GreenWeb: Melhorando a Qualidade da Informação na Web 2.0

Jussara M. Almeida[1], Marcos A. Gonçalves[1], Raquel O. Prates[1], Daniel Hasan[1], Dílson Guimarães[1], Diogo R. de Oliveira[1], Fabiano Belém[1], Flavio Figueiredo[1], Hendrickson Langbehn[1], Henrique Pinto[1], Raquel Lara[1], Saulo Ricci[1], Fabrício Benevenuto[2]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Av. Antônio Carlos 6627, Prédio do ICEx, Pampulha, Belo Horizonte, MG

[2]Departamento de Ciência da Computação – Universidade de Ouro Preto
Campus Universitário – Morro do Cruzeiro, Ouro Preto, MG

{jussara,mgoncalv,rprates,hasan,dilsonag,renno,fmuniz,flaviov,reiter,
hpinto,raqlara,saulomrr}@dcc.ufmg.br, fabricio@iceb.ufop.br

Abstract. *This paper introduces the GreenWeb framework, which aims at improving the quality of information on the Web 2.0. GreenWeb has 4 main components: (1) metrics and methods to estimate information quality; (2) profiles of users' interests and system usage; (3) strategies to communicate content quality to users; and (4) methods to detect and reduce low quality content as well as to promote high quality content. In this paper, we present the challenges and solutions already developed for each such component.*

Resumo. *Este artigo apresenta o arcabouço GreenWeb, que visa melhorar a qualidade da informação na Web 2.0. O GreenWeb consiste de 4 componentes principais: (1) métricas e métodos para estimar qualidade da informação; (2) perfis de uso e de interesse dos usuários; (3) estratégias para comunicar a qualidade de um conteúdo para os usuários; e (4) métodos para detectar e reduzir conteúdo de baixa qualidade, assim como promover conteúdo de mais alta qualidade. Neste artigo, nós apresentamos os desafios e as soluções já desenvolvidas para cada componente.*

1. Introdução

A Web 2.0 tem como ênfase facilitar a interação e a colaboração entre usuários através da criação de comunidades virtuais e do estabelecimento de plataformas de distribuição de conteúdo. Ela é marcada por um maior envolvimento dos usuários que passaram a atuar não somente como consumidores, mas também como produtores e provedores de conteúdo [Boll 2007]. Tal conteúdo, muitas vezes criado de forma colaborativa e em diferentes tipos de mídia (p.ex: áudio, video, texto), é frequentemente chamado de *mídia social*. A mídia social é tipicamente composta por um *objeto*, que representa o principal veículo de disseminação de informação na aplicação (p.ex: um vídeo no YouTube, um artigo na Wikipedia[1]), e possivelmente uma série de atributos associados.

Tipicamente, as aplicações da Web 2.0 não impõem nenhum controle editorial sobre o conteúdo gerado pelos usuários e logo não fornecem nenhuma garantia de

[1] http://www.youtube.com e http://www.wikipedia.com, respectivamente

qualidade da informação disponibilizada. Embora o conceito de "qualidade da informação" seja intuitivo, uma definição explícita do mesmo é um desafio. Nós aqui consideramos um conteúdo com qualidade se a informação associada a ele é *relevante, atende as necessidades e/ou agregue valor a serviços e aplicações para um conjunto de usuários*. Logo, o conceito de qualidade vai além de aspectos sintáticos e semânticos do conteúdo e incorpora aspectos relacionados às necessidades informacionais dos usuários e características específicas dos serviços e aplicações. Por exemplo, a qualidade de um conteúdo pode ser avaliada sob a perspectiva do seu potencial como fonte de dados para suportar serviços de informação, tais como busca, recomendação, e propaganda. Neste contexto, o foco principal está nos atributos textuais associados aos objetos (p.ex: *tags*), dado que, a despeito da existência de técnicas de recuperação de informação multimídia, a maioria dos serviços ainda utiliza apenas estes atributos textuais [Boll 2007].

Estudos recentes indicam que existe uma grande quantidade de *lixo informacional* em aplicações da Web 2.0 [Suchanek et al 2008, Figueiredo et al 2009], possivelmente devido à facilidade e liberdade com que usuários criam e disponibilizam conteúdo nestas aplicações. Tal liberdade abre oportunidade para ações maliciosas e/ou oportunistas, que resultam na introdução de conteúdo de baixa qualidade (i.e., conteúdo poluído ou simplesmente poluição) no sistema. Exemplos incluem vandalismo na Wikipédia [Potthast et al 2010] e diferentes formas de *spamming* [Benevenuto et al 2009a, Koutrika et al 2008]. Conteúdo poluído incorre em custos extras para os administradores de sistemas, afeta a eficácia de serviços de informação e compromete a paciência do usuário e sua satisfação com o sistema. Isto porque os usuários não podem facilmente identificar a poluição sem ter contato com ela, o que leva ao consumo de recursos do sistema (p.ex: largura de banda [Benevenuto et al 2009a]).

Neste contexto, este artigo apresenta a proposta de um arcabouço, denominado *GreenWeb,* que visa fundamentar o desenvolvimento de técnicas e ferramentas para melhorar a qualidade de informação na Web 2.0, contribuindo para agregar valor a várias aplicações e serviços de informação. O GreenWeb foca em três pilares principais: (1) reduzir a poluição de conteúdo, (2) aumentar a qualidade da informação disponibilizada aos usuários, e (3) manter uma relação custo-benefício favorável para usuários e administradores de sistemas. Tendo aplicações e serviços da Web 2.0 como alvo, o GreenWeb enfatiza um ambiente que vem se mostrando promissor para a disseminação de informação, para a interação e a colaboração entre as pessoas e, em última instância, para a troca de conhecimentos e experiências, o que contribui diretamente para o crescimento da sociedade.

O GreenWeb aborda aspectos relacionados a quatro dos Grandes Desafios da Pesquisa em Computação, definidos pela Sociedade Brasileira de Computação [Carvalho et al 2006]. Em sua essência, ele aborda questões relacionadas ao *acesso universal ao conhecimento* (4º desafio), uma vez que o acesso à informação de maior qualidade pode estimular a participação dos usuários nos processos de produção e de uso do conhecimento. Ele também trata de aspectos relativos à *gestão de informação, sob a perspectiva de qualidade, em grandes bases de dados multimídia* (1º desafio), explorando várias técnicas de *modelagem computacional para representar as complexas interações entre usuários e entre os usuários e o sistema* (2º desafio). Por fim, ao abordar a detecção e o combate a ações maliciosas e oportunistas, ele também visa o *desenvolvimento de sistemas seguros e escaláveis* (5º desafio).

