

Caracterizando o Uso e a Qualidade de Atributos Textuais em Aplicações da Web 2.0

Flavio Figueiredo

Dep. of Computer Science
Fed. Univ. of Minas Gerais
flaviov@dcc.ufmg.br

Jussara Almeida

Dep. of Computer Science
Fed. Univ. of Minas Gerais
jussara@dcc.ufmg.br

Edleno Moura

Dep. of Computer Science
Fed. Univ. of Amazonas
edleno@dcc.ufam.edu.br

Fabiano Belém

Dep. of Computer Science
Fed. Univ. of Minas Gerais
fmuniz@dcc.ufmg.br

Marcos Gonçalves

Dep. of Computer Science
Fed. Univ. of Minas Gerais
mgoncalv@dcc.ufmg.br

Virgilio Almeida

Dep. of Computer Science
Fed. Univ. of Minas Gerais
virgilio@dcc.ufmg.br

Henrique Pinto

Dep. of Computer Science
Fed. Univ. of Minas Gerais
hpinto@dcc.ufmg.br

David Fernandes

Dep. of Computer Science
Fed. Univ. of Minas Gerais
davidf@dcc.ufmg.br

Marco Cristo

FUCAPI - Analysis, Research
and Tech. Innovation Center
marco.cristo@gmail.com

ABSTRACT

Apesar da grande quantidade de conteúdo multimídia em aplicações Web 2.0, a maioria de seus serviços de Recuperação de Informação (RI) fazem uso apenas dos atributos textuais associados ao conteúdo (por exemplo, etiquetas ou *tags*). Entretanto, por serem tipicamente gerados pelos usuários, tais atributos não oferecem garantias de qualidade para serviços de RI. Neste trabalho, são investigadas evidências de qualidade de atributos textuais em aplicações populares da Web 2.0 em relação a três aspectos: utilização; poder descritivo e discriminativo. Para tanto, foi feita uma caracterização do uso de quatro atributos textuais (título, descrição, *tags* e comentários) nos sistemas Youtube, YahooVideo, LastFM e CiteULike. Alguns de nossos resultados, que podem ser considerados no projeto de serviços de RI na Web 2.0 são: (1) atributos textuais colaborativos, embora não sejam explorados significativamente em algumas aplicações, contêm a maior quantidade de informação quando presentes; (2) há uma diversidade significativa de informação entre os atributos textuais; e (3) o título e as *tags* dos objetos parecem ser os atributos mais promissores para serviços de RI, visto que o primeiro é quase sempre presente e tem alto poder descritivo, e o segundo, quando utilizado, apresenta alto poder descritivo e discriminativo.

Categories and Subject Descriptors

H.3.5 [Online Information Services]: Web-based services

General Terms

Experimentação, Caracterização

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WebMedia'09, October 5–7, 2009, Fortaleza, CE, Brazil.

Copyright 2009 ACM 978-1-60558-880-3/09/0010...\$10.00.

Keywords

Web 2.0, Atributos Textuais

1. INTRODUÇÃO

Aplicações Web 2.0 têm crescido significativamente em diversidade e popularidade nos últimos anos. Exemplos incluem o Youtube e o Yahoo! Video¹ (ou simplesmente YahooVideo), dois sistemas sociais de compartilhamento de vídeo, o Last.FM² (ou simplesmente LastFM), uma estação de rádio *online* que possibilita a interação social em torno de tópicos relacionados a música, e o CiteULike³, um serviço de gerência e descoberta de referências bibliográficas voltado para a comunidade acadêmica. Fazendo uso da distribuição de conteúdo gerado por usuários, do estabelecimento de redes sociais, e de colaboração, estas aplicações oferecem uma maior quantidade e diversidade de conteúdo em relação às aplicações da Web tradicional. O Youtube, por exemplo, é atualmente uma das maiores bases de dados⁴, e o segundo sistema de busca mais utilizado⁵ do mundo. Embora o conteúdo musical disponível no LastFM não seja gerado por usuários, esta aplicação é uma das estações de rádio mais populares na Internet, devido a sua organização baseada em comunidades, que permite ao usuário descrever conteúdo musical e interagir em redes sociais.

O termo “mídia social” refere-se ao conteúdo produzido por usuários e disponibilizado na Web 2.0. Este compreende tanto o conteúdo principal (denominado *objeto*), que pode ser armazenado em forma de texto, imagem, áudio ou vídeo, quanto todas as informações textuais associadas ao conteúdo (denominadas *atributos textuais*). Por ser gerada pelos usuários, não há garantias de que a mídia social ofereça a qualidade demandada por outros usuários e, em particular, por serviços de Recuperação de Informação (RI), tais como

¹<http://youtube.com> e <http://video.yahoo.com>

²<http://last.fm>

³<http://www.citeulike.org>

⁴<http://www.comscore.com/press/release.asp?press=1929>

⁵<http://bits.nytimes.com/2008/10/13/search-ads-come-to-youtube>

busca, classificação e recomendação. Embora estes serviços possam basear-se em atributos do próprio conteúdo multimídia (por exemplo, o histograma de cores de uma imagem), existem duas razões principais para não fazê-lo: (1) o uso eficaz de atributos do conteúdo multimídia, particularmente vídeos, é afetado pela sua qualidade, que em geral não é alta em aplicações Web 2.0 [3]; e (2) o custo computacional para processar mídias ricas (imagens, vídeos e áudio) e delas extrair informações úteis é relativamente alto, o que inviabiliza a sua aplicação na Web 2.0, porque esta contém um volume muito grande de dados. Por isso, na prática, a maioria dos serviços de RI em aplicações Web 2.0 exploram apenas os atributos textuais.

Neste contexto, é preciso investigar se a qualidade de atributos textuais é suficiente para subsidiar o projeto de serviços eficazes de RI. A qualidade desses atributos depende de três aspectos: (1) tais atributos devem ser explorados, no sentido de estar presente em conjunto com os objetos, e conter informação suficiente para serem utilizados por serviços de RI; (2) eles devem ser capazes de discriminar objetos para serviços como recomendação e ordenação em níveis de relevância para busca; e (3) os atributos devem capturar o conteúdo do objeto, oferecendo uma descrição razoável do mesmo, o que também é um aspecto importante para os serviços citados anteriormente⁶.

