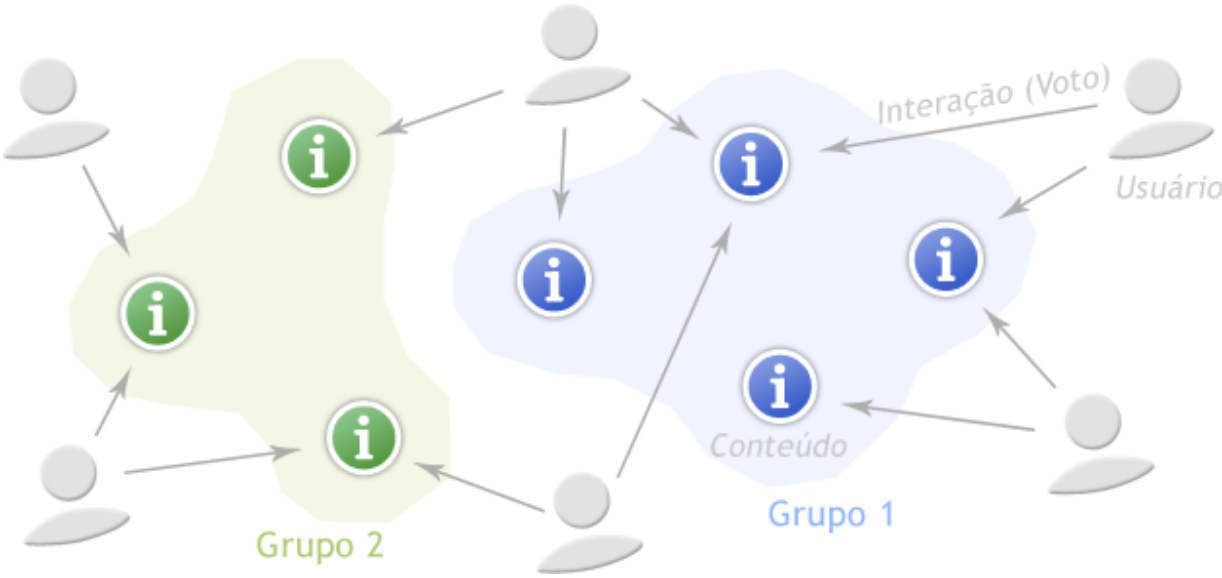

Figura 2: Filtragem colaborativa utilizada

Figura 3: Exemplo de filtragem colaborativa utilizada

2.2.2 Baseada em conteúdo

Têm-se informações sobre um determinado conteúdo e informações que possam ser relacionadas a elas sobre um usuário, podemos determinar uma relação entre usuário e conteúdo baseado nessas relações. Por exemplo, um artigo sobre programação poderia ser recomendado a estudantes de cursos ligados a computação. Esse é o princípio da recomendação baseada em conteúdo, que ao contrário da filtragem colaborativa, não utiliza relações entre usuários para determinar o conteúdo. Por esse motivo, muitas vezes a recomendação baseada em conteúdo não gera grande surpresa ao usuário, já que a relação e os meios que o sistema usou para recomendar pode ser inferida diretamente pelo usuário, mesmo que inconscientemente.

Para viabilizar recomendações baseadas em conteúdo, o uso de taxonomias é necessário para identificar classes de usuários e conteúdos que possibilitem relacioná-los. Uma classe de usuário ou conteúdo pode ser identificada por seu perfil, histórico ou atributos, com a ajuda de algoritmos de classificação. A partir desse ponto, é possível relacionar classes de pessoas e conteúdo de forma compatível.

Uma outra abordagem pode ser descrita como a utilização da folksonomia, na qual os próprios usuários definem palavras-chave para conteúdo. Nessa abordagem ele não é classificado em grupos pré-definidos, e pode ser mais difícil de definir uma relação entre usuário e conteúdo, devido à esparsabilidade de classes. Entretanto, o problema da classificação é delegado aos próprios usuários, o que contorna esse problema recorrente na taxonomia.