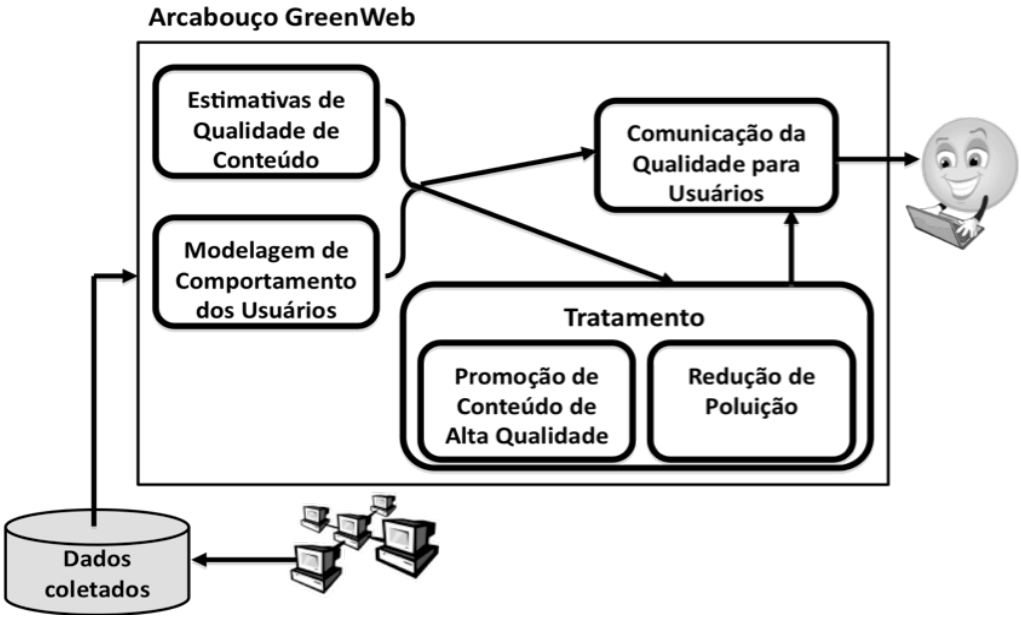

Figura 1: Componentes do Arcabouço GreenWeb

A seguir, a Seção 2 apresenta uma visão geral do arcabouço GreenWeb, seus principais componentes e desafios. As Seções 3 a 6 descrevem, mais detalhadamente, as soluções já desenvolvidas para estes componentes, como elas se posicionam frente ao estado-da-arte e os principais resultados já obtidos. A Seção 7 descreve dois protótipos desenvolvidos, enquanto conclusões e próximos passos são apresentados na Seção 8.

2. O Arcabouço GreenWeb

O arcabouço GreenWeb tem por objetivo fundamentar o desenvolvimento de soluções para agregar valor a diferentes serviços de informação da Web 2.0 a partir do combate à poluição e da promoção de conteúdo de mais alta qualidade. Para atingir esse objetivo, o GreenWeb é constituído dos seguintes 4 grandes componentes, mostrados na Figura 1:

1. **Estimativas de Qualidade do Conteúdo**: composto por técnicas e métricas para estimar a *qualidade da informação* associada a um dado conteúdo.

2. **Perfis de Uso e Interesse do Usuário**: composto por técnicas e modelos que representam aspectos relevantes sobre quem é o usuário e sobre o uso que ele faz do sistema. Informações sobre o perfil do usuário (p.ex: suas características e interesses) podem facilitar a identificação de suas necessidades informacionais e do conteúdo mais adequado (i.e., com maior qualidade) ao seu perfil. A análise de como os usuários interagem com o sistema (p.ex: as funcionalidades utilizadas, freqüência de uso, relacionamentos estabelecidos, etc.) pode ajudar na detecção de usuários maliciosos/oportunistas, que introduzem poluição no sistema [Benevenuto et al 2009a]. O conjunto de usuários analisados dependerá dos dados disponíveis (veja discussão abaixo). Por exemplo, dados coletados a partir da API de uma aplicação tipicamente são restritos a usuários que possuem conta no sistema. Por outro lado, dados coletados a partir de outras fontes, tais como servidores *proxy* ou mesmo de *logs* de acesso mantidos pela aplicação alvo podem viabilizar o estudo de uma população maior de usuários, incluindo aqueles que não têm conta no sistema.

3. **Estratégias de Tratamento**: este componente é dividido em dois subcomponentes principais que abordam o problema alvo sob perspectivas complementares: (1) Promoção de Conteúdo de Mais Alta Qualidade e (2) Detecção e Redução da Poluição no Sistema. Cada subcomponente é composto por mecanismos e técnicas voltados para atingir a estratégia definida.

4. **Comunicação da Qualidade para o Usuário**: composto por modelos e técnicas que permitam a comunicação, através da interface, da qualidade da informação sendo apresentada, apoiando o usuário na identificação de conteúdo de alta/baixa qualidade.

Os dois primeiros componentes, conjuntamente, fornecem subsídios para o desenvolvimento de soluções para promover conteúdo de mais alta qualidade e para detectar e reduzir a poluição no sistema (componente 3). As informações produzidas por todos estes componentes, por sua vez, podem subsidiar a definição, pelo componente 4, de quais aspectos devem ser comunicados a diferentes tipos de usuários. Por exemplo, o objetivo pode ser informar o usuário final sobre a qualidade de um dado conteúdo (p.ex: indicação de artigos de qualidade na Wikipédia), ou então alertar os administradores de um serviço sobre usuários fazendo uso malicioso ou oportunista do sistema (p.ex: indicar potenciais *spammers* para os administradores do YouTube).

Esta é uma descrição em alto nível do arcabouço GreenWeb, focada nos macrocomponentes principais. Cada macrocomponente é decomposto em vários subcomponentes. Por exemplo, muitas soluções propostas para os macrocomponentes exploram padrões identificados em dados coletados da aplicação alvo. Assim, faz-se necessário um subcomponente responsável pela coleta, processamento e armazenamento de dados em algum repositório (p.ex: um banco de dados).

O desenvolvimento dos componentes do GreenWeb enfrenta vários desafios, entre eles:

• Coleta, armazenamento e processamento de grandes volumes de dados: a identificação de padrões típicos de uso e perfis de usuário, bem como o desenvolvimento de técnicas para estimar qualidade de conteúdo dependem da análise de dados, obtidos de diferentes aplicações. Tais dados são também essenciais para direcionar o desenvolvimento de soluções de tratamento e de comunicação assim como para suportar a avaliação dos mesmos.