Este artigo aborda o problema acima, apresentando uma investigação inicial de evidências de qualidade de atributos textuais da Web 2.0 em relação aos três aspectos mencionados. A maioria dos trabalhos da literatura aborda apenas o atributo *tags* [8, 13, 15, 5, 9, 7, 16], não alcançando um consenso quanto a sua qualidade para serviços de RI. Para alcançar tal objetivo, foi feita uma extensa análise não apenas de *tags*, mas também de três outros atributos textuais comumente encontrados em aplicações Web 2.0, a saber, título, descrição e comentários. Além disso, foram analisadas quatro aplicações diferentes: Youtube, YahooVideo, LastFM e CiteULike, as quais são centradas em diferentes tipos de mídia e permitem diferentes níveis de colaboração entre os usuários.

Foram coletados atributos textuais de um conjunto de objetos das quatro aplicações selecionadas. Os dados coletados foram caracterizados com o objetivo de determinar a frequência de uso de cada atributo textual, bem como a quantidade e a diversidade de informação disponível neles. Os principais resultados obtidos foram: (1) alguns atributos, em particular os que são gerados colaborativamente, são significativamente inexplorados em algumas aplicações; (2) a quantidade de informação em cada atributo tende a ser maior em atributos colaborativos; (3) cada atributo textual, se presente, contribui com diferentes informações sobre o objeto; (4) atributos menores, como as *tags* e o título de um objeto, tem maior qualidade.

A principal contribuição deste artigo é a geração de um entendimento mais sólido de problemas relacionados à qualidade de atributos textuais comuns em aplicações Web 2.0. O conhecimento produzido pode auxiliar projetistas de tais aplicações no tratamento de questões como a utilidade e a importância de cada atributo como fonte de dados para tarefas de RI. Enfim, este trabalho oferece uma exploração mais detalhada que trabalhos anteriores, cobrindo diferentes

atributos textuais, em diferentes aplicações, com diferentes níveis de colaboração.

O restante do artigo é organizado da seguinte maneira: A Seção 2 sumariza trabalhos relacionados. Os atributos textuais analisados são apresentados na Seção 3. A Seção 4 discute a base de dados e a caracterização dos atributos, enquanto o seu poder descritivo e discriminativo são analisados na Seção 5. A Seção 6 contém conclusões e direções para trabalhos futuros.

2. TRABALHOS RELACIONADOS

A maioria dos trabalhos anteriores dão ênfase apenas ao atributo textual *tags*, incluindo a sua aplicação para melhorar serviços como busca, recomendação, agrupamento e indexação [8, 13, 15, 14] e a caracterização de aplicações que fazem uso deste atributo [5, 14]. No entanto, não há um consenso a respeito de sua qualidade. Algumas análises de qualidade existentes são discutidas a seguir.

Suchanek *et al* [16] mediram a qualidade de *tags* no sistema Delicious⁷ de acordo com os dicionários *online* Yago⁸ e Wordnet⁹. Os autores verificaram que apenas 46% das *tags* utilizadas pertenciam a esses dicionários e que aproximadamente metade dessas *tags* conhecidas tinham 10 ou mais significados. Foi visto também que as *tags* mais populares tendem a apresentar ruído semântico menor. Todavia, os problemas de palavras desconhecidas e polissêmicas, também levantados em [5], impõem desafios para tarefas de RI baseadas em *tags*. De forma similar, Bischoff *et al* [2] analisaram a qualidade de *tags* no contexto de busca na Web 2.0. Através da comparação das *tags* utilizadas com o conteúdo, metadados associados por especialistas e fontes externas, foi constatado que as *tags* tendem a representar bem o conteúdo dos objetos a que foram associadas e são relacionadas a termos de consultas feitas por usuários, o que também foi observado por outros autores [8].

Koutrika *et al* [7] modelaram o uso de sistemas de *tags*, separando usuários em duas categorias: *spammers* e *não-spammers*. O estudo objetivou promover o entendimento de como o *spam* afeta a busca por conteúdo através da contagem da quantidade de *spam* retornada em uma consulta. Os autores mostraram que *tags* aplicadas com maior frequência a um objeto têm maior qualidade e que uma máquina de busca que explora essa frequência como medida da qualidade de uma *tag* é menos suscetível ao *spam*.

Além da análise de *tags*, há também trabalhos que analisam comentários de usuários em *blogs* [11] e estudos da qualidade da mídia social encontrada em sistemas de pergunta e resposta [1]. Este trabalho é o primeiro esforço para analisar a qualidade de *diferentes atributos textuais* associados a um mesmo objeto multimídia em várias aplicações Web 2.0 diferentes. Além disso, diferentemente de trabalhos anteriores, foram analisados problemas relacionados aos três fatores de qualidade citados anteriormente. Em particular, foi feito uso de uma métrica de importância, previamente proposta para o contexto de atributos em páginas da Web tradicionais [4], para analisar o poder discriminativo e descritivo de atributos textuais na Web 2.0. Uma descrição detalhada sobre esta métrica é apresentada na Seção 5.

Esforços recentes para aplicar técnicas de RI multimí-

⁶Note que descrições precisas não são necessariamente adequadas para discriminar objetos, enquanto fontes de informação discriminativas podem não descrever bem o conteúdo do objeto

⁷<http://delicious.com>

⁸<http://www.mpi-inf.mpg.de/~suchanek/yago>

⁹<http://wordnet.princeton.edu>

dia em aplicações Web 2.0 também devem ser mencionados. Exemplos incluem o uso de atributos visuais e textuais para agrupar imagens similares no Flickr [12], e o uso de atributos de vídeos para anexar propagandas [10]. Entretanto, estas técnicas foram aplicadas em amostras pequenas de dados, e sua escalabilidade para a Web 2.0 ainda é um problema em aberto.

3. ATRIBUTOS TEXTUAIS NA WEB 2.0

Para realizar este trabalho, foram identificados quatro atributos textuais comumente utilizados (TÍTULO, DESCRIÇÃO, TAGS, e COMENTÁRIOS) em quatro aplicações diferentes da Web 2.0 (CiteULike, LastFM, Youtube e YahooVideo), apesar de eventualmente apresentarem denominações diferentes. O CiteULike, por exemplo, refere-se à DESCRIÇÃO como *abstract*, visto que este campo geralmente contém o resumo do artigo, e aos COMENTÁRIOS como *reviews*. Este último é denominado *shouts* no LastFM.

