EVIDÊNCIAS DE QUALIDADE DE ATRIBUTOS TEXTUAIS NA WEB 2.0

FLAVIO VINICIUS DINIZ DE FIGUEIREDO

EVIDÊNCIAS DE QUALIDADE DE ATRIBUTOS TEXTUAIS NA WEB 2.0

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientadora: Jussara Marques de Almeida
Co-Orientador: Marcos André Gonçalves

Belo Horizonte
Maio de 2010

Figueiredo, Flavio Vinicius Diniz de.

F475e Evidências de qualidade de atributos textuais na web 2.0. / Flavio Vinicius Diniz de Figueiredo. — Belo Horizonte, 2010.
xiii, 49 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais. Departamento de Ciência da Computação.

Orientadora: Jussara Marques de Almeida.
Co-Orientador: Marcos André Gonçalves.

1. Mídia social - Teses. 2. Web 2.0 - Teses.
I. Orientador. II Título.

CDU 519.6*73(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Evidências de qualidade de atributos textuais na Web 2.0

FLAVIO VINICIUS DINIZ DE FIGUEIREDO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Profa. Jussara Marques de Almeida - Orientadora
Departamento de Ciência da Computação - UFMG

Prof. Marcos André Gonçalves - Co-orientador
Departamento de Ciência da Computação - UFMG

Profa. Cristina Duarte Murta
CEFET - MG

Prof. Edleno Silva de Moura
Departamento de Ciência da Computação - UFAM

Prof. Pável Pereira Calado
Instituto de Engenharia de Sistemas e Computadores - INESC

Belo Horizonte, 18 de maio de 2010.

Agradecimentos

Antes de tudo, agradeço a minha família: Mother, Father, Dudu e Fabricio; por todo o apoio ao longo dos anos. Uma adição especial durante o mestrado foi Vinicius, que simplesmente por estar la destruindo tudo traz alegria para todos. Sou grato também por ter tido Vanessa como companheira durante todo este tempo, muitas coisas não seriam possíveis sem a ajuda dela. Agradeço também aos meus orientadores Jussara e Marcos, por me guiarem neste mundo de mestrado.

Uma lista de amigos/amigas para agradecer neste espaço seria por demais extensa. Assim, vou generalizar tudo em um único agradecimento para todos aqueles que durante momentos de: felicidade, brigas, bebidas, bares, festivais, shows, caronas, viagens, conversas (aleatórias ou filosóficas), dentre outras coisas; me guiaram para eu ser a pessoa que sou hoje.

"Arrakis teaches the attitude of the knife - chopping off what's incomplete and saying: Now it's complete because it's ended here."

(Frank Herbert)

Resumo

Com o advento da Web 2.0 os usuários da Internet deixaram de ser apenas consumidores de conteúdo para ter um papel mais ativo como criadores do mesmo. O resultado desta criação colaborativa de conteúdo, denominado de mídia social, fez com que aplicações da Web 2.0 alcançassem imensa popularidade (em quantidade de usuários e volume de dados) nos últimos anos. Porém, por não oferecer garantias de qualidade, o uso eficaz da mídia social em tarefas de Recuperação de Informação (RI) se apresenta como um desafio atualmente.

Com o intuito de prover um melhor entendimento sobre o problema supracitado, esta dissertação faz um estudo comparativo da qualidade de diferentes atributos textuais quando utilizados como subsídios para tarefas de RI, sendo um atributo textual uma região dentro da página do objeto contendo um texto com um tópico ou funcionalidade bem definida. Em especial, estudamos os atributos textuais Título, Etiquetas, Descrição e Comentários nas aplicações CiteULike, LastFM, Youtube e YahooVideo. O nosso estudo é divido em três etapas, iniciando com uma extensiva caracterização da qualidade dos atributos textuais em relações aos aspectos de qualidade: 1) uso dos atributos; 2) quantidade de conteúdo; 3) corretude sintática; 4) e por fim, capacidade descritiva e discriminativa. Após a caracterização, comparamos a qualidade dos atributos quando utilizados em duas tarefas de RI, a classificação e a recomendação. Por fim, através de um estudo com 17 voluntários, buscamos entender como os usuários percebem a qualidade dos atributos textuais.

Nossa pesquisa amplia significativamente o conhecimento da literatura, podendo ser utilizados por provedores e designers de aplicações Web 2.0 para melhorar os seus serviços de RI e desenvolver aplicações considerado os atrativos de cada atributo textual.

Abstract

The advent of the Web 2.0 changed the way users interact with web applications on the Internet. Nowadays, such users are not only passive content consumers but also content creators on the Web. The result of this collaborative form of content creation, known as social media, is one of the main factors behind the massive popularity (in amount of users and data) of the Web 2.0. One shortcoming of social media is the possible lack of quality of the content generated by users. In fact, one recent study pointed this lack of quality as one of the reasons why Information Retrieval (IR) services do not effectively explore social media yet.

In order to provide insights on the matter, this dissertation presents a comparative study of the quality of different textual features on the Web 2.0. A textual feature is a region of a web page with a well defined topic and functionality. We studied the textual features Title, Tags, Descriptions and Comments on LastFM, CiteULike, Youtube and YahooVideo. Our study consist of three different comparisons. Firstly, we compare the quality of features regarding three aspects of quality, namely: usage, amount of content, syntactic correctness, descriptive quality and discriminative quality. After our characterization, we compare textual features when applied to the IR tasks of object classification and recommendation. Lastly, a study with 17 volunteers was performed in order to compare how users perceive the quality of features.

Our work extends previous work which focus mostly on Tags. The results presented on this dissertation can be explored by Web 2.0 providers and designers in order enhance their IR services and develop Web 2.0 applications considering the benefits and shortcomings of each textual feature.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Este capítulo tem o objetivo de motivar a pesquisa apresentada nesta dissertação. Inicialmente descrevemos os conceitos de *Web 2.0* e *mídia social*, essenciais para o entendimento do nosso trabalho (Seção 1.1). Após isto, apresentamos as principais motivações que levaram a realização desta pesquisa, em seguida, descrevemos os objetivos e contribuições do nosso estudo (Seção 1.2).

1.1 Web 2.0 e a Mídia Social

Embora não exista uma definição exata para o termo Web 2.0, alguma características comuns existem entre aplicações Web 2.0 [44], estas são: a facilidade e simplicidade das aplicações, no quesito de desenvolvimento e uso; reconhecimento do valor de dados, principalmente aqueles que possam ser úteis para terceiros; modelos de lucros baseados na ideia de consumo seguindo uma distribuição de probabilidade de cauda pesada; e a inteligência colaborativa.

A natureza colaborativa ajudou aplicações Web 2.0 à alcançarem um alto nível de popularidade nos últimos anos. Alguns exemplos de sucesso são o sistema de compartilhamento de vídeos YouTube[1], o aplicativo de vídeo de maior popularidade no mundo [36, 37]. As rádios da Internet, como o LastFM[2] e o Pandora[3], estão entre as mais populares da Internet [47]. Mesmo não tendo as músicas geradas pelos usuários, estas rádios permitem que os usuários descrevam o conteúdo musical de forma colaborativa e interajam formando redes sociais. A Amazon[4] é um exemplo similar ao

[1] `http://www.youtube.com`
[2] `http://www.last.fm`
[3] `http://www.pandora.com`
[4] `http://www.amazon.com`

anterior, que suporta serviços de colaboração entre usuários, como a escrita de críticas sobre produtos na forma de texto, imagem e vídeo. Outros exemplos são os sistema de gerência pessoal de favoritos como o CiteULike[5] e o Delicious[6], assim como os portais de perguntas e respostas como o Yahoo! Answers[7]. Estes são apenas alguns de diversos casos de aplicações de sucessos que adotam o modelo Web 2.0.

O termo *mídia social*, do inglês *social media* [8], é comumente utilizado para denominar a informação que é produzida por usuários de aplicações Web 2.0. O termo foi definido em contrapartida a *mídia industrial* onde a informação é gerida por profissionais de diversas áreas específicas. O escopo da mídia social abrange todas as formas colaborativas de geração de informação, incluindo, dentre outras coisas, dados, meta-dados e a informação gerada por relações sociais. Alguns exemplos são: objetos multimídia, como os vídeos no YouTube; meta-dados sobre objetos como as etiquetas; comentários e críticas dos usuários sobre determinado assunto; e as descrições colaborativas dos usuários sobre um determinado assunto, como nos *wikis*.

1.2 Motivação, Objetivos e Contribuições

Os exemplos citados acima demonstram a importância da Web 2.0, e em consequência da mídia social, no contexto da Internet hoje em dia. O foco da nossa pesquisa é nas implicações que esta mudança de mídia industrial para mídia social trouxe para serviços computacionais que fazem uso desta informação. A Recuperação de Informação (RI) é a ciência da busca de informação em diferentes fontes de dados (e.g. banco de dados, documentos, metadados e a Web) [59]. Devido a forma não supervisionada que os usuários inserem e/ou alteram conteúdo para ser disponibilizado pelas aplicações, a mídia social não oferece garantias de qualidade para serviços de RI. Este é um dos motivos levantados por Boll [6] pelo qual as aplicações da Web 2.0 não fazem uso de técnicas de RI multimídia [48, 53], explorando de forma primordial dados textuais apenas. A autora argumenta que: (1) algoritmos complexos utilizados na recuperação de informação sobre dados multimídia não escalam para a quantidade de conteúdo na Web 2.0; (2) a falta de qualidade da mídia social é um problema para a extração de informação destes dados. Um exemplo da falta de qualidade da mídia social é o spam, gerado por atitudes maliciosas e/ou oportunistas dos usuários, um problema já identificado por trabalhos anteriores [2, 3, 32, 33, 42].

[5] `http://www.citeulike.com`

[6] `http://www.delicious.com`

[7] `http://www.answers.yahoo.com`

[8] Outros termos populares em inglês são: *user generated content (UGC)* e *consumer generated media (CGM).*

A pergunta que levantamos neste trabalho é: *Qual é a qualidade dos diferentes atributos textuais da Web 2.0 para subsidiar de tarefas de RI?* Atualmente atributos textuais são a principal fonte de informação utilizado. Um atributo textual pode ser definido como sendo uma região dentro da página Web do objeto contendo um texto com um tópico ou funcionalidade bem definida [8, 16]. Alguns exemplos de atributos textuais na Web 2.0 são as etiquetas, comentários, títulos e descrições. Estes atributos são associados a objetos da Web 2.0, onde um objeto é um pedaço de informação apresentada com algum formato multimídia (texto, áudio e vídeo). A Figura 1.1 demonstra um exemplo de uma página na Web 2.0 junto com seu objeto e atributos textuais.

Nós sumarizamos a qualidade de um atributo textual de acordo com os seguintes aspectos:

1. Um atributo textual deve estar presente (ter ao menos um termo) em uma fração significativa de objetos para ser útil. Se o uso do atributo for pouco frequente nos objetos, tarefas de RI não vão conseguir fazer uso deste;

2. além de presentes, os atributos textuais devem conter uma quantidade mínima de conteúdo para serem uteis;

3. o conteúdo em um atributo textual deve ser sintaticamente correta;

4. o conteúdo dos atributos devem ter uma qualidade mínima para *discriminar* conteúdo relevante do irrelevante em resultados de um serviço de RI;

5. os atributos devem *descrever* efetivamente o conteúdo associado aos mesmos, capturando a informação do conteúdo.

Inicialmente, caracterizamos os dados com o objetivo de extrair evidências de qualidade de acordo com os aspectos apresentados acima (uso, quantidade, corretude, poder descritivo e discriminativo). Para isto, amostramos atributos textuais associados a objetos de quatro aplicações Web 2.0 populares. Os atributos estudados foram o título, etiquetas, descrições e comentários. As aplicações foram os sistemas de compartilhamento de vídeos Yahoo! Vídeo (de agora em diante referido como YahooVideo) e YouTube, a rádio da Internet Last FM (de agora em diante referido como LastFM) e o CiteUlike, aplicação utilizada para a organização pessoal de trabalhos científicos. Cada sistema tem um objeto (vídeo, música ou texto) de foco diferente. Os resultados desta caracterização serão apresentados no Capítulo 3.

Também foram feitas caracterizações para analisar a diversidade do conteúdo entre os atributos. Nesta etapa de caracterização, nossos principais resultados foram:

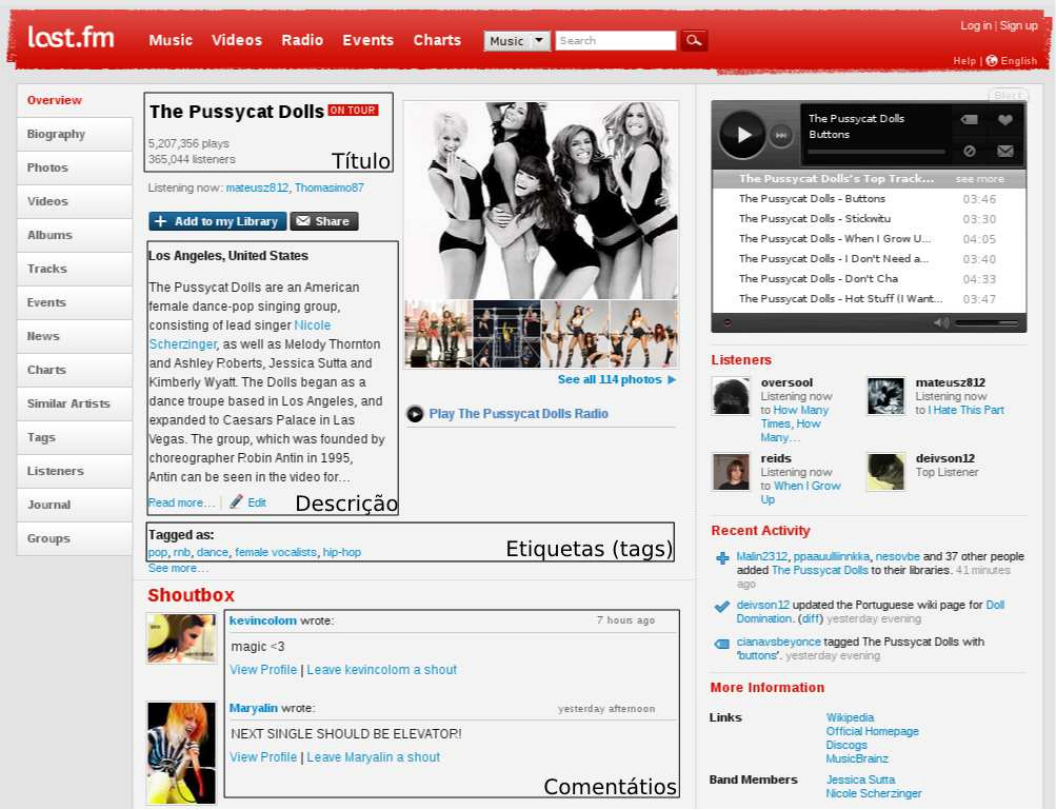


Figura 1.1: Exemplo de Atributos Textuais em uma Página Web 2.0

(1) todos os atributos, exceto o TÍTULO, são inexplorados (contém zero termo) em uma fracão significativa de objetos; (2) atributos que podem ser alterados por vários usuários, denominados de colaborativos, em geral tem mais informação; (3) atributos menores e mais utilizados, como TÍTULO e ETIQUETAS, tem capacidades descritivas e discriminativa maiores do que os outros; (4) todos os campos textuais contém uma fração significativa de palavras erradas ou com vários significados; e que (5) existe uma diversidade de conteúdo entre atributos textuais associados a um mesmo objeto, este é um indício de que cada atributo pode trazer informação nova sobre o mesmo objeto.

