UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MARCELO RAFAEL BORTH

Uma abordagem de recomendação de *tags* semânticas para sistemas baseados em *tagging*

Maringá
2011

MARCELO RAFAEL BORTH

Uma abordagem de recomendação de *tags* semânticas para sistemas baseados em *tagging*

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação

Orientador: Prof. Dr. Sérgio Roberto Pereira da Silva

Co-orientadora: Prof. Dra. Valéria Delisandra Felfrim

Maringá
2011

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá – PR., Brasil)

Borth, Marcelo Rafael
B739u Uma abordagem de recomendação de *Tags* semânticas para sistemas baseados em *Tagging* / Marcelo Rafael Borth. -- Maringá, 2011.
119 f. : il., figs., tabs., apêndice.

Orientador: Prof. Dr. Sérgio Roberto Pereira da Silva.
Co-orientadora: Profª. Drª. Valéria Delisandra Feltrim.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Tecnologia, Departamento de Informática, Programa de Pós-Graduação em Ciência da Computação, 2011.

1. *Tags* - Recomendação. 2. *Tagging* - Identificação de semântica. 3. Ontologia - Organização da informação. 4. Recuperação da informação - Recurso Web - Ontologia. 5. Sistemas de informação - *Tagging*. I. Silva, Sérgio Roberto Pereira da, orient. II. Feltrim, Valéria Delisandra, co-orient. III. Universidade Estadual de Maringá. Centro de Tecnologia. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 21.ed. 006.333

FOLHA DE APROVAÇÃO

MARCELO RAFAEL BORTH

Uma abordagem de recomendação de *tags* semânticas para sistemas baseados em *tagging*

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Departamento de Informática, Centro de Tecnologia da Universidade Estadual de Maringá, como requisito parcial para obtenção do título de Mestre em Ciência da Computação pela Banca Examinadora composta pelos membros:

BANCA EXAMINADORA

Prof. Dr. Sérgio Roberto Pereira da Silva
Universidade Estadual de Maringá – DIN/UEM

Profa. Dra. Luciana Andréia Fondazzi Martimiano
Universidade Estadual de Maringá – DIN/UEM

Profa. Dra. Vânia Paula de Almeida Neris
Universidade Federal de São Carlos – DC/UFSCar

Aprovada em: 23 de fevereiro de 2011.
Local da defesa: Sala 101, Bloco C56, *campus* da Universidade Estadual de Maringá

DEDICATÓRIA

Dedico este trabalho aos meus pais por sempre me incentivar a estudar e a lutar pelos meus sonhos.

AGRADECIMENTOS

Não podemos dar passos grandes sem contar com o apoio e o estímulo de pessoas maravilhosas que conhecemos e mantemos em nosso caminho. Por isso, quero expressar minha gratidão a todos que, direta ou indiretamente, me ajudaram a realizar este trabalho:

Sobre tudo e todos, agradeço a **Deus**. Foi dele que recebi saúde, sabedoria, discernimento, capacitação e muita força nos momentos em que precisei;

Sou eternamente grato ao meu Pai, **Valdecir Oneide Borth**, meu espelho de caráter, simplicidade, humildade, honestidade e seriedade. À minha mãe, **Fátima A. de Caldas Borth**, meu espelho de dedicação, amor, carinho, doação e alegria. Pessoas fantásticas e muito especiais que me deram o que precisei, compreensão e apoio. Por mais que eu queira expressar em palavras todo o meu amor e admiração, não seriam suficientes para agradecer a dedicação que tiveram por mim. À minha irmãzinha **Nathália Fernanda Borth**, pelo amor, carinho, companheirismo, cumplicidade e brincadeiras sempre que estive em casa;

À noiva mais amada do mundo, **Niara Silva**, por todo amor, carinho, atenção e dedicação por mim;

Ao meu orientador, **Sérgio Roberto P. da Silva**, pela confiança e por ter me guiado nessa jornada e a minha co-orientadora, **Valéria D. Feltrim**, pela ajuda durante esse período;

À minha prima **Isabel Saes** e família, **Cidinho**, **João Paulo**, **Matheus** e **Gabriel** que me acolheram com muito carinho e por um bom tempo em sua residência;

Ao **Henrique Shishido**, um grande amigo que fiz durante essa etapa. Companheiro de cativeiro, estudos, alegrias, sofrimentos e desesperos;

Ao **Walter Figueira Neto** por toda força e confiança. Você foi como um padrinho pra mim durante todo esse tempo. Sem a tua ajuda as coisas seriam bem mais difíceis;

A todos os amigos do mestrado, em especial: **Alberto Biasão**, **Aldo Sérgio de Oliveira**, **Aleksandro Montanha**, **Alexandre Huff**, **Camila Leal**, **Everton Barros**, **Gabriel Costa Silva**, **Gustavo Sato**, **Maurílio Hirata**, **Rafael Cassolato**, **Rodrigo Tomaz Pagno** e **Paulo Sabo**. Os momentos de estudos (noites em claro), descontrações e cafezinhos não seriam os mesmos sem vocês;

Aos colegas do Grupo de Sistemas Interativos Inteligentes, pela amizade, incentivos e sugestões, em especial: **Carlos Alberto Meier Basso** e **Josiane M. P. Ferreira** que colaboraram imensamente com meu objeto de pesquisa;

Aos grandes amigos que fiz e que não são da área de computação, em especial: **Amanda Fujimura** e **Bruno Fulan**;

À **Maria Inês Davanço**, secretária do mestrado. Sua competência e dedicação são exemplos para muitos; e

Ao apoio financeiro da **CAPES** e ao Programa de Pós-Graduação em Ciência da Computação (**PCC-UEM)**.

Uma abordagem de recomendação de *tags* semânticas para sistemas baseados em *tagging*

RESUMO

Devido à popularização da *Web* grandes quantidades de conteúdos tornaram-se disponíveis, principalmente em função da atual facilidade de publicação de conteúdo por usuários leigos. Essa facilidade promove um incontrolável crescimento do total de informação acessível, criando, assim, a necessidade de mecanismos para indexação de conteúdos para auxiliar o usuário na recuperação da informação. Por causa dessa sobrecarga de informação, torna-se inviável que um usuário organize suas informações sem utilizar ferramentas que facilitem o processo de recuperação. Uma forma para efetuar essa organização é por meio de categorizações a partir de um sistema baseado em *tagging*, o qual auxilia o processo de atribuição de significado a qualquer objeto acessível na *Web*. Entretanto, uma vez que as *tags* utilizadas nas categorizações são selecionadas livremente pelo usuário, sem nenhum controle ou esquema rígido a ser seguido, a utilização desse tipo de sistema pode sofrer várias limitações, principalmente quando os usuários precisam recuperar um objeto categorizado. Esse problema se agrava na medida em que o usuário utiliza o sistema, pois seu vocabulário pode tornar-se inconsistente devido a erros de digitação, a criação de termos sinônimos e polissêmicos, a introdução de formas léxicas de escrita distintas para uma mesma representação conceitual, etc., e, por consequência, muitos recursos relevantes poderão ser excluídos dos resultados das buscas. Por essa razão, nesta dissertação é proposta uma abordagem de recomendação de *tags* semânticas que visa melhorar a qualidade das categorizações realizadas pelos usuários para facilitar a recuperação de informação e, também, resolver o problema de "*cold-start*". A proposta combina três fontes de informação distintas para gerar a recomendação: (i) os elementos textuais da página *Web*, (ii) as *tags* da folksonomia relativas à página em questão, e (iii) a personomia do usuário. Deste modo, a abordagem proposta tenta personalizar as recomendações com base nas características e interesses do usuário, bem como reduzir os problemas relacionados ao processo de *tagging*. Os resultados obtidos sobre a aceitação das *tags* recomendadas, a partir da execução de alguns experimentos com usuários reais, mostraram-se favoráveis para a tarefa de categorização, o que tende a reduzir os esforços empregados na fase de recuperação dos objetos categorizados.

Palavras-chave: Recomendação de *Tags*. Ontologia. Recuperação de Informação. Sistemas Baseados em *Tagging*.

An approach to semantic tag recommendation for tagging-based systems

ABSTRACT

Due to the popularization of the *Web* large amounts of content became available, mainly due to the current ease of content publishing by nontechnical users. This easiness promotes an uncontrollable growth of the total available information, thus creating the need for mechanisms for indexing content to assist users when retrieving information. Because of this information overload, it is infeasible for a user to organize their information without using tools that facilitate the recovery process. One way to accomplish this organization is by a categorization system based on tagging, which helps the process of assigning meaning to any object accessible over the *Web*. However, since the tags used in the categorizations are freely selected by the user, without any strict control to follow, using this type of system can have various limitations, especially when users need to retrieve an object categorized. This problem gets harder as the users make use of the system, because their vocabulary can become inconsistent due to typing errors, the creation of synonyms and polysemic terms, the introduction of different lexical forms of writing for the same conceptual representation, etc., and, consequently, many relevant resources may be excluded from the search results. Therefore, in this thesis it is proposed an approach for the recommendation of semantic tags that aim to improve the quality of the categorizations made by users looking forward to facilitate the information retrieval, and also to solve the of cold-start problem. The proposal combines three different information sources to generate the recommendation: (i) the textual elements of the *Web* page, (ii) the folksonomy tags of the page in question, and (iii) the user personomy. Thus, the proposed approach tries to personalize the recommendations based on the user characteristics and interests, as well as to reduce the problems related to the process of tagging. The results for the acceptance of the recommended tags from the execution of some experiments with real users were favorable for the categorization task, which tends to reduce the efforts necessary in the recovery phase of the objects categorized.

***Keywords*:** Tag Recommendation. Ontology. Information Retrieval. Tagging-Based System.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
GSII	Grupo de Sistemas Interativos Inteligentes
HTML	*Hypertext Markup Language*
J2EE	*Java* 2 *Enterprise Edition*
PLN	Processamento de Linguagem Natural
RDF	*Resource Description Framework*
TOM	*TagOntologyManager*
UEM	*Universidade Estadual de Maringá*
URL	*Uniform Resource Locator*
WEB	*World Wide Web*
XHTML	*eXtensible HyperText Markup Language*
XML	*Extended Markup Language*

SUMÁRIO

Capítulo

1

Introdução

Nos últimos anos, a quantidade disponível de dados, conteúdos e informações é muito superior à capacidade humana de absorvê-las, interpretá-las ou gerenciá-las, resultando, assim, em uma **sobrecarga de informação** (*information overload*) (LEVY, 2008). Esse termo é amplamente utilizado para definir a quantidade excessiva de conteúdos disponíveis na *Web* que é causada, principalmente, pela facilidade de publicação, juntamente com a ausência de mecanismos que controlam a qualidade do que é publicado (LEVY, 2008) (HIMMA, 2007) (LYMAN, 2000). Em consequência disso, torna-se mais difícil para uma pessoa discernir o que é ou não relevante.

Uma alternativa para organizar e classificar a informação e, posteriormente, possibilitar o auxílio na sua recuperação está no uso de taxonomias. De acordo com o dicionário Oxford (HORNBY, 2005), taxonomia é "o processo científico de classificar coisas" ou o "ramo da ciência que trata da classificação das coisas". O emprego de uma taxonomia na classificação da informação facilita sua recuperação, pois o usuário pode especializar/generalizar sua consulta até a categoria que ele achar necessário (BREITMAN, 2005). No entanto, em ambientes nos quais a complexidade e o custo no gerenciamento de conteúdo se tornam inviáveis, como a natureza aberta e dinâmica da *Web*, não é plausível

requerer profissionais para continuamente verificar e classificar os conteúdos que são publicados. Além disso, as taxonomias possuem alguns problemas, como: a necessidade de envolver especialistas para criar as categorias para classificar os itens; a não possibilidade de adicionar um recurso em mais de uma classe sem a duplicação do mesmo; o grande esforço cognitivo por parte dos usuários no momento de classificar um novo item e; a necessidade de requerer um especialista para adicionar uma nova classe quando um item a ser classificado pertencer a uma classe que ainda não existe. Breitman afirma que até mesmo quando as taxonomias são muito especializadas elas se tornam confusas ao usuário, tanto na classificação quanto na recuperação da informação.

Outro meio para organizar a informação é utilizar palavras-chave ou rótulos de texto (*tags*) para categorizá-las. Em torno de 2002, vários sistemas começaram a utilizar *tags* com o objetivo de organizar, descrever e atribuir significado aos conteúdos disponíveis na *Web* (SMITH, 2008). Assim, surgiram os sistemas baseados em *tagging*, os quais permitem que os usuários categorizem os recursos a partir de *tags* (livremente selecionadas) associadas a qualquer objeto acessível na *Web* (*URLs*[1], fotos, vídeos, etc.), como iniciativa no processo de atribuição de significado e organização da informação. Nesses sistemas, o grupo de *tags* e objetos de um usuário compõem sua **personomia** (HOTHO *et al.*, 2006), a qual tem por objetivo organizar as informações em um espaço pessoal, refletindo o vocabulário do usuário, suas preferências, interesses, conhecimentos, etc. O conjunto de personomias disponibilizadas para uma comunidade de usuários caracteriza uma **folksonomia** (WAL, 2005) (MATHES, 2004). Atualmente, essas técnicas são utilizadas por diversos sistemas *online* para organizar os dados dos usuários de uma forma rápida e prática (BEARMAN *et al.*, 1998) (CHUN *et al.*, 2005) (MILLER, 2005), concedendo-os, assim, um papel mais ativo. Desse modo, para sistemas que utilizam essa forma de organização, não se torna necessário selecionar profissionais ou especialistas no assunto para organizar o conteúdo, pois é adotado o princípio de que se alguém está produzindo e publicando algum conteúdo essa pessoa está apta a organizar e atribuir significado a ele da forma que melhor lhe convém (RIDDLE, 2005). Como resultado, o processo de *tagging* se torna uma alternativa interessante e viável para um ambiente aberto e mutável como a *Web*.