Os atributos textuais podem ser classificados de acordo com o nível de colaboração permitido pela aplicação. Em particular, os atributos textuais analisados podem ser ou *colaborativos* ou *restritivos*. Atributos colaborativos podem ser alterados ou criados por qualquer usuário, enquanto os restritivos podem ser modificados apenas por um usuário, tipicamente o proprietário do objeto, ou seja, quem o submeteu ao sistema. Esta propriedade é aqui chamada de *permissão de anotação*, uma generalização do termo “*tagging rights*”, previamente utilizado [9].

O atributo TAGS é colaborativo no CiteULike, LastFM e YahooVideo, visto que qualquer usuário pode adicionar *tags* aos objetos existentes nestas aplicações. No Youtube, ao contrário, apenas o proprietário do vídeo pode adicionar *tags* a ele. Enquanto restritiva no YahooVideo e Youtube, a DESCRIÇÃO tem natureza colaborativa no LastFM e no CiteULike. No LastFM, usuários podem alterar informações sobre um artista ou música de maneira semelhante a um *wiki*. No CiteULike, usuários podem prover diferentes resumos a uma mesma publicação, apesar de que verificou-se que isso é muito raro no conjunto de dados coletado. Nas quatro aplicações, o TÍTULO é restritivo e COMENTÁRIOS são colaborativos.

Algumas aplicações podem automaticamente preencher alguns desses campos no momento da submissão. O Youtube, por exemplo, adiciona o nome do arquivo de vídeo como seu TÍTULO e como TAGS, caso estes dados não sejam fornecidos. Contudo, é possível que os usuários removam *tags* adicionadas automaticamente quando um título for fornecido. No CiteULike, o usuário pode requisitar ao sistema a extração de TÍTULO e DESCRIÇÃO a partir de várias bibliotecas digitais. Entretanto, em todas as aplicações, exceto uma, usuários podem mudar todos os campos, respeitando sua permissão de anotação. A exceção é o LastFM, no qual o TÍTULO (ou seja, o nome do artista) é automaticamente inserido via software reprodutor de mídia, baseando-se nos metadados dos arquivos de música.

4. CARACTERIZAÇÃO DE ATRIBUTOS TEXTUAIS

Nessa seção, é caracterizado o uso dos quatro atributos textuais selecionados das quatro aplicações. O objetivo é levantar evidências de quais atributos são mais utilizados pelos usuários (Seção 4.1), a quantidade e a semântica da informação em cada atributo (Seção 4.2) e a diversidade de

informação entre os atributos (Seção 4.3), todos problemas relacionados à utilidade dos atributos para serviços de RI eficazes. Como será visto na Seção 5, a caracterização dará suporte aos resultados encontrados para o poder descritivo e discriminativo dos atributos.

Foram implementados coletores para amostrar atributos textuais de cada aplicação. Para Youtube, YahooVideo e LastFM, os coletores utilizam *snowball sampling* [6]. Cada coletor iniciou com uma certa quantidade de objetos como sementes (pelo menos 800). Para cada objeto em consideração, são retornados todos os objetos relacionados (seguindo links fornecidos pelas aplicações), armazenando-os em uma fila para coleta futura. No Youtube e YahooVideo, essas sementes foram selecionadas entre os vídeos mais populares. No LastFM, elas consistem nos artistas associados às *tags* mais populares. A amostra coletada a partir do CiteULike consiste dos objetos listados em um *snapshot* publicado pelo *site*.

Os coletores do Youtube, YahooVideo e LastFM foram executados por aproximadamente duas semanas, respectivamente em Julho, Setembro e Outubro de 2008. O *snapshot* do CiteULike foi coletado em Setembro de 2008. No total, foram coletados dados relativos a: 678.614 artigos do CiteULike; 193.457 artistas do LastFM; 227.562 e 211.081 vídeos do YahooVideo e Youtube, respectivamente. A partir deste momento, tais informações sobre um objeto serão referenciadas como objeto apenas.

4.1 Utilização de Atributos Textuais

	TÍTULO	TAGS	DESC.	COMENT.
CiteUL.	0.53% (R)	8.26% (C)	51.08% (C)	99.96% (C)
LastFM	0.00% (R)	18.88% (C)	53.52% (C)	54.38% (C)
Yahoo	0.15% (R)	16.00% (C)	1.17% (R)	99.88% (C)
Youtube	0.00% (R)	0.06% (R)	0.00% (R)	23.36% (C)

Table 1: Porcentagem de instâncias vazias (C = atributo colaborativo, R = atributo restritivo).

A Tabela 1 mostra, para cada atributo e aplicação, a porcentagem de instâncias vazias, ou seja, objetos sem termo algum no campo correspondente ao atributo. A permissão de anotação de cada atributo é mostrada entre parênteses.

A fração de instâncias vazias é muito maior para atributos colaborativos. De fato, alguns desses atributos, tais como COMENTÁRIOS, em todas as aplicações, exceto Youtube, e DESCRIÇÃO no LastFM e CiteULike são pouco utilizados. Mesmo o Youtube, com milhões de usuários, tem uma fração significativa de COMENTÁRIOS vazios, o que é consistente com o observado em *blogs* [11]. O atributo TAGS, foco da maioria dos esforços para melhorar serviços de RI, também está ausente em 16% dos objetos coletados do YahooVideo e quase 20% dos objetos coletados do LastFM. Somente o atributo TÍTULO, restritivo em todas as aplicações, está presente em praticamente todos os objetos.

Estes resultados podem ser explicados parcialmente pelas restrições impostas pelas aplicações, tais como o caso de TÍTULO e TAGS no Youtube, que podem ser preenchidos automaticamente. No entanto, nenhuma das aplicações exige o preenchimento do campo DESCRIÇÃO, por exemplo. Ainda assim, este atributo tem uma presença muito maior no Youtube e YahooVideo, onde ele é restritivo. De fato, no YahooVideo, ele está mais presente do que TAGS, um atributo colaborativo.

	TÍTULO				TAGS				DESCRIÇÃO				COMENTÁRIOS			
	DA	μ	CV	max	PA	μ	CV	max	PA	μ	CV	max	PA	μ	CV	max
CiteULike	R	7.5	0.40	73	C	4.0	1.33	194	C	65.2	0.5	3890	C	51.9	1.42	449
LastFM	R	1.8	0.47	23	C	27.4	1.49	269	C	90.1	1.06	3390	C	110.2	3.55	22634
YahooVideo	R	6.3	0.39	16	C	12.8	0.52	52	R	21.6	0.71	141	C	52.2	2.51	4189
Youtube	R	4.6	0.43	36	R	10.0	0.60	101	R	40.4	1.75	2071	C	322.3	1.94	16965

Table 2: Tamanho do vocabulário de instâncias de atributos textuais não-vazias.