Após a caracterização dos dados, nós avaliamos a qualidade dos atributos textuais quando utilizados em duas tarefas comuns de RI: a classificação e a recomendação de etiquetas. A primeira tarefa de RI analisada, a classificação, foi escolhida pois tem aplicação em diversos serviços como a construção de diretórios Web e o ranqueamento de resultados de busca em níveis de relevância [30]. Esta tarefa também nos foi útil pois ela possibilita uma verificação automatizada da qualidade, dado que objetos das aplicações YouTube e Yahoo Vídeo já contém classes assinaladas aos mesmos. No caso do LastFM, objetos podem ter classes extraídas de outras páginas da Web como o dire-

tório musical All Music Guide [9]. Nossos experimentos de classificação foram realizados considerando o texto dos atributos textuais como as características dos objetos nas Web 2.0, esta é uma estratégia comum para classificar documentos textuais [30]. No nosso caso, consideramos seis estratégias diferentes de classificação, as quatro primeiras fazem uso de cada atributo textual isoladamente, enquanto as duas últimas são diferentes formas de combinar o conteúdo dos quatro atributos. Os nossos resultados de classificação foram condizentes com nossa caracterização e mostram que: (1) quando considerado isoladamente, ETIQUETAS é o atributo mais promissor para a tarefa de classificação, obtendo os melhores resultados; (2) a combinação do conteúdo dos quatro atributos obtém resultados melhores do que o uso de atributos isolados, isto confirma que cada atributo contém informação relevante sobre os objetos; e por fim, (3), a qualidade dos atributos é afetada pela quantidade de informação, pois embora TÍTULO tenha sido o melhor atributo em quesito de poder descritivo e discriminativo, a sua baixa quantidade de informação fez com que este atributo obtivesse piores resultados de classificação. Os resultados destas avaliações são apresentados no Capítulo 4.

Motivado pelo problema de recomendação de etiquetas [7, 22, 52, 54], que visa aumentar a quantidade e qualidade de etiquetas de um certo objeto (ou usuário dependendo do contexto), nós caracterizamos evidências de como o conteúdo existente nos campos TÍTULO, DESCRIÇÕES e COMENTÁRIOS podem melhorar a qualidade do atributo ETIQUETAS. Para isto, usamos como sugestões de novos termos para o atributo ETIQUETAS, termos dos outros atributos. Os três atributos a serem recomendados têm seus termos ordenados de acordo com sua capacidade descritiva e/ou discriminativa. Acreditamos que os atributos obtém maior similaridade as ETIQUETAS são os mais promissores para a recomendação. Os resultados deste estudo também condizem com a nossa caracterização demonstrando que, neste caso, TÍTULO obtêm os melhores resultados, sendo o melhor atributo para a recomendação de etiquetas. DESCRIÇÃO e COMENTÁRIOS são menos eficazes porém tem a capacidade de recomendar mais termos, devido ao maior tamanho destes atributos.

Os nossos resultados de classificação, recomendação de etiquetas são discutidos com base nos resultados da caracterização para os diferentes aspectos de qualidade. Nossos estudos mostram evidência de que a capacidade descritiva é mais importante para a recomendação de etiquetas e para o estudo com usuários. Para a classificação, a quantidade de conteúdo é o aspecto de qualidade mais importante. Notamos que nem todos aspectos de qualidade são igualmente importantes para diferentes tarefas. Classificadores estado da arte, como *support vector machines* (SVM), obtém melhores

[9] http://www.allmusic.com

resultados com uma maior quantidade de dados pois combinam informações de vários termos para tomar decisões. A capacidade descritiva é mais importante para a recomendação, termos mais descritivos são mais relacionados com o conteúdo de um objeto. Também observamos que uma capacidade descritiva e/ou discriminativa é de pouco uso se o atributo estiver ausente na maioria dos objetos ou não possuir conteúdo suficiente.

Por fim, realizamos um experimento para levantar evidências de qualidade de acordo com a percepção dos usuários. Para isto, foi desenvolvido um sistema onde usuários avaliavam a qualidade descritiva dos atributos de acordo com vídeos do YouTube. Dezessete voluntários fizeram uso do sistema e avaliaram a qualidade de atributos associados a dez vídeos. Os resultados deste experimentos novamente reforçam a nossa caracterização, indicando que atributos pequenos como ETIQUETAS e TÍTULO tem maior qualidade.

Salvo poucas exceções [39, 41, 43], a maioria dos trabalhos da literatura focam em um atributo textual apenas, em especial as etiquetas [5, 10, 20, 35, 45, 51, 52, 54, 56], ou estudam poucos objetos em uma única aplicação [39]. Este trabalho amplia os resultados da literatura comparando centenas de milhares atributos diferentes em mais de uma aplicação. Os resultados desta pesquisa ampliam significativamente o conhecimento da literatura, abordando não apenas uma fonte textual e fazendo este estudo em diferentes aplicações. Os resultados obtidos podem ser utilizados por provedores e designers de aplicações Web 2.0 para melhorar os seus serviços de RI e desenvolver aplicações considerado os atrativos de cada fonte.

No próximo capítulo, Capítulo 2, apresentaremos uma discussão dos trabalhos relacionados ao nosso. As conclusões desta dissertação serão apresentadas no Capítulo 5.

Capítulo 2

Embasamento Teórico

Neste capítulo serão discutidos os trabalhos relacionados com a pesquisa realizada. Inicialmente discutimos sobre o que são atributos textuais na Web 2.0. Após isto, fazemos um levantamento de trabalhos relacionados às aplicações Web 2.0 que estudamos. Apresentamos também um levantamento dos conceitos de RI necessários para o entendimento desta dissertação. Em seguida, apresentamos uma discussão sobre os trabalhos que abordam o problema de qualidade em páginas da Web tradicional e Web 2.0. Por fim, discutimos os trabalhos relacionados à etiquetagem por ETIQUETAS ser o atributo textual da Web 2.0 mais estudado atualmente.

2.1 Atributos Textuais na Web 2.0

Como já mencionado, foram identificados quatro atributos textuais comumente utilizados (TÍTULO, ETIQUETAS, DESCRIÇÃO e COMENTÁRIOS) em quatro aplicações diferentes da Web 2.0 (CiteULike, LastFM, YouTube e YahooVideo). Embora referenciamos os atributos textuais por estes nomes, em algumas aplicações os mesmos apresentam denominações diferentes, porém com o mesmo significado. O CiteULike, por exemplo, refere-se a 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.

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 [38, 58].

Etiquetas é colaborativo no CiteULike, LastFM e YahooVideo, visto que qualquer usuário pode adicionar etiquetas aos objetos existentes nestas aplicações. No YouTube, ao contrário, apenas o proprietário do vídeo pode adicionar e alterar as etiquetas de um vídeo. 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 disto ser uma prática incomum. Nas quatro aplicações, o título é restritivo e comentários são colaborativos. A Tabela 2.1 resume as permissões de anotações para os atributos textuais estudados.

	título	etiquetas	descrição	comentários
CiteULike	restritivo	colaborativo	colaborativo	colaborativo
LastFM	restritivo	colaborativo	colaborativo	colaborativo
YahooVideo	restritivo	colaborativo	restritivo	colaborativo
YouTube	restritivo	restritivo	restritivo	colaborativo

Tabela 2.1: Permissões de Anotação para cada Atributo Textual.

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 etiquetas, caso estes dados não sejam fornecidos. Contudo, é possível que os usuários removam *etiquetas* 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.

2.2 Aplicações Web 2.0

2.2.1 CiteULike

O CiteULike [13] é um serviço de gerência e descoberta de referências bibliográficas voltado para a comunidade acadêmica. De forma similar ao Delicious, o aplicativo é o que pode ser chamado de um gerenciador de favoritos. Neste, usuários postam seus

artigos e livros acadêmicos favoritos e anotam os mesmos com etiquetas, formando assim uma biblioteca pessoal. A aplicação também possibilita aos usuários discutirem os trabalhos em grupos de discussão e postarem opiniões sobre os trabalhos. Desde 2007 o aplicativo disponibiliza parte da sua base de dados para uso público. Esta inclui informações sobre quais usuários postaram, quais artigos e as respectivas etiquetas[1].

2.2.2 LastFM

LastFM é uma rádio da Internet fundada em 2002. Com mais de 30 milhões de usuários ativos,de acordo com a própria aplicação, a rádio é considerada uma das mais populares rádios da Internet atualmente [47]. O aplicativo oferece para seus usuários serviços de descoberta e recomendação de músicas através de tecnologia Audioscrobbler[2]. O sistema também possibilita aos usuários a capacidade de interagir em grupos de discussões de música ou sobre artistas específicos, formar redes sociais de amizade e descobrir outros usuários com gostos musicais em comum.

Diversos artigos já fizeram estudos LastFM ou do Audioscrobbler. Em especial mencionamos aqueles voltados para as áreas de: recomendação de conteúdo [10] ou de como usar as etiquetas e informações sociais do LastFM para melhorar a classificação de artistas em gêneros musicais [26].

2.2.3 YouTube e YahooVideo

O YouTube é o mais popular sistema de compartilhamento de vídeos atualmente. Em 2006 foi averiguado que a aplicação recebia centenas de milhares de novos vídeos por dia e servia milhões para espectadores [36]. Em 2007 o YouTube obteve sétimo lugar em um estudo sobre as 10 maiores base de dados do mundo, neste mesmo ano um outro estudo demonstrou que o YouTube é a aplicação de vídeos mais popular nos Estados Unidos [37]. Recentemente foi averiguado que a máquina de busca que apoia a aplicação é a segunda mais popular do mundo [37].

O YahooVideo é o aplicativo de distribuição de vídeos da empresa Yahoo. O domínio `yahoo.com` terceiro domínio mais acessado da Internet. Em 2008 foi constatado que o YahooVideo era o segundo aplicativo de vídeo mais popular dos Estados Unidos [31].

Diversos trabalhos já estudaram os dois aplicativos, em especial o YouTube, em diferentes contextos. Em especial, mencionamos os trabalhos de Cha et al. [9], Gill et al. [19], Halvey et al. [23] e Kang et al. [31] apresentam diferentes caracterizações

[1] `http://www.citeulike.org/faq/data.adp`
[2] `http://www.audioscrobbler.net`

do aplicativo abrangendo a quantidade e popularidade de vídeos no sistema, o uso de anotações e outros atributos textuais como também analises de trafego e acesso.

Nosso trabalho amplia o conhecimento gerado pelos trabalhos supracitados apresentando uma caracterização das quatro aplicações voltada para o entendimento da qualidade da informação contida nas mesmas, em especial quando esta informação é utilizada por serviços de RI.

2.3 Recuperação de Informação

2.3.1 Importância de Atributos Textuais

Diversos estudos exploram a ideia de computar a importância de um bloco (região específica em uma página da Web, que podem ser mapeadas para os atributos textuais) para melhorar a qualidade de tarefas RI na Web [8, 16]. Tais métricas são importantes para ajudar projetistas de aplicações Web a criar bons atributos, acompanhar a utilidade dos existentes e utilizá-los adequadamente para o subsídio de serviços RI. Neste trabalho, fazemos uso do método proposto por Fernandes et al [16] para avaliar a capacidade descritiva e discriminativa de diferentes atributos textuais. De forma análoga as métricas de RI, como o $TF \cdot IDF$ [49], que mensuram a importância de um termo para um documento, tais métricas capturam a importância de diferentes atributos textuais.

A capacidade discriminativa está relacionada a capacidade de um termo . Por exemplo, ela ajuda algoritmos de classificação a identificar automaticamente a categoria de um documento, ou contribuem para que sistemas de busca melhor ordenem os documentos retornados por uma consulta de acordo com sua relevância. A capacidade descritiva captura como o atributo textual está relacionado ao objeto multimídia, por exemplo, se os termos de um atributo não forem capazes de capturar o conteúdo do objeto estes termos têm um efeito negativo para RI, pois, por exemplo, ao retornar este objeto como relevante para uma pesquisa, o conteúdo do objeto não será relacionado com os termos da consulta. 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 de serviços RI.

Durante o resto da dissertação vamos nos referir a: f como uma instância de um atributo e a F como a coleção de todas as instâncias f; o como um objeto na coleção O; e por fim, t como um termo. Letas maiúsculas denotam conjuntos, como uma coleção de atributos, as minúsculas denotam instâncias.

2.3.1.1 Poder Discriminativo

Com o objetivo de capturar o poder discriminativo de um termo em um atributo, fazemos uso da métrica *Inverse Feature Frequency*, IFF, definida na Equação 2.1. Esta 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. O IFF é uma alteração da métrica IDF, onde o IDF considera o objeto como um todo e o IFF apenas um atributo.

$$IFF(t, F) = \log(\frac{|F|}{freq(t, F)}), \tag{2.1}$$

onde $freq(t, F)$ é a quantidade de instâncias de um atributo onde o termo t aparece.