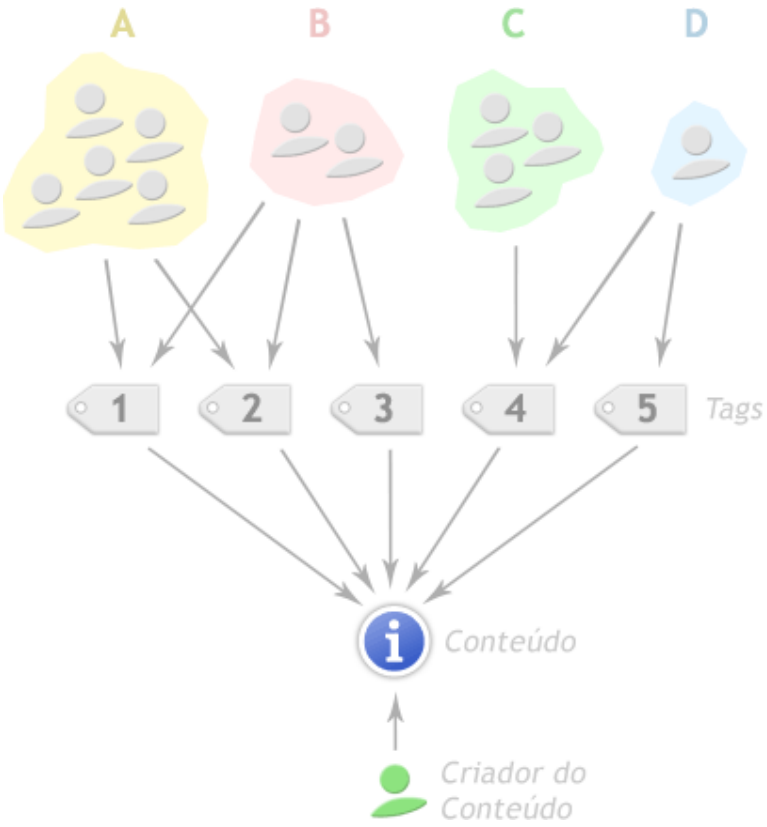

Figura 4: Folksonomia

3. Protótipo Rec6

Em setembro de 2006 foi criado o projeto Rec6 com o intuito de estudar sistemas de recomendação de notícias baseados em votos já em contato direto com o usuário final. Em outubro de 2007, o sistema já dispunha de mais de 5.800 usuários cadastrados e 57.000 notícias inseridas. O modelo conceitual do projeto foi baseado no americano Digg, projeto de muito sucesso nos EUA, que recomenda milhões de notícias baseadas no mesmo tipo de votação.

3.1 Introdução

Para a elaboração desse primeiro protótipo, foi escolhida como objeto de recomendação a filtragem colaborativa, utilizando-se a variante descrita no item 2.2.1.

O Rec6 baseia-se basicamente em duas operações principais: inserção e votação de notícias. A inserção de notícias é feita por meio de *hyperlinks*, a partir de uma *URI* válida. Uma vez inserida no sistema, a notícia fica disponível em uma fila de entrada, onde pode ser votada (recomendada) por usuários para que seja anexada à página principal. Dessa forma, a partir dos votos de usuários, é possível determinar quais as notícias mais relevantes para determinado público.

Figura 5: Captura da tela do Rec6

A interface do Rec6 é dividida em sete categorias (Tecnologia, Blogosfera, Economia e Negócios, Gestão e Marketing, Acadêmico e Educação, Mundo e Política, Entretenimento) e pode ser descrita em dois grandes grupos, que assim serão referenciados em todo esse trabalho: "capa" e "mais novas". Na "capa", a página padrão de uma categoria, são apresentadas as notícias ordenadas por sua relevância, após a influência de votos dos usuários. Na "mais novas" são apresentadas notícias por ordem de inserção (da mais nova a mais velha), formando uma lista de notícias que poderão então ser promovidas para a capa.

Cada notícia é inserida no sistema por um usuário, que fornece dados para sua identificação (título, *URI* e descrição) e para sua classificação (categoria e palavras-chave). A partir da categoria definida pelo usuário, a notícia é inserida no respectivo grupo. A partir das palavras-chave (folksonomia) o sistema será capaz de classificar as notícias em ferramentas externas, como a comunidade virtual Via6.

3.2 Tecnologia

O Rec6 foi criado sob plataforma LAMP (Linux, Apache, Mysql e PHP). A escolha pelo uso de ferramentas abertas envolveu aspectos de custo, escalabilidade e documentação, amplamente disponível para essa plataforma.

A arquitetura dos servidores é baseada em *webfarms*. A aplicação tem disponível dois servidores web, dois servidores de banco de dados, um servidor de imagem e um servidor virtual (Linux VS).

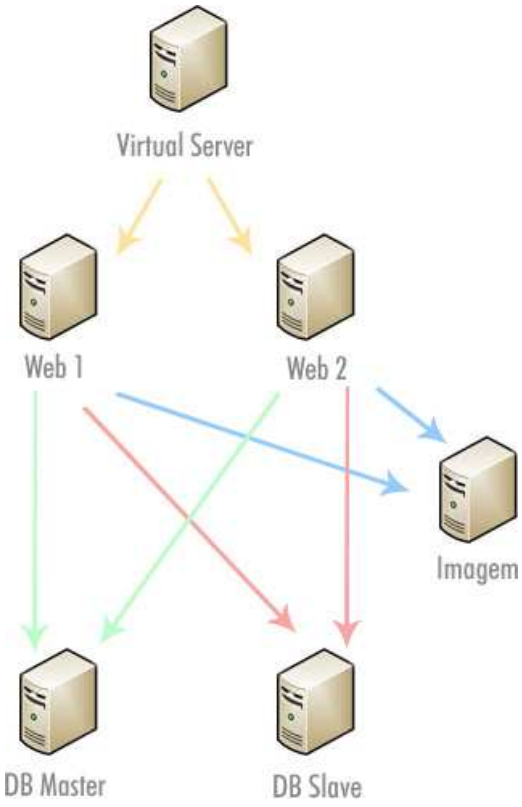

Figura 6: Arquitetura dos servidores

A tabela de versões e distribuições utilizada nos servidores está disponível abaixo:

Responsabilidade	Software	Versão/Distribuição
S.O.	Linux	Debian 4
Web Server	Apache	2.0
Banco de dados	Mysql	5.1
Interpretador	PHP	5.1
Monitoração	Cacti/MRTG	0.86

Tabela 2: Responsabilidades e softwares da arquitetura de servidores

3.2.1 Padrões de design de software

Para o desenvolvimento do protótipo, utilizamos os seguintes padrões de design de software:

I) MVC

O padrão de design de software MVC (*Model-View-Controller*) tem como objetivo dividir a aplicação em três camadas separadas. São elas:

***Model* -** Representa o domínio de uma informação em que a aplicação opera. A classe é responsável pela persistência das informações.

View - Renderiza o *Model*, é responsável pela interface do usuário.

Controller - Processa e responde aos eventos do usuário. Invoca alterações no *Model* e convoca chamadas do *View*.

Dessa maneira, obtém-se uma independência entre a manipulação dos dados e a interface do sistema, tornando possível o desenvolvimento dos mesmos de forma separada.