Nota-se, portanto, que é necessário prover incentivos aos usuários para que eles utilizem com mais frequência os atributos colaborativos, mesmo quando os usuários têm interesse no objeto. Esse é um problema previamente levantado para sistemas de *tagging* [9].

Portanto, o uso de TAGS, DESCRIÇÃO ou COMENTÁRIOS como uma única fonte de dados para serviços de RI, independentemente de suas permissões de anotação, pode não ser eficaz, devido à falta de informação em uma parcela significativa de objetos.

4.2 Quantidade e Semântica da Informação

Foram analisadas a quantidade e as propriedades semânticas da informação disponível nos atributos textuais. Como esta análise é dependente de idioma, foi dada ênfase aos objetos de língua inglesa, aplicando um filtro simples que descarta objetos com menos do que três *stopwords*¹⁰ do idioma inglês em seus atributos textuais.

Após remover objetos com idioma diferente do inglês e instâncias de atributos vazias, restou um número diferente de objetos para análise. Esse número é maior que 150.000 para TÍTULO, DESCRIÇÃO e TAGS em todas as aplicações, exceto LastFM, para o qual ele excede 86.500 objetos. Nossos dados filtrados também incluem 152.717, 76.627, 6.037 e 76 objetos com COMENTÁRIOS no Youtube, LastFM, YahooVideo e CiteULike, respectivamente. Também foram removidos termos contendo caracteres não alfanuméricos e foi aplicado o algoritmo de Porter¹¹ para remoção de afixos. Ao final, foram eliminadas *stopwords*, porque estas palavras trazem pouca, ou nenhuma, informação semântica sobre os objetos.

Primeiramente, foi analisado o número de termos distintos (tamanho do vocabulário) em cada instância de atributo, o que representa a quantidade de informação disponível para uso em serviços de RI. A Tabela 2 apresenta a média μ , coeficiente de variação CV (razão entre desvio padrão e média), e valores máximos para os tamanhos do vocabulário. As colunas PA mostram as permissões de anotação.

Em geral, o atributo TÍTULO tem instâncias com o menor vocabulário, seguido por TAGS, DESCRIÇÃO e COMENTÁRIOS. Além disso, com poucas exceções, instâncias de atributos colaborativos tendem a ter vocabulários muito maiores em média, mas com grande variabilidade (CV). Por exemplo, a Tabela 2 mostra que instâncias de COMENTÁRIOS (quando presentes) têm o maior vocabulário, em média (exceto no CiteULike) exibindo um coeficiente de variação muito elevado. Similarmente, instâncias de DESCRIÇÃO no CiteULike e LastFM, e de TAGS no LastFM, também colaborativos, têm vocabulários maiores (tanto em média quanto nos valores percentuais), do que instâncias dos mesmos atributos

no Youtube, onde eles são restritivos. Entretanto, COMENTÁRIOS carregam muito menos informação no YahooVideo e CiteULike, onde são pouco utilizados (vide Seção 4.1). Comparando estes resultados com estudos prévios da aplicação Delicious, já havia sido verificado que o tamanho médio do vocabulário de instâncias de TAGS é próximo de 100 [16, 8], muito maior do que foi mensurado no presente trabalho, possivelmente devido ao fato de que uma entrada no Delicious é uma página que pode ter conteúdo mais amplo do que um objeto apenas.

Para entender se objetos mais populares contêm mais informações que outros, foi quantificada a correlação, medida pelo coeficiente de Pearson ρ , entre o tamanho do vocabulário em cada atributo e a popularidade do objeto, medida em número de visualizações no Yahoo e Youtube, número de ouvintes no LastFM e número de usuários que submeteram o mesmo artigo ao CiteULike. Em geral, foram encontradas correlações positivas para atributos colaborativos, tais como COMENTÁRIOS no LastFM ($\rho=0.5$) e Youtube ($\rho=0.24$), TAGS no CiteULike ($\rho=0.23$) e LastFM ($\rho=0.41$), e DESCRIÇÃO no LastFM ($\rho=0.25$). Uma exceção é o atributo TAGS no YahooVideo ($\rho=0.003$). Ao contrário de usuários do CiteULike e LastFM, que se sentem incentivados a atribuir *tags* para organizar suas bibliotecas pessoais de publicações e artistas, usuários do YahooVideo não parecem ter um incentivo para tal.

Em suma, atributos restritivos estão presentes na maioria dos objetos em todas as quatro aplicações, mas alguns dos atributos colaborativos são pouco explorados. Contudo, quando estes estão presentes, eles provêm uma quantidade maior de informação, especialmente em objetos mais populares. Ambas as características utilização e tamanho do vocabulário são relevantes para serviços de RI eficazes, porém o tamanho pode ser afetados por outros fatores como ruído que será discutido na Seção 5. Outros aspectos importantes serão discutidos a seguir.

Foi analisada também a popularidade de termos em cada atributo, dada pelo número de instâncias que contêm o termo. As Figuras 1(a-d) mostram que as distribuições são de cauda pesada, como observado para *tags* em [8, 2]. Para todos os atributos textuais, um número muito grande de termos aparecem apenas em uma pequena fração das instâncias, logo podem ser úteis para discriminação de objetos.

Também foram investigadas as propriedades semânticas do vocabulário de cada atributo. Como em [16], foram utilizados os dicionários *online* WordNet e Yago, mas os termos foram analisados em sua forma original no singular. Como observado para *tags* em [16], uma fração significativa (15-39%, em média) dos termos em todos os atributos e aplicações não foram encontrados nos dicionários.