Para capturar a qualidade discriminativa de um atributo como um todo, fazemos uso da métrica *Average Inverse Feature Frequency* $AIFF$. Esta representa a média dos valores do IFF de todos os termos que ocorrem em todas as instâncias ,f, do atributo F, ou seja:

$$AIFF(F) = \frac{\sum_{t \in F} IFF(t, F)}{|F|}, \tag{2.2}$$

onde $|F|$ é o tamanho do vocabulário completo do atributo F, ou seja, o tamanho do conjunto representando pela união de todos os termos de todas as instâncias de F.

2.3.1.2 Poder Descritivo

Para avaliar o poder descritivo de um atributo, fazemos uso da métrica denominada *Term Spread*, $TS(t, o)$:

$$TS(t, o) = \sum_{f \in o} i, \text{ onde } i = \begin{cases} 1 & \text{se } t \in f \\ 0 & \text{cc.} \end{cases} \tag{2.3}$$

A ideia por trás do TS é 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 *Feature Instance Spread*, $FIS(f, o)$, é definido como o valor médio dos valores $TS(t, o)$ para uma instância atributo de um objeto. Essa métrica trata-se de uma

heurística para estimar o grau de relacionamento de uma instância de atributo com o conteúdo do objeto em si.

$$FIS(f, o) = \sum_{t \in f} \frac{TS(t, o)}{|f|}. \tag{2.4}$$

onde $|f|$ é o tamanho do vocabulário completo de uma instância atributo f.

Para capturar o poder descritivo de um atributo em relação a coleção de objetos, fazemos uso do valor médio de todos os FIS na coleção de objetos. Este valor média é capturado pela métrica *Average Feature Spread*, $AFS(F)$:

$$AFS(F) = \frac{\sum_{o \in O} FIS(f, o)}{|O|} \tag{2.5}$$

onde $|O|$ é o tamanho de toda a coleção.

2.3.2 Recuperação de Informação Multimídia

O já discutido artigo publicado pela autora Boll [6] levanta o questionamento de como a pesquisa em RI multimídia (ver [48] e [53] para mais informações sobre tais técnicas) e as pequisas de Web 2.0 podem ser utilizadas para a criação de novos serviços e aplicações. Consideramos nosso trabalho como um novo resultado nesta direção. Neste trabalho apresentamos uma das primeiras análises da qualidade de diferentes atributos textuais da Web 2.0 para RI quando considerados como uma alternativa ao processamento dos objetos multimídia que os mesmos descrevem.

Embora ainda não tenha sido aplicada em larga escala, existem trabalhos que exploram como aplicar técnicas de RI multimídia para a Web 2.0. Alguns exemplos são os estudos de Rege et al. [46], Mei et al. [40] e Chen et al. [10], que fazem uso de atributos visuais e textuais para agrupar imagens similares no Flickr. O trabalho de San Pedro et al. [50] que também faz uso de atributos visuais e textuais para buscar imagens atrativas também no Flickr.

2.4 Qualidade de Informação

2.4.1 Qualidade de Dados

Embora não exista uma métrica padrão para definir a qualidade de um documento ou atributo textual, alguns trabalhos descrevem conjuntos de fatores desejáveis para definir a qualidade de bancos de dados. O trabalho de Strong et al. [55] define quatro

categorias de aspectos de qualidade desejáveis em um banco de dados. A Tabela 2.2 apresenta os aspectos de qualidade definidos pelos autores.

Categoria	**Aspecto**
QD Intrínseca	Acurácia, Objetividade, Credibilidade, Reputação
QD Acessibilidade	Acessibilidade, Segurança de acesso
QD Contextual	Relevância, Valor Agregado, Temporalidade, Completude, Quantidade de Dados
QD Representacional	Facilidade de interpretação e entendimento, Representação consistente e coesa

Tabela 2.2: Classes e atributos de qualidade de dados.

Os autores também definem um problema de qualidade de dados como sendo qualquer dificuldade que abrange duas ou mais categorias. Embora não tratamos diretamente com bancos de dados, nosso trabalho explora a qualidade de atributos textuais com relação ao uso, quantidade, corretude (de acordo com bases semânticas), poder descritivo e discriminativo. Uso e quantidade de conteúdo se encaixam nos aspectos de completude e quantidade de dados da categoria Contextual, enquanto corretude poder descritivo e discriminativo podem ser vistos tanto na categoria Intrínseca de acordo com os aspectos acurácia e credibilidade, ou na categoria Representacional se encaixando no aspecto de facilidade de interpretação e entendimento. Portanto, consideramos o problema de qualidade dos atributos textuais como um problema de qualidade de dados de forma mais ampla.

2.4.2 Qualidade de Informação na Web 2.0

Devido à sua natureza colaborativa, a possível falta de qualidade da informação na Web 2.0 já é um problema conhecido e estudado. Inicialmente mencionamos um estudo recente que, de forma similar ao nosso, compara a qualidade de diferentes atributos textuais no Flickr [39]. A autora faz uma inspeção manual de diferentes fotografias de um mesmo ponto turístico na Espanha comparando a qualidade dos atributos TÍTULO, ETIQUETAS e DESCRIÇÃO. A principal questão levantada pelo trabalho é se as ETIQUETAS, por ser o atributo textual mais estudado atualmente, é a fonte de informação mais confiável. A autora afirma que, devido à natureza organizacional das etiquetas, estas não são confiáveis no sentido de descrever bem o contexto da imagem (e.x. explicar que a imagem faz parte das fotos de férias de um casal) sendo TÍTULO e DESCRIÇÃO mais confiáveis neste aspecto. Este resultado é contraditório com os nossos, pois concluímos que ETIQUETAS é o melhor atributo textual de acordo com as heurísticas de qualidade,

aplicabilidade para tarefas de classificação e por avaliação manual dos atributos. Notamos que a base de dados analisada pela autora consiste de centenas de imagens e atributos textuais associados às mesmas imagens. Em contrapartida, o nosso trabalho estuda diferentes aspectos de qualidade com coleções de dados mais diversas e mais abrangentes, pois capturam características de quatro aplicativos diferentes. É importante ressaltar também que este trabalho lida com uma mídia diferente (imagens) e toda a análise foi feita de forma manual. No nosso caso fazemos diferentes estudos de caracterização, tarefas de RI e estudos com usuários.

Diversos trabalhos [11, 18, 25] debatem sobre problemas de qualidade da Wikipedia. Neste aplicativo, qualquer pessoa pode inserir e alterar o conteúdo de artigos formando assim uma enciclopédia colaborativa. Nesta enciclopédia, não apenas o vocabulário e corretude gramatical dos artigos é questionável, como também a acurácia histórica dos artigos. O estudo controverso de 2005 afirma que a corretude da Wikipedia é comparável com à Enclyclopedia Britannica [18]. Esta afirmação foi o foco de um debate entre Nature e editores da Britannica, e a discussão pode ser vista na página Web do artigo [18]. Embora o nosso trabalho não trate da Wikipedia e que não podemos debater sobre qual dos dois partidos está correto, este debate serve como um exemplo de que o problema de qualidade da mídia social ainda é uma questão em aberto.

Existem também diversos estudos que exploram a qualidade de respostas em aplicações de Perguntas-e-Respostas, como o Yahoo Answers, ou em fóruns de discussão [1, 4, 29]. Em um aplicativo desta natureza, um usuário posta perguntas e recebe respostas de outros. Os usuários podem então atribuir notas às respostas indicando qual foi a melhor para a respectiva dúvida. Usando como exemplo o trabalho de Agichtein et al [1], os autores classificaram manualmente um conjunto de dados de respostas indicando se a informação da mesma era de *baixa, média or alta* qualidade. Fazendo uso de atributos sociais e textuais relacionados às respostas, os autores fazem uso de um classificador. Atributos textuais incluem a corretude sintática e gramatical do texto, enquanto os sociais são derivadas de análises do grafo de interconexões entre usuários e objetos no aplicativo. Usando classificadores, especificamente *stochastic gradient boosted trees* [17], os autores concluem que é possível identificar conteúdo de qualidade usando os atributos sociais e textuais do sistema. Nosso trabalho é complementar aos citados, pois também buscamos atributos textuais indicadores de qualidade, porém focamos em diferentes aplicativos e apenas em características dos atributos textuais. Um outro trabalho faz uma análise similar no Wikipedia [11], buscando artigos de qualidade.

2.5 Etiquetagem

Como mencionado, a maioria dos trabalhos anteriores dão ênfase apenas ao atributo textual *etiquetas*, incluindo a sua aplicação para melhorar serviços como busca, recomendação, agrupamento e indexação [35, 51, 52, 54] e a caracterização de aplicações que fazem uso deste atributo [20, 52]. No entanto, não há um consenso a respeito de sua qualidade, como levantado pelo já discutido trabalho da Marshall [39].

Algumas evidências de qualidade são levantadas por trabalhos anteriores. Suchanek et al. [56] mensuram a qualidade de *etiquetas* no sistema Delicious[3] de acordo com duas bases semânticas *online* Yago[4] e Wordnet[5]. Os autores verificaram que apenas 46% das *etiquetas* utilizadas pertenciam a esses dicionários e que aproximadamente metade dessas *etiquetas* conhecidas tinham 10 ou mais significados. Foi visto também que as *etiquetas* mais populares tendem a apresentar ruído semântico menor (por serem mais comum nas bases semânticas). Todavia, os problemas de palavras desconhecidas e polissêmicas [20], impõem desafios para tarefas de RI baseadas em *etiquetas*.

De forma similar, Bischoff et al. [5] analisaram a qualidade de etiquetas no contexto de busca na Web 2.0. Através da comparação das etiquetas utilizadas com o conteúdo (página na Web), metadados (contidos no HTML) e fontes externas (revistas especializadas). Foi constatado que etiquetas tendem a representar bem o conteúdo dos objetos a que foram associadas e são relacionadas a termos de consultas feitas por usuários. Esta segunda constatação também foi observada por outros autores [35].

Koutrika et al. [32,33] modelaram o uso de sistemas de etiqueta, separando usuários em duas categorias: *spammers* e *não-spammers*. O estudo objetivou promover o entendimento de como o *spam* afeta a busca por conteúdo. Os autores mostraram que etiquetas 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 etiqueta é menos suscetível ao *spam*.

Além da análise de etiquetas, existem trabalhos que caracterizam o uso de comentários em aplicações Web 2.0, especificamente em Blogs. Mishne et al [42, 43] estudam as características do conteúdo de comentários e como combater problemas de Spam nos comentários. Os resultados dos autores, no quesito de caracterização do conteúdo dos comentários, têm resultados bastante similares ao nosso. Vemos isto como um indicativo alguns padrões de uso de cometários tende a ser comum independente da aplicação sendo estudada.

[3] `http://delicious.com`

[4] `http://www.mpi-inf.mpg.de/~suchanek/yago`

[5] `http://wordnet.princeton.edu`

Capítulo 3

Coleta e Caracterização de dados

Neste capítulo caracterizamos as evidências de qualidade dos quatro atributos textuais nas quatro aplicações estudadas. Este estudo tem como objetivo comparar a qualidade relativa dos quatro atributos quando utilizados como subsídios de tarefas de RI. Nesta seção, fazemos uso da mesma notação que foi utilizada no Capítulo 2.

Inicialmente discutimos como foi feita a coleta de dados (Seção 3.1). Após isto, apresentamos os resultados da caracterização de diferentes evidências de qualidade dos quatro atributos textuais nas quatro aplicações estudadas (Seção 3.2).

3.1 Coleta de Dados

Para realizar nosso estudo, implementamos coletores de dados na Web para cada aplicação. Para o LastFM, YouTube e YahooVideo, nossos coletores seguem a abordagem de amostragem denominada de *snowball sampling strategy* [21]. Nesta, cada coletor começa com uma lista inicial de objetos a serem coletados, denominados de sementes. Para cada objeto coletado, a estratégia snowball enfileira para coleta futura os *links* para objetos relacionados. No nosso caso, pelo menos 800 sementes foram utilizadas por aplicação. No YouTube e YahooVideo as sementes foram escolhidas como sendo os vídeos mais populares da aplicação. No LastFM, iniciamos a coleta com artistas que foram anotados com as etiquetas mais utilizadas pelos usuários. No nosso estudo, consideramos como um objeto do LastFM como sendo páginas de artistas com as músicas dos mesmos.

No caso do CiteULike, uma lista artistas ou etiquetas populares não é provida pela aplicação. Também não são providos links para objetos relacionados. Por isto, amostramos aleatoriamente objetos da base pública [1] provida pela aplicação.

[1] `http://www.citeulike.org/faq/data.adp`

Os coletores do YouTube, YahooVideo e LastFM foram executados por aproximadamente duas semanas em Julho, Setembro e Outubro de 2008, respectivamente. Os dados do CiteULike foram amostrados da base de Setembro de 2008. Abaixo segue a quantidade de objetos coletados por aplicação:

- 678,614 artigos do CiteULike;
- 193,457 páginas de artistas do LastFM; e
- 227,562 e 211,081 vídeos do YahooVideo and YouTube, respectivamente.

Para as análises de classificação, também coletamos as classes dos objetos para o YouTube, YahooVideo e LastFM. Tanto o YouTube quanto o Yahoo vídeo permitem que usuários façam upload de vídeos em categorias pré-determinadas pela aplicação, a categoria de cada objeto também foi coletada para estas duas aplicações. Cada categoria equivale a uma classe na classificação. O LastFM, por sua vez, não oferece tais categorias para os artistas, por isto,recorremos à fontes externas para coletar os gêneros musicais dos artistas, informação que usamos como classes no experimento. Recorremos ao All Music Guide[2], aplicativo Web que auto denomina ter a maior base de dados sobre músicas no mundo, para coletar os gêneros dos artistas. Foi possível amostrar classes para 5536 artistas, um valor significativamente mais alto do que o estudos feitos por trabalhos passados [10].

De forma similar ao LastFM, o CiteULike não provê uma listagem de categorias para os artigos postados no mesmo. Como não temos conhecimento de uma listagem confiável para as diversas áreas científicas, cujos objetos da nossa coleta fazem parte, deixamos a análise de qualidade em relação à classificação para o CiteULike como trabalho futuro.

No total, 20 categorias existem para o YahooVideo e 15 para o YouTube. No All Music Guide, encontramos 9 gêneros musicais para os artistas do LastFM. Tais categorias e gêneros, agora referidos como classes, estão listados na Tabela 3.1.