• Padrões dinâmicos e heterogêneos: tipicamente, a forma como as pessoas utilizam os sistemas e os conteúdos criados e acessados por elas variam conforme suas preferências pessoais e foco de uso. Tais padrões também tendem a variar com o tempo. Logo, as soluções desenvolvidas devem lidar com alta heterogeneidade e dinamicidade. As variações temporais implicam na necessidade de realizar coletas freqüentes a fim de manter o repositório de dados atualizado;

• Diferentes perspectivas de qualidade: a qualidade da informação associada a um conteúdo pode variar dependendo do usuário (grupo de usuários) alvo, da aplicação ou serviço (classe de aplicações ou serviços), ou ainda do uso feito da informação. Fundamentalmente, ela depende da necessidade informacional do usuário, que, por sua vez, pode variar dependendo do tipo de aplicação e serviço. Além disto, características da aplicação ou serviço tais como o tipo de mídia usado para disseminar informação e o seu público alvo podem também afetar a percepção de qualidade. Logo, as melhores estratégias de promoção de conteúdo, redução de poluição e comunicação de qualidade

podem depender da aplicação/serviço alvo, suas funcionalidades e detalhes da interface. Mais ainda, usuários da Web 2.0 cada vez mais utilizam serviços (p.ex: busca) para recuperação de informação e organização de seu conteúdo. Assim, a eficácia destes serviços no atendimento das necessidades dos usuários é também importante. Logo, a qualidade de um conteúdo pode ainda ser analisada sob a perspectiva do seu potencial como fonte de dados para suportar serviços de informação mais eficazes. Neste caso, ela pode depender do tipo de serviço: por exemplo, um conteúdo pode ter qualidade para suportar um serviço de busca, mas não um serviço de classificação de conteúdo [Almeida et al 2010]. Em suma, as diferentes perspectivas de qualidade implicam na sua alta dependência do domínio e contexto considerados. Logo, não é possível desenvolver uma solução única para todos os contextos, exigindo, pois, instanciações.

Logo, nós analisamos a aplicabilidade do arcabouço GreenWeb desenvolvendo soluções específicas para diferentes contextos. Vale ressaltar que nem sempre todos os componentes precisam ser instanciados, o que mostra a flexibilidade do arcabouço. Isto também aponta para a necessidade de uma investigação sobre quais componentes são de maior interesse, o que depende do contexto específico, da perspectiva de qualidade sob análise e da relação custo-benefício associada ao desenvolvimento. A seguir, apresentamos uma breve descrição das soluções já desenvolvidas para alguns contextos específicos. Em particular, discutimos soluções para estimar qualidade da informação (Seção 3), comunicar qualidade (Seção 4), promover conteúdo de qualidade (Seção 5) e reduzir poluição (Seção 6). Os modelos de perfis e padrões de uso já desenvolvidos foram aplicados na detecção de usuários poluidores, sendo, pois, discutidos na Seção 6. Apesar da existência de soluções alternativas para alguns dos problemas específicos discutidos a seguir, nós não temos ciência de nenhuma proposta de arcabouço abordando, de forma unificada, as várias perspectivas do problema de melhorar a qualidade da informação na Web 2.0. Esta é a principal contribuição deste artigo.

3. Estimativas de Qualidade da Informação

Estimar a qualidade da informação associada a um dado conteúdo é uma tarefa complexa pelo alto grau de subjetividade e pela inevitável necessidade de considerar aspectos relativos tanto aos usuários quanto à aplicação e ao contexto do estudo. Nesta seção, nós discutimos as soluções propostas para dois contextos específicos.

3.1. Qualidade de Atributos Textuais para Recuperação de Informação

Serviços de informação na Web 2.0 exploram majoritariamente atributos textuais como fontes de dados [Boll 2007]. Entretanto, embora existam vários estudos sobre os padrões de uso de *tags* [Heymann et al 2010, Santos-Neto et al 2010, Sigurbjornsson and van Zwol 2008] e sua qualidade para suportar busca, recomendação e classificação de objetos [Clements et al 2010, Schenkel et al 2008, Ramage et al 2009], os resultados obtidos não apontam um consenso. Enquanto alguns concluem que *tags* têm boa qualidade [Bischoff et al 2008], outros evidenciam problemas como *tag spamming* [Koutrika et al 2008] e uso frequente de termos sem significado ou com múltiplos significados [Suchanek et al 2008], que impactam negativamente os serviços de informação. Mais ainda, a maioria dos estudos anteriores focou apenas em *tags*, negligenciando o uso potencial de outros atributos, tais como título e descrição.

Considerando o foco em serviços de informação, nós argumentamos que um atributo textual de alta qualidade deve: 1) conter uma quantidade de conteúdo suficiente para ser útil; 2) prover uma boa descrição do conteúdo, o que é importante para serviços que exploram a semântica dos objetos (p.ex: recomendação); e 3) poder distinguir o objeto de outros para tarefas como separar os objetos em classes semânticas ou em níveis de relevância para uma dada consulta. Embora cada um destes três aspectos – quantidade de conteúdo, poder descritivo e poder discriminativo – esteja relacionado à qualidade de um atributo, eles não são igualmente importantes para todos os serviços [Almeida et al 2010]. Por exemplo, um bom poder discriminativo é importante para serviços de classificação, enquanto um bom poder descritivo pode ser mais importante para serviços de recomendação. Além disto, alguns serviços, tais como classificação, podem se beneficiar mais da presença de uma maior quantidade de conteúdo.

Assim, em [Figueiredo et al 2009] nós realizamos uma extensa caracterização da qualidade, considerando os três aspectos acima, de 4 atributos – título, *tags*, descrição e comentários – em 4 aplicações – YouTube, YahooVideo, LastFM e CiteULike[2]. Nossa análise foi feita em amostras com mais de 200.000 objetos (e seus atributos) coletados de cada aplicação. A quantidade de conteúdo foi estimada pelo número de termos distintos presentes em cada atributo associado a cada objeto analisado. Para estimar os poderes descritivo e discriminativo, optamos pelo uso de métricas heurísticas que, apesar de aproximadas e invariavelmente conterem limitações, podem ser computadas a baixo custo em grandes bases de dados. As heurísticas usadas são adaptações de um modelo de recuperação de informação em páginas Web estruturadas [Moura et al 2010].

Nós estimamos o poder descritivo de um termo *t* contido em um atributo *f* de um objeto *o* pelo *espalhamento* de *t* em *o*, definido como o número de atributos associados a *o* que contêm *t*. O poder descritivo de *f* é estimado pelo espalhamento médio de todos os termos de *f*. A intuição é que termos que aparecem em vários atributos associados ao mesmo objeto têm uma maior chance de serem relacionados ao seu conteúdo. Por exemplo, se o termo "Sting" aparece em 4 dos atributos de um objeto (espalhamento = 4), há uma alta chance de que ele seja relacionado ao famoso cantor.