Um dos fatores de sucesso dos sistemas baseados em *tagging* é a falta de controle imposta ao usuário na escolha do vocabulário a ser empregado na categorização. Em vista disso, as palavras-chave podem ser associadas de forma que não causem uma sobrecarga

1*Uniform Resource Locator*

cognitiva ao usuário. Entretanto, o processo de livre atribuição de termos pode gerar redundância e ambiguidade, uma vez que a associação de uma *tag* a um objeto pode ser feita de várias maneiras devido a sinônimos, polissemia (GOLDER *et al.*, 2006), formas léxicas de escrita distintas (singular e plural), erros ortográficos (MATHES, 2004) (WU, 2006), abreviações vagas e palavras-chave cujos significados não estão diretamente relacionados ao recurso (CÔGO e DA SILVA, 2008), diferentes níveis de precisão e alta quantidade de dados individuais (SOOD *et al.*, 2007), ampliando ainda mais as dificuldades de gerenciamento de informações presentes nas personomias na medida em que o conjunto de *tags* utilizadas no sistema aumenta. Outro problema que afeta os usuários em qualquer sistema baseado em *tagging*, porém, não está diretamente relacionado à maneira de como os usuários interagem com o sistema e sim aos seus aspectos cognitivos, é que os usuários estão sujeitos a esquecerem facilmente os rótulos atribuídos aos recursos. Por essa razão, no momento em que o usuário precisa categorizar algum objeto semelhante ou tentar recuperar um objeto categorizado, nem sempre ele se lembrará quais foram as *tags* utilizadas nas categorizações.

Um vocabulário desorganizado, isto é, um vocabulário que possui termos inconsistentes, produz impactos negativos nas personomias. Em consequência dessa falta de coerência entre os termos empregados na categorização de um objeto, muitos recursos relevantes poderão ser excluídos dos resultados das buscas nos sistemas baseados em *tagging*, pois em uma eventual tentativa de recuperar um objeto a presença de termos polissêmicos, de *tags* sinônimas, de variantes no singular e plural poderão retornar recursos que não condizem com o significado desejado. Isso acontece devido às regras determinadas pelos algoritmos de busca, pois muitos dos sistemas realizam a comparação da *string* informada pelo usuário com a escrita das *tags* lexicamente e não semanticamente. Guy *et al.* (2006) afirmam que para reduzir os problemas apontados anteriormente, os sistemas que aplicam a técnica de *tagging* deveriam fornecer mecanismos para auxiliar o usuário na categorização e na recuperação da informação. Assim, muitas das dificuldades enfrentadas pelos usuários poderiam ser minimizadas pelos mecanismos de ajuda empregados pelos sistemas.

Uma alternativa para reduzir esses problemas, é desenvolver um mecanismo de detecção de inconsistências nas palavras-chave que foram utilizas como *tags* (CÔGO e DA SILVA, 2008) e sugerir modificações dessas ou das que possam dificultar a busca de um recurso. Em consequência, os termos considerados inconsistentes de cada categorização estariam sendo trocados por outros de forma mais conveniente. Sugerir correções de possíveis erros ortográficos, unificar separadores de termos compostos e padronizar o uso de singular ou plural nas *tags* podem ser exemplos para esse detector de inconsistências. No entanto, esse

caminho solucionará apenas os problemas para os recursos previamente categorizados e, por isso, se o sistema mantiver apenas essa alternativa para auxiliar o usuário na tentativa de manter seu vocabulário organizado, esse processo precisará ser executado periodicamente para as categorizações posteriores. Outro meio para reduzir os problemas citados e melhorar a recuperação da informação é realizar uma recomendação de *tags* que sejam bem aceitas pelos usuários no momento da categorização de uma página *Web*. No entanto, para evitar os problemas de sinonímia e polissemia é necessário que as *tags* recomendadas considerem a semântica dos termos, ou seja, que sejam *tags* semânticas – essa é a solução desenvolvida neste trabalho.

O processo de recomendação de *tags* funciona como um mediador entre o usuário e o sistema, pois ao invés de atribuir as *tags* automaticamente de forma implícita, ele permite que o usuário identifique os termos relevantes para um determinado recurso e, então, selecione os mais apropriados. Vários dos sistemas baseados em *tagging* existentes possuem um serviço que auxilia o usuário no processo de categorização, recomendando um conjunto de termos para um determinado objeto. Esse serviço de recomendação de *tags* pode ser desenvolvido de várias formas: pela exibição de instruções para escolha das *tags*; pela sugestão de *tags* similares que foram informadas por outros usuários; pela sugestão de sinônimos dos termos selecionados pelos usuários; apontando erros ortográficos nas *tags*; dentre outros. Além disso, o processo de *tagging* quando bem realizado, pode reforçar o vocabulário sem que a informação cognitiva seja prejudicada. Portanto, a recomendação pode ajudar em propósitos como: consolidar um vocabulário entre os usuários; fornecer uma opinião que não estava à espera ou ao alcance do usuário sobre o recurso; acelerar o processo de categorização; aumentar a eficácia da recuperação da informação mediante a consistência no vocabulário utilizado (JÄSCHKE *et al.*, 2007).

De modo geral, pode-se dizer que o problema da recuperação de informação dos sistemas baseados em *tagging* também está relacionado ao esquecimento das palavras-chave utilizadas na categorização. Isso pode ocorrer devido às informações que competem por espaços em nossa memória, pois cada vez que lembramos algo enfraquecemos outras memórias já armazenadas em nosso cérebro (IZQUIERDO, 2004), dificultando, assim, a lembrança dos termos que foram utilizados. Além do mais, Anderson (1995) afirma que, os seres humanos possuem dificuldades cognitivas em tarefas que envolvem a lembrança de termos que são previamente associados a um objeto, o que afeta a tarefa da recuperação de informação.

Uma possibilidade para reduzir esses problemas seria utilizar uma **recomendação de *tags* semânticas** no momento da categorização de um recurso, conforme sugerido por Adrian *et al*. (2007) e Marchetti (2007). Outros autores também (JÄSCHKE *et al.*, 2007) (SYMEONIDIS, 2008) apresentam soluções semelhantes a essas, embora sigam abordagens não semânticas as quais podem ser eficazes caso o sistema não precise de um tratamento tão rígido de classificação e de recuperação. Um problema das abordagens que não aplicam semântica entre as *tags* é que se torna mais difícil determinar, automaticamente, qual o tipo de relação que um par de *tags* possui entre si (ex.: generalização, especialização), deixando em aberto o problema da ausência de relações entre as *tags*. Por essa razão, Basso *et al.* (2009) acreditam que seria melhor gerar uma ontologia a partir dos dados da personomia do usuário, os quais estejam estruturados a partir de alguma fonte ontológica de informação (e.g. *WordNet*[2], *ConceptNet*[3] e *DBPedia*[4]), e sugerir conceitos semânticos no momento da categorização de um objeto, em vez de termos que não possuem relações e significados explícitos. Isso auxiliaria os usuários na redução do esforço cognitivo para recuperar a informação desejada, pois as *tags* sugeridas já estariam associadas a um conceito da ontologia, possibilitando, assim, a categorização/recuperação de recursos a partir de conceitos alternativos que as relações semânticas, como *hasAlternative*[5], *isA*[6] *e kindOf*[7], oferecem.

Dentre alguns trabalhos relacionados, o trabalho de Symeonidis (2008) propõe um *framework* para recomendação de *tags* que modela os três principais elementos existentes nos sistemas baseados em *tagging*: o usuário, o objeto e as *tags* presentes na personomia. No entanto, nos sistemas baseados em *tagging* que precisam de soluções mais rigorosas para classificar a informação essa não seria uma boa opção, pois sua recomendação não associa um sentido a *tag*, o que não traz uma solução para os problemas de sinonímia e polissemia. Por outro lado, Adrian *et al.* (2007) propõem uma abordagem de recomendação de *tags* semânticas para documentos, utilizando uma ontologia e serviços *Web*. Nesse trabalho, os documentos são primeiramente analisados e convertidos em RDF (*Resource Description Framework*) para a interpretação do algoritmo; em seguida, são extraídos tópicos de cada documento utilizando serviços *Web*; então, é realizado o processamento da ontologia para o

[2] Um grande banco de dados léxico eletrônico com relações formais entre seus termos. Disponível em http://wordnet.princeton.edu/
[3] Ontologia que tenta mapear o "senso-comum" dos seres humanos para tornar acessível por computadores. Disponível em: http://conceptnet.media.mit.edu/
[4] É um projeto que permite extrair informações estruturadas presentes no projeto *Wikipedia*. Disponível em: http://dbpedia.org/
[5] Relaciona duas *tags* que possuam o mesmo sentido. Ex.: "*plane*" é um termo alternativo para "*airplane*".
[6] Relaciona uma *tag* mais específica com uma mais abrangente. Ex.: um "carro" é um "veículo".
[7] É a relação contrária de *isA*.

tópico, obtendo objetos similares; e, por fim, é realizada a recomendação com base na ontologia gerada. Esse trabalho é interessante por ser um dos primeiros a tratar da recomendação de *tags* semânticas analisando termos do próprio documento.

Dado o espaço de problemas apresentado, este trabalho tem como propósito recomendar *tags* semânticas (baseadas em ontologia) que procurem melhorar a qualidade do vocabulário do usuário. O critério de qualidade empregado neste estudo está relacionado ao percentual das *tags* recomendadas que são identificadas como relevantes e aceitas pelos usuários. Assim, deseja-se reduzir os problemas citados que estão associados ao resultado do processo de *tagging* e à recuperação de informação. Para isso, esta proposta visa criar uma recomendação de *tags* semânticas que permite a sugestão de palavras-chave aos usuários com o objetivo de agregar termos que possam melhorar a qualidade de sua personomia e que possam ser relevantes na recuperação de informação. Esta abordagem se baseia em três fontes de informação distintas: (i) a **personomia** do usuário; (ii) as *tags* da **folksonomia** do sistema baseado em *tagging* e; (iii) os termos do **recurso *Web*** (este trabalho trata apenas páginas *Web* textuais). Dessa forma, utilizando essas três fontes de informação para a recomendação, este estudo agregará também outras vantagens, pois será possível identificar e recomendar conceitos relevantes da página *Web* que não foram encontrados na personomia ou na folksonomia e, principalmente, porque o usuário dificilmente ficará sem uma recomendação de *tags*, evitando o problema de "*cold-start*" (MALTZ, 1995), o qual está relacionado à dificuldade de se fazer recomendações para um recurso que ainda não foi categorizado pelos usuários ou que seja menos popular no sistema, não lhe permitindo que seja inferida qualquer característica social. Para validar a abordagem, foram realizados alguns experimentos com usuários para identificar o percentual de aceitação das *tags* recomendadas, os quais mostraram bons resultados, conforme descrito no Capítulo 6.

Esta dissertação está organizada da seguinte forma: no Capítulo 2, são apresentadas as principais bases teóricas para o entendimento deste estudo, descrevendo os principais conceitos envolvidos em *tagging*, folksonomia, ontologia, recomendação de *tags* e alguns trabalhos relacionados. No Capítulo 3 são mostradas as principais vantagens das três fontes de informação utilizadas para a recomendação desta abordagem. No Capítulo 4, é apresentada a metodologia para realizar a recomendação de *tags*. No Capítulo 5, é mostrada a arquitetura desenvolvida da aplicação de recomendação de *tags* e alguns aspectos de implementação. No Capítulo 6, são detalhados e apresentados os experimentos realizados, juntamente com os respectivos resultados a partir de sua execução com usuários reais. E, por fim, no Capítulo 7 são apresentadas as conclusões e as considerações finais do presente estudo.

Capítulo

2

Tagging, Ontologia e Recomendação de *Tags*

Devido à popularização da *Web* grandes quantidades de conteúdos tornaram-se disponíveis, principalmente em função da atual facilidade de publicação de conteúdo por usuários leigos. Essa facilidade promove um incontrolável crescimento do total de informação acessível, gerando uma sobrecarga de informação a qual torna difícil ao usuário recuperar a informação desejada. Além disso, devido a dificuldades técnicas e aos elevados custos de implementação, torna-se impraticável ter peritos qualificados para controlar e avaliar todo o conteúdo que é publicado na *Web*. Essa falta de alternativas para garantir a qualidade e a organização da informação resulta na necessidade de criação de mecanismos para organizar os conteúdos de forma a auxiliar o usuário quando da recuperação da informação. Atualmente, essa recuperação de informação é executada por alguns sistemas especializados como o Google[8], Yahoo[9], Bing[10], dentre outros. Esses sistemas possuem algoritmos complexos que analisam as páginas *Web* com o intuito de retornar os melhores resultados aos usuários. Nesse ambiente, o

[8] Disponível em: http://www.google.com
[9] Disponível em: http://www.yahoo.com
[10] Disponível em: http://www.bing.com

processo de *tagging* torna-se uma alternativa aos possíveis indexadores tradicionais (STURTZ, 2004), pois ele é realizado de forma coletiva, representando uma iniciativa para ajudar no processo de organização e atribuição de significado aos conteúdos acessíveis na *Web*.

Nesse capítulo são apresentadas as bases teóricas para os principais tópicos abordados no decorrer deste trabalho. Inicialmente, é explicado o processo de *tagging*, destacando seus pontos fortes e fracos e, como ele pode ajudar a atingir os objetivos citados no Capítulo 1. Além disso, são apresentados os problemas que os sistemas baseados em *tagging* enfrentam na recuperação da informação e, como solução, é mostrado como as ontologias podem ser utilizadas para fornecer uma recomendação de *tags*. Por fim, são discutidos alguns trabalhos relacionados referente à abordagem de recomendação de *tags* proposta nesta dissertação.

2.1. O Processo de *Tagging*

Enquanto organizar e classificar a informação usando uma taxonomia é uma tarefa que demanda tempo e alto custo cognitivo, utilizar *tags* (rótulos livres) para a mesma finalidade simplifica essa tarefa (SMITH, 2008). O processo de *tagging*, também conhecido por categorização ou anotação, pode ser definido como a prática de associar arbitrariamente rótulos de texto, palavras-chave ou marcações[11] para descrever, organizar ou atribuir algum tipo de significado aos conteúdos disponíveis na *Web*, tais como: páginas *Web*, fotos, vídeos, localizações em mapas, postagens de *blogs*, etc. O objetivo da técnica é facilitar o gerenciamento das informações pessoais de um usuário e auxiliar na recuperação desses objetos quando necessário (SMITH, 2008).