3.2 Caracterização de Atributos Textuais

Nesta seção, caracterizamos diferentes aspectos de qualidade dos quatro atributos textuais nas quatro aplicações estudadas. Nossa caracterização foi feita a partir das nossa bases coletadas. Neste estudo o uso de classes não foi necessário, portanto utilizamos a

[2] `http://www.allmusic.com/`

Tabela 3.1: Classes dos Objetos.

Aplicação	Classes
LastFM	Blues, Classical, Country, R & B, Electronica, Jazz, Pop/Rock, Rap, World
YahooVideo	Action, Animals, Art & Animation, Commercials, Entertainment & TV, Family, Food, Funny Videos, Games, Health & Beauty, How-to, Movies & Shorts, Music, News & Politics, People & Vlogs, Products & Tech., Science & Environment, Sports, Transportation, Travel
YouTube	Autos & Vehicles, Comedy, Education, Entertainment, Film & Animation, Gaming, Howto & Style, Music, News & Politics, Nonprofits & Activism, People & Blogs, Pets & Animals, Science & Technology, Sports, Travel & Events

base completa do LastFM e CiteULike. A análise que será apresentada visa responder as seguintes questões:

Seção 3.2.1: Qual a fração de objetos com instâncias de atributos textuais vazia?

Uso de atributos textuais é de suma importância para RI. Atributos subexplorados, ou seja, que contém conteúdo vazio (0 termos) em uma fração significativa de objetos, provavelmente não serão uma fonte de informação confiável para RI.

Seção 3.2.2: Qual a quantidade de conteúdo por instância não vazia de atributos textual?

Se presente (não vazia), um atributo textual deve prover conteúdo suficiente sobre objetos para ser uma fonte de informações útil.

Seção 3.2.3: Qual a corretude sintática e propriedades semânticas dos atributos textuais?

É também importante a corretude e propriedades semânticas dos termos nos atributos textuais. Por exemplo, um termo que é um substantivo tem mais chances de ser útil para os usuários do que um advérbio. Em contraste, uma palavra pode pertencer a mais de uma classe semântica e ter vários significados no dicionário (efeito denominado de polissemia), afetando também sua qualidade. Nesta análise, verificamos se os termos dos atributos estão corretos de acordo com duas bases semânticas Wordnet [15] and Yago [57].

Seção 3.2.4: Qual a capacidade descritiva e discriminativa dos atributos textuais?

Para ser uma fonte de informação eficaz, um atributo textual necessita oferecer uma descrição correta do objeto que este está anotando. O atributo deve também ser capaz de diferenciar este objeto de outros objetos da coleção.

Seção 3.2.5: Quão diverso é o conteúdo e informações entre atributos textuais?

Além das evidências de qualidade acima, também caracterizamos a diversidade de conteúdo entre os blocos textuais. Tais resultados indicam se o conteúdo de diversos atributos podem ser combinados de forma eficaz. Para esta análise, fizemos usos das métricas de Jaccard [27], que mensuram a fração de termos em comum entre dois campos.

O resumo dos nossos resultados de caracterização será apresentado na Seção 3.2.6.

3.2.1 Uso de Atributos Textuais

A Tabela 3.2 mostra, para cada atributo e aplicação, a percentagem de instâncias vazias, ou seja, objetos sem termo algum no campo correspondente ao atributo. A permissão de anotação de anotação de cada atributo é mostrada entre parênteses, sendo C para colaborativo e R para restritivo.

	TÍTULO	ETIQUETAS	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)

Tabela 3.2: Percentagem de instâncias vazias (C = atributo colaborativo, R = atributo restritivo).

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 e sistemas de compartilhamento de imagens [39,41]. O atributo ETIQUETAS, 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 pelas restrições impostas pelas aplicações, tais como o caso de título e etiquetas no YouTube, que podem ser preenchidos automaticamente. No entanto, nenhuma das aplicações exige o preenchimento do campo descrição e, ainda assim, este atributo tem uma presença muito maior no YouTube e YahooVideo, onde ele é restritivo. De fato, no YahooVideo, as descrições são mais presentes do que as etiquetas, um atributo colaborativo. Tal foi algo inesperado, visto que por razões de esforço é mais fácil prover poucas palavras como etiquetas do que algumas sentenças como uma descrição.

Os resultados indicam que uso de etiquetas, descrição ou comentários como uma única fonte de dados para serviços de RI, independentemente de suas permissões de anotação, podem não ser eficaz, devido à falta de informação em uma parcela significativa de objetos. Vemos os nossos resultados como motivadores para a necessidade que é 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 etiquetagem [38].

3.2.2 Quantidade de Conteúdo

Foram analisadas a quantidade de conteúdo 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*[3] 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 etiquetas em todas as aplicações, exceto para o LastFM, onde o valor 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 Porter[4] para remoção de afixos. Ao final, foram eliminadas as stopwords, pois estas palavras trazem pouca, ou nenhuma, informação semântica sobre os objetos.

Inicialmente nós caracterizamos o número de termos distintos, referido como tamanho do vocabulário, em cada instância de atributo. Estes dados representam a

[3] *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`

[4] `http://tartarus.org/~martin/PorterStemmer/`

quantidade de conteúdo disponível para uso em serviços de RI. A Tabela 3.3 apresenta a média μ, coeficiente de variação CV (razão entre desvio padrão e média), e valores máximos para os tamanhos do vocabulário.

	TÍTULO			ETIQUETAS			DESCRIÇÃO			COMENTÁRIOS		
	μ	CV	max	μ	CV	max	μ	CV	max	μ	CV	max
Citeul.	7.5	0.40	73	4.0	1.33	194	65.2	0.5	3890	51.9	1.42	449
Lastfm	1.8	0.47	23	27.4	1.49	269	90.1	1.06	3390	110.2	3.55	22634
Yahoov.	6.3	0.39	16	12.8	0.52	52	21.6	0.71	141	52.2	2.51	4189
Youtub.	4.6	0.43	36	10.0	0.60	101	40.4	1.75	2071	322.3	1.94	16965

Tabela 3.3: Tamanho do vocabulário de instâncias de atributos textuais não-vazias.

Em geral, o atributo TÍTULO tem instâncias com o menor vocabulário, seguido por ETIQUETAS, 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 e gráficos mostram que instâncias de COMENTÁRIOS (quando presentes) têm o maior vocabulário (exceto no CiteULike) exibindo um coeficiente de variação muito elevado. Similarmente, instâncias de DESCRIÇÃO no CiteULike e LastFM, e de ETIQUETAS 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 3.2.1). Comparando estes resultados com estudos prévios da aplicação Delicious, já havia sido verificado que o tamanho médio do vocabulário de instâncias de ETIQUETAS é próximo de 100 [35, 56], 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. Outro fator pode ser a popularidade do Delicious como aplicativo de etiquetagens. Notamos que os nossos resultados são consistentes com trabalhos passados que observaram TÍTULO, ETIQUETAS e DESCRIÇÃO no Flickr [39].

Embora seja colaborativo no CiteULike, o atributo ETIQUETAS tende a ter menos conteúdo do que o atributo TÍTULO. Nas outra três aplicações, TÍTULO é o menor atributo textual. Isto provavelmente dar-se ao fato de que nomes de artigos e livro científicos são geralmente longos, um outro fator complementar é que o vocabulário de etiquetas (números de termos únicos) tende a se estabilizar com o tempo [20, 35]. Ainda assim, existem objetos que alcançam até 194 termos nas etiquetas. Também diferente das outras aplicaçoes, o atributo DESCRIÇÃO no CiteULike é o maior dentre os quatro. Possivelmente isto é devido ao fato de que, neste caso, as descrições são os

resumos dos artigos que também tender a ser longos, tendo assim mais conteúdo do que o pouco explorado comentários.

No LastFM, título tende ter valores muito baixos. Isto é esperado pois, neste caso, os títulos dos objetos são nomes de artistas, algo que contém poucos termos. Em comparação, as etiquetas tendem a carregar bem mais conteúdo com um valor médio de 27 termos e chegando a um máximo de 269 termos. Comentários é o atributo textual com mais conteúdo alcançando até 22.000 termos, notamos que este valor é tendencioso para artistas populares que geram mais discussões entre usuários (mais resultados abaixo). O atributo Descrição é maior no LastFM do que nas outras aplicações, possivelmente pelo fato de ter uma natureza colaborativa onde usuários descrevem artistas de forma colaborativa em um wiki.

O YahooVideo e YouTube demostram tendências similares entre si. Porém, a quantidade de conteúdo dos campos descrição and comentários tende a ser maior no YouTube, possivelmente devido a maior audiência. Em contraste, a quantidade de conteúdo em título e etiquetas é maior no YahooVideo. Acreditamos que a diferença em etiquetas é devido ao maior grau de colaboração nesta aplicação, enquanto as diferenças no título possivelmente refletem padrões de uso diferentes.

Para entender melhor como conteúdo é inserido em atributos textuais, mensuramos a correlação entre a popularidade de objetos e a quantidade de termos em cara atributo. Identificar se objetos populares tender a ter mais (ou possivelmente menos) conteúdo é uma questão relevante para RI, pois talvez usuários necessitem de incentivos para inserir conteúdo em objetos não populares. Para este fim, fizemos uso do coeficiente de Pearson (ρ) [28] calculado entre a quantidade de termos nos atributos e a popularidade dos objetos. Popularidade de objetos foi mensurado com base no número de exibições do vídeo no YahooVideo e YouTube, número de ouvintes do artista no LastFM e pelo número pessoas que marcaram o artigo como favorito no CiteULike.

Apenas para os atributos colaborativos as correlações mensuradas podem ser consideradas significativas. Estes casos foram: comentários no LastFM (ρ=0.5) e YouTube (ρ=0.24), etiquetas no CiteULike (ρ=0.23) e no LastFM (ρ=0.41) e por fim descrição no LastFM (ρ=0.25). Duas exceções foram o atributo descrição no CiteULike (ρ=0.006) e etiquetas no YahooVideo (ρ=0.003). A primeira exceção pode ser explicada pelo fato de que o atributo descrição no CiteULike contém os resumos dos artigos. A segunda exceção pode ser explicada pela falta de incentivos necessárias para que os usuários coloquem etiquetas nos objetos. Observamos que no CiteULike e LastFM os usuários organizam coleções pessoais de artigos ou músicas com as etiquetas, no YahooVideo a etiquetagem é colaborativa mas não existe este conceito de uma biblioteca pessoal, os usuários podem colocar etiquetas em qualquer

Tabela 3.4: Percentual de termos com pelo menos um significado.

	TÍTULO	ETIQUETAS	DESCRIÇÃO	COMENTÁRIOS
CiteULike	84.7%	61.3%	83.1%	63.1%
LastFM	61.0%	74.1%	71.9%	63.0%
YahooVideo	78.2%	79.0%	80.0%	71.0%
YouTube	73.3%	72.4%	72.9%	64.5%

vídeo porém sem um retorno explícito.

3.2.3 Propriedades Semânticas

Golder e Hubberman [20] levantam questões de como sistemas de etiquetagem podem ser afetados pelos sinônimos (várias palavras com um mesmo significado) ou pela polissemia (uma única palavra com várias significados). Além disto, palavras sem significado podem trazer problemas para serviços de RI que fazem uso de etiquetas. Esta analise aborda a qualidade dos atributos textuais de acordo com a classe sintáticas dos termos dos mesmos e seus significados.

Assim como em [56], nesta seção, analisamos as propriedades semânticas do vocabulário dos atributos. Nosso objetivo é levantar evidências de qualidade em relação a corretude das palavras de acordo com bases semânticas Wordnet e Yago (de agora em diante referidas como *dicionário*). O Wordnet é um banco de dados léxico de palavras da língua inglesa, nesta base de dados as palavras (substantivos, verbos, adjetivos etc) são agrupados em grupos de sinônimos (denominados de *synsets*). Estes grupos são conectados com base nas similaridades semânticas e sintáticas das palavras. Por exemplo, a palavra "cadeira" por ser um assento será relacionada com a palavra "sofa". O Yago é uma base de dados similar que relaciona nomes próprios, nesta base de dados por exemplo a atriz "Marilyn Monroe" estará relacionada com a atriz "Jane Fonda" pois as duas são atrizes americanas. Assim como na Seção 3.2.2, esta analise foca apenas em palavras da língua inglesa.

Os resultados demonstrados na Tabela 3.4, apresentam a o percentual de palavras com pelo menos um significado conhecido no dicionário para cada atributo. Este valor visa capturar a quantidade de lixo (palavras desconhecidas) nos atributos. Assim como foi observado para o atributo ETIQUETAS [56], uma fração significativa de termos (entre 15% e 39%) em *todos* os atributos não estão presentes no dicionário. Este valor é maior para COMENTÁRIOS (entre 29% e 37%), provavelmente devido ao fato deste atributo ter mais termos do que os outros e sofrer de problemas de spam [42].

A Tabela 3.5 demonstra a distribuição dos termos conhecidos (aqueles com pelo

Tabela 3.5: Percentual de termos em cada classe gramatical.

	Atributo	Substântivo	Verbo	Adjetivo	Adverbio	Nomes
CiteULike	título	66.7%	17.2%	27.2%	2.3%	3.6%
	etiquetas	76.3%	16.1%	17.6%	0.6%	5.7%
	descrição	54.7%	20.8%	33.2%	6.1%	3.6%
	comentários	56.6%	21.5%	28.2%	6.4%	5.7%
LastFM	título	72.4%	23.8%	19.4%	3.4%	15.0%
	etiquetas	65.3%	26.8%	41.0%	8.8%	12.7%
	descrição	63.6%	23.4%	31.5%	8.5%	8.7%
	comentários	61.2%	28.3%	30.5%	11.1%	10.2%
YahooVideo	título	80.1%	28.3%	22.7%	5.5%	7.8%
	etiquetas	76.2%	23.1%	20.4%	5.0%	9.4%
	descrição	68.2%	30.9%	24.2%	7.1%	6.7%
	comentários	66.7%	32.3%	30.8%	11.0%	8.9%
YouTube	título	76.1%	25.4%	25.3%	4.4%	10.9%
	etiquetas	77.8%	26.8%	20.7%	3.5%	13.0%
	descrição	67.6%	27.4%	27.0%	7.3%	9.8%
	comentários	61.9%	29.0%	28.9%	9.1%	8.8%

menos um significado no dicionário) em diferentes classes gramaticais. Note que os valores de cada linha na tabela pode não ter uma soma equivalente a 100%, isto ocorre devido ao fato de que uma mesma palavra poder ser categorizada em múltiplas classes gramaticais (e.x., a palavra "dance" da língua inglesa é um verbo e o substantivo ao mesmo tempo). Os resultados mostram que um alto percentual (55%-80%) dos termos são substantivos. Tais termos podem ser consideradas bons descritores de conteúdo por trazerem mais informação do que outras classes. A fração de nomes próprios, que também são bons descritores, também é alta.