Para estimar o poder discriminativo de um termo *t* contido em um atributo *f* de um objeto *o*, foi proposta a heurística Freqüência Inversa nos Atributos (FIA), baseada na métrica IDF, amplamente usada em recuperação de informação [Baeza-Yates and Ribeiro-Neto 2011]. FIA estima o poder discriminativo de *t* pelo inverso da freqüência de *t* em todas as instâncias do atributo *f* na coleção de objetos. A intuição é que termos que ocorrem em muitas instâncias de um dado atributo são pouco discriminativos. Por exemplo, a ocorrência do termo "music" no título de um vídeo do YouTube é pouco discriminativa se o mesmo ocorre nos títulos de vários outros vídeos.

Os principais resultados da caracterização são: (1) todos os atributos, **exceto título**, estão ausentes (i.e., com nenhum conteúdo), em uma fração não desprezível dos objetos coletados e, logo, podem **não** ser eficazes como fontes únicas de dados, já que, neste caso, os serviços não atingiriam muitos objetos; (2) considerando atributos não vazios, atributos criados e editados colaborativamente tendem a conter mais conteúdo se comparados àqueles editados somente pelo usuário que criou o objeto; (3) título e *tags*, tipicamente com menos conteúdo, têm, em geral, melhores poderes descritivo e

[2] http://video.yahoo.com, http://last.fm , http://www.citeulike.org, respectivamente.

discriminativo, seguidos de descrição e comentários; (4) todos os atributos contêm uma grande quantidade de termos sem significado (lixo) ou então com muitos significados.

As métricas desenvolvidas assim como os resultados da caracterização têm norteado o desenvolvimento de serviços de informação mais eficazes. Em [Figueiredo et al 2009], nós mostramos o uso destas métricas na avaliação da qualidade dos atributos para suportar classificação automática de objetos, concluindo que *tags* é o melhor atributo isolado graças ao seu bom poder discriminativo e quantidade de conteúdo razoável. Também concluímos que a combinação de múltiplos atributos pode trazer benefícios, uma vez que eles tendem a contribuir com diferentes informações sobre o objeto. Nós também temos explorado as métricas propostas na proposição de mecanismos de recomendação mais eficazes, conforme será discutido na Seção 5.

3.2. Qualidade de Artigos da Wikipédia

A estimativa da qualidade de artigos na Wikipédia e em outros ambientes colaborativos similares são essenciais para garantir a confiança do leitor no conteúdo ao qual ele está sendo exposto. Apesar de estudos indicando que certos artigos da Wikipédia têm uma qualidade similar àqueles da Enciclopédia Britânica [Giles 2005] e de esforços no sentido de definir critérios qualitativos para estimar essa qualidade [Dondio et al 2006, Santos e Prates 2010], tais soluções são baseadas em análise manual, e portanto não escalam frente ao volume e à velocidade com que o conteúdo é atualizado. Assim, soluções automáticas para produzir estimativas de qualidade [Dondio et al 2006, Rassbach et al 2007] são necessárias. Tais estimativas podem ser usadas como indicadores de documentos que necessitam revisão, para identificar vandalismo ou para recomendar artigos baseados em sua qualidade estimada.

Para abordar este problema, nós propusemos um método automático para estimativa de qualidade, tratando-o com um problema de *regressão* [Dalip et al 2009]. Ou seja, nós estimamos a qualidade dos artigos na Wikipédia como um valor numa escala contínua de qualidade, fazendo uso de regressão baseada em Máquinas de Vetores de Suporte (*Support Vector Machines*) [Vapnik 1995]. Nossa principal contribuição nesse trabalho foi um estudo detalhado de várias características dos artigos como fontes de evidência e seu impacto na estimativa da qualidade. As seguintes características, algumas das quais foram propostas por nós, foram consideradas:

- Características associadas à revisão do artigo, tais como: número de revisões, quantidade de mensagens de discussão, número de revisões feitas por usuários casuais e por usuários especialistas, estabilidade das revisões etc.
- Características de rede: número de *links* de entrada e saída, *PageRank* [Baeza-Yates and Ribeiro-Neto 2011], coeficiente de agrupamento, etc.
- Características de texto, que por sua vez foram divididas em quatro subgrupos:
    - Características relativas ao tamanho do artigo;
    - Características relativas à estrutura do artigo, tais como: número de seções, número e cobertura das citações, tamanho médio das seções;
    - Características de estilo, tais como: tamanho do maior e do menor parágrafo, uso de pronomes, advérbios;
    - Características de facilidade de leitura, baseadas no número, tamanho e

distribuição de palavras, sentenças e sílabas, usadas para estimar o grau de educação necessário para se entender um artigo.

Experimentos realizados com uma amostra da Wikipédia usando uma escala de qualidade proposta pelo próprio sistema demonstraram que o uso do método de aprendizado proposto junto com o conjunto de características analisadas apresentou melhores resultados do que as melhores abordagens disponíveis na literatura [Dondio et al 2006, Rassbach et al 2007]. As características relativas à estrutura do texto foram as mais eficazes, sendo também as mais fáceis de computar. Os melhores resultados foram obtidos combinando essas características com as de revisão.

4. Comunicação da Qualidade aos Usuários

Como discutido na Seção 2, a qualidade de um conteúdo depende do contexto no qual o usuário fará uso da informação associada. Nesta seção, apresentamos a investigação feita para o contexto específico da Wikipédia. Propostas anteriores de interfaces que apóiam os usuários na sua inferência sobre a qualidade do artigo [Pirolli et al 2009, Krieger et al 2009, Chevalier et al 2010] focam em uma melhor visualização de informações já disponibilizadas pela Wikipédia, tais como histórico, discussões, tamanho do artigo ou da página de discussão. Em contraste, a nossa proposta é utilizar informações novas, tais como as características discutidas na Seção 3.2.

Embora a análise das características propostas na Seção 3.2 tenha tido bons resultados, o nosso objetivo não é filtrar ou classificar conteúdo, mas ser capaz de informar ao usuário qual a qualidade do artigo sendo acessado, para que ele possa decidir se ele atende ou não suas necessidades. Assim, para gerar uma proposta da comunicação a ser feita, a primeira etapa envolveu uma análise semiótica da Wikipédia para identificar se (e quais) considerações sobre qualidade dos artigos são feitas pelo sistema, e como estas são apresentadas aos usuários. A partir desta análise, identificou-se que a Wikipédia adota 9 estratégias que visam obter (ou encorajar) artigos de maior qualidade [Santos e Prates 2010]. As estratégias são classificadas em 2 categorias: (1) ações tomadas pelos administradores da Wikipédia e (2) ações disponíveis aos usuários para que melhorem a qualidade do conteúdo. Identificou-se também vários problemas na forma em que estas estratégias são apresentadas aos usuários.