A fração é maior ainda para COMENTÁRIOS, sendo, em média, 29% no YahooVideo e 37% nas outras aplicações. Dos termos conhecidos, foi encontrada uma grande fração (55%-80%) de substantivos, que tipicamente são bons descritores de conteúdo. Finalmente, foi observado algum grau

¹⁰ *Stopwords* são palavras muito frequentes em um texto, como artigos, pronomes e conjunções, que não carregam informação semântica. A lista de *stopwords* utilizada está disponível em: http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words

¹¹ <http://tartarus.org/~martin/PorterStemmer/>

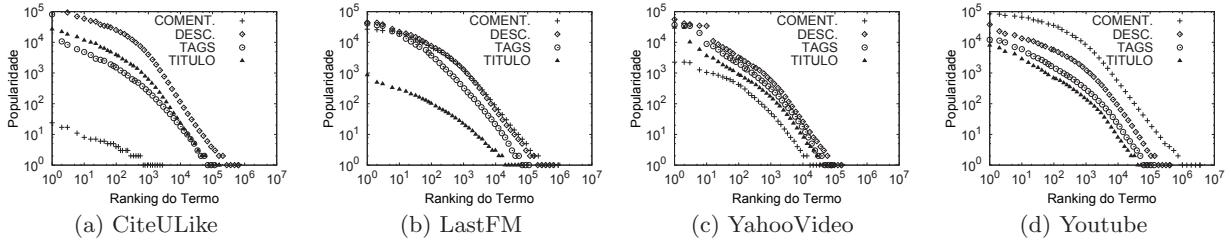


Figure 1: Distribuição da Popularidade dos Termos

de polissemia em todos os atributos e aplicações. A fração de termos com 10 ou mais significados varia de 6% a 20%. Isso é verdade também para DESCRIÇÃO e COMENTÁRIOS, que tipicamente contêm sentenças completas e não palavras isoladas, como em TAGS. Esse resultado mostra que o problema de polissemia afeta não apenas as *tags*, mas também os outros atributos textuais encontrados na Web 2.0.

4.3 Diversidade da Informação

Foi investigado se diferentes atributos contribuem com informações (termos) diferentes sobre os o objeto associado, conceito denominado de diversidade da informação. Com este resultado, pode-se afirmar que diferentes atributos trazem mais informação para descrever um objeto na Web 2.0, sendo assim uma motivação para utilizar todos os atributos. Como na Seção 4.2, foi dada ênfase ao idioma inglês, utilizando dados filtrados pela remoção de afixos. Foi quantificada a similaridade entre os atributos associados a um mesmo objeto, em termos de co-ocorrência, utilizando o coeficiente de Jaccard como métrica de similaridade. Dados dois conjuntos de itens T_1 e T_2 , o Coeficiente de Jaccard é calculado como:

$$J(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \quad (1)$$

A similaridade entre cada par de instâncias de atributo textual foi comparada, usando como entrada conjuntos dos N termos de maior peso $TF \times IFA$ em cada instância. O componente TF é igual à frequência interna do termo na instância de atributo. O componente *Inverso da Frequência no Atributo* (IFA) é uma pequena variação da métrica de RI IDF (*Inverse Document Frequency*) computada no escopo do atributo, ou seja, considerando todas as instâncias de um atributo como uma “coleção de documentos separada”. Dado o atributo A com n instâncias, e um termo t que ocorre em pelo menos uma instância de A , o IFA de um termo t em A é definido como:

$$IFA(t, A) = \log \frac{n}{f(t, A)}, \quad (2)$$

onde $f(t, A)$ é o número de instâncias de A em que t ocorre.

A Tabela 3 mostra os Coeficientes de Jaccard médios entre atributos para $N=5$ termos e, entre parênteses, quando todos os termos de cada atributo são considerados. Resultados para $N=15, 30$ e 45 estão no intervalo entre esses valores. Parece haver maior similaridade entre atributos restritivos (por exemplo, TÍTULO e DESCRIÇÃO no YahooVideo e Youtube), visto que o mesmo usuário é o responsável por atribuir termos a esses campos, tendendo a usar os mesmos

	CiteULike	LastFM	Yahoo	YouTube
TÍTULO×TAGS	0.08 (0.09)	0.03 (0.01)	0.30 (0.33)	0.22 (0.25)
TÍTULO×DESC.	0.19 (0.09)	0.17 (0.03)	0.23 (0.20)	0.16 (0.15)
TÍTULO×COMENT.	0.08 (0.05)	0.09 (0.02)	0.07 (0.03)	0.09 (0.02)
TAGS×DESC.	0.07 (0.02)	0.04 (0.04)	0.19 (0.20)	0.14 (0.14)
TAGS×COMENT.	0.03 (0.03)	0.03 (0.03)	0.05 (0.03)	0.07 (0.02)
DESC.×COMENT.	0.10 (0.07)	0.06 (0.03)	0.04 (0.03)	0.06 (0.03)

Table 3: Similaridade média (Coeficiente de Jaccard) entre instâncias de atributo textual não vazias.

termos. A exceção é TAGS no YahooVideo, que, apesar de colaborativo, tem similaridade maior que 0.19 com TÍTULO e DESCRIÇÃO. Todavia, os coeficientes de Jaccard são todos abaixo de 0.34. Logo, a co-ocorrência de termos importantes entre os atributos não é frequente, e cada atributo textual frequentemente traz informação nova sobre o objeto.

Dois pontos não levantados pelo coeficiente de Jaccard precisam ser investigados a fundo: (1) se esse pequeno vocabulário compartilhado tem termos espalhados na maioria dos atributos, o que pode ser um indicio de que estes termos são descritivos em relação ao conteúdo do objeto; e (2) o poder discriminativo relativo dos termos contidos nesses atributos, que contribui, de certa forma, com o poder discriminativo do atributo como um todo. O primeiro é relacionado a qualidade descritiva e o segundo a discriminativa, tais propriedades serão investigadas a seguir.

5. AVALIANDO O PODER DESCRITIVO E DISCRIMINATIVO DOS ATRIBUTOS

A capacidade de discriminar um objeto de outros e descrever o seu conteúdo de forma precisa são propriedades desejáveis em atributos utilizados em tarefas de RI. A primeira propriedade está relacionada ao poder de separar objetos relevantes, por exemplo, ela ajuda algoritmos de classificação a identificar automaticamente a categoria de um objeto, ou contribuem para que sistemas de busca melhor ordenem os objetos retornados por uma consulta de acordo com sua relevância. A segunda captura como o atributo textual está relacionado ao objeto multimídia, por exemplo, se a palavra “Sting” aparece nos atributos textuais de um vídeo é esperado que este vídeo tenha alguma relação com o cantor. Assim, é importante prover métricas para avaliar essas duas propriedades e com isso ajudar projetistas de aplicações Web 2.0 a criar bons atributos, acompanhar a utilidade dos existentes e utilizá-los adequadamente para o subsídio

	CiteULike			LastFM			YahooVideo			Youtube		
	EIA	IFA	I	EIA	IFA	I	EIA	IFA	I	EIA	IFA	I
TÍTULO	1.916	11.773	22.557	2.653	10.723	28.448	2.262	10.724	24.258	2.536	11.227	28.472
TAGS	1.629	11.635	18.953	1.328	10.112	13.429	1.868	10.495	19.605	2.073	11.135	23.083
DESCRIÇÃO	1.126	11.953	13.459	1.211	10.583	12.816	1.511	10.777	16.284	1.721	11.259	19.377
COMENTÁRIOS	-	-	-	1.207	10.623	12.822	-	-	-	1.124	11.517	12.945

Table 4: Valores de \overline{EIA} , \overline{IFA} e I para Recursos Textuais.

de serviços RI.