Nós também mensuramos o percentual de termos com mais de k significados; onde k maior do que 1. Este valor captura o nível de polissemia dos termos. A Tabela 3.6 demonstra este valor para $k = \{5, 10\}$. Os valores variam entre 15% e 33% para $k = 5$ e entre 5% e 17% para $k = 10$. Consideramos estes valores como sendo altos, indicando que a polissemia afeta todos os atributos textuais, independente de aplicação. Isto indica que a polissemia pode impactar serviços de RI independente de aplicação e atributo explorado. Observamos que tal impacto pode ser reduzido se algum tipo de análise contextual do texto for empregado. Isto é possível para atributos como descrição e comentários que contém o texto em sentenças e não palavras isoladas como o etiquetas.

Por fim, notamos que o nosso dicionário, composto de duas bases de dados semânticas, não consegue cobrir todos os significados existentes para os termos utilizados em aplicações de mídia social. Em especial, mesmo fazendo da Wikipedia (no caso do

Tabela 3.6: Percentual de termos com pelo menos k significados.

	k=5				k=10			
	TÍTULO	ETIQ.	DESC.	COME.	TÍTULO	ETIQ.	DESC.	COME.
CiteULike	26.95%	15.23%	28.47%	21.81%	9.53%	5.20%	11.15%	9.93%
LastFM	17.54%	25.65%	25.51%	23.34%	7.38%	11.48%	12.10%	10.98%
YahooV.	28.05%	24.71%	33.59%	31.19%	15.59%	11.68%	16.14%	15.99%
YouTube	23.66%	22.16%	26.98%	23.68%	9.93%	8.13%	12.70%	11.12%

Yago), algumas gírias mais recentes da Internet provavelmente não vão existir no dicionário. Também notamos que analisar termos sem nenhum contexto como fazemos tende a aumentar valores encontrados como aqueles de polissemia. Independente destes fatores, podemos concluir que: (1) Existe uma fração significativa de termos que podem ser vistos como lixo, não existindo no dicionário; (2) a maioria dos termos são bons descritores de conteúdo, sendo substantivos ou nomes próprios; e por fim, (3) quando analisamos termos de forma independente, problemas de ambiguidade tem um impacto nos termos independente de atributo textual.

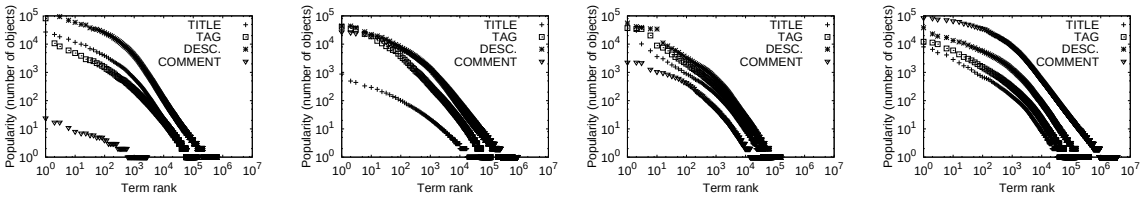
3.2.4 Capacidade Descritiva e Discriminativa

Atributos textuais da Web 2.0 geralmente estão contidos em locais (definidos como blocos) diferentes das páginas Web. Por exemplo, no YouTube, os comentários dos vídeos são apresentados próximos uns dos outros dentro de um bloco comum. Nesta seção fazemos das métricas *Average Inverse Feature Frequency* ($AIFF$) e *Average Feature Spread* (AFS) para averiguar a capacidade descritiva e discriminativa dos atributos textuais. Lembramos que as duas métricas já foram definidas no Capítulo 2. Lembramos também que a métrica $AIFF$ captura o poder discriminativo de um atributo, enquanto a AFS captura o poder descritivo.

Nós computamos os valores das duas métricas $AIFF$ e AFS para cada atributo textual em cada aplicação. Novamente, fizemos apenas de atributos não vazios e da língua inglesa com remoção de afixos. Dado que apenas uma pequena fração atributos *comentários* são não vazios no YahooVideo e CiteULike, nós descartamos este atributo nestas análises. A Tabela 3.7 demonstra os valores computados de $AIFF$ e AFS. Todos os valores são estatisticamente de acordo com o Teste-t [28] com 90% de confiança. Por estes valores podemos ver que a métrica AFS provê um ranking consistente dos atributos em todas as quatro aplicações. Título é o atributo mais descritivo, seguido de etiquetas, descrição e comentários [5]. Em compensação, a métrica $AIFF$

[5]É importante salientar que o AFS pode ser tendencioso para atributos com menos termos. Para miminizar este viés computados o AFS apenas com os 5 termos com maior FS, esta nova análise

apresenta pouca diferença entre os atributos, fato que discutiremos a seguir.



(a) CiteULike (b) LastFM (c) YahooVideo (d) YouTube

Figura 3.1: Distribuição de Popularidade dos Termos

Tabela 3.7: Valores de AFS e $AIFF$. Intervalos com 90% de confiança foram omitidos, estes exibem erros abaixo de 1%

	CiteULike		LastFM		YahooVideo		YouTube	
	AFS	$AIFF$	AFS	$AIFF$	AFS	$AIFF$	AFS	$AIFF$
TÍTULO	1.91	11.04	2.65	10.09	2.23	10.16	2.53	10.29
ETIQUETAS	1.63	10.95	1.41	9.52	1.86	9.86	2.07	10.23
DESCRIÇÃO	1.22	10.78	1.28	9.75	1.58	10.06	1.77	11.24
COMENTÁRIOS	-	-	1.27	9.61	-	-	1.19	10.33

Para entender os motivos pelo qual a métrica $AIFF$ não é capaz to claramente distinguir um atributo de outros, nós plotamos as distribuições de popularidade dos termos para cada atributo em cada aplicação. A Figura 3.1 demonstra no eixo-x o ranking de popularidade de cada termo, sendo o primeiro ponto do gráfico o ponto mais popular. O eixo-y demonstra em quantos objetos o termo aparece. Pelos gráficos podemos ver que a popularidade dos termos segue distribuições de cauda pesada, existindo alguns poucos termos muito populares e vários termos pouco populares. Uma consequência deste fato é que os muitos termos com baixa popularidade vão aumentar os valores de $AIFF$ para todos os atributos. Estudos passados [24, 45] demonstram que métricas de ranking de termos baseadas no inverso da frequência na coleção, assim como o IFF, super-enfatizam termos únicos e não populares. O problema é que tais termos não são úteis para algumas tarefas de RI. Dois exemplos são a classificação e a clusterização, pois não é possível generalizar grupos de objetos semanticamente significativas a partir de termos não populares (estes termos pertecem a pouquíssimo objetos).

Nós recomputamos os valores de $AIFF$ considerando apenas termos que aparecem em mais de 50 objetos[6]. Os resultados apresentados na Tabela 3.8 demonstram

trouxe resultados qualitativamente similares aos apresentados. O mesmo vale para o $AIFF$.

[6]Fizemos uso de outros limiares como: 10, 50, 100 e 1000. Todos estes limiares alcançaram resul-

Tabela 3.8: Valores de $AIFF$ (ignorando termos populares). Intervalos com 90% de confiança foram omitidos, estes exibem erros abaixo de 1%

	CiteULike	LastFM	YahooVideo	YouTube
TITLE	7.84	6.97	7.05	7.27
TAGS	7.60	6.39	6.86	7.14
DESC.	7.01	6.12	6.74	6.89
COMM.	-	6.08	-	6.65

uma distinção mais clara entre os atributos. Desta forma, de acordo com a heurística $AIFF$ desconsiderando termos raros, a ordenação de capacidade discriminativa dos atributos é a mesma da capacidade descritiva. Título é o atributo mais discriminativo, seguido de etiquetas, descrição e comentários.

3.2.5 Diversidade de Conteúdo

Também investigamos se diferentes atributos contribuem com conteúdo (termos) diferentes sobre um mesmo objeto. 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 3.2.2, 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:

$$J(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \tag{3.1}$$

A similaridade entre cada par de instâncias de atributo textual foi comparada, usando como entrada conjuntos dos N termos de maior peso $TS \times IFF$ em cada instância.

A Tabela 3.9 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 termos. A exceção é etiquetas no YahooVideo, que, apesar de colaborativo, tem similaridade maior que 0.19 com título

tados similares aos apesentados com 50. Também experimentamos filtrar os termos muito populares (inicio da curva) por serem muito gerais, porém os valores das métricas não se alteraram significativamente.

Tabela 3.9: Similaridade média entre instâncias de atributos não vazias. Intervalos com 90% de confiança foram omitidos, estes exibem erros abaixo de 0.3%

	CiteULike	LastFM	Yahoo	YouTube
TÍTULO×ETIQUETAS	0.13 (0.09)	0.07 (0.01)	0.52 (0.33)	0.36 (0.25)
TÍTULO×DESC.	0.31 (0.09)	0.22 (0.03)	0.40 (0.20)	0.28 (0.015)
ETIQUETAS×DESC.	0.13 (0.02)	0.13 (0.04)	0.43 (0.20)	0.32 (0.14)
TÍTULO×COME.		0.12 (0.02)		0.14 (0.02)
ETIQUETAS×COME.		0.10 (0.03)		0.17 (0.02)
DESC.×COME.		0.18 (0.03)		0.16 (0.03)

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.

3.2.6 Resultados da Caracterização

Os nossos estudos de caracterização podem ser sumarizados nos quatro resultados principais. Primeiramente, dos quatro atributos textuais, três (ETIQUETAS, DESCRIÇÃO e COMENTÁRIOS) tem uma fração não desprezível de instâncias vazias em pelo menos duas aplicações. Devido a isto, tais atributos podem ter impactos negativos em serviços de RI, pois não será possível explorar tal atributo nas instâncias vazias. No geral, atributos restritivos aparentam ser mais explorados pelos usuários do que os colaborativos. Segundo, a quantidade de conteúdo tender a ser maior em atributos colaborativos. Terceiro, mesmo sendo composto em maioria de termos descritivos (sinônimos), todos os atributos textuais aparentam sofrer por problemas de ambiguidade. Por fim, atributos menores e mais utilizados, como TÍTULO e DESCRIÇÃO, tem maior capacidade descritiva e discriminativa.

No próximo capítulo nos averiguamos a qualidade relativa dos atributos textuais quando aplicados para duas tarefas de RI: a classificação de objetos e a recomendação de etiquetas. Também realizamos um estudo com 17 usuários, onde os mesmos avaliaram manualmente a qualidade relativa dos atributos textuais. Os resultados dos estudos de tarefas de RI e de inspeção manual dos atributos serão discutidos com base nos nossos resultados de caracterização.

Capítulo 4

Experimentos com Serviços de RI e Usuários

Neste capítulo, apresentamos três estudos diferentes que visam entender a qualidade dos atributos textuais quando aplicada em duas diferentes tarefas de RI: classificação e recomendação de etiquetas. Apresentamos também um estudo como 17 voluntários onde avaliamos a qualidade dos atributos para usuários finais de aplicações Web 2.0.

4.1 Classificação de Objetos

Formalmente, o problema de classificação pode ser definido da seguinte forma: dado um conjunto de objetos de treino, $\{(x_1, y_1), ..., (x_n, y_n)\}$, produza uma função $f : X \rightarrow Y$ que associa cada objeto $x \in X$ para a sua classe correta $y \in Y$. Esta função é o modelo utilizado para classificar outros objetos. Podemos testar a eficácia de um classificador com um conjunto de objetos diferente do teste, denominado de treino, cujas classes já são conhecidas.

4.1.1 Modelo de Representação de Objetos

Para realizar experimentos de classificação, inicialmente modelamos o nossos objetos em um espaço vetorial denominado de *vector space model (VSM)* [49]. Neste, cada termo do vocabulário total dos atributos pode ser visto como uma dimensão do espaço vetorial. Uma instância de um objeto (o) será então um vetor (V), com dimensionalidade ($|V|$) igual ao vocabulário da coleção. Os termos de uma instância são representados por um valor real diferente de zero no vetor. Este valor deve, idealmente, capturar

o grau de relação deste termo com a instância. Para a atribuição destes valores nos vetores, fizemos uso das seguintes estratégias de ponderação de termos:

TS : o peso de um termo t e uma instância de objeto o é igual ao espalhamento de t em o $TS(t, o)$. Heuristicamente, este valor captura o poder descritivo do termo nesta instância (ver a Seção 2.3.1).

IFF : o peso de um termo t em uma instância o é igual ao logaritmo do inverso da sua frequência em um atributo $IFF(t, F)$. Este valor captura o poder discriminativo do termo.

TS $\times$ **IFF** : o peso de um termo t é dado pelo produto $TS(t, o) \times IFF(t, F)$, capturando assim tanto o poder descritivo quanto o discriminativo.

TF : o peso do termo t é dado pela frequência ($TF(t, o)$) do termo naquela instância.

TF $\times$ **IFF** : o peso do termo t é pelo produto $TF(t, o) \times IFF(t, F)$.

As últimas duas estratégias nós permitiu comparar a heurísticas TS com a mais utilizada TF. Além de definir a ponderação dos termos, precisa-se também definir quais termos compõem os vetores:

Título : apenas os termos do título de o constituem o vetor V, isto é, $V = V_{titulo}$. Formalmente, o vetor é definido como sendo: $V_{titulo} = \langle p_{t1}, p_{t2}, ..., p_{tn} \rangle$, sendo p_i é o peso de um termo t, representado na posição i do vetor, onde $t \in f$ e $f =$ título.

Etiquetas : de forma análoga ao acima, apenas os termos das etiquetas de o constituem o vetor V, isto é, $V = V_{etiquetas}$.