O *tagging* permite a criação descentralizada de *metadados* (dados para descrever dados), definindo conceitos e desempenhando o trabalho que seria realizado apenas por especialistas (RUSSELL, 2005). Esse processo é caracterizado basicamente por três elementos (MATHES, 2004) (RUSSELL, 2005) (SHEN *et al.*, 2005) (SMITH, 2008), ilustrados pela Figura 1: o usuário – o responsável por realizar a categorização; o objeto – o recurso que é categorizado; e as *tags* – as palavras-chave associadas ao objeto com a finalidade de fornecer uma descrição ao mesmo. A fim de exemplificar esse processo, considere um usuário que tem uma página *Web* (o objeto) o qual ele deseja categorizar. Para fazer isso, ele utiliza uma ou várias palavras-chaves (as *tags*) do próprio vocabulário que representa o conteúdo da página para associar ao objeto, com a intenção de recuperá-lo

[11] Utilizado para organizar os emails pelo Gmail. Disponível em: http://mail.google.com

posteriormente. Por essa razão, o *tagging* se diferencia de uma taxonomia que "classifica" a informação. Dessa forma, quando se usa a palavra *tagging*, há preferência por alguns autores (STRUTZ, 2004) (MATHES, 2004) (RUSSEL, 2005) (RIDDLE, 2005) em se utilizar o termo categorização, pois a palavra "categorizar" sugere um esquema menos rígido de organização do que "classificar". Em uma categorização, um objeto pode possuir vários significados relacionado a ele que podem ser adequados para várias categorias. Ao contrário, em uma classificação, geralmente é aplicada uma única categoria para cada objeto. Diante desse contexto, uma página *Web*, por exemplo, pode ser identificada por várias categorias: esporte, copa do mundo, seleção, futebol, etc. Isso demonstra que apenas um termo nem sempre é capaz de classificar um objeto com total precisão.

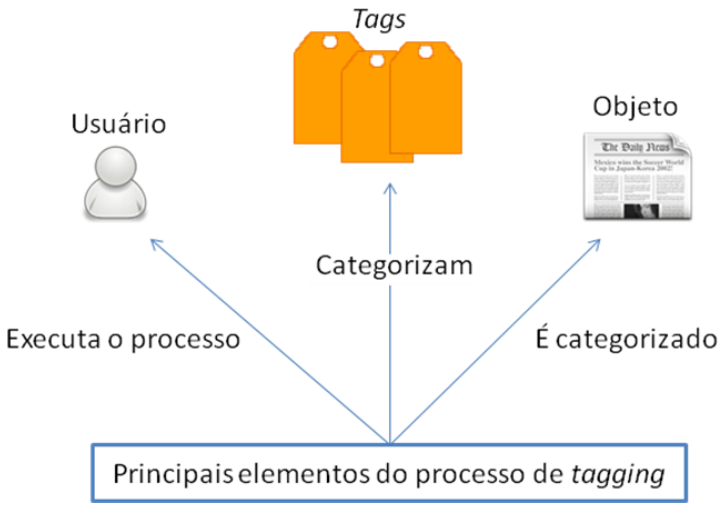

Figura 1: Os três principais elementos do processo de tagging.

Uma das principais características do processo de *tagging* é a liberdade que o usuário detém para realizar a categorização dos objetos, pois não há controle algum ou qualquer esquema rígido a ser respeitado, tornando o sistema fácil de ser utilizado. Dessa forma, os usuários apenas atribuem as *tags* que melhor lhes convém para referenciar o objeto. Pereira e Da Silva (2008) ressaltam que a liberdade concedida ao categorizar um recurso reduz os esforços cognitivos do usuário, facilitando a organização da informação comparada aos esforços utilizados para exercer uma classificação específica dentro de uma estrutura hierárquica como a taxonomia, visto que, na classificação é necessário especificar e/ou generalizar a categoria até que ela seja encontrada. Além disso, Mathes (2004) afirma que a categorização representa uma mudança fundamental ao configurar processos que não derivam de especialistas, possibilitando que a terminologia e precisão da informação se evidenciem na medida em que o volume de informação aumente, uma vez que serão os usuários que estarão

categorizando e atribuindo significados às informações por meio de suas visões, conhecimentos e culturas distintas. Alguns dos sistemas baseados em *tagging* mais conhecidos atualmente pela comunidade de usuários *Web* podem ser visualizados na Tabela 1.

Tabela 1: Alguns sistemas baseados em tagging disponíveis online.

Sistema	**Tipo de recurso**	***URL***
Delicious	Páginas *Web*	http://delicious.com
Flickr	Imagens	http://flickr.com
YouTube	Vídeos	http://youtube.com
Technorati	Postagens de *blogs*	http://technorati.com
Citeulike	Bibliografias	http://citeulike.org

De acordo com Smith (2008), os usuários possuem algumas motivações para utilizar *tags*, tais como: (i) facilidade – adicionar *tags* requer o mínimo investimento de tempo e cognição; (ii) simplicidade – basta apenas informar as palavras-chave para iniciar a organização da informação; (iii) flexibilidade – utilizar *tags* pode ser adaptado a qualquer situação, a qualquer propósito e a qualquer tipo de informação; (iv) extensibilidade – um termo já adicionado anteriormente não restringe o usuário de informar qualquer outro termo; e (v) agregabilidade – um recurso pode estar relacionado a diversas palavras-chave com vários conceitos distintos, ao contrário da taxonomia.

Conforme afirmam Al-Khalifa *et al*. (2007), geralmente, as *tags* podem ser classificadas em três tipos básicos, conforme a intenção do usuário no instante de cada categorização:

- ***Tags* fatuais**: identificam fatos a respeito do recurso, por exemplo, "*java*", "*photo*" e "*programming*";
- ***Tags* subjetivas**: exprimem a opinião dos próprios usuários a respeito do recurso, por exemplo, "*beautiful*" e "*funny*"; e
- ***Tags* pessoais**: são relacionadas a uma interpretação pessoal do usuário e, muitas vezes, são utilizadas para referências ou gerenciamento de suas tarefas, por exemplo, "*toread*" e "*mycar*".

Nos sistemas baseados em *tagging*, o grupo de *tags* e objetos utilizados pelo usuário compõem sua personomia, isto é, um conjunto de palavras-chaves contidas no próprio vocabulário que são utilizadas para organizar informação, juntamente com um conjunto de objetos. Essa personomia reflete o vocabulário, cultura, interesses, costumes, etc. do próprio usuário. Assim, uma personomia pode ser entendida como um conjunto de relações

estabelecidas entre as *tags* e os objetos por meio de um usuário em específico, conforme ilustra a Figura 2.

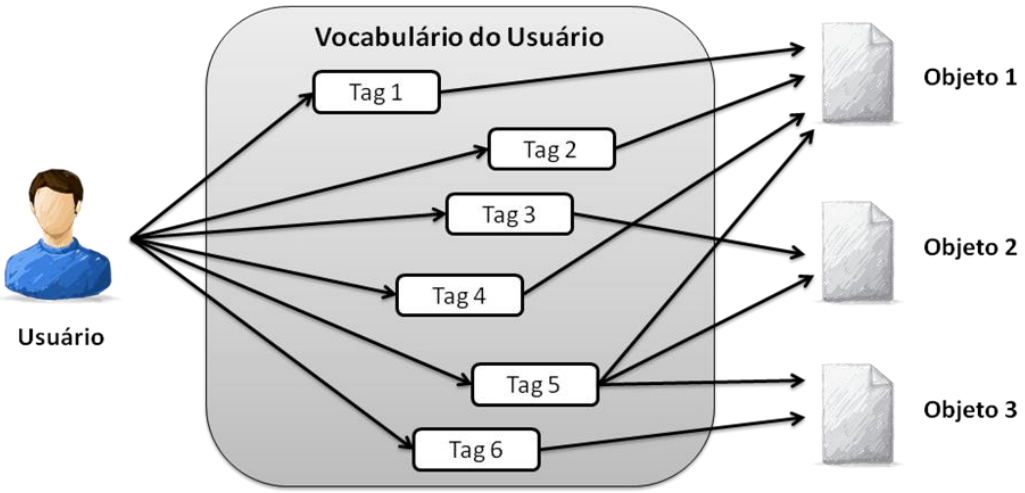

Figura 2: Ilustração de uma personomia de um usuário em um sistema baseado em tagging.

Uma das principais características quanto ao uso das *tags* e a criação da personomia do usuário é que não há a intervenção ou controle rígido no momento da escolha das *tags*. Além disso, os usuários podem atribuir quantos termos acharem necessários. A vantagem desse processo de atribuição de *tags* é justificável pelo fato de que as pessoas utilizam o próprio vocabulário para adicionar significado ao objeto, sem a necessidade de entender categorias taxonômicas, que muitas vezes podem ser complexas.

2.2. Folksonomias

Em vários casos, sistemas adotam a estratégia de permitir que os usuários compartilhem suas personomias com outros usuários. O compartilhamento de várias personomias é conhecido como **folksonomia**. Em 2004, Thomas Vander Wal criou o neologismo *folksonomy*[12], que representa a junção das palavras "*folks*"[13] e "*taxonomy*"[14], caracterizando um conjunto de personomias disponibilizadas socialmente para uma comunidade de usuários (WAL, 2005) (MATHES, 2004). Dessa forma, o ato de categorizar um recurso deixa de ser realizado apenas por um usuário (o criador do conteúdo), possibilitando que esse seja realizado por uma comunidade de usuários. A folksonomia surgiu como uma evolução do processo de *tagging*, embora na prática, essa técnica não traga nada de inovador. A diferença fundamental é que a

[12] Folksonomy: do inglês, folksonomia.
[13] Folks: do inglês, pessoas.
[14] Taxonomy: do inglês, taxonomia.

folksonomia trabalha com um conjunto de personomias compartilhadas entre os usuários do sistema.

Segundo Wal (2005), a folksonomia – resultado do vocabulário coletivo – está dividida em dois tipos: larga (*broad folksonomy*) e estreita (*narrow folksonomy*). O autor argumenta que na folksonomia larga muitas pessoas categorizam o mesmo objeto com suas próprias *tags*, como exemplo desse tipo há o sistema *Delicious* (um *website* gerenciador de *bookmarks* social que permite a categorização de conteúdos que não são do próprio usuário). Por outro lado, temos a folksonomia estreita que pode ser compreendida como uma única pessoa ou um pequeno grupo de usuários categorizando um objeto, como exemplo desse tipo há o sistema *Flickr* (um *website* de armazenamento e compartilhamento de imagens fotográficas no qual os usuários utilizam rótulos para categorizar as imagens pertencentes a eles mesmos). Portanto, para o mecanismo da folksonomia estreita apenas o próprio usuário categoriza seus conteúdos, não sendo permitido categorizar qualquer outro objeto que não lhe pertença, enquanto que na folksonomia larga os usuários podem categorizar conteúdos que são disponibilizados por outros usuários (RIDDLE, 2005) (RUSSEL, 2005) (SHEN *et al.*, 2005) (STURTZ, 2004).

A Figura 3 ilustra um esquema que representa um sistema baseado em folksonomia. Nessa imagem, os três usuários (Usuário 1, Usuário 2 e Usuário 3) utilizam um conjunto de *tags* (*Tag* 1, *Tag* 2, ..., *Tag* N) para categorizar os objetos (Objeto 1, Objeto 2 e Objeto 3). Também é possível notar na ilustração as interligações entre suas personomias, pois no momento em que dois usuários utilizam uma mesma *tag* para referenciar um determinado objeto, eles estão construindo implicitamente uma relação entre si.

No sistema *Delicious* (YAHOO, 2003) muitas pessoas podem categorizar os mesmos objetos com os termos de seu próprio vocabulário e compartilhar com outros usuários. Uma pesquisa realizada pelo *KDnuggets* (KDNUGGETS, 2007) mostra que o *Delicious* aparece como o gerenciador de *bookmarks* mais utilizado na *Web* com 25% de uso pelos usuários, seguido pelo *Digg*[15] e *Slashdot*[16] (12,9% e 9,5%, respectivamente). A principal distinção entre o sistema *Delicious* e os demais gerenciadores de *bookmarks* está no foco direcionado aos usuários, permitindo-lhes que adicionem *tags* para a organização dos *bookmarks* em um âmbito social, caracterizando a folksonomia. Assim, os sistemas baseados em folksonomia podem refletir em tempo real qualquer mudança de característica do ponto de vista social para

[15] Disponível em: http://digg.com/
[16] Disponível em: http://slashdot.org/

um recurso *Web* (MATHES, 2004) (HALPIN *et al.*, 2006) (SHEN *et al.*, 2005), o que não acontece em nenhum outro tipo de sistema.