No passo seguinte, investigou-se na literatura propostas de indicadores quantitativos e qualitativos relativos a artigos da Wikipédia. A partir desta investigação, fez-se uma proposta da comunicação a ser feita aos usuários sobre a qualidade que consiste de **indicadores quantitativos**, que permitem que se tenha sempre uma avaliação atualizada de cada artigo. No entanto, estes indicadores apenas não são suficientes, pois eles se baseiam normalmente em aspectos estruturais do texto ou da interação sobre o texto, e o significado associado a ele pode não ser único em alguns contextos. Por exemplo, a cobertura de um artigo (distribuição de citações ao longo artigo) considera que quanto mais referências e melhor distribuídas, melhor a qualidade do artigo. No entanto, a qualidade das referências citadas não é considerada. Mais ainda, pode ainda não haver muitas referências para um assunto inovador, o que não implica que o texto tenha baixa qualidade. Assim, para cada indicador quantitativo, deve-se ter também **explicações qualitativas** associadas a ele. Estas explicações devem no mínimo definir o aspecto sendo considerado (e.g. fator de cobertura) e o impacto esperado disso na qualidade do artigo. Por fim, devem ser disponibilizadas **visualizações** que permitam uma visão mais detalhada ou mesmo complementar ao indicador quantitativo. Por

exemplo, para o caso do fator cobertura, pode-se mostrar visualmente como está a cobertura de cada seção. Vale ressaltar que a proposta foi feita para a Wikipédia, mas acredita-se que ela seja válida para outras enciclopédias colaborativas.

A proposta foi implementada no protótipo do GreenWiki (ver Seção 7). Foi feita uma primeira avaliação deste protótipo em ambiente controlado, incluindo observação da interação de 9 usuários com o sistema e entrevistas sobre suas experiências de uso [Pereira 2011]. A avaliação mostrou que os usuários conseguiram entender e utilizar sem dificuldades o painel disponibilizado. Além disso, a partir da interação, os usuários passaram a ter uma maior preocupação e uma visão mais crítica sobre aspectos de qualidade. Eles perceberam que a métrica quantitativa apenas não seria suficiente, pois poderia gerar falsos positivos, e que a informação complementar (i.e., explicações qualitativas e visualizações) era relevante para a avaliação sobre a qualidade do artigo.

5. Promoção de Conteúdo de Alta Qualidade

A promoção de conteúdo de qualidade pode ser abordada de várias maneiras. Serviços de busca podem ser otimizados para levar em consideração estimativas de qualidade na ordenação dos resultados de uma consulta, enquanto métodos de recomendação de conteúdo podem incluir métricas de qualidade como um de seus critérios. Nosso foco, até então, tem sido na recomendação de *tags* de qualidade. Este foco é motivado por: (1) *tags* são amplamente exploradas por vários serviços de informação, e 2) nossa caracterização da qualidade de atributos textuais apontam *tags* como um atributo promissor para esta tarefa (vide Seção 3.1). Logo, o nosso objetivo é desenvolver métodos para sugerir *tags* de qualidade para um dado objeto (i.e., termos relacionados ao seu conteúdo), visando melhorar a qualidade deste atributo e, indiretamente, a eficácia de serviços que dele dependam.

Os métodos de recomendação de *tags* existentes exploram tipicamente [Lipczak et al 2009, Menezes et al 2010, Sigurbjornsson and van Zwol 2008]: (1) regras de associação para inferir padrões de co-ocorrência de termos com *tags* previamente associadas ao objeto alvo; (2) termos extraídos de múltiplos atributos e (3) métricas de qualidade (p.ex: freqüência, entropia) para filtrar termos irrelevantes e promover termos com maior qualidade. Entretanto, a maioria dos métodos existentes explora no máximo duas destas três dimensões. Em [Belém et al 2010], nós desenvolvemos soluções que exploram as três dimensões conjuntamente. Nós estendemos métodos baseados em padrões de co-ocorrência para incluir tanto *tags* previamente atribuídas aos objetos quanto termos extraídos de outros atributos (título e descrição). Todos estes termos são então ordenados quanto à qualidade (ou relevância) para a tarefa de recomendação. Para tanto, utilizamos várias métricas heurísticas que tentam capturar a qualidade de um termo para um objeto alvo. O problema de recomendação então se reduz a projetar uma função que combina as métricas para ordenar os termos candidatos por qualidade.

As funções desenvolvidas são extensões de soluções disponíveis na literatura [Menezes et al 2010, Sigurbjornsson and van Zwol 2008], que se diferenciam por incluir métricas de poder descritivo, particularmente a métrica espalhamento. Avaliamos as soluções propostas, em um total de oito estratégias, utilizando bases de dados reais coletadas do YouTube, YahooVideo e LastFM. Assim como em trabalhos anteriores, nossa avaliação foi automatizada, utilizando parte das *tags* já atribuídas ao objeto alvo como gabarito: apenas termos do gabarito são considerados relevantes.

A Tabela 1 mostra alguns dos resultados obtidos, em termos da precisão nas 5 primeiras posições da ordenação, ou seja, em termos da fração dos termos nas 5 primeiras posições das recomendações que foram considerados relevantes conforme gabarito. A tabela mostra resultados médios e intervalos de confiança de 95%. Sum+ [Sigurbjornsson and van Zwol 2008], LATRE [Menezes et al 2010] e CTTR [Lipczak et al 2009], soluções consideradas estado-da-arte, exploram um subconjunto das três dimensões mencionadas acima. Mostramos resultados apenas para as 2 melhores heurísticas, LATRE+TS e SUM+TS, omitindo as demais por questões de espaço. LATRE+TS, estende LATRE, baseado somente em padrões de co-ocorrência, para incluir a métrica espalhamento (*Term Spread* ou TS*)* e também para extrair termos candidatos de múltiplos atributos textuais. Sum+TS estende o Sum+, baseado em co-ocorrência e em algumas métricas de qualidade, de forma similar. Note que a nossa melhor heurística produz melhorias de até 26% sobre o estado-da-arte, graças ao uso da métrica espalhamento e da exploração de múltiplos atributos textuais.