Descrição : apenas os termos da descrição de o constituem o vetor V, isto é, $V = V_{descricao}$.

Comentários : apenas os termos dos comentários de o constituem o vetor V, isto é, $V = V_{comentarios}$.

Conjunto-de-Termos : assim como é feito em algoritmos tradicionais de RI, os quatro atributos são considerados como um único documento, sem considerar a estrutura da página em blocos. Neste caso, um termo presente em mais de um atributo é considerado na mesma posição do vetor V. Formalmente, o vetor é definido como sendo: $V_{conj.} = \langle p_{t1}, p_{t2}, ..., p_{tn} \rangle$, sendo p_i é o peso de um termo t,

representado na posição i do vetor, onde $t \in o$. Neste caso, o uso da métrica IFF equivale ao uso da métrica mais comum IDF.

Concatenação : os vetores V_{titulo}, $V_{descricao}$, $V_{etiqueta}$ e $V_{comentarios}$ são concatenados em único vetor $V_{conc.}$. Ou seja: $V_{conc.} = \langle V_{comm.}, V_{desc.}, V_{tag}, V_{title} \rangle$. Neste caso, um termo t que aparece em mais de um atributo tem posições, e provavelmente pesos, diferentes no vetor. A ideia desta estratégia é considerar os termos de todos os atributos textuais, porém sendo cada um uma fonte de informação diferente dos outros.

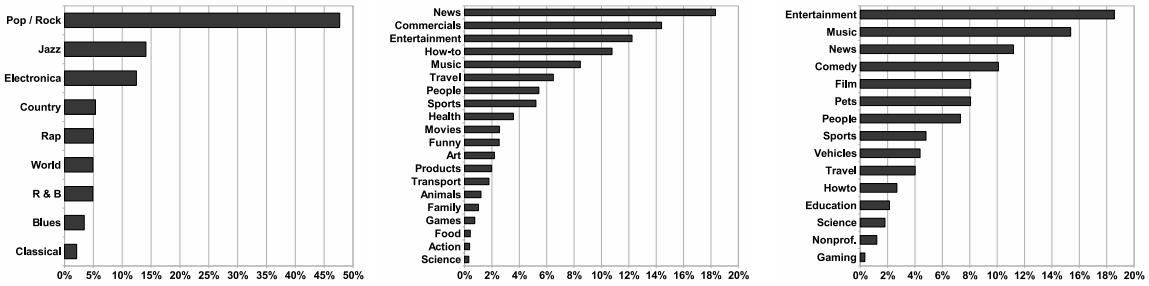
Independente de estratégia de peso ou termos que constituem os vetor, em todos os casos os vetores V são normalizados para que as normas sejam $||V|| = 1$. As duas últimas estratégias são motivadas pelos resultados da Seção 3.2.5, estas nos permitem comparar a eficácia dos atributos quando utilizados de forma isolada com a combinação de vários atributos.

4.1.2 Definição dos Experimentos

Nossos experimentos de classificação fizeram uso de um algoritmo do tipo *Support Vector Machine (SVM)* com um núcleo (*kernel*) linear. Não apenas as SVMs são consideradas o estado-da-arte no quesito de algoritmos de classificação, o uso do núcleo linear se provou bastante eficaz e eficiente para a classificação automática de um grande número de documentos de texto [30]. Mais especificamente, utilizamos a ferramenta Liblinear que implementa tal algoritmo [14]. Nestes estudos, fazemos uso dos objetos com classes coletadas do LastFM, YahooVideo e YouTube (ver Seção 3.1). Assim como na Seção 3.2.2, utilizamos apenas instâncias de atributos não vazias, da língua inglesa e com remoção de afixos. Novamente desconsideramos o atributo COMENTÁRIOS no YahooVideo. Além do mais, como demonstrado na Figura 4.1[1], algumas classes de objetos contém poucos objetos, sendo sub populadas. Nos nossos experimentos, desconsideramos todas as classes com menos de 2.2% dos objetos totais para cada aplicação. Com isto removemos 8 e 4 classes para o YahooVideo e YouTube, respectivamente.

Para cada aplicação, selecionamos de forma aleatória e uniforme dez amostras de 5000 objetos. Devido ao menor número de objetos com classes, no LastFM apenas uma amostra foi utilizada. Nossos experimentos de classificação foram constituídos de uma validação cruzada de dez partes. Nesta, dividimos cada amostra em dez partes iguais. Cada rodada do classificador faz uso de nove partes (90%) dos objetos para treino,

[1]Os nomes das classes na Figura 4.1 foram reduzidos por questões de espaço. A Tabela 3.1 apresenta os nomes completos das classes.



(a) LastFM (b) YahooVideo (c) YouTube

Figura 4.1: Distribuição de Objetos nas Classes.

gerando o modelo do classificador, e uma parte (10%) de teste[2]. Cada uma destas partes em cada uma das amostras foi configurada de acordo com cada modelo vetorial definido acima (combinação de uma estratégia de ponderação de termos com um tipo de vetor) descritos acima. Ao utilizar cada uma das dez partes como teste, teremos então dez diferentes resultados de classificação para uma única amostra, totalizando cem resultados diferentes para cada modelo vetorial.

Para cada experimento de classificação, fizemos uma busca pelo melhor valor do parâmetro custo (C) do SVM. Diferentes valores de custo afetam significativamente o resultado final de classificação. Testamos os valores de 0.01, 0.1, 1, 10 e 100. Na nossa busca, encontramos que o valor padrão do custo ($C = 1$) como sendo o melhor na maioria dos casos, portanto apresentamos apenas os resultados com este valor.

Nós averiguamos a qualidade dos atributos utilizando duas métricas de avaliação de classificadores: Micro-F1 e Macro-F1. Sendo $TP(c)$, $FP(c)$ e $FN(c)$ a quantidade de resultados (com base no treino) verdadeiro positivos, falso positivos e falso negativos que o classificador retornou para uma classe c. Podemos definir a Precisão do classificador como sendo a fração de todos os objetos que foram atribuídos para a classe correta, isto é:

$$P(c) = \frac{TP(c)}{TP(c) + FP(c)} \tag{4.1}$$

Por outro lado, a Revocação é fração de objetos retornados como corretos que realmente pertenciam a classe c:

$$R(c) = \frac{TP(c)}{TP(c) + FN(c)} \tag{4.2}$$

As duas métricas são complementares. A Precisão indica se o classificador é

[2]Este processo é denominado do *10-fold cross validation* em inglês

capaz de atribuir uma fração alta de objetos para a classe correta c, em contrapartida, a Revocação indica se o classificador foi capaz de atribuir à classe c uma fração alta dos objetos realmente são desta classe. É comum resumir o valor das duas métricas de forma conjunta através da média harmônica entre Precisão e Revocação, esta nova métrica é denominada de $F1(c)$:

$$F1(c) = \frac{2 \cdot P(c) \cdot R(c)}{P(c) + R(c)} \tag{4.3}$$

As métricas definidas acima são válidas para uma classe apenas. Para avaliar a eficácia do classificador em várias classes, pode-se usar uma das duas estratégias abaixo:

Macro F1: A métrica *Macro F1* é definida como sendo o valor médio do $F1(c)$ por todas as classes.

Micro F1: As métricas Precisão e Revocação são reformuladas para utilizar valores de TP, FP e FN iguais a soma destes valores por todas as classes, ao invés dos valores para um classe apenas. Então, Micro-F1 pode ser calculado usando os novos valores de Precisão e Revocação.

O uso da validação cruzada com 10 partes já foi demonstrado ser a forma mais eficaz de identificar se diferentes estratégias de classificação produzem resultados estatisticamente diferentes [12].

4.1.3 Resultados

Os resultados de classificação foram analisados como a média das métricas Micro-F1 e Macro-F1 considerando todas as partes e amostras para cada aplicação. Considerando um intervalo de confiança de 90%, um erro máximo de 2% foi obtido. Nós apresentamos apenas os resultados obtidos com o melhor valor do parâmetro custo ($C = 1$). Iniciamos nossa discussão comparando os diferentes atributos textuais quando ponderados com a métrica $TS \times IFF$. As Tabelas 4.2 e 4.1 demonstram os valores de Macro-F1 e Micro-F1 para os esquemas TS e $TS \times IFF$, respectivamente, para cada atributo ou combinação de atributos.

Iniciamos a discussão dos resultados comparando a qualidade dos atributos, após isto discutiremos o impacto dos diferentes esquemas de ponderação. Se considerarmos apenas os resultados dos atributos quando utilizados de forma isolada, as ETIQUETAS claramente é o atributo textual nas quatro aplicações, tanto no Micro-F1 quanto no

Tabela 4.1: Valores de Macro and Micro F1 para o modelo $TS \times IFF$. Intervalos com 90% de confiança foram omitidos, estes exibem erros abaixo de 10%. Os melhores resultados, assim como empates estatísticos, são apresentados em negrito.

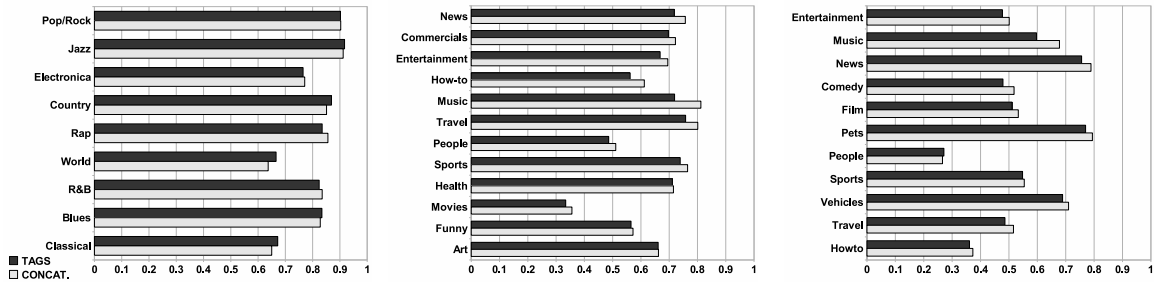
	Micro F1						Macro F1					
	TIT.	ETIQ.	DESC.	COM.	CONJ.	CONC.	TIT.	ETIQ.	DESC.	COM.	CONJ.	CONC.
Lastfm	0.52	**0.86**	0.81	0.69	**0.86**	**0.87**	0.20	**0.80**	0.70	0.46	**0.79**	**0.79**
Yahoo	0.57	0.66	0.63	-	**0.70**	**0.70**	0.52	0.63	0.57	-	**0.66**	**0.65**
Yout.	0.43	0.57	0.46	0.50	**0.62**	**0.62**	0.40	0.55	0.42	0.44	0.58	0.59

Tabela 4.2: Valores de Macro and Micro F1 para o modelo TS. Intervalos com 90% de confiança foram omitidos, estes exibem erros abaixo de 10%. Os melhores resultados, assim como empates estatísticos, são apresentados em negrito.

	Micro F1						Macro F1					
	TIT.	ETIQ.	DESC.	COM.	CONJ.	CONC.	TIT.	ETIQ.	DESC.	COM.	CONJ.	CONC.
Lastfm	0.53	**0.87**	0.80	0.67	0.85	**0.87**	0.19	**0.80**	0.70	0.43	0.78	**0.81**
Yahoo	0.56	0.66	0.61	-	0.68	**0.70**	0.52	0.62	0.54	-	**0.64**	**0.66**
Yout.	0.43	0.58	0.44	0.48	0.59	**0.62**	0.40	0.56	0.41	0.42	0.55	**0.58**

Macro-F1. Este resultado é consistente com a nossa caracterização pelos seguintes motivos: (1) de acordo com as heurísticas AFS e $AIFF$, o atributo ETIQUETAS tem uma boa capacidade descritiva e discriminativa, tendo estes valores apenas pouco abaixo daqueles obtidos pelo atributo TÍTULO; e (2) as ETIQUETAS têm pelo menos duas vezes a quantidade de conteúdo do TÍTULO, beneficiando assim a eficácia deste atributo para a classificação. Uma maior quantidade de conteúdo (termos) já foi demonstrado como sendo um fator importante para a eficácia tarefas de classificação [30, 34], porém não é o único fator que afeta tal eficácia. Acreditamos que este é o principal motivo pelo qual o atributo ETIQUETAS obteve os piores resultados, mesmo este atributo sendo o de maior capacidade descritiva e discriminativa.

Os resultados também demonstram que o atributo COMENTÁRIOS é melhor do que o DESCRIÇÃO no YouTube. Acreditamos que, mesmo COMENTÁRIOS obtendo valores menores de AFS (poder descritivo) e valores próximos de $AIFF$ (pode discriminativo) em comparação ao atributo DESCRIÇÃO, instâncias de COMENTÁRIOS tem, em média, 8 vezes mais termos do que as DESCRIÇÕES. O principal motivo para este resultado é a importância da quantidade de conteúdo para a tarefa de classificação. Em contrapartida, no LastFM foi averiguado que o atributo DESCRIÇÃO obteve melhores resultados de classificação do que os COMENTÁRIOS. Embora não seja consistente com o anterior, acreditamos que a natureza colaborativa na forma de wiki do atributo DESCRIÇÃO faz com que o mesmo tem conteúdo semântico de alta qualidade, um fato já observado em outros aplicativos wiki como a Wikipedia [25]. Este fator social não é capturado por nossas métricas e é um possível tópico para trabalhos futuros.



(a) LastFM (b) YahooVideo (c) YouTube

Figura 4.2: Valores de F1 para ETIQUETAS and CONCATENAÇÃO.

Os resultados de diversidade de conteúdo (ver a Seção 3.2.5) demonstrou que existe conteúdo distinto entre os atributos textuais. Os resultados quando combinamos os atributos (Concatenado e Conjunto-de-Termos) foram melhores do que o uso de atributos isolados no YouTube e YahooVideo. No caso do LastFM, os resultados apresentam-se estatisticamente iguais ao melhor atributo (ETIQUETAS). Complementar à expansão natural da quantidade de conteúdo, a eficácia das combinações de atributos pode ser interpretado como evidência de que cada atributo textual é uma fonte de dados com informações diferentes que são, possivelmente, complementares [45].