Para avaliar a qualidade de atributos textuais, foram inicialmente adotadas duas métricas previamente propostas para avaliar o poder descritivo e discriminativo de atributos de páginas da Web tradicionais [4], adaptando-as para atributos textuais de objetos da Web 2.0 (Seção 5.1). Também foi mensurada a entropia da informação nos atributos (Seção 5.2), como uma alternativa à análise do poder discriminativo. Por fim, foi realizado um experimento com usuários (Seção 5.3) onde foi mensurado como os usuários percebem os atributos em relação ao objeto multimídia, sendo assim uma outra forma de identificar o poder descritivo.

Nestas análises, foram utilizados os quatro conjuntos de dados filtrados, considerando apenas objetos com todas as instâncias de atributos textuais não-vazias. Dada a fração desprezível de instâncias não-vazias do atributo COMENTÁRIOS no YahooVideo e CiteULike, este atributo foi desconsiderado nessas duas aplicações. Onde não explicitamente citado, os valores médios apresentados foram analisados com intervalos de confiança de 95%, e não houve diferença estatística entre as médias pois o erro era desprezível (no máximo 1.5%).

5.1 Métricas de Importância de Atributos

Para tentar capturar o poder discriminativo de um atributo, utilizou-se a métrica IFA , definida na Equação 2. Ela baseia-se na hipótese de que termos mais raros na coleção são mais discriminativos. Por exemplo, enquanto a ocorrência do termo “vídeo” em um TÍTULO de um objeto do Youtube traz pouca informação a respeito do seu conteúdo, a ocorrência do termo “Sting” pode ser mais útil para diferenciá-lo de outros objetos. Assim, para capturar a qualidade discriminativa de um atributo como um todo, foi definida a *Média do Inverso da Frequência no Atributo*, $\overline{IFA}(A)$, como a média dos valores do IFA de todos os termos que ocorrem em todas as instâncias do atributo A , ou seja:

$$\overline{IFA}(A) = \frac{\sum_{t \in A} IFA(t, A)}{V_A}, \quad (3)$$

onde V_A é o tamanho do vocabulário completo do atributo A , ou seja, a união de todos os termos de todas as instâncias de A .

Por outro lado, para avaliar o poder descritivo de um atributo, foi proposta uma métrica denominada *Espalhamento*. Sendo, uma instância $a \in A$ um atributo textual de um determinado objeto o , o grau de espalhamento, $E(t, a)$ de um termo t em um dado objeto o é definido como o número de instâncias de atributos que contém t . Esta métrica baseia-se na hipótese de que quanto mais atributos contiverem um termo t , mais t estará relacionado ao conteúdo de o . Por exemplo, se o termo “Sting” aparece em todos os atributos de um dado vídeo, há uma alta probabilidade de que o vídeo seja relacionado ao cantor.

O *Espalhamento de uma Instância de Atributo* ($EIA(a)$)

é definido como o valor médio dos valores $E(t, a)$. Essa métrica mede o quanto os termos de uma dada instância de um atributo estão relacionados ao conteúdo de instâncias de outros atributos associadas ao mesmo objeto. Portanto, trata-se de uma heurística para estimar o grau de relacionamento de uma instância de atributo com o conteúdo do objeto em si.

Levando-se em consideração o conjunto de todas as instâncias, o espalhamento produz resultados que podem estimar o quanto, em média, um atributo está conectado aos objetos que eles representam. Logo, foi computado o *Espalhamento do Atributo* ($\overline{EIA}(A)$) como a média dos valores do $EIA(a)$ de todas as instâncias do atributo A , ou seja:

$$\overline{EIA}(A) = \frac{\sum_{a \in A} EIA(a)}{N_A}, \quad (4)$$

onde N_A é o número de instâncias de A .

Finalmente, os valores de \overline{EIA} e \overline{IFA} são usados para estimar a *Importância do Atributo* (I), que mede a qualidade de um atributo considerando a capacidade discriminativa e descritiva. A importância I de um atributo A é dada por:

$$I(A) = \overline{EIA}(A) \times \overline{IFA}(A) \quad (5)$$

As métricas \overline{IFA} , \overline{EIA} e I foram aplicadas para comparar a qualidade dos atributos. Como mostra a Tabela 4, os valores do \overline{EIA} fornecem uma ordenação consistente dos atributos para todas as quatro aplicações. Atributos menores, TÍTULO e TAGS, têm maior qualidade que os maiores DESCRIÇÃO e, quando considerados, COMENTÁRIOS. Isso é esperado porque os atributos menores têm a funcionalidade de resumir o conteúdo do objeto em poucas palavras, que são repetidas nos outros atributos.

Em contrapartida, os valores do \overline{IFA} não mostram uma distinção clara entre os atributos de uma aplicação, tendo uma diferença relativa menor que o \overline{EIA} . Isso deve-se ao fato de que as distribuições de cauda pesada da popularidade de termos (Figura 1) aumentam o poder discriminativo de todos os atributos de forma similar. Este fator pode ajudar a explicar o porquê de valores absolutos do \overline{IFA} serem muito maiores que os encontrados para os bem comportados sítios de notícias e fóruns na Web tradicional [4]. Logo, a quantidade não desprezível de ruído em todos os atributos analisados afeta o seu poder discriminativo, de acordo com a heurística \overline{IFA} .

Tendo em vista que termos raros afetam a distinção entre os valores do \overline{IFA} de cada atributo, estes foram computados desconsiderando termos raros (com frequência absoluta menor ou igual a N). Foram testados limiares no intervalo $1 < N < 1000$ e isso mostrou que, em geral, a retirada dos termos muito raros aumenta levemente a diferença entre os valores \overline{IFA} de cada atributo. A Tabela 5 mostra estes valores para $N = 50$.