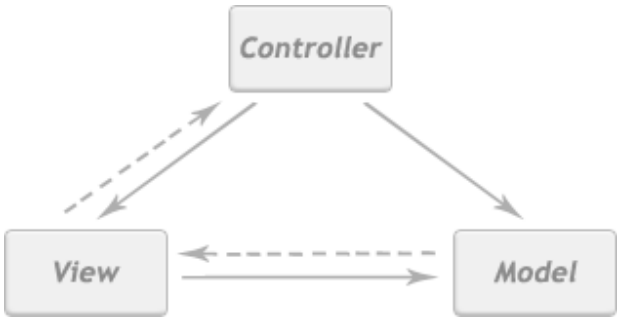

Figura 7: Diagrama que explica a relação entre o Model, o View e o Controller. As linhas pontilhadas representam uma ligação indireta, enquanto que as linhas não pontilhadas representam uma ligação direta.

II) *Factory*

O padrão de design de software *factory* provê uma fábrica de objetos de classes que implementam uma mesma interface. Assim, quando precisamos de uma instância dessa classe, a *factory* retorna a instância da mesma de forma dinâmica. Além disso, outra vantagem desse padrão é a não necessidade do conhecimento da classe de implementação, basta que o desenvolvedor conheça apenas a interface.

III) *Singleton*

O padrão de design de software *singleton* possibilita a garantia da unicidade da instância de uma classe. Assim, um objeto que utiliza o padrão *singleton*, sempre será instanciado através de uma função estática desse mesmo objeto. Essa função estática ficará responsável pela garantia da existência e unicidade da instância da classe.

3.2.2 Disponibilidade e Monitoração

Em sistemas web em produção a necessidade de alta disponibilidade é alta, de modo que os usuários possam ser servidos de forma contínua e transparente a falhas. Para garantir alta disponibilidade do sistema, é necessário que:

1. O sistema permita a detecção de falhas ou níveis de alerta (monitoração)
2. O sistema corrija ou permita corrigir em tempo aceitável tais falhas.

A arquitetura desenvolvida para esse protótipo conta com um servidor virtual (Linux VS) capaz de balancear proporcionalmente a carga dos clientes para os servidores web. O Linux VS é um ponto de entrada único para as requisições, tornando toda a estrutura de servidores transparente para o usuário final. Ele mantém uma tabela de servidores reais e detecta a queda de algum deles, através de um sistema de *heartbeat*. Nesse caso, automaticamente retira esse servidor da tabela até que o mesmo volte a responder. Entretanto, tal estrutura gera um ponto falho no sistema: o próprio servidor Linux VS. Portanto, um novo servidor deve detectar a queda do Linux VS e assumir seu papel em caso de falha, também através de um sistema de *heartbeat*.

O repositório de dados utiliza a arquitetura *master-slave*. Por esse motivo o servidor *master* de banco de dados é um ponto falho no protótipo. A queda de tal servidor causaria deixaria a aplicação fora do ar. Entretanto, o mesmo pode ser contornado rapidamente em tempo hábil através do servidor *slave* de banco de dados, que assumiria seu papel. Uma solução para esse problema seria transformar o servidor *slave* em *master*, adotando a arquitetura *master-master*.

A monitoração de níveis de alerta é feita através da análise de consumo de CPU, memória, transferência de rede e número de processos em fila. Através do Cacti[12], um *frontend* do MRTG, é possível gerar gráficos que acompanhem tais indicadores.

[12] http://cacti.net/

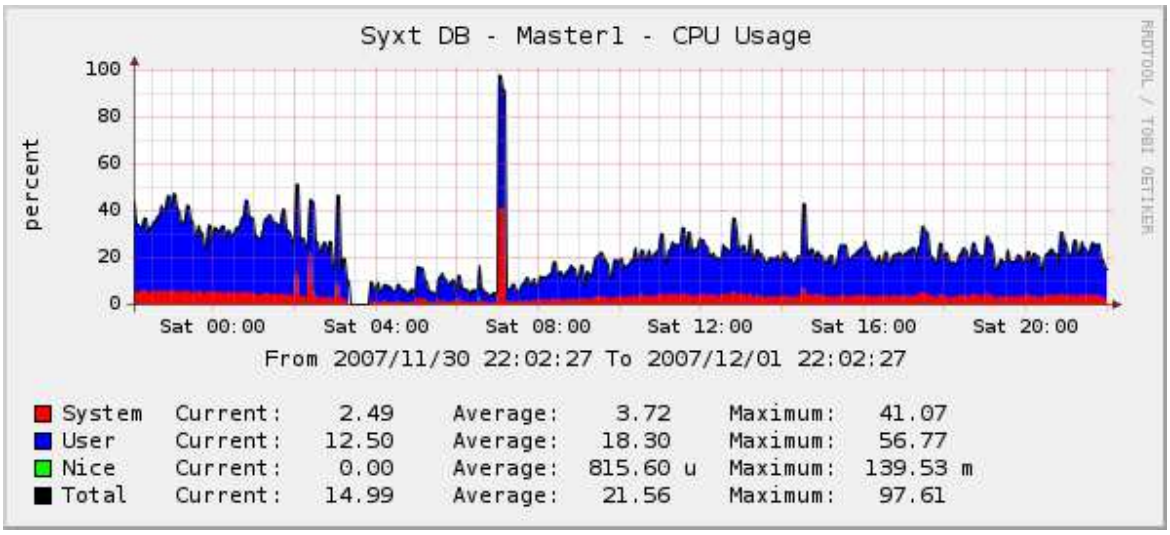

Figura 8: Gráfico de uso do processador do servidor Web 1

3.3 Implementação

O sistema segue três passos básicos:

- **Usuários cadastram notícias**

Um usuário que tenha lido alguma notícia interessante pode cadastrá-la no Rec6. Para isso basta ele ter uma conta no site e clicar no link "Enviar notícia". Ao acessar a página, será solicitado o link da notícia, o título, a descrição, a categoria e três *tags* (palavras-chave). Com essas informações o sistema já tem a habilidade de classificar a notícia na categoria correta, deixá-la disponível para buscas e encontrar notícias relacionadas.