Tabela 1: Precisão nas 5 Primeiras Posições da Recomendação: Valores Médios e Intervalos de Confiança de 95% (Melhores Resultados em Negrito)

Estratégia		LastFM	YahooVideo	YouTube
Estado-da-Arte	Sum+	0.411 ± 0.001	0.484 ± 0.003	0.245 ± 0.002
	LATRE	0.405 ± 0.001	0.608 ± 0.003	0.285 ± 0.004
	CTTR	0.260 ± 0.001	0.465 ± 0.004	0.376 ± 0.002
Novas soluções heurísticas	Sum+TS	**0.418 ± 0.002**	0.674 ± 0.003	**0.475 ± 0.002**
	LATRE+TS	0.411 ± 0.001	**0.716 ± 0.003**	0.467 ± 0.003

6. Redução de Poluição

Diversas formas de poluição já foram detectadas em vários contextos: vandalismo na Wikipédia [Potthast et al 2010], lixo informacional em atributos textuais [Suchanek et al 2008], *spamming* em atributos textuais [Koutrika et al 2008]. O contexto escolhido aqui foi a poluição em Sistemas de Compartilhamento de Vídeos Online (SCVOs), com foco no sistema YouTube. Em particular, focamos em um recurso do YouTube ainda pouco investigado, as **vídeo-respostas**, que são vídeos postados como respostas a outros vídeos. O nosso interesse nasceu de uma investigação prévia que evidenciou a exploração deste recurso por dois tipos de usuários poluidores [Benevenuto et al 2009b]. ***Spammers*** são usuários que postam vídeos não relacionados em resposta a vídeos populares visando aumentar a visibilidade de seus próprios vídeos. **Promotores** são usuários que postam um grande número de vídeos, na sua maioria não relacionados, em resposta ao seu próprio vídeo, visando inflar, artificialmente, os contadores internos mantidos pelo YouTube (p.ex: número de vídeo-respostas) a fim de que seu vídeo venha a ser promovido para a primeira página da aplicação na lista de mais respondidos.

Assim, nós desenvolvemos um método automático para detectar *spammers* e promotores no YouTube [Benevenuto et al 2009a]. Note que o nosso foco não é a detecção de conteúdo poluído, mas ao contrário, a detecção de potenciais poluidores (*spammers* ou promotores). A idéia é que o método possa auxiliar administradores do sistema no esforço de detecção. Cabe a eles decidir sobre que políticas aplicar para reduzir a poluição gerada por eles (p.ex: suspensão de conta, e-mail de alerta, etc).

A primeira tarefa foi o desenvolvimento de um modelo do comportamento típico dos usuários, incluindo usuários poluidores e usuários legítimos, visando tentar diferenciá-los em um segundo momento. O comportamento de um usuário foi modelado por um conjunto de características que expressam seu comportamento no que tange o uso feito do sistema [Benevenuto et al 2009a]. As características são categorizadas em três grupos. **Características dos vídeos do usuário** capturam propriedades específicas dos vídeos postados e respondidos pelo usuário, tais como: duração média, números de visualizações e de comentários, número de vezes que o vídeo foi selecionado como favorito, números de honrarias e de links externos. **Características individuais** incluem número de amigos, número de vídeos postados, número de vídeos assistidos, número de vídeos adicionados como favoritos, número de vídeo-respostas postados e recebidos, tempo médio entre postagens, etc. Por fim, as **características das redes sociais** estabelecidas com outros usuários via interações de vídeo-resposta incluem coeficiente de agrupamento, *betweeness* (ou centralidade), reciprocidade, e assortatividade.

Nós coletamos dados do YouTube referentes a 829 usuários, que foram pré-classificados em 641 legítimos, 157 *spammers*, e 31 promotores. Esta pré-classificação exigiu a avaliação manual de mais de 20.000 vídeos. Para a detecção automática dos usuários nas três classes de usuários, utilizamos dois algoritmos de aprendizado de máquina supervisionado considerados estado-da-arte: SVM [Vapnik 1995] e Lazy Associative Classifier (LAC) [Veloso et al 2006]. Algoritmos de aprendizado supervisionado "aprendem" padrões a partir de um conjunto de treino (previamente rotulado nas classes corretas) para aplicá-los em um conjunto de teste (a ser classificado). Assim, utilizamos SVM e LAC para aprender padrões de combinações das características descritas acima que maximizassem a identificação dos usuários nas respectivas classes. Realizamos uma bateria de experimentos com a coleção de 829 usuários, utilizando um processo de validação cruzada com 5 partições (4 partições para treino do algoritmo e uma para teste, com 25 repetições).

A Tabela 2 mostra resultados obtidos usando o classificador SVM. Ela mostra as porcentagens de usuários de cada classe **(classe real)** que foram classificados como promotores, *spammers* e legítimos **(classe predita)**. Os resultados são valores médios de 25 execuções. Intervalos de confiança de 95%, omitidos por clareza, indicam um erro máximo de 5% sobre as médias reportadas. Estes resultados indicam que o algoritmo de aprendizado foi capaz de detectar corretamente quase todos os usuários promotores e legítimos, classificando aproximadamente 60% dos *spammer*s corretamente, enquanto cerca de 40% deles foram considerados usuários legítimos. Uma análise mais aprofundada desses resultados revelou que os *spammers* não detectados tinham um comportamento dual, ora agindo como *spammers* ora agindo como usuários legítimos, tornando a identificação automática muito difícil. Como trabalho futuro, pretendemos investigar estratégias para melhorar a eficácia da detecção de *spammers*. Uma possível abordagem seria tratar separadamente diferentes perfis de um mesmo usuário. De qualquer forma, considerando que a intenção é utilizar o método proposto como ferramenta de investigação para a implementação de políticas específicas de redução de poluição, muito provavelmente incluindo uma análise manual dos usuários suspeitos detectados, consideramos os resultados obtidos bastante satisfatórios.

Contudo, apesar dos bons resultados, métodos supervisionados têm a desvantagem de necessitar de dados de treinamento para o aprendizado dos padrões. No caso de SCVOs, isso pode ser muito custoso, pois pode envolver a verificação e

rotulação de milhares de vídeos. Visando reduzir este custo, nós desenvolvemos recentemente uma abordagem semi-supervisionada para o problema [Langbehn et al 2010]. Nessa abordagem, as características de vídeo, individuais e de redes são particionadas em "visões" distintas, e um classificador é treinado com um conjunto bem pequeno de instâncias de treino contendo apenas os atributos relativos às características de cada visão, gerando assim três classificadores diferentes. Esses classificadores são então aplicados a um conjunto de dados não rotulados. Quando esses classificadores concordam quanto à classificação de uma mesma instância desse conjunto com uma alta confiança, essa instância é incorporada ao treino. A idéia é, portanto, explorar as múltiplas visões para expandir um conjunto de treino originalmente reduzido e melhorar a classificação. Nós investigamos várias estratégias para combinar os classificadores gerados para cada visão. Os nossos melhores resultados indicam que conseguimos reduzir a necessidade de treino em até 80% mantendo as taxas de acerto do classificador em níveis muito próximos aos apresentados na Tabela 2 [Langbehn et al 2010].