Portanto, de acordo com a métrica \overline{IFA} filtrada, os valores

	CiteULike	LastFM	YahooVideo	Youtube
TÍTULO	7.311	6.644	6.674	7.126
TAGS	7.595	5.997	6.548	6.997
DESC.	7.028	5.833	6.371	6.739
COMENT.	-	5.908	-	6.649

Table 5: Valores de \overline{IFA} desprezando termos com frequência absoluta menor que 50

de TÍTULO e TAGS são melhores do que aqueles de *descrição* e *comentários*.

5.2 Entropia da Informação

Como alternativa à métrica \overline{IFA} , pode-se fazer uso do cálculo de entropia da informação. No contexto de classificação de textos, a entropia mensura o quanto um termo é indicador de uma categoria. Objetos com termos mais concentrados em uma determinada categoria tem maiores indícios de serem desta categoria. É esperado que os resultados da entropia sejam consistentes com aqueles do \overline{IFA} , pois ambos mensuram a qualidade discriminativa de termos em cada atributo.

As aplicações de compartilhamento de vídeo estudadas, YahooVideo e Youtube, provêm categorias para que os usuários possam atribuir aos vídeos que eles submeteram. “Notícias e Política”, “Ciência e Tecnologia” e “Música” são exemplos de categorias presentes em ambos os sistemas¹². Essa informação foi utilizada para calcular a entropia dos termos em relação a estas categorias. Para calcular a entropia, precisa-se primeiramente calcular a concentração de um termo em um dada categoria, definida por:

$$h(t, c, A) = \frac{f(t, c, A)}{f(t, A)}, \quad (6)$$

onde $f(t, c, A)$ é a frequência do termo t na categoria c no atributo textual A , e $f(t, A)$ é a frequência de t em toda a coleção de instâncias do atributo A , ou seja, $f(t, A) = \sum_{c \in C} f(t, c, A)$, onde C é o conjunto de categorias. A entropia ($H(T)$) é definida como:

$$H(t, A) = - \sum_{c \in C} h(t, c, A) \log h(t, c, A) \quad (7)$$

Quanto menor o valor da entropia do termo, maior será a sua a qualidade discriminativa. O valor mínimo para a entropia é zero, que ocorre quando $h(c, t, A) = 1$, ou seja, todos as ocorrências de t pertencem à categoria c .

É provável que atributos que contêm, em geral, termos com baixa entropia permitem discriminar melhor objetos em categorias, tendo assim uma qualidade discriminativa maior. Para testar isso, foram comparados os valores médios das entropias dos termos de cada atributo textual.

Observou-se que, de maneira semelhante ao que ocorreu com o \overline{IFA} , a entropia foi afetada pela presença de termos muito raros. Por exemplo, quando um termo é único na coleção de objetos (portanto pertence a uma única categoria), sua entropia é igual a zero. De fato, verificou-se que isso ocorre para mais de 65% dos termos em todos os atributos textuais.

Assim, também foi feita uma análise desprezando termos raros. Ambos os resultados são representados na Tabela

¹²As listas completas de categorias do Youtube e do YahooVideo podem ser encontradas respectivamente em <http://video.yahoo.com/> e <http://www.youtube.com/browse>

6. Os valores das colunas rotuladas com (1) são as médias de valores da entropia de todos os termos, enquanto os das colunas rotuladas com (2) são as médias dos valores da entropia dos termos com frequência maior que 50¹³. De acordo com os valores da tabela, obtêm-se uma mesma ordenação de qualidade que os valores do $\overline{IFA}(A)$ quando se consideram todos os termos. Quando os termos raros são desprezados, as entropias de TAGS e TÍTULO são maiores do que as de DESCRIÇÃO e COMENTÁRIOS, de forma similar ao $\overline{IFA}(A)$. Porém, é importante ressaltar que, no caso do $\overline{IFA}(A)$, TÍTULO obteve uma maior qualidade do que o atributo TAGS.

A diferença entre os resultados existe pois a entropia mensura a capacidade discriminativa do atributo em um conjunto de categorias, enquanto o $\overline{IFA}(A)$ mensura a dispersão dos termos independente de categoria. Assim, os termos em TAGS são mais populares que os de TÍTULO (o que explica o seu menor $\overline{IFA}(A)$), porém menos espalhados entre categorias, o que é mensurado pela entropia.

	Youtube		YahooVideo	
	H (1)	H (2)	H (1)	H (2)
TÍTULO	0.32	1.70	0.32	1.48
TAGS	0.34	1.61	0.36	1.63
DESC.	0.30	1.85	0.29	1.78
COMENT.	0.24	2.01	-	-

Table 6: Valores da entropia considerando todos os termos (1) e filtrando os não populares (2).

É importante ressaltar que são usuários que atribuem categorias aos objetos de acordo com a lista pré-determinada pela aplicação, portanto essa categorização pode estar sujeita a erros devido ao comportamento dos usuários ou à dificuldade de enquadrar um objeto em uma categoria. Porém, a nossa avaliação parte da premissa de que, a despeito da subjetividade associada à atribuição de uma categoria a um vídeo e eventuais erros por parte dos usuários, a maioria dos vídeos, em ambos sistemas, têm categorias que refletem, razoavelmente bem, os conteúdos dos objetos.

5.3 Experimento com Usuários

Como uma outra forma de verificação do poder descritivo, foi realizado um experimento com usuários, no qual estes foram solicitados a assistir a vídeos selecionados do Youtube e dar notas aos quatro atributos textuais de acordo com a sua capacidade de descrever os vídeos assistidos. Este experimento é comparado com a qualidade descritiva dos atributos, capturados pela métrica \overline{EIA} . Dado o custo desse tipo de experimento, foi dada ênfase somente ao Youtube, visto que este tem a maior audiência e trata com o tipo de mídia mais rico e desafiador, o vídeo.

Um total de 17 voluntários participaram do experimento. Os voluntários selecionados foram alunos do curso de Ciência da Computação da Universidade Federal do Amazonas não diretamente ligados à esta pesquisa. Foi desenvolvida uma aplicação simples para mostrar aos voluntários apenas o vídeo e o texto dos quatro atributos textuais analisados, sem mostrar quais eram os atributos. Dez dos mais populares vídeos da amostra coletada do Youtube, com duração entre 4 e 10 minutos, foram assistidos por cada participante, totalizando 170 avaliações. A escolha do intervalo de duração

¹³Outros limiares foram testados, obtendo-se resultados semelhantes

foi baseada em evidências empíricas de experimentos iniciais que mostraram ser suficiente para os usuários compreenderem o conteúdo do vídeo.