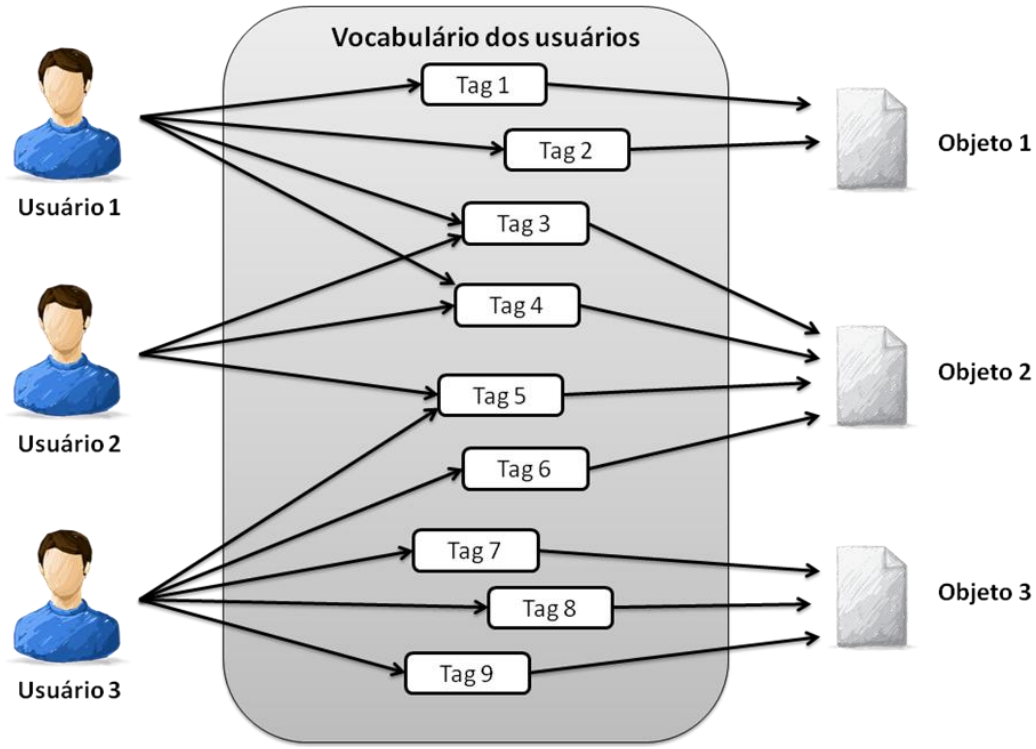

Figura 3: Modelo de um sistema baseado em folksonomia.

Utilizar o aspecto social para o processo de recuperação de informação pode trazer resultados ainda mais relevantes, pois os sistemas baseados em folksonomia permitem que sejam recuperadas informações as quais em um sistema tradicional nunca seriam apresentadas aos usuários (RUSSEL, 2005). Desse modo, os sistemas que adotam esse mecanismo possibilitam que os usuários construam um vocabulário representativo dos objetos, definindo, assim, grande parte do conhecimento e/ou comportamento dos objetos. É por isso que pesquisadores como (HALPIN *et al.*, 2007) (WU *et al.*, 2006) (SHEPTSEN *et al.*, 2008) apontam a folksonomia como uma técnica que é adotada pelos sistemas por ser conveniente e intuitiva em sua forma de uso, visto que são os próprios usuários que a criam. Wu (2006) reforça essa afirmação e vai além ao mencionar que a folksonomia pode ser a chave para o desenvolvimento da *Web* semântica.

2.3. Problemas Enfrentados na Recuperação da Informação em Sistemas Baseados em *Tagging*

Nos dias atuais, enfrentamos alguns problemas típicos nos sistemas de busca convencionais – problemas os quais são independentes dos tipos de recursos que estão sendo aplicados – como, o efeito da sinonímia[17] e da polissemia[18]. A presença de sinonímia pode omitir objetos relevantes dos resultados ao usuário, uma vez que as palavras-chave utilizadas na *string* de busca podem não ser as mesmas presentes nos documentos. Já a presença de polissemia pode causar confusão nas pesquisas no instante de selecionar os conteúdos resultantes, uma vez que as palavras-chave utilizadas na *string* de busca podem estar presentes no documento, entretanto, apresentando significados distintos do esperado. Esses fatos refletem diretamente nos resultados oferecidos pelos sistemas de busca, que tendem a retornar recursos irrelevantes, exigindo que o usuário faça uma filtragem manual dos resultados que realmente traduzam seu interesse. O esforço empregado na filtragem pode ser alto de acordo com a precisão da classificação dos resultados oferecidos pelo sistema de busca (HARDTKE, 2009) (BRUSILOVSKY, 2009) (MICARELLI, 2007) (PANT, 2003).

O fato das categorizações utilizarem *tags* proporciona certa liberdade ao usuário, pois ele escolhe as *tags* de forma livre. Entretanto, essa liberdade traz algumas limitações quando é necessário recuperar um objeto. No processo de *tagging*, o usuário pode aplicar uma única *tag* para vários objetos ou atribuir várias *tags* ao mesmo objeto. Em consequência disso, os resultados de uma pesquisa podem conduzir a itens que não seriam de interesse do usuário, devido a ruídos gerados pelos problemas de polissemia e ambiguidade decorrentes da total liberdade nas categorizações (MATHES, 2004). Além disso, a liberdade da escolha das *tags* acarreta em outros problemas como: erros ortográficos, uso de singulares e plurais para identificar o mesmo tipo de objeto (ex: *car*, *cars*). Também é possível encontrar palavras em formatos diferentes, siglas, gêneros, números, abreviações vagas e termos cujos significados não estão diretamente associados ao recurso.

Essa natureza não controlada dos sistemas baseados em *tagging* é caótica e sofre com os problemas da falta de organização e ambiguidade no processo de gerenciamento do vocabulário dos usuários. Wu (2006) afirma que a ausência de qualquer controle diante das categorizações realizadas pelos usuários implica em uma tendência na criação de informações inúteis do ponto de vista coletivo. Por isso, a falta de critério na escolha e utilização das *tags*

[17] São palavras diferentes que apresentam o mesmo significado.
[18] A polissemia é um fenômeno em que uma palavra possui múltiplos significados.

traz desvantagens. Por exemplo, a *tag* "rosa" pode ser aplicada em diferentes contextos. Ela pode ser usada tanto como etiqueta em uma foto de uma rosa – a flor; em um texto que se refere à pantera cor-de-rosa – personagem de desenho animado; ou ainda para uma reportagem sobre a tendência e elegância em usar camisas rosa. O uso de vocabulário controlado é adotado em vários sistemas justamente para tentar suprir problemas dessa natureza. De acordo com Shen (2005), Sturtz (2004) e Mathes (2004), os sistemas deveriam desenvolver recursos que permitam usufruir melhor dos benefícios oferecidos pela categorização, pois, como a folksonomia é originada a partir das personomias compartilhadas, ela é considerada uma especialização do processo de *tagging* e, consequentemente, herda os mesmos problemas.

2.4. Recomendação de *Tags*

Sistemas de recomendação são tipos especiais de sistemas que atuam na filtragem de informação que seja de interesse do usuário (METEREN, 2000), cujo objetivo é reduzir a sobrecarga de informação que fica exposta ao usuário. Os sistemas de recomendação auxiliam na indicação de recursos baseando-se em características individuais ou intermediadas por outras pessoas. Para que uma recomendação possa sugerir resultados úteis, torna-se necessário realizar uma avaliação comportamental do usuário no sistema. O comportamento de um usuário pode ser derivado de várias formas, tais como: informações de *logs*, visualização dos objetos, categorizações, dentre outros. Essas informações servem para que sejam encontrados um conjunto de usuários, termos ou recursos que representem interesses similares.

Muitos sistemas baseados em *tagging* facilitam alguns processos para o usuário, por exemplo, fornecendo recomendação de *tags* no momento da categorização. Essa tarefa pode aliviar os usuários do esforço de pensar nas melhores palavras-chave e associar diferentes facetas da informação que o usuário possa ter diante de um recurso. A recomendação de *tags* apresenta uma série de benefícios que podem servir para vários propósitos, pois além de reduzir problemas de polissemia, sinonímia, dados individuais dos usuários, erros de ortografia, ela acelera o processo de categorização efetuado pelo usuário. Em diversos sistemas baseados em *tagging*, as pessoas frequentemente usam *tags* como, "*web*2.0", "*web*2", "*web*-2.0", "*web_2.0*" e "*web*20" – todas essas formas tem breves variações léxicas, sendo usadas para representar o mesmo significado. Na maioria dessas palavras, essas variações são desnecessárias e uma recomendação de *tags* auxiliaria o usuário a selecionar o

termo mais adequado e até mesmo o termo que permitiria recuperar a informação de forma mais eficaz, reduzindo esses tipos de problemas. Conforme afirma Jäschke (2007), quando o sistema sugere um termo as chances de um usuário categorizar um recurso tornam-se ainda maiores. Além disso, o fato de sugerir relembra o usuário sobre o que se trata o recurso, facilitando, assim, a consolidação de um vocabulário entre os usuários.

Symeonidis (2008) argumenta que a preocupação fundamental com um sistema de recomendação baseado em *tagging* é tornar a folksonomia relativamente "estável" a partir de seu uso e seu tempo de vida. Entende-se por "estável" o momento em que os usuários tenham desenvolvido um senso comum dos recursos categorizados, isto é, gerado uma quantidade considerável de informações para esses recursos, oferecendo, assim, um equilíbrio no vocabulário utilizado nos objetos do sistema. Essa estabilidade do vocabulário tem duas consequências antagônicas. Se por um lado ela melhora a recuperação de informação, pois o vocabulário passar a ser bem conhecido, por outro lado, um sistema pode empregar a recomendação de *tags* baseando-se apenas nos dados da folksonomia e a eficácia da sua recomendação pode ser prejudicada, pois se um objeto ainda não foi categorizado o usuário ficará sem a recomendação de *tags* do sistema (problema esse denominado de "*cold-start*") ou, até mesmo, se o objeto tiver sido categorizado poucas vezes, poderá haver uma baixa precisão na recomendação, visto que o objeto foi avaliado por poucos usuários.

Smith (2008) apresenta uma divisão de três categorias para a recomendação de *tags*: (i) *tags* previamente utilizadas – termos que estão presentes na personomia do usuário; (ii) *tags* populares – termos utilizados frequentemente por outros usuários (folksonomia); e (iii) *tags* recomendadas – termos que o usuário deveria considerar populares, *tags* recentemente utilizadas ou *tags* geradas por meio de uma recomendação. Além disso, existem outros meios de recomendar *tags*, por exemplo, sugerindo os termos mais frequentes que estão presentes no texto de um documento *Web*; ou também, usando as estruturas formais que as ontologias oferecem. A utilização dessa última técnica pode ser interessante para a representação dos recursos que são utilizados no processo de recomendação, uma vez que a partir de uma ontologia é possível saber com mais detalhes as relações de uma *tag* com as outras, visto que, em uma categorização a única relação que as *tags* possuem entre si é a de co-ocorrência[19].

[19] Quando duas *tags* são utilizadas em conjunto em uma mesma categorização.

2.5. Ontologias

Ontologia é uma palavra originária na filosofia que é usada para representar uma visão geral do mundo e a organização dos seres. Uma ontologia permite descrever e organizar a informação a partir de um grupo de palavras, tornando-as um vocabulário livre de ambiguidade e passível de formalismo. Na tentativa de auxiliar o processamento e a representação da informação na *Web*, as ontologias vêm sendo empregadas como um mecanismo que permite oferecer consistência na sua representação em um ambiente aberto, heterogêneo e ubíquo como a Internet.

Na Ciência da Computação, uma das definições mais comuns encontradas na literatura é feita por Gruber (1993). O autor define ontologia como "uma especificação formal explícita de uma conceitualização compartilhada". Há outras definições importantes, como a descrita por Chandrasekaran *et al.* (1999), os quais afirmam que uma ontologia tem uma visão de um artefato designado a usos específicos que permite capturar o domínio de conhecimento de uma forma genérica, fornecendo um entendimento sobre o que está sendo explorado; e a descrita por Mizoguchi (1993), que define uma ontologia como sendo um modelo com vocabulários/conceitos para a construção de sistemas artificiais. Qualquer uma dessas citações pode ser direcionada às áreas de gestão do conhecimento e inteligência artificial com o intuito de fornecer uma representação de algum domínio de conhecimento.

Neste trabalho é adotado como base a definição descrita por Gruber (1993). Entendemos que uma ontologia identifica conceitos diante de uma área de conhecimento em específico, tornando-a formal de tal maneira que os conceitos se tornam explícitos a partir de suas relações, a ponto de que seja passível de ser processada ou interpretada por máquinas. Além disso, as ontologias não são apenas para as máquinas, pois possuem relações entre a percepção humana sobre a realidade das coisas e os modelos computacionais (HEPP, 2007). O autor afirma ainda que é crucial que os seres humanos entendam a especificação da ontologia. Atualmente, aplicações do ramo de recuperação de informação, Processamento de Linguagem Natural (PLN), gerenciamento do conhecimento, *Web* semântica, *e-commerce*, integração de informações entre sistemas, dentre outras, utilizam ontologias como parte de uma abordagem de sistemas de informação inteligentes (FENSEL, 2004).

Assim, uma ontologia é um modelo de dados que representa um conjunto de conceitos dentro de um determinado domínio e seus respectivos relacionamentos (ECHARTE *et al.*, 2004). Os conceitos denotam os significados de uma palavra e as relações representam o tipo de conexão que os conceitos possuem entre si. Na composição de uma ontologia, também é

comum encontrar um conjunto de axiomas e instâncias, que são respectivamente regras definindo sentenças verdadeiras diante de um domínio e os próprios dados representados pela ontologia. Ontologias fornecem a estruturação de um domínio, bem como um vocabulário em comum da área, possibilitando relações com diferentes níveis de formalidade.

Ontologias podem possuir relações taxonômicas e não taxonômicas, e possuir restrições quanto às propriedades aplicadas às entidades em um domínio de conhecimento específico. Um exemplo de uma ontologia pode ser visualizado na Figura 4. Na ilustração podem ser observados os vários conceitos e os relacionamentos entre eles, os quais devem ser formalizados para se obter uma ontologia completa. Dessa forma, é possível que uma ontologia modele rigorosamente um domínio de conhecimento, estabelecendo um conjunto de conceitos e relações (ECHARTE *et al.*, 2007).

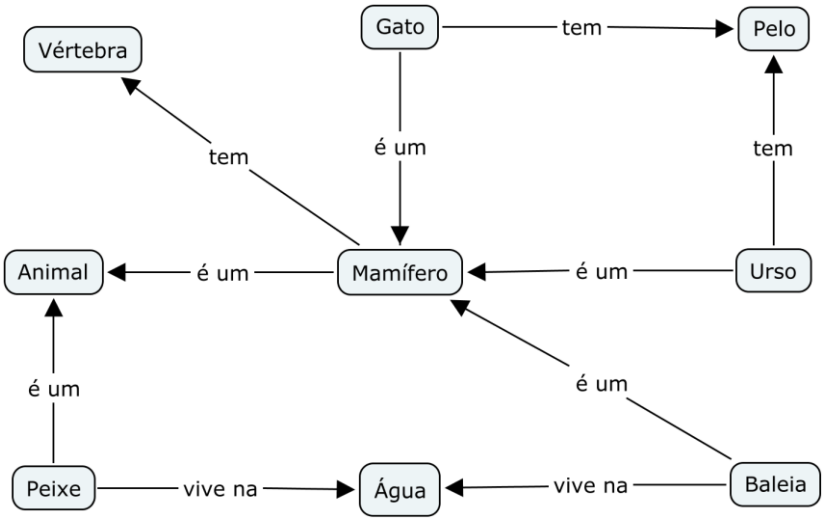

Figura 4: Um exemplo de uma ontologia com alguns conceitos e relações.