Tabela 2: Classificação de Usuários do YouTube usando o Classificador SVM

Classe Real	Classe Predita		
	Promotor	*Spammer*	Legítimo
Promotor	96.13%	3.87%	0%
Spammer	1.40%	56.99%	41.91%
Legítimo	0.31%	5.02%	94.66%

7. Protótipos

Como prova de conceito, nós instanciamos as técnicas e métodos descritos nas Seções 3-6 em dois protótipos: GreenMeter e GreenWiki[3]. GreenMeter é uma ferramenta para estimar a qualidade de *tags* e para recomendar *tags* em aplicações da Web 2.0. Ela utiliza as métricas de qualidade e os métodos de recomendação apresentados nas Seções 3.1 e 5. A Figura 2 mostra uma tela do protótipo desenvolvido para a aplicação LastFM[4], sendo aplicado na página da artista "Nina Simone": a qualidade de cada *tag* é mostrada com uma cor, em uma escala de vermelho (pior qualidade) a verde (melhor qualidade). O medidor indica a qualidade média de todas as *tags* da nuvem de *tags*. A figura também mostra as *tags* recomendadas pelo GreenMeter, "jazz", "soul", e "piano", que parecem descrever bem a famosa pianista e cantora.

GreenWiki é uma ferramenta para apresentar aos usuários indicadores da qualidade dos artigos da Wikipédia, seguindo a proposta de comunicação da qualidade apresentada na Seção 4 e as estimativas discutidas na Seção 3.2. Na sua versão atual, o GreenWiki implementa dois indicadores: cobertura (distribuição das citações ao longo do artigo) e estabilidade (número de edições feitas no artigo em um determinado período sobre número de edições total do artigo). Maiores cobertura e estabilidade indicam um artigo com melhor sua qualidade. Na interface foram acrescentados um medidor para cada indicador, seguindo o mesmo padrão do GreenMeter. Ao clicar no medidor, outra tela é aberta mostrando as explicações e as visualizações associadas. A

[3] http://sites.google.com/site/greenmeterdemo/ e http://www.dcc.ufmg.br/projetos/greenwiki/mediawiki.

[4] LastFM ordena as *tags* pela sua popularidade, mostrando *tags* mais populares em fontes maiores.

Figura 3 mostra a tela do GreenWiki com os indicadores apresentados no canto superior direito e a tela que será aberta ao seu clicar no indicador de estabilidade.

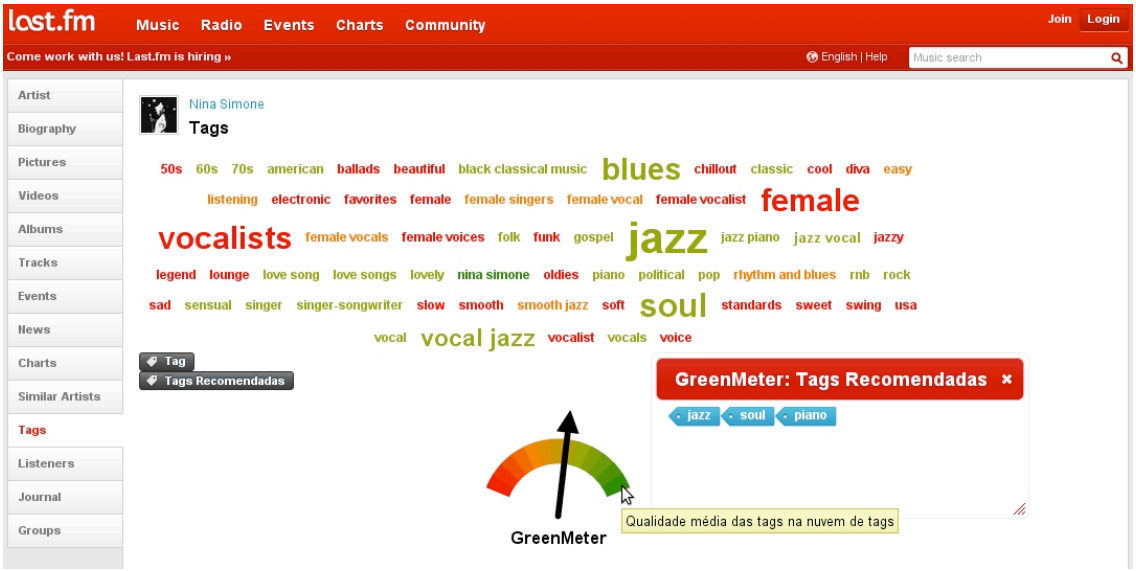

Figura 2: GreenMeter: Estimador de Qualidade e Recomendador de Tags (estudo de caso: LastFM)

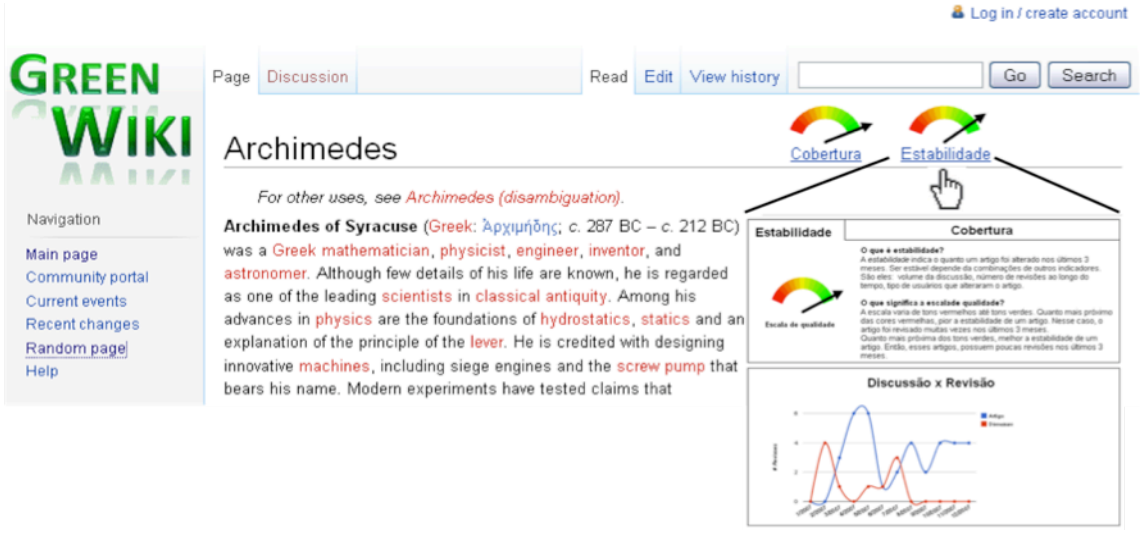

Figura 3: GreenWiki: Estimador de Qualidade de Artigos do Wikipédia

8. Considerações Finais

Este artigo apresentou o GreenWeb, um arcabouço que visa melhorar a qualidade da informação em aplicações e serviços da Web 2.0 através da promoção de conteúdo de maior qualidade e da redução de conteúdo poluído. Soluções desenvolvidas para 3 instanciações do arcabouço foram apresentadas: (1) qualidade de atributos textuais na Web 2.0; (2) qualidade de artigos da Wikipédia e (3) qualidade de vídeos no YouTube.