Com o vídeo, foi mostrado o conteúdo completo do TÍTULO, TAGS, e DESCRIÇÃO bem como os 5 comentários mais recentes, devido à verbosidade desse atributo. O enunciado dado aos usuários foi: “Associe a cada fragmento de texto apresentando uma nota baseada em sua capacidade de descrever o conteúdo deste vídeo”. As notas possíveis eram: (0) O texto não está relacionado com o conteúdo deste vídeo; (1) Partes do texto estão relacionadas ao vídeo mas, em geral, ele não o descreve bem; (2) O texto está relacionado ao vídeo, descrevendo-o.

Com 95% de confiança, as notas associadas por usuários a TÍTULO, TAGS, DESCRIÇÃO e COMENTÁRIOS foram 1.62 ± 0.09 , 1.57 ± 0.08 , 1.44 ± 0.1 , e 0.89 ± 0.09 , respectivamente. Logo, de acordo com a percepção do usuário, TÍTULO e TAGS são os atributos com a maior qualidade, estatisticamente sem diferenças entre suas notas. Isto é de certa maneira consistente com os resultados da aplicação da métrica I , que claramente aponta estes dois atributos como os melhores em termos de poder discriminativo e descritivo. Note que a percepção de usuário é muito subjetiva e pode ser influenciada não apenas pelo poder descritivo, mas também pela quantidade e diversidade de informação, que, em última instância, impacta o poder discriminativo dos atributos. A DESCRIÇÃO é classificada em uma posição intermediária, ainda com uma nota alta. COMENTÁRIOS têm a nota mais baixa, visto que os usuários geralmente o utilizam para debater e não necessariamente descrever o conteúdo, uma tendência também observada em *blogs* [11].

Apesar desta nota mais baixa poder ter sido influenciada pelo número limitado de COMENTÁRIOS mostrados aos voluntários, os resultados estão consistentes com as heurísticas propostas e com os resultados de entropia, ambos computados considerando *todos* os comentários.

6. CONCLUSÕES

Este trabalho provê uma análise de evidências da qualidade de atributos textuais para serviços de recuperação de informação em aplicações Web 2.0. Para isto, foram coletados dados de quatro atributos textuais (TÍTULO, DESCRIÇÃO, TAGS e COMENTÁRIOS) associados a objetos de quatro aplicações Web 2.0 (CiteULike, LastFM, YahooVideo e YouTube), estendendo trabalhos anteriores.

A caracterização do uso dos quatro atributos revelou que os colaborativos, incluindo TAGS, são significativamente explorados em algumas aplicações, embora, quando presentes, tendem a conter maior quantidade de informação do que atributos não-colaborativos, tais como TÍTULO. Também foi encontrada uma diversidade significativa de informação disponível entre os atributos. Além disso, uma quantidade não desprezível de termos raros e desconhecidos afeta a qualidade dos quatro atributos em todas as quatro aplicações.

A avaliação do poder discriminativo e descritivo dos atributos foi realizada por meio de métricas heurísticas de importância, cálculo de entropia e experimentos com usuários. Os três experimentos mostraram que TÍTULO e TAGS parecem ser os atributos mais promissores para exploração em serviços de RI. O primeiro apresenta evidências de qualidade descritiva e discriminativa nas heurísticas e experimento com o usuário, o segundo contém evidências de qualidade mais forte quando se considera a entropia, sendo este um melhor

discriminador de objetos em categorias. Porém, TAGS são afetadas pela baixa utilização em alguma aplicações.

7. AGRADECIMENTOS

O presente trabalho foi realizado com o apoio do UOL (www.uol.com.br), através do Programa UOL Bolsa Pesquisa, processo número 20090215103600. A pesquisa também teve apoio do Instituto Brasileiro de Ciência e Tecnologia para Pesquisa Web (MCT/CNPq/INCT 573871/2008-6), o projeto REBU (MCT/CNPq/CTInfo 550995/2007-2) e o projeto Infoweb (MCT/CNPq/CTInfo 550874/2007-0).

8. REFERENCES

- [1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding High-Quality Content in Social Media. In *Proc. WSDM*, 2008.
- [2] K. Bischoff, F. Claudiu-S, N. Wolfgang, and P. Raluca. Can All Tags Be Used for Search? In *Proc. CIKM*, 2008.
- [3] S. Boll. MultiTube—Where Web 2.0 and Multimedia Could Meet. *IEEE Multimedia*, 14(1), 2007.
- [4] D. Fernandes, E. de Moura, B. Ribeiro-Neto, A. da Silva, and M. Gonçalves. Computing Block Importance for Searching on Web Sites. In *Proc. CIKM*, 2007.
- [5] S. Golder and B. Huberman. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2), 2006.
- [6] L. A. Goodman. Snowball Sampling. *Annals of Math. Statistics*, 32(1), 1961.
- [7] G. Koutrika, F. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating Spam in Tagging Systems. In *Proc. AIRWeb*, 2007.
- [8] X. Li, L. Guo, and Y. Zhao. Tag-based Social Interest Discovery. In *Proc. WWW*, 2008.
- [9] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, Tread. In *Collaborative Web Tagging Workshop (WWW'06)*, 2006.
- [10] T. Mei, L. Yang, X. Hua, H. Wei, and S. Li. VideoSense: a Contextual Video Advertising System. In *Proc. ACM Multimedia*, 2007.
- [11] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *Proc. Workshop on Weblogging Ecosystem (WWW'06)*, 2006.
- [12] M. Rege, M. Dong, and J. Hua. Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering. In *Proc. WWW*, 2008.
- [13] R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. Parreira, and G. Weikum. Efficient Top-k Querying Over Social-Tagging Networks. In *Proc. SIGIR*, 2008.
- [14] B. Sigurbjornsson and R. van Zwol. Flickr Tag Recommendation Based on Collective Knowledge. In *Proc. WWW*, 2008.
- [15] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W. Lee, and C. Giles. Real-Time Automatic Tag Recommendation. In *Proc. SIGIR*, 2008.
- [16] F. Suchanek, M. Vojnovic, and D. Gunawardena. Social Tags: Meaning and Suggestions. In *Proc. CIKM*, 2008.