- **Usuários votam em notícias que julgam interessantes**

Um usuário que visita o site com o objetivo de buscar algum conteúdo, lê alguma notícia sugerida pelos outros usuários e caso ache que a notícia é relevante, dá um voto a ela.

- **Sistema filtra e ordena as notícias por relevância**

A partir dos votos dados as notícias, o sistema filtra e ordena as notícias por sua relevância. Com isso, os visitantes têm disponível uma lista de notícias relevantes pronta para ser lida.

Figura 9: Esquema da implementação do sistema

Todas as notícias da ferramenta possuem dois atributos: votos e o valor da sua relevância. Os votos de uma notícia é um número inteiro que representa exatamente o número de votos que ela recebeu. Cada voto aumenta em certo valor a relevância da notícia dependendo do seu peso e suas características, podendo, assim, ocorrer de um voto não aumentar a relevância.

A ordenação das notícias mais relevantes é feita pela relevância e essa listagem é exibida na capa de uma categoria. No "mais novos" são listadas as notícias inseridas recentemente.

3.4 Estatísticas

- Usuários cadastrados: 5.846
- Votos cadastrados: 187.184
- Notícias cadastradas: 57.861
- Média de usuários únicos por mês: 238.741 (referente a Outubro/2007)

Observação: dados referentes ao período: 05/09/2006 a 31/10/2007.

4. Abordagem e desafios da filtragem colaborativa

A abordagem utilizada para esse protótipo, a variação descrita em 2.2.1, levanta alguns desafios a serem enfrentados, comuns a outros tipos de filtragens colaborativas e também próprios dessa abordagem. Em especial, o desafio principal é oferecer um ou mais métodos que ajudem a determinar a relevância de uma notícia x, considerando que a ação do tempo influi diretamente sobre ela: uma notícia muito relevante há um ano atrás pode ter pouca relevância nos dias de hoje. Portanto, para considerar relevância de uma notícia, é essencial considerar o momento t em que ela foi divulgada. Também é necessário considerar fatores como influência social, vantagem cumulativa, votos em grupo com o intuito de promover notícias com interesses pessoais e outros problemas comuns em filtragem colaborativa clássica que podem ser eventualmente aplicáveis a essa variação.

4.1 Normalização Relevância vs. Tempo

Para considerar a relevância de uma notícia é necessário considerar o momento em que ela foi divulgada, além do contexto em que ela se insere. Em um ambiente com um grande número de notícias sendo inseridas, é interessante que a notícia x perca mais relevância por tempo e, em ambientes com um menor número de notícias sendo inseridas, perca menos; normalizando a rotatividade da capa de modo proporcional ao número de notícias inseridas. Essa foi a abordagem utilizada no protótipo do Rec6.

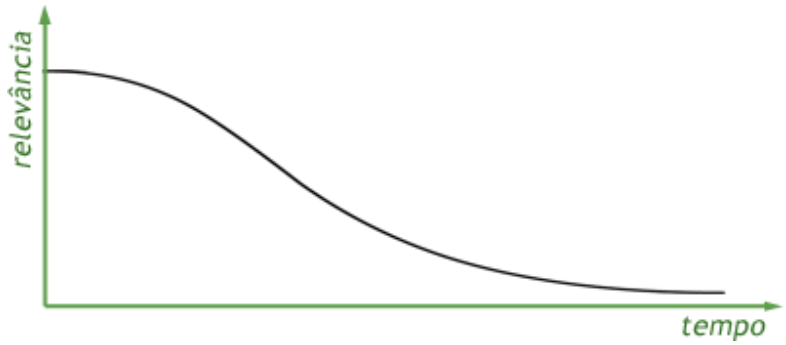

Figura 10: Gráfico da relevância de uma notícia através do tempo

Dessa forma, podemos definir que a relevância rx de uma notícia x é multiplicada por um índice m (0 < m < 1) a cada 24 horas. Ou seja, a cada 24 horas:

```
rx = rx * m
```

Se considerarmos essa execução automática, realizada em intervalos fixos de 10 minutos teremos:

```
m = u^((24*60)/10)  =>  u = raiz(m, 144)

rx = rx * (u)
```

O desafio, portanto é definir o índice m adequado a cada contexto (figura 11).

Para definir m, primeiramente foi definido o índice m0 = 0,5, correspondente a 100 notícias inseridas em 24 horas. A partir daí, o comportamento sugerido para m seria o definido pela reta vermelha da figura 11.