Como trabalho futuro, pretendemos estender o arcabouço para outros contextos, incluindo: detecção de usuários vândalos na Wikipédia e de *spammers* em atributos textuais assim como o uso das métricas de qualidade propostas no projeto de serviços classificação automática de objetos e em serviços de recomendação de conteúdo.

Pretendemos ainda estender nossas soluções para conteúdo multimídia explorando técnicas de processamento e de recuperação de informação especializadas.

Agradecimentos

Este trabalho é desenvolvido como parte do Instituto Nacional de Ciência e Tecnologia para Web (MCT/CNPq proc.53.3871/2008-6), com o apoio do CNPq e da FAPEMIG.

Referências

Almeida, J. M., Gonçalves, M. A., Figueiredo, F., Belém, F. and Pinto, H. (2010) "On the Quality of Information for Web 2.0", In: IEEE Internet Computing, v. 14.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011) "Modern Information Retrieval", Addison-Wesley Professional, second edition.

Belém, F., Martins, E., Almeida, J. M., Gonçalves, M. A. and Pappa, G. (2010) Exploiting Co-Ocurrence and Information Quality Metrics to Recommend Tags in Web 2.0 Applications", In: Proc. ACM CIKM.

Benevenuto, F. , Rodrigues, T., Almeida, V., Almeida, J. and Gonçalves, M. (2009) "Detecting Spammers and Content Promoters in Online Video Social Networks", In: Proc. ACM SIGIR.

Benevenuto F., Rodrigues T., Almeida, V. Almeida, J. and Ross, K. (2009) "Video Interactions in Online Video Social Networks", ACM TOMCCAP, 5(4) article 30.

Bischoff, K., Claudiu-S, F., Wolfgang, N. and Raluca, P. (2008) "Can All Tags Be Used for Search?", In: Proc. ACM CIKM.

Boll, S. (2007) "MultiTube – Where Web 2.0 and Multimedia Could Meet", In:.IEEE Multimedia, 14(1).

Carvalho, A., Brayner, A. Loureiro, A., Furtado, A. et al. (2006) "Grandes Desafios da Computação – 2006 a 2016", Relatório disponível em http://www.sbc.org.br .

Chevalier, F.; Huot, S. & Fekete, J.-D. (2010). WikipediaViz: Conveying article quality for casual Wikipedia readers. 2010 IEEE Pacific-Vis, pp. 49--56.

Clements, M., de Vries, A. P. and Reinders, M. (2010) "The Task Dependent Effect of Tags and Ratings on Social Media Access", In: ACM TOIS, 28(4).

Dalip, D., Gonçalves, M., Cristo, M. and Calado, P. (2009) "Automatic Quality Assessment of Content Created Collaboratively by Web Communities. A Case Study of Wikipedia", In: Proc. JCDL.

Dondio, P., Barrett, S., Weber, S., and Seigneur, J. (2006) "Extracting Trust from Domain Analysis: A Case Study on the Wikipedia Project", Autonomic and Trusted Computing. Springer Berlin / Heidelberg.

Figueiredo, F., Belém, F., Pinto, H., Almeida, J. M., Gonçalves, M. A. , Fernandes, D., Moura, E. and Cristo (2009) "Evidence of Quality of Textual Features on the Web 2.0", In: Proc. ACM CIKM.

Giles J. (2005) "Internet Encyclopaedias Go Head to Head", In: Nature 438, 7070.

Krieger, M.; Stark, E. M. & Klemmer, S. R. (2009) "Coordinating tasks on the commons" In: Proc. CHI.

Heymann, P. Paepcke, A. and Garcia-Molina, H. (2010) "Tagging Human Knowledge", In Proc. ACM WSDM.

Koutrika, G., Effendi, F., Gyöngyi, Z., Heymann, P. and Garcia-Molina, H. (2008) "Combating spam in tagging systems: An evaluation", In: ACM TWEB 2(4).

Langbehn, H., Ricci, S., Gonçalves, M. A., Almeida, J., Pappa, G. and Benevenuto, F. (2010) "A multi-view approach for detecting spammers and content promoters in online video social networks", In: Journal of Information and Data Management v.1.

Lipczak, M., Hu, Y., Kollet, Y. and Milios, E. (2009). "Tag Sources For Recommendation In Collaborative Tagging Systems", In: Proc. PKDD.

Menezes, G., Almeida, J. Belém, F., Gonçalves, M., Lacerda, A., Moura, E., Pappa, G., Veloso, A. and Ziviani, N. (2010) "Demand-Driven Tag Recommendation", In: Proc. PKDD.

Moura, E., Fernandes, D., Ribeiro-Neto, B., Silva, A. and Gonçalves, M. (2010) "Using Structural Information to Improve Search in Web Collections", In: JASIST, 61.

Pereira, R. L. dos Santos. (2011) "Qualidade de Artigos na Wikipedia para seus Usuários – Análise e Proposta da Interação", Dissertação de mestrado, DCC/UFMG.

Pirolli, P.; Wollny, E. & Suh, B. (2009) "So you know you're getting the best possible information", In: Proc. CHI.

Potthast, M., Stein, B. and Holfeld, T. (2010) "Overview of the 1st International Competition on Wikipedia Vandalism Detection", In: Notebook Papers of CLEF 2010 LABs and Workshops.

Ramage, D., Heymann, P., Manning, C. and Garcia-Molina, H. (2009) "Clustering the Tagged Web", In: Proc. WSDM.

Rassbach, L., Pincock, T., and Mingus, B. (2007) "Exploring the Feasibility of Automatically Rating Online Article Quality. http://upload.wikimedia.org/wikipedia/wikimania2007/d/d3/RassbachPincockMingus07.pdf.

Santos, R. L. e Prates, R. O. (2010) "Estratégias para Comunicar Qualidade na Wikipédia", In: Proc: IHC.

Santos-Neto, E., Figueiredo, F. Figueiredo, Almeida, J., Mowbray, M., Gonçalves, M. and Ripeanu, M. (2010) "Assessing the Value of Contributions in Tagging Systems", In:Proc. IEEE 2nd International Conference on Social Computing, 2010.

Schenkel, R., Crecelius, T. , Kacimi, M., Michel, S., Neumann, T., Parreira, J.X. and Weikum, G. (2008), "Efficient Top-k Querying Over Social-Tagging Networks", In: Proc. SIGIR.

Sigurbjornsson, B. and van Zwol R. (2008) "Flickr Tag Recommendation Based on Collective Knowledge", In: Proc. WWW.

Suchanek, F. M., Vojnovic, M. and Gunawardena, D. (2008) "Social Tags: Meaning and Suggestions", In: Proc. ACM CIKM.

Vapnik, V. (1995) "The Nature of Statistical Learning Theory", Springer.

Veloso, A., Meira Jr, W. and Zaki, M. (2006) "Lazy Associative Classification", In: Proc. ICDM.