A partir do uso de ontologias em sistemas, torna-se evidente uma de suas vantagens, a simplicidade do uso do vocabulário para representar o domínio de conhecimento. Esses conceitos presentes no vocabulário podem ser relacionados por relações do tipo: *is-a*, *has-a*, etc. O vocabulário utilizado possui como sua sustentação os conceitos, evitando que sejam interpretados de maneira ambígua pelas aplicações semânticas (BREITMAN, 2005). Além do mais, as ontologias podem ser estruturas eficientes para auxiliar na navegação do usuário, podendo fornecer uma exibição diferenciada dos conceitos na recuperação de informação.

2.6. O Uso de Ontologias em Sistemas baseados em *Tagging*

De modo geral, pode-se dizer que parte dos problemas na recuperação de informação em sistemas baseados em *tagging* é devido aos usuários precisarem "lembrar" os termos que foram associados ao recurso no momento da categorização, tarefa para a qual os seres humanos possuem dificuldades cognitivas (ANDERSON, 1995). O autor ainda afirma que após as pessoas processarem uma mensagem linguística, elas lembram apenas do sentido/contexto dos fatos e não exatamente das palavras utilizadas. Isso afeta a recuperação de informação, sendo que uma possibilidade para ajudar na solução desse problema é por meio do uso de ontologias.

Utilizando ontologias no processo de recomendação de *tags* é possível melhorar a organização da informação e sua posterior recuperação, como é discutido em Basso *et al*. (2009). Os autores afirmam ainda que as relações semânticas presentes em uma ontologia podem tornar o processo de recuperação de informação menos custoso para o usuário, por disponibilizar outros meios para acessar a mesma informação. Dessa forma, a partir de seu uso é possível gerar um acesso diferenciado na visualização da informação. Por exemplo, o usuário pode generalizar/especializar uma *tag* até o conceito que ele deseja. Desse modo, no momento de recuperar um recurso categorizado é possível pesquisar por *tags* "específicas" e/ou "genéricas" que representam um conceito em particular ou, até mesmo, utilizar as relações semânticas entre os conceitos para encontrar um conceito sinônimo, ou derivado, ao informado. Por exemplo, uma pesquisa realizada pela palavra "*airplane*", pode retornar recursos categorizados com as *tags* "*plane*" e "*aero plane*", conforme exemplo da Figura 5. Assim, é possível evitar que o usuário realize várias pesquisas para encontrar o que deseja.

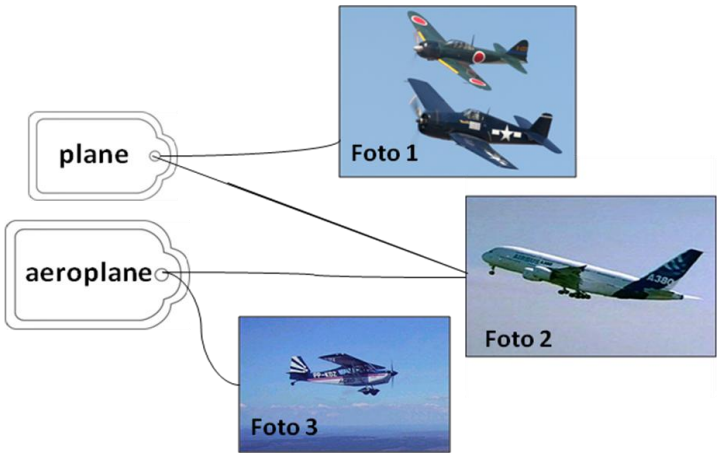

Figura 5. Benefícios das relações semânticas na recuperação da informação.

Com o uso de ontologias nos sistemas baseado em *tagging* é possível que uma palavra-chave seja representada por um conceito, deixando de ser interpretada apenas pela sua forma léxica. Isso significa que as *tags* associadas a um objeto terão significados formais fornecidos por uma fonte de informação. Dessa maneira, trabalhos que usam uma ontologia para a recomendação de *tags* empregam um *mashup*[20] com alguma ontologia pré-existente para gerar as relações semânticas entre os conceitos. Para realizar essa tarefa há opções como, a *WordNet*, a *ConceptNet*, a *DBPedia*, dentre outras. Essas são bases de dados que possibilitam acessar conceitos e suas relações e, a partir disso, auxiliar na geração das ontologias.

As principais vantagens em se usar ontologias em um sistema baseado em *tagging* são: (i) a partir da personomia é possível inferir por meio da navegação na ontologia as principais áreas de interesse do usuário; (ii) a possibilidade de melhorar a recuperação de informação usando relações semânticas entre as *tags* utilizadas na categorização, permitindo pesquisas a partir de conceitos, em vez de apenas pela sua forma léxica; e (iii) a possibilidade de se usar a ontologia para realizar a recomendação de *tags*. Essa última é o foco deste trabalho. Assim, a próxima seção mostra a importância do uso de uma abordagem de recomendação de *tags* baseada em ontologia.

2.6.1. A Importância de um Sentido Associado à *Tag* para a Recuperação de Informação

Para demonstrar o problema das *tags* sem semântica em um sistema baseado em *tagging*, considere um usuário que está navegando em vários *websites* à procura de um lugar para passar suas férias. Esse usuário encontra um *website* interessante que fala sobre *Jakarta* – a capital e maior cidade da Indonésia, situada na ilha de *Java* – e, consequentemente, salva a respectiva página *Web* em seu sistema baseado em *tagging* favorito com as seguintes *tags*: *Jarkarta*, *Indonesia* e *Java*. Em outro momento, esse usuário acessa *websites* relacionados à programação *Java* e encontra uma página relevante sobre a plataforma *J2EE* (*Java* 2 *Enterprise Edition*) e decide salvar com as *tags*: *Application*, *J2EE* e *Java*. Após alguns dias, o usuário decide recuperar a página cujo assunto está relacionado às suas férias na cidade de *Jakarta* e no momento de digitar a palavra-chave no campo de pesquisa, informa o termo *Java*, pois se lembrou apenas da ilha de *Java* e não do nome da cidade. Assim, como não há

[20] É uma combinação de N fontes de dados incorporadas de forma personalizada em apenas uma fonte de informação.

semântica relacionada às *tags* utilizadas nas categorizações, o termo informado na pesquisa se tornou polissêmico e, como consequência, o resultado da pesquisa exibirá todos os recursos cujas categorizações possuem a *tag Java* associada – para muitos usuários isso pode significar de dezenas a centenas de recursos. Dessa maneira, a página relacionada às férias na ilha de *Java* poderá estar no meio de muitas outras páginas relacionadas à linguagem de programação *Java*, dificultando o acesso ao conteúdo desejado. Por essa razão, o usuário terá um esforço cognitivo maior para discernir quais os recursos do resultado são relevantes.

Caso esse sistema usasse uma abordagem de recomendação de *tags* semânticas, e tivesse uma interface que permitisse ao usuário selecionar o significado da *tag* (conceito), o usuário poderia não ter sido exposto a uma sobrecarga de informação, pois a pesquisa realizada poderia retornar apenas os recursos relacionados a ilha de *Java*. Uma solução para esse problema seria utilizar uma recomendação de *tags* semânticas, juntamente com uma interface como a ilustrada na Figura 6. Essa interface ilustra os conceitos da ontologia conforme o usuário digita a *tag* na caixa de texto, exibindo os conceitos, juntamente com sua descrição semântica. Dessa forma, na categorização de um *website* o usuário poderia informar a *tag* e selecionar um conceito à respectiva *tag*. Para o caso citado acima, seria possível digitar a *tag* "*java*" e selecionado o conceito *island*. Assim, os *websites* relacionados a "*object-oriented programming language*" não seriam exibidos juntamente com os *websites* relacionados a ilha de Java, resolvendo o problema das *tags* homônimas utilizadas por esse usuário. Além disso, caso necessário, o usuário poderia ignorar a lista sugerida e informar outras *tags* cujos significados não estão presentes na ontologia, ou seja, *tags* sem um conceito relacionado.

Figura 6. Interface alternativa para a categorização/busca de um recurso Web em um sistema baseado em tagging.

Para criar uma abordagem de recomendação de *tags* baseada em ontologia é necessário um modelo ontológico que permita modelar o domínio de conhecimento que o sistema abrange. Atualmente, não há um modelo ontológico padrão para representar e gerar a

estrutura dos termos presentes nos sistemas baseados em *tagging*. Por isso, cada sistema adota um modelo que melhor se aplica ao seu respectivo domínio da aplicação como, as abordagens de Knerr (2006), Echarte *et al*. (2007) e Basso *et al*. (2009). As duas primeiras definem modelos ontológicos para o processo de *tagging*, enquanto a última estende os modelos de Knerr e Echarte *et al*. para atribuir significado às *tags* e relacioná-las umas às outras usando relações semânticas. Essa abordagem, denominada *TagOntologyManager* (*TOM*), é o modelo ontológico utilizado neste trabalho.

2.7. O *TagOntologyManager* e a *WordNet*

O *TOM* (BASSO *et al*., 2009) define um modelo ontológico que permite emergir[21] uma ontologia a partir dos dados de *tagging* contidos na personomia de um usuário. Conforme Basso *et al*., esse modelo ontológico expressa o conhecimento de como o processo de *tagging* deveria ser modelado com atributos e relações semânticas entre as *tags* para ajudar na recuperação da informação de um sistema baseado em *tagging*. As principais vantagens na utilização desse modelo ontológico é que ele possibilita pesquisas avançadas, recomendação de termos e visualização de dados, tudo por meio de conceitos. Esse modelo ontológico permite conceder significado às *tags* utilizadas nas categorizações e relacioná-las com outras *tags* usando relações semânticas. Essa abordagem é baseada na emergência de uma ontologia a partir da personomia do usuário, embora também possibilite emergir uma ontologia a partir de outros tipos de fontes de informação baseadas em termos como, por exemplo, o conteúdo de uma página *Web* e os termos de uma folksonomia de um sistema baseado em *tagging*.

Basicamente, o processo do *TOM* é composto por três passos. Inicialmente, realiza-se o processamento léxico das *tags*, no qual elas são normalizadas para que possam ser reconhecidas como conceitos semânticos em uma fonte de informação semântica e, além disso, são criadas as relações não-semânticas da ontologia; na sequência, é realizada a atribuição de sentido das *tags*. Nesse passo, é feita uma escolha do sentido das *tags* consideradas ambíguas (*i.e*. as *tags* que retornem mais de um sentido obtido na fonte de informação semântica); e, por fim, é feita a criação de estruturação semântica, no qual são obtidas as relações semânticas para associar *tags* e termos auxiliares.

Na Figura 7 pode ser vista uma ilustração da estrutura gerada (*Output*) pelo *TOM* a partir das *tags* de uma categorização (*Input*) com as palavras-chave "*java*",

[21] O termo emergir (ou emergência) citado neste trabalho, também é conhecido como instanciação ou população de ontologias.

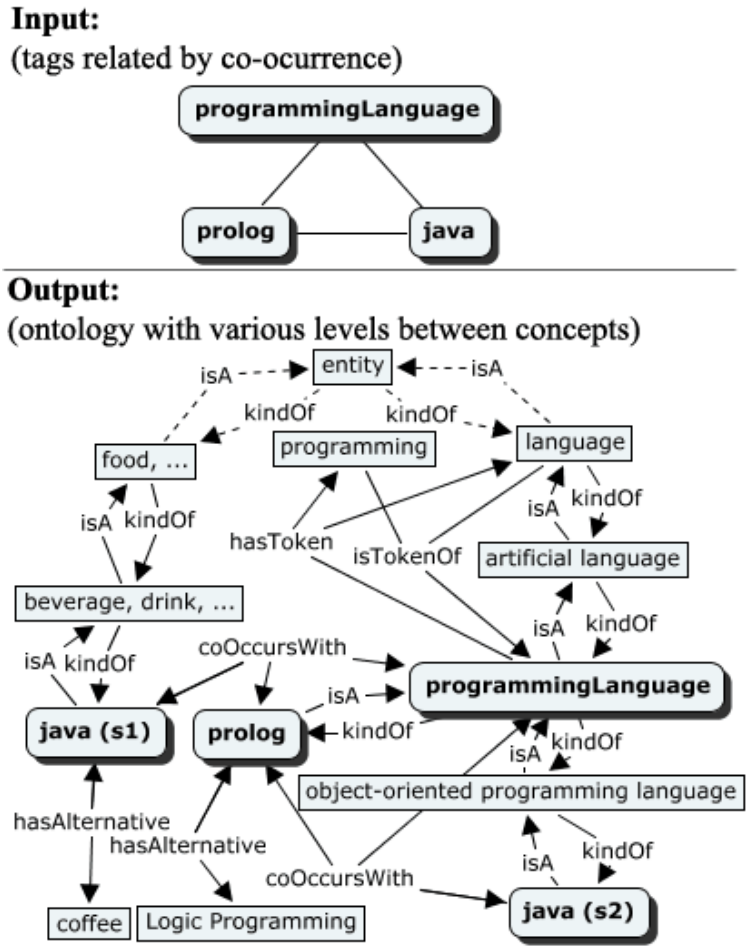

Figura 7: Um exemplo da ontologia gerada pelo TOM a partir da entrada de dados onde a única relação é a co-ocorrência (BASSO et al., 2009).