Os resultados também mostram que, na maioria dos casos, os resultados, tanto de Micro-F1 e Macro-F1, para a aplicação LastFM é significativamente maior do que nas outras aplicações. Isto é possivelmente devido ao fato de que as a atribuição de classes nos objetos do LastFM é feita por especialistas da indústria musical, assim garantindo um maior nível de consistência nas classes. No YouTube e YahooVideo, os próprios usuários atribuem as classes dos vídeos, podendo um vídeo pertencer a apenas uma classe. Este problema é acentuado quando consideramos o YouTube e YahooVideo tem um maior número de classes com, possível, sobreposição semântica. Isto torna ainda mais difícil a escolha da classe para um vídeo. Por exemplo, um vídeo de comédia pode ser corretamente inserido nas classes "Entertainment" ou "Funny Videos". A Figura 4.2 demonstra os resultados de classificação por classe (F1) para ETIQUETAS and CONCATENAÇÃO. Podemos claramente visualizar que o classificador obtêm resultados maiores nas categorias LastFM, sendo maior que 0.65 para todas as classes. Além disso, os resultados são mais igualmente distribuído entre as classes. Em contrapartida, no YouTube e YahooVideo, os valores de F1 tem valores menores, sendo 0.27 e 0.33 os piores resultados respectivamente.

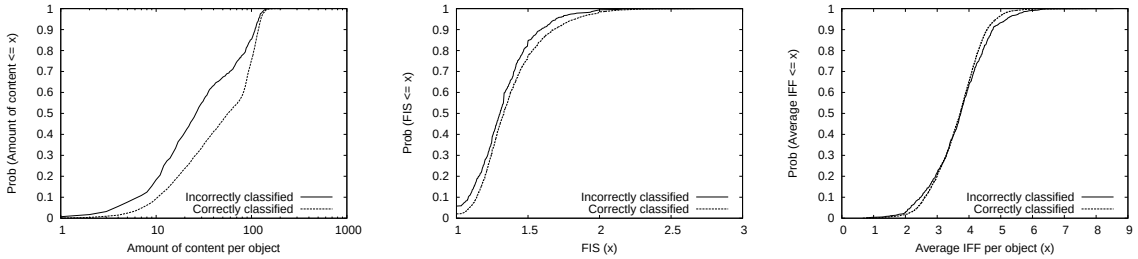
Por fim, analisamos o impacto das diferentes estratégias de ponderação de termos. Nossos resultados demonstraram que o impacto das diferentes estratégias é mínimo.

Este resultado, surpreendentemente, não é consistente com trabalhos passados [24, 45] que indicam que métricas com base no IDF degradam a eficácia de algumas tarefas de RI, isto ocorre devido ao ruído e falta de informação dos termos raros. Entretanto, devemos considerar dois aspectos dos nossos experimentos: (1) primeiramente, estratégias de classificação com base em SMVs são robustas a tais termos, sendo capaz de filtrar os mesmos como uma seleção de atributos [30]; (2) os nossos objetos contém poucos termos quando comparados a páginas web ou outros documentos, foco dos estudos anteriores, acreditamos que este motivo fez com que a quantidade de conteúdo fosse o fator mais impactante nos nossos experimentos.

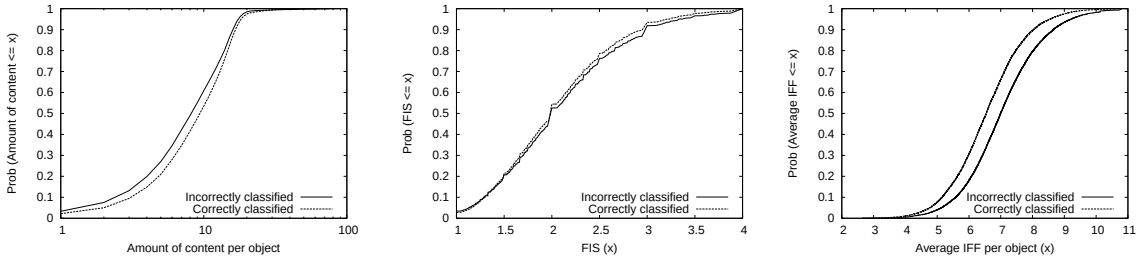
Para entender melhor estes resultados, nós comparamos as distribuições cumulativas de probabilidade da quantidade de conteúdo, poder descritivo e poder discriminativo para os objetos classificados corretamente e incorretamente. Por exemplo, a Figura 4.3 mostra um comparativo dos objetos classificados corretamente e incorretamente do atributo etiquetas no LastFM e YouTube. Pelos gráficos, podemos ver uma distinção clara apenas para a comparação entre a quantidade de conteúdo (Gráfico 4.3 (a)), sendo a diferença entre o poder descritivo não visualmente perceptível, enquanto o poder discriminativo aparenta degradar os resultados, tendo objetos classificados incorretamente valores de IFF menores.

Com o objetivo de comparar as distribuições para todas as aplicações, atributos e combinações de atributo fizemos uso do teste Kolmogorov-Smirnov e para o Test-T [28]. O primeiro compara se as curvas das distribuições podem ser vistas como estatisticamente diferentes, enquanto o segundo verifica se as médias são diferentes. Os dois testes foram realizados com um nível de confiança de 0.90. Os resultados demonstraram diferenças estatísticas entre os objetos corretos e incorretos existe apenas para a quantidade de conteúdo. A única exceção foi título no LastFM que, como já mencionado, é um caso especial devido ao conteúdo deste atributo (nome de artistas). Interpretamos este resultado como um suporte a nossa hipótese de que, das nossas métricas de qualidade, a quantidade de conteúdo é a mais importante para a classificação. Em outros cenários de classificação, estudos passados obtiveram o mesmo resultado [34].

De forma sucinta, podemos concluir que: 1) etiquetas, quando presente, é o melhor atributo para ser utilizado de forma isolada para a tarefa de classificação; 2) a combinação de atributos pode trazer melhoria na eficácia do classificador, isto ocorre devido a presença de termos distintos e, de certa forma, complementares entre os atributos; 3) devido ao fato de a atribuição de classes no LastFM ser feita por pessoas mais qualificadas, a classificação se torna mais eficaz; 4) diferentes formas de ponderar termos não traz diferenças nos resultados de classificação, porém existe evidência de



(a) Quantidade de conteúdo, LastFM (b) FIS, LastFM (c) IFF Médio, LastFM



(d) Quantidade de conteúdo, YouTube (e) FIS, YouTube (f) IFF Médio, YouTube

Figura 4.3: Impacto de cada aspecto de qualidade nos tesultados de classificação para o atributo ETIQUETAS ponderado por $TS \times IFF$.

que a quantidade de conteúdo é uma aspecto impactante nos resultados.

Aproveitamos a discussão do impacto dos diferentes aspectos de qualidade para relembrar que nem todo aspecto é igualmente importante para diferentes tarefas de RI. Na próxima seção extraímos mais evidências de qualidade para outra tarefa de RI, a recomendação de etiquetas. Neste caso, o poder descritivo aparece como o aspecto de maior importância.

4.2 Recomendação de Etiquetas

Nesta seção focamos em uma outra tarefa de RI, a recomendação de etiquetas. Esta tarefa consiste em aumentar a quantidade de conteúdo relevante no atributo etiquetas. Escolhemos focar no atributo ETIQUETAS pois: 1) este foi o melhor atributo isolado de acordo com os resultados e classificação; 2) este é o atributo de maior interesse para a comunidade, sendo o mais estudo e existindo diversos trabalhos na área de recomendação [7,22,52,54]; e, (3) ampliar o conteúdo de outros atributos não é factível dado que estes são compostos de sentenças completas, enquanto ETIQUETAS é composto de termos isolados.

Em uma tarefa de recomendação de etiquetas três problemas existem: 1) recomendar a maior quantidade de termos relevantes; 2) minimizar a quantidade de termos irrelevantes (lixo) a serem recomendados; 3) espera-se que os termos relevantes sejam recomendados antes dos irrelevantes, caso contrário o ranking de relevância pode ser visto como incorreto. No nosso estudo, fazemos uso de uma estratégia simples para a recomendação de etiquetas, que seria recomendar os termos relevantes dos outros atributos textuais. Em outras palavras, o conteúdo do atributo ETIQUETAS será expandido com termos dos outros atributos. Um termo é considerado relevante se o mesmo for relacionado ao objeto cujo atributo ETIQUETAS está sendo expandido. No nosso esquema de recomendação, fazemos uso dos modelos vetoriais definidos Seção 4.1. Novamente, os vetores TÍTULO, DESCRIÇÃO e COMENTÁRIOS serão poderados usando as métricas TF, TS, IFF, $TF \times IFF$, $TS \times IFF$[3]. No nosso esquema, definimos como termos relevantes todos os termos do atributo que será usado como recomendação (definido como sugestão). Os termos serão recomendados ordenados pelo peso, ou seja, a primeira recomendação será o termo com maior peso de acordo com a estratégia de ponderação.

Para a avaliação, assumimos que o atributo ETIQUETAS é vazio. Para a recomendação ($f_{sugestao}$) em instância ($f_{etiquetas}$), um termo recomendado será considerado relevante se o mesmo existir na nossa coleta. Ou seja, uma recomendação eficaz deve inserir no atributo ETIQUETAS os termos que aparecem na nossa coleta. A eficácia da sugestão de acordo com a quantidade de termos relevantes e lixo recomendado pode ser capturada pelas métricas Precisão, Revocação e *Mean-Average-Precision* (MAP). Neste contexto, a métrica de Precisão captura a fração de termos relevantes de foram sugeridos, ou seja:

$$P(f_{sugestao}) = \frac{|f_{etiquetas} \bigcap f_{sugestao}|}{|f_{sugestao}|} \tag{4.4}$$

A Revocação captura a fração de termos que deveriam ser recomendados, ou seja:

$$R(f_{sugestao}) = \frac{|f_{etiquetas} \bigcap f_{sugestao}|}{|f_{etiquetas}|} \tag{4.5}$$

As duas métricas acima capturam a quantidade de termos relevantes e irrelevantes de uma recomendação. Estas, podem ser novamente combinadas nas métrica $F1$. As métricas acima não capturam se a ordenação de relevância dos termos é correta, para isto fazemos uso da MAP. Esta métrica não só considera se os termos recomendados são relevantes, como também a posição dos mesmos na ordenação. Em outras palavras, se termos relevantes forem recomendados mais cedo o valor de MAP serão maiores.

[3]Neste caso a métrica TS não considera o atributo ETIQUETAS no seu cálculo

Para computar MAP, precisamos inicialmente computar as métricas *Precision-at-k* ($P@k$) e *Average Precision* (AP). $P@k$ é a fração de termos relevantes que aparecem até uma certa posição (k) do vetor:

$$P@k(f_{sugestao}, k) = \frac{|f_{etiquetas} \bigcap Rank(f_{sugestao}, k)|}{k} \tag{4.6}$$

onde $Rank(f_{sugestao}, k) = \{x_i \in f | 0 < i \leq k\}$, ou seja, os k primeiros termos da sugestão f. Por sua vez, AP é a média dos valores de $P@K$ para cada os valores de k que contém uma sugestão relevante:

$$AP(f_{sugestao}) = \frac{1}{|f_{etiquetas} \bigcap f_{sugestao}|} \sum_{k=0}^{|f_{sugestao}|} P@k(f_{sugestao}, k) \cdot rel(k) \tag{4.7}$$

onde $rel(k)$ é um função binária que indica se o termo na posição k é relevante. MAP é então definido como a média dos valores de AP para todos os objetos:

$$MAP(F) = \frac{1}{|F|} \sum_{f \in F} AP(f) \tag{4.8}$$

Na Tabela 4.3 mostramos valores de Precisão e Revocação para cada atributo textual. Mostramos também valores de $F1$, novamente sendo a média harmônica entre as duas métricas anteriores. Como podemos ver pela tabela, os valores de Precisão são maiores para atributos com pouco conteúdo, como título. Em todos os casos, o segundo melhor atributo é a descrição seguido dos comentários, quando considerado. Quando analisamos a Revocação, os resultados se alteram sendo título o pior atributo. Comentários é o segundo pior atributo, sendo sempre menor ou igual aos valores de descrição. Quando consideramos a combinação das duas métricas (F1), avaliando se as recomendações são relevantes e com pouco lixo, a ordenação é consistente com nossas métricas AFS e $AIFF$, sendo título o melhor atributo, seguido de descrição e, quando considerado, os comentários. Porém titulo é novamente afetado pelo seu menor tamanho, obtendo os piores valores de Precisão.

Para avaliar as diferentes estratégias de ponderação, fazemos uso da métrica MAP. A Tabela 4.4 apresenta os valores de MAP para os diferentes atributos e estratégias de ponderação. Pela tabela podemos ver que, independente da ponderação, título é o melhor atributo, sendo sempre maior ou igual aos valores da descrição. O atributo Comentários, quando considerados, apresentam os piores valores. Este valor é consistente com os anteriores e novamente com as métricas AFS e $AIFF$. Percebe-se que, pelos valores, ordenações com base na métrica IFF tende a obter

Tabela 4.3: Resultados de Recomendação de Etiquetas Avaliados por Precisão, Revocação e F1. Intervalos com 90% de confiança foram omitidos, estes exibem erros abaixo de 2%.

	CiteULike		LastFM			YahooVideo		Youtube		
	TIT.	DESC.	TIT.	DESC.	COM.	TIT.	DESC.	TIT.	DESC.	COM.
Prec.	0.14	0.03	0.23	0.07	0.05	0.63	0.28	0.55	0.22	0.03
Revoc.	0.22	0.38	0.01	0.21	0.10	0.39	0.44	0.31	0.38	0.37
$F1$	0.15	0.05	0.02	0.08	0.06	0.45	0.32	0.36	0.22	0.04

os piores resultados, possivelmente devido ao fato de que esta métrica considera a importância o termo em toda a coleção e não a sua importância para um objeto. Heurísticas descritivas, TS e TF, apresentam os melhores resultados, sendo, em alguns casos (YahooVideo e YouTube), os valores de TS significativamente maiores. Vemos este resultado como evidência de que o poder descritivo apresenta-se como aspecto mais importante para a tarefa de recomendação.

Tabela 4.4: Resultados de Recomendação de Etiquetas por MAP. Intervalos com 90% de confiança foram omitidos, estes exibem erros abaixo de 2%.