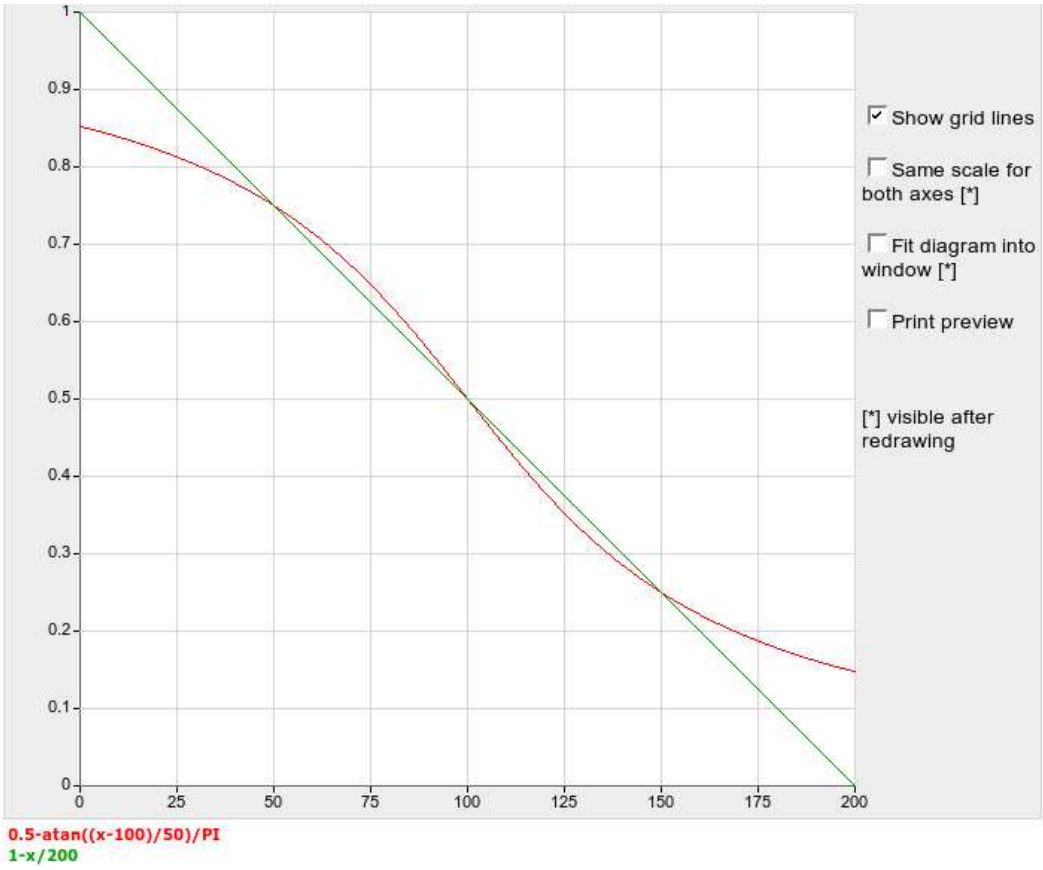

Figura 11: Normalização Relevância vs. Tempo

4.2 Vantagem cumulativa e influência social

Quando um usuário comum entra no Rec6 ele é direcionado para a seção das notícias "Mais Populares" de uma categoria específica. Nesta seção, encontram-se as notícias mais relevantes. Para que outras sejam visualizadas, é necessário que o usuário navegue pelas outras páginas da seção ou selecione outro tipo de visualização: o das notícias "Mais Novas". Verificamos que esta navegação ocorre em um percentual baixo dos usuários (tabela 3).

Seção	Número médio de visitas à seção (por dia)
Mais Populares (capa)	6750
Mais Novos	1322

Tabela 3: Número médio de visitas por seção

Porcentagem de usuários que acessam a Capa e terminam a navegação*:* 84%

Fonte: Google Analytics[13]

Por este motivo, a quantidade de votos feitos na seção "Mais Populares" é maior do que a feita nas na seção "Mais Novos", diminuindo, assim, a chance de uma notícia considerada de boa qualidade que não se encontra na seção "Mais Populares" ter votos suficientes para chegar a ela.

Este problema também sofre influência do caso *early-rater* da abordagem clássica da filtragem colaborativa baseada no usuário, que no nosso caso, é tratada de um modo diferente (ver item 4.4.3). Neste caso, ela fica limitada a uma página de pouca visualização do site, o que acaba sendo prejudicial.

Podemos também analisar outro problema menos impactante que ocorre quando um usuário sofre influência de uma notícia muito votada. Notícias com muitos votos têm maior chance de receber votos do que notícias com poucos votos. Este é um problema de vantagem cumulativa devido à influência social. Deste modo, como as notícias "Mais Populares" são as que possuem maior relevância e em geral possuem mais votos, elas acabam sofrendo esta influência e recebem mais votos.

Figura 12: Influência social

Para amenizar esta situação, cada voto dentro do sistema deve ter o peso proporcional ao peso da página em que ele for feito.

No Rec6, implementamos um sistema que calcula quantos votos uma notícia recebe em média na seção "Mais Populares" e na seção "Mais Novos". A partir destas proporções, conseguimos verificar quantos votos a mais uma notícia da seção "Mais Populares" recebe, fazendo com que um voto em uma notícia fora dela tenha um peso proporcionalmente maior ou menor. Abaixo podemos ver os cálculos efetuados.

[13] http://www.google.com/analytics/

`vmp` = Votos que uma notícia recebe em média por dia na seção "Mais Populares"

`vmn` = Votos que uma notícia recebe em média por dia na seção "Mais Novos"

Os votos de uma notícia que está na capa pode ser calculado da seguinte maneira:

```
vcapa = vmp + vmn
```

Isso ocorre, pois uma notícia que está na seção "Mais Populares", também pode ser encontrada na seção "Mais Novos".

Agora, os votos de uma notícia que está nos mais novos é calculado da seguinte maneira:

```
vncapa = vmn
```

Uma notícia que se encontra na seção "Mais Novos", não está obrigatoriamente na seção "Mais Populares". Existe um valor mínimo de relevância que possibilita uma notícia ser exibida na capa. Abaixo deste valor, ela é apenas visualizada na seção "Mais Novos".

Por fim, a proporção de uma categoria é calculada assim:

```
Proporção = (vmp + vmn) / vmn = 1 + (vmp / vmn)
```

Observação: Os cálculos acima não são feitos em tempo real (no momento do voto). Existe uma rotina que efetua os cálculos das proporções de cada categoria a cada 24 horas. Ela foi criada para evitar o problema de Escalabilidade (ver item 4.4.5) e deixar o processo de votação mais ágil.