"*programmingLanguage*", e "*prolog*". Ao emergir a ontologia pelo *TOM*, pode-se obter uma lista com vários conceitos que não estavam presentes nas *tags* usadas na categorização e várias relações semânticas que podem ajudar na recuperação de informação. Além disso, Basso *et al.* (2009) acreditam que seria melhor gerar um modelo ontológico e sugerir conceitos semânticos no momento da categorização, pois isso auxiliaria os usuários, reduzindo o esforço cognitivo para recuperar as informações desejadas. Para isso, basta o sistema salvar o termo selecionado pelo usuário juntamente com o seu respectivo sentido. Isso evitaria uma posterior desambiguação dos termos presentes na personomia, pois os termos selecionados já estariam associados a um conceito da ontologia. Essa sugestão é que nos leva a propor a utilização da ontologia para a tarefa de recomendação de *tags*.

Um benefício desse modelo ontológico é que ele possui relações semânticas, como, *hasAlternativeTag*, *isPartOf*, *hasPart*, *isA*, *kindOf*, e também relações não semânticas, como, *hasToken*, *isTokenOf*, *coOccurswith*. Dessa forma, por meio desse modelo ontológico é possível fazer uma busca mais abrangente, retornando objetos categorizados com *tags* mais

específicas. Por exemplo, uma busca pelo termo "*vehicle*" poderia retornar objetos categorizados com a *tag* "*airplane*", "*bike*", "*car*", etc. Por outro lado, as relações não-semânticas podem auxiliar na obtenção do contexto e das relações semânticas entre as *tags*. Logo, a presença dessas relações no modelo ontológico facilita o propósito da recomendação de *tags* semânticas, pois tornar-se possível generalizar/especializar uma *tag* até o ponto em que o usuário deseja por meio da hierarquia presente no modelo. Para adquirir essas relações formais entre os termos, o *TOM* realiza um *mashup* com a *WordNet* (MILLER, 1995), visto que essa possibilita obter informações importantes sobre os conceitos e suas relações hierárquicas. Maiores detalhes sobre o processo de emergência da ontologia e a desambiguação de sentido das *tags* podem ser vistos em Basso *et al*. (2009).

2.7.1. A WordNet

Para que as relações de uma ontologia sejam obtidas, uma das opções seria acoplar a aplicação a uma base de dados externa. Atualmente, uma das alternativas para criar essa ligação é utilizar a *WordNet*. Esse é um banco de dados léxico eletrônico que surgiu em 1984, baseado em um experimento linguístico na Universidade de Princeton (MILLER, 1995), cuja estrutura surgiu de teorias neurolinguísticas da memória léxica humana (WORDNET, 2006) (FELLBAUM, 1998). Inicialmente, a *WordNet* era utilizada apenas para a língua inglesa, no entanto, há esforços de adaptação e criação de extensões para outros idiomas, como o português (DIAS-DA-SILVA, 2008). Embora interrompido, outro projeto interessante era a *EuroWordNet*, pois visava construir um banco de dados léxico para vários idiomas por meio de suas relações correspondentes baseando-se na estrutura da *WordNet* (VOSSEN, 1998). A *WordNet* se difere de outras bases de dados léxicas, pois os termos se dispõem em classes gramaticais dentro de um grupo de sinônimos cognitivos denominados de *synsets* (abreviação do termo em inglês "*synonym sets*"). Cada *synset* expressa um conceito distinto, o qual é associado de acordo com seu significado semântico, formando uma rede de conceitos (WORDNET, 2006). Suas relações e estruturas são construídas manualmente por especialistas, justamente para possibilitar o tratamento computacional.

Quando ocorre a sinonímia, uma palavra pode ser efetivamente substituída por outra sem efeito colateral (JURAFSKY *et al*., 2000). Nesse caso, quando duas palavras são sinônimas elas estão presentes no mesmo *synset*. Dentre as instâncias presentes na base de dados, a relação de sinonímia é criada pela regra: "um conceito C1 é sinônimo do conceito C2 apenas se ambos pertencerem ao mesmo *synset*". Logo, uma representação baseada na

WordNet nada mais é do que conceitos que possuem relações formais entre si, formando uma relação semântico-conceitual. Assim, um *synset* pode ser interseccionado por outros *synsets* por meio de uma relação hierárquica, possibilitando a exploração de cada relação.

Os *synsets* estão associados entre si de acordo com as relações semânticas de hiperonímia[22], hiponímia[23], holonímia[24], meronímia[25], dentre outros. Além disso, os relacionamentos entre os conceitos e suas relações formam um grafo que proporcionam para cada conceito uma definição semântica. Os *synsets* estão divididos em quatro tipos de classes gramaticais, sendo elas: substantivos, verbos, adjetivos e advérbios. Os dados da Tabela 2 exibem as classes gramaticais com as suas respectivas quantidades na base de dados atual (*WORDNET*, 2010) e as principais relações semânticas entre os *synsets* (MILLER, 1995).

Tabela 2. Características das classes gramaticais da WordNet.

Classe Gramatical	**Palavras**	***Synsets***	**Principais relações semânticas entre os *synsets***
Substantivos	117.798	82.115	Sinonímia, antonímia, hiperonímia, hiponímia, holonímia, meronímia
Verbos	11.529	13.767	Sinonímia, antonímia
Adjetivos	21.479	18.156	Sinonímia, antonímia
Advérbios	4.481	3.621	Sinonímia, antonímia
Total	**155.287**	**117.659**	

A Figura 8 ilustra as relações entre os conceitos que são provenientes de uma pesquisa realizada na *WordNet* a partir da palavra "*car*". Embora o resultado da pesquisa pela palavra "*car*" tenha retornado cinco *synsets*, a ilustração exibe apenas dois *synsets* (por questões de estética). Nessa ilustração pode ser observado que para o "*synset 1"* existem vários conceitos de especialização (*hyponym*), sua superclasse (*hypernym*) e alguns conceitos que fazem parte do conceito "*car*" (*meronym*). Além disso, o "*synset 2"* exibe um conceito que foi identificado como holonímia (*holonym*).

[22] É a relação que conecta a superclasse a uma subclasse específica. Por exemplo, "*automotive vehicle*" é uma superclasse de "*car*".
[23] É o oposto de hiperonímia. É a relação que conecta a classe subordinada à sua superclasse. Por exemplo, "*car*" é uma subclasse de "*automotive vehicle*".
[24] É o conceito que representa o conjunto de tudo. Por exemplo, "*car*" é o todo (ou holônimo) de "*car window*", "*car door*" e outros conceitos.
[25] É o conceito que representa parte de um todo. Por exemplo, "car window" é parte de "*car*".

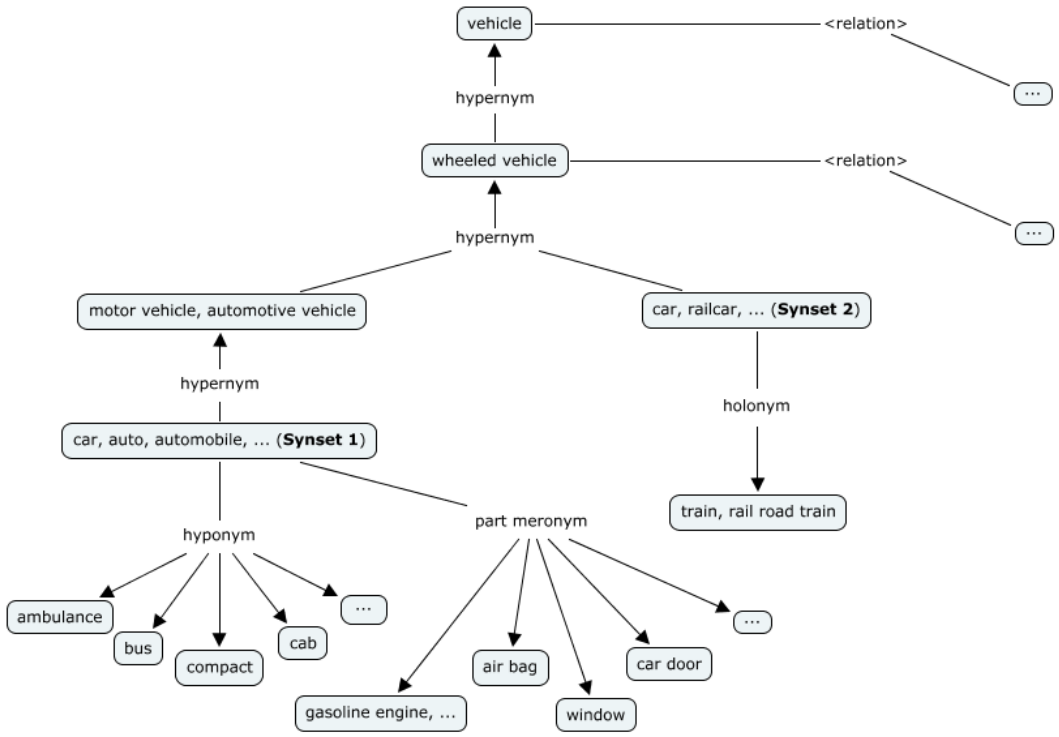

Figura 8. Algumas relações que podem ser obtidas a partir da WordNet, provenientes de uma busca pela palavra "car".

2.8. Algumas Propostas para Recomendação de *Tags*

Atualmente, há alguns sistemas disponíveis *online* que realizam recomendação de *tags*, por exemplo, *Delicious*, *Bibsonomy*[26], *Fuzzzy*[27], *ZigTag*[28], dentre outros. Como exemplo de um sistema baseado em *tagging* que realiza recomendação de *tags* há o *Delicious* – atualmente é o sistema mais popular disponível que possibilita o gerenciamento e organização de *bookmarks*. Esse sistema, como pode ser observado na Figura 9, utiliza duas formas de recomendação de *tags*. A primeira com palavras-chave de outros usuários para realizar as recomendações (folksonomia), sendo denominada de "*Popular Tags*". A segunda pela intersecção entre as *tags* da folksonomia e as *tags* da personomia do usuário, sendo denominada de "*Recommended Tags*". Ambas as recomendações nesse sistema são geradas a partir da frequência com que as *tags* são utilizadas pelos usuários.

O sistema *Delicious* utiliza uma característica interessante no seu processo de recomendação que é a utilização dos dados da folksonomia e da personomia, uma vez que as

[26] Gerenciador de *bookmarks* para páginas *Web* e publicações científicas. Disponível em: http://www.bibsonomy.org/
[27] Gerenciador de *bookmarks* para páginas *Web* e números de ISBN. Disponível em: http://www.fuzzzy.com/
[28] Um gerenciador de *bookmarks* social para recursos *Web*. Disponível em: http://www.zigtag.com/

informações da folksonomia denotam um panorama social sobre o que os usuários estão pensando sobre uma respectiva *URL* e, a personomia fornece um direcionamento às *tags* mais comuns do próprio usuário. Assim, embora seja interessante utilizar a folksonomia como foco principal em uma recomendação, se um *bookmark* não tiver sido categorizado ainda ou, até mesmo, tiver poucas categorizações, um usuário do sistema poderá ficar sem recomendação ou, possivelmente, poderá ter uma recomendação com baixo percentual de aceitação.

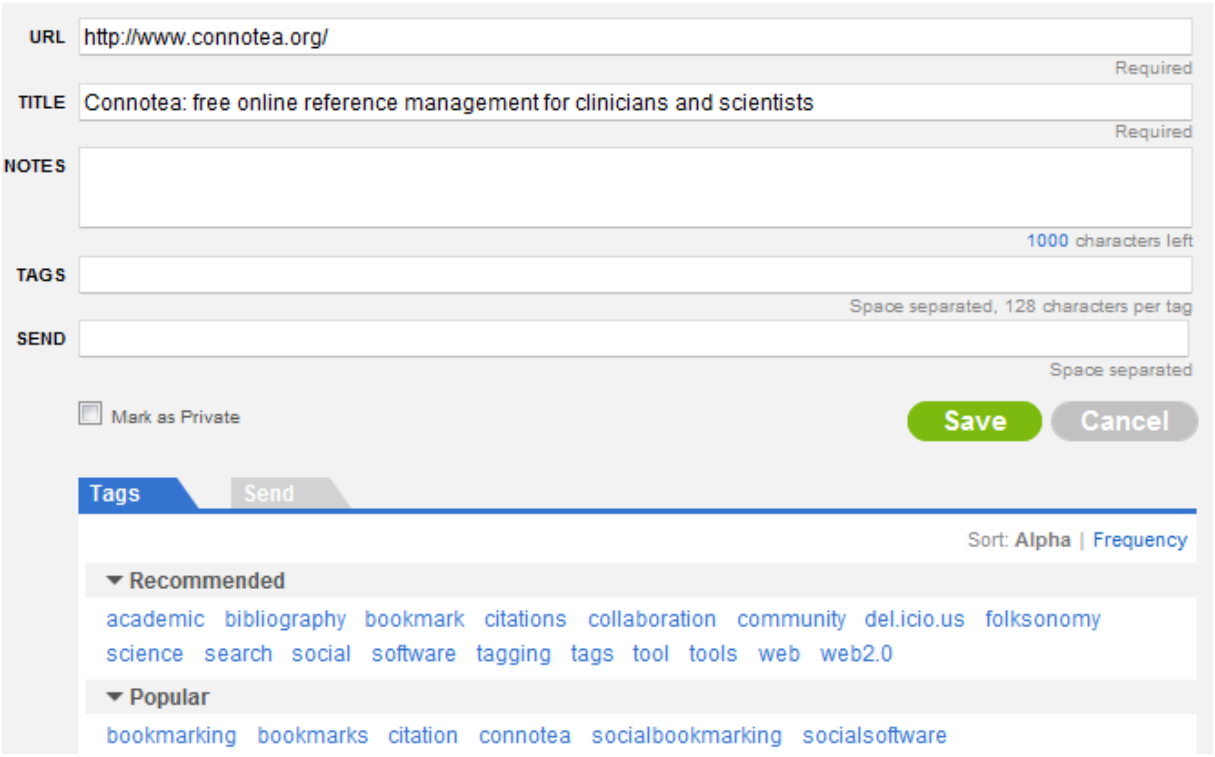

Figura 9. Recomendação de tags do Delicious (Adaptado do sistema Delicious).