	CiteULike		LastFM			YahooVideo		Youtube		
	TIT.	DESC.	TIT.	DESC.	COM.	TIT.	DESC.	TIT.	DESC.	COM.
TF	0.24	0.23	0.24	0.22	0.16	0.73	0.49	0.61	0.38	0.20
TS	0.25	0.24	0.24	0.21	0.20	0.78	0.63	0.71	0.51	0.41
IFF	0.24	0.09	0.24	0.10	0.09	0.70	0.43	0.66	0.34	0.10
$TF \times IFF$	0.25	0.25	0.24	0.16	0.14	0.71	0.51	0.66	0.40	0.27
$TS \times IFF$	0.26	0.23	0.24	0.15	0.15	0.77	0.58	0.72	0.50	0.31

Desta análise podemos concluir que o título é o melhor atributo para recomendação, provavelmente devido a maior similaridade deste com as etiquetas (Seção 3.2.5), como também pelo fato de que os dois tem propósitos similares, resumir o conteúdo do objeto em poucas palavras. Mesmo sendo o melhor atributo, este sofre um impacto devido ao seu menor tamanho e, em consequência, baixa Revocação. Um outro achado foi que a capacidade descritiva apresentou-se como aspecto importante para a recomendação, onde mesmo tendo valores baixos de Revocação, o título é mais eficaz quando consideramos a métrica F1. A nossa análise apenas indica se os termos que já conhecemos (aparecem na coleta das etiquetas são relevantes. Analisar a relevância dos outros termos sugeridos é um possível tópico para trabalho futuro.

4.3 Experimentos com Usuários

Como uma outra forma de verificação da qualidade dos atributos, realizamos um experimento com usuários. Neste, cada usuário foi solicitado a assistir à vídeos selecionados do YouTube e atribuir 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 AFS. 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.

Um total de 17 voluntários participaram do experimento, estes 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. Foram exibidos dez dos mais populares vídeos da amostra coletada do YouTube, onde cada tinha uma duração entre 4 e 10 minutos. Cada usuário assistiu todos os dez vídeos, totalizando 170 avaliações. A escolha do intervalo de duração e de vídeos populares foi baseada em um outro estudo inicial onde selecionamos vídeos aleatórios. Neste piloto, muitos vídeos de curtíssima duração (abaixo de 5 segundos) e de baixa qualidade visual fizeram com que os usuários não compreendessem 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 a verbosidade desse atributo [4]. 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 etiquetas 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 FS, que claramente aponta estes dois atributos como os melhores em termos de capacidade descritiva. 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 in-

[4] Notamos que a página principal de um objeto na Web 2.0 geralmente não torna visível todos os comentários, mas sim apenas os mais recentes.

formação. 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* [43].

Apesar desta nota mais baixa poder ter sido influenciada pelo número limitado de comentários mostrados aos voluntários, os resultados estão consistentes com a heurística FS que considera *todos* os comentários.

Capítulo 5

Conclusões e Trabalhos Futuros

O advento da Web 2.0 alterou a forma que usuários interagem com aplicações da Web. Hoje em dia, tais usuários tem um papel mais ativo como criadores de conteúdo e não apenas consumidores passivos. O resultado desta criação colaborativa de conteúdo, denominado de mídia social, fez com que aplicações da Web 2.0 alcançassem imensa popularidade (em quantidade de usuários e volume de dados) nos últimos anos. A mídia social abrangem não só apenas os objetos multimídia de uma aplicação (como vídeos no YouTube), como também os meta-dados comumente associados a estes objetos (e.x. as etiquetas). Porém, por não oferecer garantias de qualidade, o uso da eficaz da mídia social em tarefas de RI ainda se apresenta como um desafio para o desenvolvimento de serviços de RI eficazes.

Nesta dissertação foram investigadas questões relacionadas a qualidade de atributos textuais em aplicações da Web 2.0. Para melhor entender a qualidade dos atributos, foram amostrados dados de TÍTULO, DESCRIÇÃO, ETIQUETAS and COMENTÁRIOS associados com objetos do CiteULike, LastFM, YahooVideo e YouTube. Iniciamos nosso estudo com uma caracterização que levantou evidências de qualidade dos atributos no quesito dos seguintes aspectos: 1) uso dos atributos; 2) quantidade de conteúdo; 3) corretude sintática; 4) e por fim, capacidade descritiva e discriminativa. Nossos resultados de caracterização mostraram que atributos colaborativos, incluindo o mais estudado pela literatura, as etiquetas, estão ausentes (zero termo) em uma fração não desprezível de objetos. Entretanto, quando presentes, estes tendem a prover mais conteúdo do que os atributos restritivos, como o TÍTULO. Também descobrimos que atributos menores, como o TÍTULO e ETIQUETAS, tem uma maior capacidade descritiva e discriminativa de acordo com as heurísticas adotadas. Também mostramos que todos os atributos sofrem de problemas relacionados à ambiguidade dos termos, algo previamente analisado em aplicações de etiquetagem. Por fim, analisamos a diversidade de conteúdo

entre atributos e demonstramos que cada atributo traz conteúdo novo sobre um mesmo objeto.

Seguido da classificação, nós avaliamos a qualidade dos atributos textuais quando utilizados em duas tarefas de RI: a classificação e a recomendação de etiquetas. Nossos resultados de classificação mostraram que, quando presente, ETIQUETAS é o atributo textual de maior qualidade quando utilizado de forma isolada. Porém, a ausência deste atributo em uma fração não desprezível de objetos é um fator a ser considerado. Os resultados de classificação também demonstraram que o atributo TÍTULO, embora provido de uma alta capacidade descritiva e discriminativa, foi o pior nos resultados de classificação, sendo severamente impactado pela baixa quantidade de conteúdo. Também mostramos que resultados de classificação podem ser melhorados através da combinação de atributos, indicando que cada atributo traz não apenas conteúdo novo, como também relevante sobre os objetos. Por fim, nossos estudos demonstram que, nos nossos experimentos, a quantidade de conteúdo foi o aspecto de qualidade de maior impacto para a classificação.

Em contrapartida, na nossa análise de recomendação de etiquetas, o TÍTULO apresentou-se como o atributo de maior qualidade. Embora ainda seja impactado pelo pouco conteúdo, capturado pela Revocação, o atributo apresentou os melhores resultados de Precisão, MAP e F1. Neste caso, não apenas o resultado é consistente com as métricas $AIFF$ e AFS, como também mostramos que para a recomendação, o poder descritivo surge como aspecto mais importante.

Também foi realizado um estudo com 17 voluntários onde estes avaliaram a qualidade dos atributos textuais. Os resultados deste estudo mostrou que os usuários perceberam atributos menores, como o TÍTULO e as ETIQUETAS, como de maior qualidade no quesito de capturar o conteúdo de vídeos do YouTube. Novamente ressaltamos que os aspectos de qualidade levantados no nosso estudo não são igualmente importantes para diferentes tarefas de RI e/ou percepção do usuário. Embora pareçam inconsistentes, diferentes aspectos de qualidade afetaram os resultados de classificação, recomendação e estudo com usuários, causando assim as divergências entre os resultados.

Ressaltamos também que o conhecimento gerado nesta dissertação pode ser de grande uso para desenvolvedores de aplicações Web 2.0. Através de estudos de qualidade em amostras de dados, tais desenvolvedores podem fazer uso dos nossos aspectos de qualidade e heurísticas para avaliar a qualidade dos atributos existentes na sua aplicação. Este conhecimento pode então levar a decisões importantes em como serão feitas as tarefas de RI em uma aplicação Web 2.0 e, possivelmente, como melhorar a qualidade dos dados para RI, algo que vemos como trabalho futuro.

Referências Bibliográficas

[1] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding High-Quality Content in Social Media. In *Proc. WSDM*, 2008.

[2] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *SIGIR*, 2009.

[3] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross. Identifying Video Spammers in Online Social Networks. 2008.

[4] J. Bian, Y. Liu, E. Agichtein, and H. Zha. A Few Bad Votes Too Many? Towards Robust Ranking in Social Media. In *Proc. AIRWeb*, 2008.

[5] K. Bischoff, F. Claudiu-S, N. Wolfgang, and P. Raluca. Can All Tags Be Used for Search? In *Proc. CIKM*, 2008.

[6] S. Boll. MultiTube–Where Web 2.0 and Multimedia Could Meet. *IEEE Multimedia*, 14(1), 2007.

[7] A. Byde, H. Wan, and S. Cayzer. Personalized tag recommendations via tagging and content-based similarity metrics. In *Proc. ICWSM*, 2007.

[8] D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Block-based Web Search. In *Proc. SIGIR*, 2004.

[9] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. IMC*, 2007.

[10] L. Chen, P. Wright, and W. Nejdl. Improving music genre classification using collaborative tagging data. In *Proc. WSDM*. ACM New York, NY, USA, 2009.

[11] D. Dalip, M. Gonçalves, M. Cristo, and P. Calado. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In *Proc. JCDL*. ACM New York, NY, USA, 2009.

[12] T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[13] K. Emamy and R. Cameron. Citeulike: A Researcher's Social Bookmarking Service. *Ariadne*, 51, 2007.

[14] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. LIBLINEAR: A Library for Large Linear Classification. *J. of Mach. Learn. Res.*, 9, 2008.

[15] C. Fellbaum and I. NetLibrary. *WordNet: An Electronic Lexical Database*. MIT Press USA, 1998.

[16] 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.

[17] J. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4), 2002.

[18] J. Giles. Special Report: Internet Encyclopedias Go Head to Head. *Nature*, 438(15), 2005.

[19] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube traffic characterization: a view from the edge. In *Proc. IMC*, 2007.

[20] S. Golder and B. Huberman. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, 32(2), 2006.

[21] L. A. Goodman. Snowball Sampling. *Annals of Math. Statistics*, 32(1), 1961.

[22] Z. Guan, J. Bu, Q. Mei, C. Chen, and C. Wang. Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In *Proc. SIGIR '09*, New York, NY, USA, 2009. ACM.

[23] M. Halvey and M. Keane. Analysis of online video search and sharing. In *HT*, 2007.

[24] T. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. 2002.

[25] M. L. E. Hu, A. Sun, H. Lauw, and B. Vuong. Measuring article quality in wikipedia: models and evaluation. In *Proc. CIKM*, 2007.

[26] X. Hu and J. Downie. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *ISMIR*, 2007.

[27] P. Jaccard. Étude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37, 1901.

[28] R. Jain. *The art of computer systems performance analysis*. Wiley, 1991.

[29] J. Jeon, W. Croft, J. Lee, and S. Park. A Framework to Predict the Quality of Answers with Non-textual Features. In *Proc. SIGIR*, 2006.

[30] T. Joachims, C. Nedellec, and C. Rouveirol. Text categorization with support vector machines: learning with many relevant. In *Europ. Conf. on Machine Learning*. Springer, 1998.

[31] X. Kang, H. Zhang, G. Jiang, H. Chen, X. Meng, and K. Yoshihira. Understanding internet video sharing site workload: a view from data center design. *Journal of Visual Communication and Image Representation*, 2009.

[32] G. Koutrika, F. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina. Combating Spam in Tagging Systems. In *Proc. AIRWeb*, 2007.

[33] G. Koutrika, F. Effendi, Z. Gyongyi, P. Heymann, and H. Garcia-Molina. Combating spam in tagging systems: An evaluation. 2008.

[34] L. Larkey. Automatic essay grading using text categorization techniques. In *Proc. SIGIR*. ACM New York, NY, USA, 1998.

[35] X. Li, L. Guo, and Y. Zhao. Tag-based Social Interest Discovery. In *Proc. WWW*, 2008.

[36] A. Lipsman. comScore Data Confirms Reports of 100 Million Worldwide Daily Video Streams from YouTube.com in July 2006, `http://www.comscore.com/press/release.asp?press=1023`.

[37] A. Lipsman. YouTube Continues to Lead U.S. Online Video Market with 28 Percent Market Share, According to comScore Video Metrix, `http://www.comscore.com/press/release.asp?press=1929`.

[38] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Position Paper, Tagging, Taxonomy, Flickr, Article, Toread. In *Collaborative Web Tagging Workshop (WWW'06)*, 2006.

[39] C. Marshall. No Bull, No Spin: A comparison of tags with other forms of user metadata. 2009.

[40] T. Mei, L. Yang, X. Hua, H. Wei, and S. Li. VideoSense: a Contextual Video Advertising System. In *Proc. ACM Multimedia*, 2007.

[41] G. Mishne. Using blog properties to improve retrieval. *Proc. of ICWSM 2007*, 2007.

[42] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proc. AIRWeb*, 2005.

[43] G. Mishne and N. Glance. Leave a reply: An analysis of weblog comments. In *Proc. Workshop on Weblogging Ecosystem (WWW'06)*, 2006.

[44] T. O'reilly. What is web 2.0. *Design patterns and business models for the next generation of software*, 30:2005, 2005.

[45] D. Ramage, P. Heymann, C. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Proc. WSDM*. ACM New York, NY, USA, 2009.

[46] 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.

[47] W. Rotham. A Radio Station Just for You , 2007.

[48] Y. Rui, T. Huang, S. Mehrotra, and M. Ortega. A relevance feedback architecture for content-based multimedia information retrieval systems. In *Proc. CBAIVL*, 1997.

[49] G. Salton and M. McGill. *Introduction to Modern Retrieval*. McGraw-Hill Book Company, 1983.

[50] J. San Pedro and S. Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *Proceedings of the 18th international conference on World wide web*, pages 771–780. ACM, 2009.

[51] 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.

[52] B. Sigurbjornsson and R. van Zwol. Flickr Tag Recommendation Based on Collective Knowledge. In *Proc. WWW*, 2008.

[53] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE TPAMI*, 2000.

[54] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W. Lee, and C. Giles. Real-Time Automatic Tag Recommendation. In *Proc. SIGIR*, 2008.

[55] D. Strong, Y. Lee, and R. Wang. Data quality in context. *Commun. ACM*, 40(5), 1997.

[56] F. Suchanek, M. Vojnovic, and D. Gunawardena. Social Tags: Meaning and Suggestions. In *Proc. CIKM*, 2008.

[57] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *Proc. WWW*, 2007.

[58] T. Vander Wal. Explaining and Showing Broad and Narrow Folksonomies. *Personal InfoCloud.com*, 2005.

[59] R. Yates and B. Neto. Modern information retrieval. *New York, Addision Wesley*, 1999.