Quando um voto é efetuado na seção "Mais Novos", ao invés de incrementarmos a relevância da notícia com o valor exato de um voto, incrementamos seu valor com o número obtido no cálculo da proporção multiplicado pelo valor do voto.

4.3 Métodos para inviabilizar votos inválidos

Para que a variação desenvolvida extraia o máximo de relevância dos votos (recomendações), é necessário que todo voto passe por um processo de validação e se o mesmo for aprovado, então, sua relevância é computada.

Definição: Um voto da notícia x pelo usuário u (cadastrado no sistema) é válido se e somente se o mesmo não é duplicado ou forjado, ou seja, um voto é válido se e somente se todas as afirmações abaixo são verdadeiras para ele:

a) não existe outro voto de x feito por u no sistema;
b) não existe outro voto de x partindo do mesmo IP de u;
c) IP de u não pertence a uma lista de IPs banidos e/ou representa um *proxy* público;
d) u não está agindo em grupo para promover notícias com interesses pessoais.

Caso uma das afirmações acima não seja válida, a relevância do voto é descartada automaticamente pelo sistema.

Figura 13: Votos inválidos

Da definição acima, temos que o sistema não irá computar votos duplicados (a e b) ou falsos (c) ou de usuários que estão agindo em grupo para promover notícias com interesses pessoais (d).

No caso dos votos duplicados, basta uma verificação rápida na base de dados para a verificação das afirmações a e b. Quanto aos votos falsos, basta gerar uma base com os principais *proxies* existentes na Internet e verificar se a recomendação vem do IP de algum deles. Já para afirmação d, temos um problema um pouco mais complexo.

4.3.1 Usuários agem em grupo para promover notícias com interesses pessoais

Um desafio da variação desenvolvida é o problema dos usuários que agem em grupo para promover notícias com interesses pessoais nessa promoção. Esse problema de confiabilidade da recomendação, que também é informalmente conhecido como "panelinha", pode ser exemplificado da seguinte maneira:

Suponha que um usuário u1 deseja promover a notícia x1. Então, o usuário u1 irá convidar todos os seus contatos (u2, u3, u4, ..., ui) para recomendar sua notícia no sistema. Com isso, x1 poderá vir a ser promovida mesmo não tendo relevância.

Agora, suponha que o usuário u2 deseja promover a notícia x2. Então, o usuário u2 irá convidar todos os seus contatos (u1, u3, u4, y1, y2, ... yi) para recomendar sua notícia no sistema. Com isso, x2 poderá vir a ser promovida mesmo não tendo relevância.

Não encontramos referência para esse problema em outros estudos e com isso desenvolvemos nosso próprio método para diminuir a ocorrência de notícias sem relevância sendo promovidas.

Método desenvolvido:

Para mostrar o método desenvolvido, inicialmente, precisamos definir o grau de relacionamento entre dois usuários.

Definição: Seja o grau de relacionamento av aplicado a dois usuários (u1 e u2) determinado da seguinte maneira:

$$av = A/B * 100$$

Onde, A é o número de notícias votadas por u2 que também foram votadas por u1 nos últimos z dias e; B é o número de notícias votadas por u1 nos últimos z dias.

Em outras palavras, o grau de relacionamento entre u1 e u2 será determinado pela porcentagem de notícias votadas por u1 que também foram votadas por u2 (note que av aplicada a u1 e u2 pode ser diferente do valor de av aplicada a u2 e u1).

Assim, seja u o usuário votante, x a notícia a ser votada, V o grupo dos usuário que já votaram em x e Amin um limitante inferior do grau de relacionamento para que a relevância do voto seja descartada.

- Então, para cada v pertencente a V, determina-se a relação av entre u e v;
- Se av > Amin, para qualquer v, então, o voto de u não atua sobre a relevância de x.

Explicando com outras palavras, temos que ao votar, o método analisa o grau de relacionamento entre o usuário votante e os usuários que já votaram na notícia. E caso, o grau de relacionamento seja maior que um limitante inferior quando calculado entre o usuário votante e algum votante prévio, a relevância do voto do usuário atual é descartada.

Problemas de desempenho

Com a implementação do método desenvolvido para impedir o problema dos usuários que agem em grupo para promover notícias com interesses pessoais, obtém-

se um problema de desempenho do sistema. Como o sistema é obrigado a calcular o grau de relacionamento do usuário votante com todos os outros usuários que já votaram na notícia a cada voto, então, é necessário que o algoritmo seja otimizado.

Para isso, uma solução viável, é recalcular e armazenar diariamente os graus de relacionamento entre os usuários para todos os usuários. Dessa maneira, restringimos a falta de desempenho a um horário programado, que pode ser agendado para o horário em que o sistema é menos requisitado, e que, portanto, não afetará diretamente a grande maioria dos usuários.

Com essa otimização, nosso método sofreria uma pequena mudança:

Seja u o usuário votante, x a notícia a ser votada, V o grupo dos usuário que já votaram em x e Amin um limitante inferior do grau de relacionamento para que a relevância do voto seja descartada.

- Então, para cada v pertencente a V, lê-se a relação av entre u e v armazenada;
- Se av > Amin, para qualquer v, então, o voto de u não atua sobre a relevância de x.

Penalizações

Todo usuário, que tem um voto invalidado por estar agindo em grupo para promover notícias com interesses pessoais, recebe uma punição. Esta punição refere-se a perda de 10% da relevância de seu voto a cada voto invalidado por esse motivo.

Nota: Todo usuário do sistema, que não possui 100% da relevância de seu voto, recebe um acréscimo diário de 3% da relevância de seu voto.