Uma abordagem interessante que realiza recomendação de *tags* semânticas é a utilizada pela companhia *ZigTag Inc*. Essa abordagem permite que os usuários escolham as palavras-chave a partir de *tags* com significados específicos, permitindo identificar e associar a *tag* com um conceito formal e não apenas associá-las pela sua forma léxica. Para exemplificar, na Figura 10 temos a categorização de uma página relacionada ao animal *jaguar* a partir do sistema *ZigTag*, na qual informamos no campo de texto a palavra "*jaguar*". Pode ser observado que o sistema exibe uma lista de termos, permitindo que seja selecionada uma *tag* referenciando-a a um conceito explícito como, uma "marca de carro", um "animal", etc. Comparando com o sistema *Delicious*, o *ZigTag* possui como diferencial o emprego de *tags* como conceitos. Contudo, mesmo com a sua principal característica que é permitir associar *tags* semânticas em uma categorização, pode haver uma redução de informações relevantes no resultado das suas recomendações, pois ele não analisa o conteúdo das páginas *Web*. Dessa

forma, a recomendação pode conter *tags* que não estejam relacionadas ao conteúdo do recurso, como é o caso das *tags* "Jaguar (Alec Empire)" e "Jaguar (Atari Jaguar)" recomendadas por esse sistema, conforme ilustrado na Figura 10.

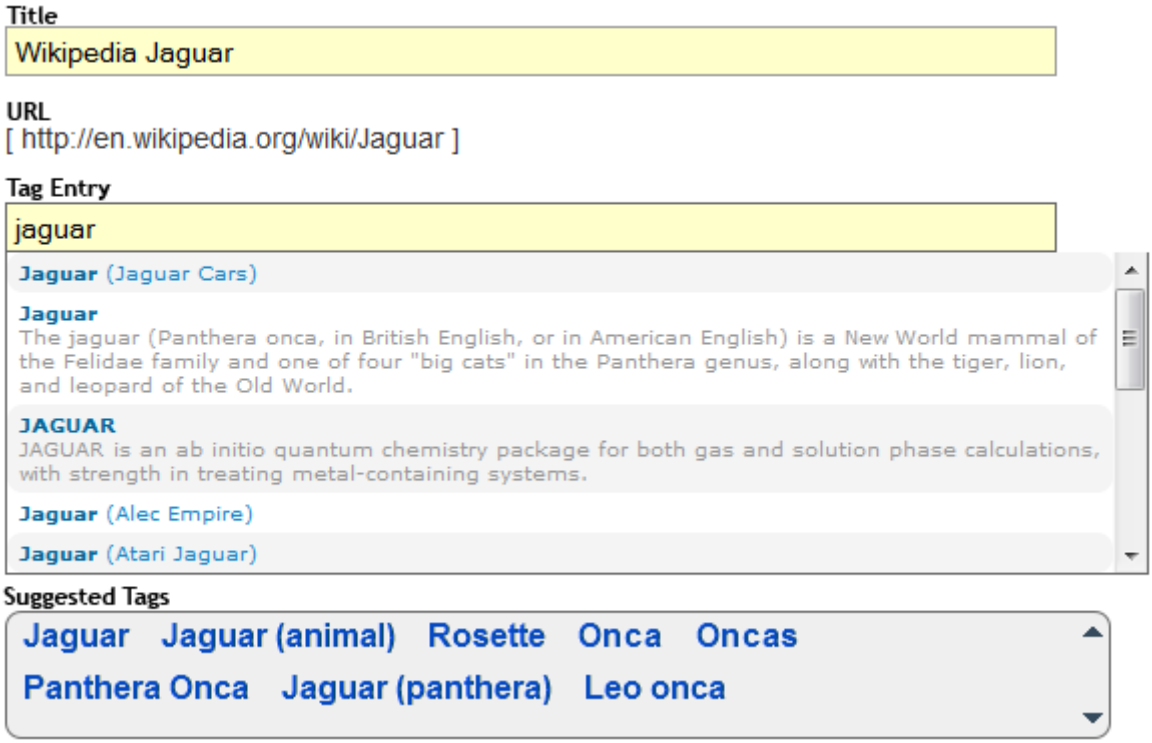

Figura 10: Ilustração de uma Recomendação de tags do ZigTag (Adaptado do sistema ZigTag).

Na tentativa de melhorar a recuperação de informação em sistemas baseados em *tagging*, algumas pesquisas desenvolvidas são exibidas na Tabela 3. Nessa tabela podem ser observadas as principais informações adotadas pelos autores para o desenvolvimento das respectivas recomendações de *tags*, como: tipos de recurso para o qual é efetuada a recomendação ("Tipo de Recurso"); se a recomendação é efetuada por meio da Filtragem Colaborativa[29] ("FC") e/ou Filtragem Baseada em Conteúdo[30] ("FBC"); e, qual(is) técnica(s) a abordagem de recomendação de *tags* utiliza para gerar a recomendação ("Fontes Utilizadas").

Como exemplo de abordagens semânticas, o trabalho de Adrian *et al.* (2007) realiza a recomendação de *tags* para documentos utilizando serviços *Web*. Nesse trabalho, foi criado um processo de normalização dos documentos para o formato RDF; em seguida, são extraídos

[29] Filtra informações baseando-se na correlação entre outras pessoas que apresentam características e interesses similares.
[30] Filtra informações baseando-se no conteúdo residente nos recursos disponíveis, permitindo identificar informações e características presentes nos documentos.

os tópicos dos documentos utilizando serviços *Web* e processada uma ontologia para cada tópico, obtendo objetos similares entre si; e, por fim, é gerado um "*schema"* de visualização das recomendações de *tags* a partir de uma ontologia. Pelo fato da tarefa de recomendação de *tags* semântica ser um processo recente, a principal contribuição desse trabalho é exatamente a criação do processo de recomendação utilizando serviços da *Web* 2.0. Uma vantagem desse trabalho é que ele realiza a tarefa de recomendação analisando o conteúdo do documento que será categorizado, sendo possível direcionar e focar em *tags* que estão presentes no próprio documento. Há também o trabalho de Fountopoulos (2007), cuja principal característica é constituir uma ontologia com os principais conceitos do objeto a ser categorizado, os quais são gerenciados pelos próprios usuários. Essa abordagem deriva os conceitos e relações semânticas por meio das informações contidas na folksonomia. Dessa forma, o sistema é apenas um mecanismo intermediador que garante a consistência do vocabulário empregado pelos usuários. O fato de recomendar semanticamente *tags* a partir da folksonomia é uma característica positiva em um processo de recomendação, porém, quando um objeto não estiver sido categorizado ainda, o usuário ficará sem recomendação.

Tabela 3. Algumas propostas para recomendação de tags semânticas.

Autores	**Tipo de Recurso**	**FC**	**FBC**	**Fontes Utilizadas**
Contag: A Semantic Tag Recommendation System (**Adrian *et al.*, 2007**)	Documentos e *Bookmarks*		X	Ontologia
RichTags: A Social Semantic Tagging System (**Fountopoulos, 2007**)	Documentos	X		Ontologia
SemKey: A Semantic Collaborative Tagging System (**Marchetti *et al.*, 2007**)	*Bookmarks*	X		*WordNet* / *Wikipedia*

Por fim, outro trabalho interessante é a abordagem proposta por Marchetti *et al*. (2007). Nesse trabalho é explorada uma junção entre as principais características da *WordNet* e da *Wikipedia*. A primeira possui um vocabulário rico e extremamente estruturado a partir de suas relações. A segunda possui uma rica quantidade de informações com suas respectivas descrições, contudo, possui um baixo acoplamento de relações entre os conceitos presentes na sua base de dados. Dessa forma, Marchetti *et al*. realizam uma junção de ambas, unindo a boa estruturação da *WordNet* com a quantidade excessiva de dados da *Wikipedia* para gerar a recomendação.

Diferente dos trabalhos citados, este trabalho apresenta uma proposta para recomendação de *tags* semânticas para categorização de recursos *Web* a partir da análise de três fontes de informação: a personomia do usuário, as *tags* da folksonomia do sistema baseado em *tagging* e a página *Web*. Dessa forma, esta abordagem não fica dependente de uma única fonte de informação para realizar a recomendação como a folksonomia e a personomia, pois mesmo sem dados nessas fontes para um respectivo recurso é possível recomendar *tags* aos usuários pela análise do conteúdo da página *Web*. De modo geral, este trabalho assemelha-se à recomendação de *tags* do sistema *ZigTag*, embora para o desenvolvimento deste estudo é analisado o conteúdo das páginas *Web*, bem como explorado tecnologias e fontes de informações distintas. A importância de cada fonte de informação para a abordagem desta recomendação de *tags* é mostrada no próximo capítulo.

Capítulo

3

A Importância da Personomia, da Folksonomia e da Página *Web* como Fontes de Informação

A proposta de recomendação de *tags* deste trabalho se baseia em três fontes de informação: (i) a personomia do usuário, (ii) a folksonomia de um sistema baseado em *tagging* e, (iii) o conteúdo textual de uma página *Web*. A razão pela qual foi decidido gerar recomendações de *tags* a partir da combinação dessas três fontes é devido ao fato de que cada uma possui uma importância específica no processo de recomendação. A personomia permite identificar as **preferências do indivíduo** – o histórico do usuário; as *tags* da folksonomia do sistema baseado em *tagging* fornecem uma visão geral sobre o que os outros usuários estão pensando sobre o recurso – um **panorama colaborativo** e; o conteúdo do recurso *Web* contém as propriedades e as características sobre o que está sendo categorizado – **o contexto e o real motivo** da categorização.

Para exemplificar e mostrar a importância das três fontes de informação no processo de recomendação de *tags* em um sistema baseado em *tagging* é discutido quatro diferentes cenários que podem ocorrer em uma eventual categorização. Os cenários mostrados a seguir têm por objetivo simular a tarefa de categorização.

3.1. Possíveis Cenários para a Recomendação de *Tags* em uma Categorização

Considerando as três fontes de informação empregadas neste trabalho, identificamos quatro diferentes cenários que podem acontecer a um usuário em uma categorização. Para cada cenário pode ser considerado que o recurso *Web* estará sempre disponível para ser analisado, pois esse é o elemento principal de uma categorização, variando, assim, apenas entre ter ou não uma quantidade suficiente de dados na personomia e na folksonomia do sistema baseado em *tagging* para inferir informação sobre os mesmos.

Cenário 1: Um usuário com informações em sua personomia e um sistema baseado em *tagging* com informações na folksonomia para um respectivo recurso.

Nesse cenário todas as fontes de informação anteriormente citadas estão disponíveis (a personomia, a folksonomia e o recurso *Web*). Em virtude disso, esse é considerado o melhor cenário para a tarefa de recomendação, pois é possível analisar o vocabulário do usuário, o ponto de vista da comunidade de usuários e o conteúdo do recurso *Web*. A Figura 11 (a) ilustra esse cenário destacando o conjunto de informações que acreditamos ser adequado (representado em cinza) para o processamento de uma recomendação de *tags*. Dessa forma, é possível considerar o que cada fonte possui de maior relevância e realizar a recomendação com base nessas informações.

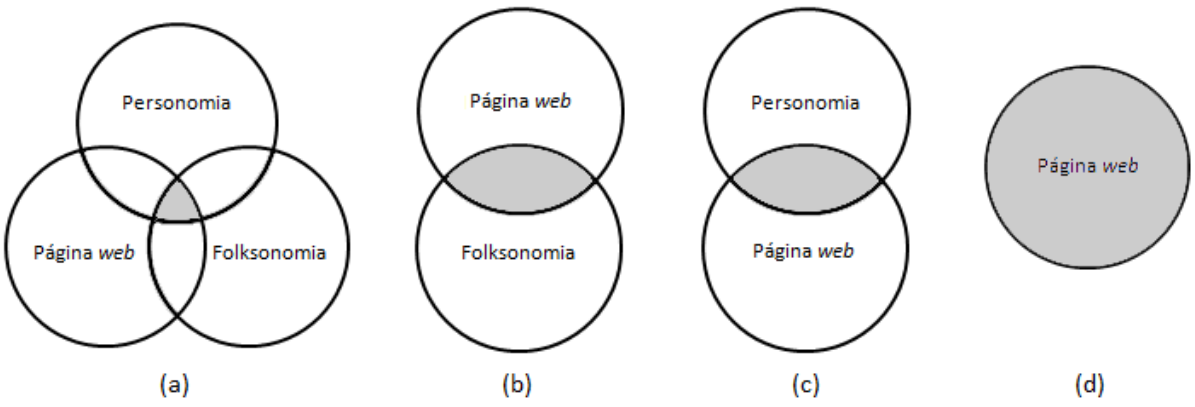

Figura 11. Recursos disponíveis e o conjunto de dados que consideramos adequados para os quatro possíveis cenários de uma recomendação de tags em um sistema baseado em tagging.

Para exemplificar esse cenário, considere que um usuário com várias *tags* em sua personomia – conforme ilustra a Figura 12 (a) – está categorizando a página *Web* "*Java (Programming Language)*" da *Wikipedia* que, por sua vez, possui várias informações em sua

folksonomia, conforme ilustra a Figura 13. Por essa razão, no momento de gerar a recomendação de *tags*, o sistema terá disponível três fontes de informação para análise: a personomia com 1009 *tags* e vários recursos categorizados pelo usuário, várias *tags* da folksonomia relativas ao respectivo recurso e o conteúdo da página *Web*. Essa disponibilidade fornece aos sistemas uma quantidade maior de informações para serem analisadas.