4.3.2 Problemas resolvidos

Com a validação de um voto feita dessa maneira, eliminamos/reduzimos drasticamente os seguintes problemas do sistema:

- Criação de usuários falsos por uma mesma pessoa para votar em sua própria notícia;
- Criação de scripts/robôs que utilizam um mesmo IP ou IP de *proxies* mais conhecidos para efetuar votos em uma notícia;
- Criação de "panelinhas" entre os usuários para promoção das próprias notícias.

4.3.3 Problemas que ainda precisam ser resolvidos

Com o bloqueio dos votos a partir de um mesmo IP, acabamos por bloquear o voto de pessoas diferentes mas que estão "atrás" de um mesmo IP, por exemplo, em uma empresa onde existe apenas um computador (um IP) que serve Internet aos outros computadores. Além disso, em alguns casos, o usuário ainda pode conseguir IPs diferentes desconectando-se e conectando-se novamente a seu provedor.

Também, é praticamente impossível ter uma lista completa de todos os *proxies* existentes na Internet e, portanto, não podemos garantir que todos os *proxies* estão bloqueados com esse método.

4.4 Outros desafios da filtragem colaborativa

Abaixo serão listados desafios comumente encontrados na filtragem colaborativa clássica que também podem ser analisados na abordagem utilizada na ferramenta Rec6 e como esses problemas foram abordados.

4.4.1 Novo usuário (*Cold Start*)

Quando um usuário é novo no sistema, ainda não existem avaliações realizadas por ele, portanto, seu perfil de avaliações está vazio.

Na ferramenta Rec6, a recomendação de conteúdo relevante não é baseada nas avaliações prévias de um usuário somente, mas sim, de vários usuários dentro de uma categoria. Deste modo, o desafio enfrentado diante de um novo usuário é encontrar a categoria adequada a ser exibida, com suas notícias mais relevantes.

Uma possível solução para este problema é exigir no cadastro alguma informação do usuário que auxilie na identificação de uma categoria adequada para ser exibida no acesso inicial, por exemplo, sua profissão, interesses pessoais, idade, etc.

Outra solução é a criação de uma seção que englobe todas as categorias da ferramenta, com as notícias mais relevantes de cada uma. Esta seria a seção visualizada pelos novos usuários.

Uma funcionalidade já implementada no sistema que auxilia na exibição de uma categoria relevante para o usuário é de sempre direcioná-lo para a última categoria visitada por ele. Assim, tentamos garantir que a categoria exibida no acesso inicial à ferramenta é de seu interesse.

4.4.2 Ovelha Negra (*Gray Sheep*)

Quando o usuário possui gostos bastante raros fica difícil recomendar algo de seu interesse.

No Rec6 o usuário está condicionado a visualizar notícias de um mesmo tipo quando está dentro de uma categoria. Não existe uma solução prática para usuários com gostos raros, já que a ferramenta abrange somente as notícias das categorias que ela possui.

4.4.3 *Early-rater*

Quando um novo item surge no sistema, não é possível recomendá-lo a algum usuário até que outro usuário o avalie.

Na ferramenta Rec6, quando uma notícia é enviada ela já recebe uma primeira avaliação do usuário que a enviou. Porém, esta situação é tratada de outra maneira no Rec6: as novas notícias que surgem no sistema ficam limitadas à seção "Mais Novas" aguardando por avaliações, enquanto a maior parte dos usuários acessa somente a seção "Mais Populares" da ferramenta.

Uma maneira de amenizar este problema é incentivar o acesso à seção "Mais Novas" da ferramenta, aumentando a possibilidade de notícias receberem votos suficientes para serem exibidas nas primeiras páginas da seção "Mais Populares" e, logo, serem recomendadas aos usuários.

4.4.4 Avaliações Esparsas

Quando o sistema possui poucos usuários para muitos itens, as avaliações tornam-se esparsas e fica difícil encontrar usuários similares para serem feitas recomendações.

No Rec6, em caso de avaliações esparsas, a lista de notícias recomendadas ("Mais Populares") fica menos rotativa, já que poucos usuários estão votando nas notícias do sistema. Este problema é amenizado através de duas funcionalidades implementadas no Rec6 (Normalização Relevância vs. Tempo; Vantagem Cumulativa e Influência Social) que são explicadas nas seções 4.1 e 4.2 deste documento.

4.4.5 Escalabilidade

Quando o volume de usuários, itens e avaliações é muito grande, o sistema que faz o cálculo das relações entre os usuários de forma on-line pode chegar a executá-lo com um tempo de resposta muito elevado.

A natureza do sistema de recomendação baseado em votos diminui a necessidade de processamento, pois não precisa realizar cálculo de relações entre usuários, já que essa relação é definida pelo próprio usuário ao visitar uma categoria. Isso torna o sistema computacionalmente mais viável e escalável, embora possa tornar maior o grau de incerteza da recomendação.

4.4.6 Super-especialização

Este problema refere-se ao fato do sistema só conseguir recomendar itens muito semelhantes àqueles que o usuário já avaliou. Por exemplo, um usuário do *MovieLens*[14] que tem o perfil formado basicamente de filmes de guerra, receberá recomendações, em grande parte, de outros filmes de guerra. Isto pode tornar-se um problema, uma vez que os interesses dos usuários tendem a apresentar mudanças com o passar do tempo [BAL 97] [ADO 2005a].

O Rec6 utiliza um sistema de filtragem colaborativa baseado no grupo a que o usuário pertence e não baseado no usuário isoladamente. Isso significa que a recomendação de notícias é separada por categorias, por exemplo, Tecnologia. Deste modo, ao entrar em uma categoria, o usuário está ciente de que verá somente notícias relacionadas àquela categoria. Qualquer mudança de interesse de um usuário requer o acesso a outra categoria compatível com seu novo interesse.

[14] http://movielens.umn.edu/

4.4.7 Falta de surpresa na recomendação (*Serendipity*)

Itens que não se relacionam com o perfil de um usuário podem nunca ser recomendados.

Na ferramenta Rec6, um usuário que possui um perfil e acessa categorias específicas nunca receberá recomendações de notícias de outras categorias. Neste caso, ele nunca poderá ser surpreendido com uma notícia que eventualmente seja de seu interesse. Isso ocorre devido à variação desenvolvida para o Rec6, que utiliza categorias para agrupar notícias do mesmo assunto.

4.4.8 O conteúdo de alguns tipos de dados ainda não pode ser analisado

O conteúdo de alguns tipos de dados, como vídeo, imagem e som, ainda não pode ser analisado com total precisão. Isso ocorre devido à falta de tecnologia para a análise desses tipos de dados.

Um método para facilitar a análise do conteúdo de um tipo de dado é através de descrições textuais, como por exemplo, palavras-chave. Na ferramenta Rec6, para toda notícia inserida é exigido o preenchimento de três palavras-chave (*tags*) que facilitam a definição do conteúdo da notícia. Apesar desta funcionalidade, este problema não afetaria diretamente as notícias inseridas no sistema, já que elas são compostas apenas de textos que são facilmente analisados.

Abaixo, os problemas listados foram identificados na variação desenvolvida para o Rec6 e podem não ser aplicados a outros sistemas de recomendação.

4.4.9 Inserção de conteúdo impróprio ou inválido

Em sistemas em que o conteúdo é gerado principalmente por usuários, é muito comum o surgimento de itens que não correspondem com o objetivo do sistema ou que são inválidos. No caso do Rec6, itens como pornografia e violência não são permitidos.

Para isso, foi criado um sistema de banimento de domínios e denúncia de notícias. Qualquer usuário do sistema pode denunciar uma notícia como duplicada, *spam*, duvidosa, inadequada, com link quebrado ou que viola direitos autorais. As notícias denunciadas ficam marcadas para que outros usuários sejam desencorajados a votar nelas.

Figura 14: Denúncia de uma notícia

Outro método desenvolvido foi o banimento de domínios. Qualquer usuário que tentar inserir uma noticia de um domínio que tenha sido banido devido a uma grande quantidade de denúncias ou inserção de conteúdo impróprio, não conseguirá fazê-lo.

4.4.10 Usuário comum não entende facilmente um sistema de recomendação

É muito comum um usuário novo não entender como funciona um sistema de recomendação, porém é fundamental que os usuários do sistema entendam-no e colaborem para torná-lo mais relevante a todos.

No Rec6, existe um documento chamado "Instruções de Uso" com uma explicação de como funciona a ferramenta, como inserir notícias e como votar corretamente.

5. Conclusões

A busca por informação relevante na internet está deixando de ser uma tarefa fácil e corriqueira, devido ao enorme aumento de informações disponíveis, potencializado pelo advento da Web 2.0. Além dos veículos de mídia tradicionais, os próprios usuários criam e compartilham conteúdo através de blogs, vídeos, compartilhamento de *links*, perfis em redes sociais, comentários, fóruns e e-mails.

Sistemas de recomendação têm o objetivo de otimizar ferramentas de busca e sugerir conteúdo durante a navegação do usuário. O objetivo desse trabalho foi estudar esses sistemas e propor um sistema de recomendação de notícias baseado em votações ativas dos usuários.

Como resultado, um sistema protótipo, o Rec6, foi criado e colocado em produção, com mais de 5 mil usuários. Com ele, pudemos implementar um sistema de recomendação de notícias baseado em votos dos usuários, descobrir e combater seus principais problemas.

Esse tipo de abordagem de filtragem colaborativa consegue desviar de problemas encontrados em outras abordagens, como o *early-rater*; mas traz outros desafios, principalmente ligados à influência de grupos de usuário com interesses próprios. O grande desafio foi minimizar esses problemas.

Entretanto, dentre todos os trabalhos realizados, pesquisados e constatados em sistemas já estabelecidos, não foi possível encontrar um método definitivo para resolver esses problemas. É possível minimizá-los de forma eficiente para a maioria dos casos, que é o que se propôs e foi apresentado nesse trabalho.

Portanto, o desafio futuro é trabalhar nesses problemas e expandir o universo de trabalho para um ambiente aberto, como em um sistema de busca. Para isso, é necessário um estudo de mineração e classificação de dados.

6. Bibliografia

Applying user's opinion relevance in a Recommender System to researchers - Sílvio César Cazella - http://www.inf.unisinos.br/%7Ecazella/papers/VersaoFinal2006TeseSilvioCazellahomologacao.pdf

Paul Resnick and Hal R. Varian. Recommender systems. Commun. ACM, Março 1997. - http://portal.acm.org/citation.cfm?id=245108.245121

Voss, Jakob (2007). "Tagging, Folksonomy & Co - Renaissance of Manual Indexing? - http://arxiv.org/abs/cs/0701072

What Is Web 2.0 - http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Tim O'Reilly (2006-07-17). Levels of the Game: The Hierarchy of Web 2.0 Applications. O'Reilly radar. Retrieved on 2006-08-08. http://radar.oreilly.com/archives/2006/07/levels_of_the_game.html

Social Networks and Social Information Filtering on Digg - K Lerman 2006 - http://arxiv.org/abs/cs/0612046

Design Patterns: Elements of Reusable Object-Oriented Software - Gamma, E., R. Helm, R. Johnson, and J. Vlissides - Addison-Wesley, 1995

Folksonomia - http://pt.wikipedia.org/wiki/Folksonomia - Wikipédia