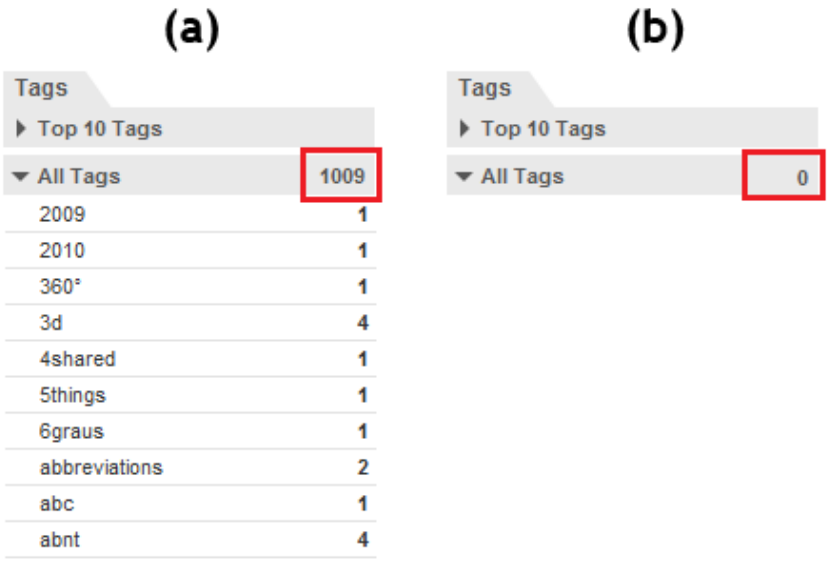

Figura 12. Exemplo da quantidade de tags de um usuário tem em sua personomia (Adaptado do sistema Delicious): a) um usuário com muitas tags em sua personomia; e b) um usuário sem tags em sua personomia.

Figura 13. Exemplo das tags utilizadas por usuários do sistema Delicious para categorizar a página "Java (Programming Language)" da Wikipedia (Adaptado do sistema Delicious).

Na maioria das vezes, mesmo tendo essas três fontes disponíveis para o processamento, a maioria dos sistemas realiza a recomendação de *tags* apenas pela análise da personomia e da folksonomia ou, simplesmente, pela folksonomia, não se beneficiando do conteúdo do recurso *Web*, como é o caso do sistema *Delicious*. Neste estudo, o sistema *Delicious* é tomado como o sistema base para se obter diversas comparações, pois é um dos sistemas baseados em *tagging* com mais tempo de uso pelos usuários, é muito utilizado pelas pessoas e muitos outros estudos já adotaram o *Delicious* como sistema base. Além disso, ele possui uma *API* (*Application Programming Interface*) que permite realizar várias operações com os dados de um usuário, juntamente com uma boa documentação (não completa, mas bem compreensível).

Cenário 2: Um usuário sem informações em sua personomia e um sistema baseado em *tagging* com informações na folksonomia para um respectivo recurso.

Uma vez que há informações disponíveis na folksonomia para o recurso a ser categorizado é possível gerar uma recomendação eficaz do ponto de vista social, pois o recurso já foi avaliado e categorizado pelos usuários do sistema. Embora para esse cenário não seja possível personalizar a recomendação pelo vocabulário do usuário, pois não há informações na sua personomia, a recomendação poderá ser ainda mais eficaz se combinar os dados da folksonomia com os dados do recurso *Web*. O conjunto adequado de informações para o processamento da recomendação nesse cenário (representados em cinza) está ilustrado na Figura 11 (b).

Para exemplificar esse cenário, considere que um usuário sem informações em sua personomia – conforme ilustra a Figura 12 (b) – está categorizando a página *Web* "*Java (Programming Language)*" da *Wikipedia* que, por sua vez, possui várias informações em sua folksonomia, conforme ilustra a Figura 13. Nessa situação, o sistema *Delicious* usa apenas as informações da folksonomia para gerar a recomendação, não se beneficiando do conteúdo do recurso *Web*.

Cenário 3: Um usuário com informações em sua personomia e um sistema baseado em *tagging* sem informações na folksonomia para um respectivo recurso.

Esse cenário requer que o sistema baseado em *tagging* realize o processamento do conteúdo da página *Web* para gerar a recomendação de *tags*, pois, caso contrário, não é possível realizar essa tarefa, visto que a simples análise dos termos contidos na personomia do usuário não garante que as características do recurso sejam identificadas. Além disso, a partir

da análise do recurso *Web* é possível evitar o problema de "*cold-start*" nos sistemas que dependem das informações contidas na folksonomia. Entretanto, se o conteúdo do recurso *Web* for analisado, é possível personalizar a sugestão de *tags* ao vocabulário do usuário, baseando-se nas informações contidas em sua personomia.

Para exemplificar esse cenário, considere que um usuário com informações em sua personomia – conforme ilustra a Figura 12 (a) – está categorizando a página pessoal de um pesquisador qualquer, porém, esse *website* ainda não foi categorizado no sistema baseado em *tagging* que o usuário está utilizando. Nesse caso, o processo de recomendação de *tags* estará restrito a utilizar apenas as informações da personomia e da página *Web*. Logo, em sistemas que adotam a folksonomia como uma fonte de informação obrigatória para a tarefa de recomendação, como é o exemplo do sistema *Delicious*, não haverá recomendação de *tags* para o usuário. O conjunto adequado de informações para o processamento da recomendação de *tags* nesse cenário (representados em cinza) está ilustrado na Figura 11 (c).

Cenário 4: Um usuário sem informações em sua personomia e um sistema baseado em *tagging* sem informações na folksonomia para um respectivo recurso.

Esse cenário é o mais desfavorável para que uma aplicação realize a recomendação de *tags*, pois há apenas o recurso *Web* para ser analisado. Na maioria dos sistemas baseados em *tagging* disponíveis *online* não é possível gerar uma recomendação para esse cenário, pois geralmente esses sistemas interpretam apenas a personomia e a folksonomia. Pelo fato do recurso *Web* que está sendo categorizado não estar inserido em uma folksonomia, ou seja, os usuários presentes no sistema não terem realizado categorizações sobre esse recurso, a única alternativa para fornecer uma recomendação de *tags* seria pelo processamento do conteúdo do recurso. Logo, abordagens de recomendação como a utilizada pelo sistema *Delicious* deixam de recomendar *tags* aos seus usuários. Por essa razão, acreditamos que o recurso *Web* seja bastante útil e possa beneficiar as recomendações, principalmente quando ele ainda não foi categorizado ou teve poucas categorizações. As fontes de informação disponíveis para esse cenário estão ilustradas na Figura 11 (d).

Para concluir a discussão sobre os quatro cenários supracitados, é interessante comparar esses cenários com a abordagem de recomendação de *tags* proposta neste trabalho, pois assim será possível perceber que este estudo possibilita a recomendação para todos os cenários citados, e não apenas para alguns, como acontece na maioria dos sistemas baseados em *tagging*. O fato desta abordagem analisar as três fontes de informação, em especial, o

conteúdo do recurso *Web* garante que, para *websites* com informações textuais, esta proposta pode fornecer recomendações para qualquer situação, não dependo dos dados da personomia e/ou da folksonomia. A Figura 11 (a) mostra o conjunto de dados que se pretende combinar neste trabalho para gerar as recomendações de *tags*. Até o presente momento não encontramos nenhum estudo que realize a recomendação de *tags* utilizando essas três fontes de informação em conjunto. Cada tipo de fonte possui uma justificativa para estar sendo utilizada na recomendação. Essas justificativas são detalhadas a seguir.

3.2. A Importância da Personomia

A personomia pode agregar uma vasta diversidade de conhecimento sobre o indivíduo, visto que em uma categorização o usuário expressa, a partir das *tags*, seu conhecimento, preferência e terminologia relativo ao conteúdo de cada recurso. A partir da análise de uma personomia em um sistema baseado em *tagging* é possível direcionar a recomendação de *tags* para se aproximar do vocabulário do usuário, oferecendo termos de acordo com as suas necessidades (SHEPTSEN *et al.*, 2008) (ZHAO *et al.*, 2008) (RASHID *et al.*, 2002) (WETZKER *et al.*, 2009). Logo, é possível que seja realizada uma aprendizagem dos interesses do usuário na qual o sistema se torna capaz de entender suas preferências.

As informações que podem ser obtidas a partir de uma personomia geralmente são (i) os próprios recursos categorizados, (ii) as informações derivadas do recurso como, as *tags*, o seu título, um texto descritivo, (iii) as iterações realizadas dentro do sistema, (iv) as avaliações de itens no sistema, dentre outras. Essas informações são representativas dos interesses de um usuário, pois a liberdade fornecida por um sistema baseado em *tagging* permite fortalecer a idiossincrasia no espaço pessoal de cada indivíduo, o que geralmente é benéfico, pois se trata de uma característica comportamental individual ou de um grupo de usuários. Neste trabalho, foi optado por considerar os dados que os usuários utilizam nas categorizações, sendo eles: a *URL* da página *Web* e o conjunto de *tags* utilizadas nas categorizações. Esse conjunto de dados foi escolhido por possibilitar o acesso ao vocabulário do usuário sem ter que solicitar qualquer ação/tarefa que interfira na sua navegação.

3.3. A Importância da Folksonomia

De acordo com Sturtz (2004) e Mathes (2004), possivelmente, o benefício mais relevante da folksonomia é sua capacidade de refletir o vocabulário dos usuários, reduzindo, assim, os

problemas associados à avaliação da relevância e utilidade de recursos recentes no sistema que ainda não foram vistos (ZHAO *et al.*, 2008) (ADOMAVICIUS *et al*., 2005). Em geral, a folksonomia expõe um panorama social, o qual um único usuário jamais conseguiria visualizar sozinho. Dessa forma, utilizar os dados da folksonomia para recomendar *tags* pode ser uma boa alternativa, pois o usuário está categorizando um recurso que outras pessoas da comunidade já categorizaram e, consequentemente, pode haver um interesse em comum entre eles, caso contrário ele não estaria realizando a categorização para um mesmo recurso, o que garante a utilidade das *tags* da folksonomia (ZHAO *et al*, 2008) (ADOMAVICIUS *et al.*, 2005).

As idiossincrasias presentes em uma folksonomia podem beneficiar a recuperação de informação, pois elas representam termos alternativos e interessantes dos usuários (serendipismo[31]) (RUSSEL, 2005) (MATHES, 2004) (STURTZ, 2004). Entretanto, mesmo combinando os dados da folksonomia com os dados da personomia em uma recomendação de *tags*, termos relevantes relacionados às características da página *Web* podem ficar excluídos da recomendação. De fato, isso pode acontecer porque a maioria das abordagens de recomendação de *tags* atuais não são direcionadas ao conteúdo do recurso *Web*, mas para a *URL* em si, visto que elas não utilizam nenhum tipo de análise do conteúdo do recurso *Web*.

3.4. A Importância da Página *Web*

O recurso *Web* – nesse caso, a página *Web* – é o principal elemento de uma categorização. Para cada recurso há vários fatores influenciando o contexto em que ele é utilizado. Isso pode ser determinado pela forma na qual o vocabulário é empregado pelo autor da página com o objetivo de expor o conteúdo ao leitor. Neste trabalho, o **contexto** é representado pelo conjunto de palavras-chave mais representativas das características e propriedades do recurso *Web*. Desse modo, para trabalhar com o contexto de um recurso é necessário extrair as palavras-chave contidas no texto e identificar os termos mais relevantes relacionados ao seu conteúdo. Em uma página *Web* o conteúdo pode estar disponível de várias formas, tais como: textual, animações em *flash*, imagens, vídeos, etc. Neste trabalho é considerado apenas o conteúdo textual.

A tarefa de obtenção dos termos para representar o conteúdo de uma página *Web* não é trivial, uma vez que a página pode conter divergências em seu conteúdo, tais como: erros de digitação; sinonímia; polissemia; flexão de termos; parte da página que não se refere ao

[31] Do inglês, *serendipity*, designa a arte de fazer descobertas felizes e não-intencionadas.

conteúdo da página, por exemplo: cabeçalho, rodapé, colunas de menu, propagandas, dentre outros. Em virtude disso, para possibilitar uma melhora na identificação dos termos de uma página, foi necessário analisar algumas características presentes no conteúdo das páginas *Web*.

3.4.1. Análise do Conteúdo de Páginas *Web* em relação à Base de Dados da *WordNet*

Análises relacionadas às personomias de usuários que, consequentemente, refletem aos dados da folksonomia, já foram realizadas em Basso *et al*. (2009). Logo, coube a este trabalho analisar os termos presentes nas páginas *Web*. O objetivo da análise foi tentar identificar algum padrão comum dos termos nos textos, bem como a taxa de reconhecimento desses termos na *WordNet*. A análise envolveu 238 páginas *Web*, selecionados aleatoriamente a partir de categorizações no sistema *Delicious*. As páginas eram da língua inglesa, uma vez que esse é o idioma básico adotado pela *WordNet*. Além disso, foram analisadas páginas de diferentes assuntos e níveis de conhecimento como medicina, engenharia, ciência da computação, etc. Em geral, os resultados mostraram que 76% dos termos presentes nas páginas *Web* foram identificados pela *WordNet* e para os 24% restantes não foi encontrado nenhum significado.

Para os termos reconhecidos, 82,5% foram identificados como substantivos; 7,9% eram verbos; 8,3% eram adjetivos e 1,3% eram advérbios, conforme ilustra a Figura 14. Esse resultado foi semelhante ao experimento realizado por Basso *et al*. (2009) para as *tags* contidas nas personomias dos usuários em sistemas baseados em *tagging*. Como existe um número significativo de verbos, adjetivos e advérbios no conteúdo das páginas *Web*, e considerando que esses também podem ser representativos do contexto do recurso, foi adaptado o algoritmo proposto por Basso *et al*. para também tentar identificar verbos, adjetivos e advérbios, visto que até o momento esse algoritmo realizava a identificação apenas de substantivos. Dessa maneira, em nossas ontologias, foi possível ter um aumento no reconhecimento dos termos em aproximadamente 17,5%.

Os termos não reconhecidos, por sua vez, foram detectados como: erros de digitação, datas, números, caracteres especiais, abreviações, termos novos adotados nas diversas áreas existentes, nomes de empresas, projetos e sistemas existentes. Em áreas que têm uma rápida ascensão do vocabulário, como acontece na ciência da computação, são encontrados

problemas como: (i) acrônimos que não constam na *WordNet* como: *pdf*[32], *api*, *soa*[33], *rss*[34], *xml*[35], etc.; (ii) termos novos que são adotados nos textos como: *ebooks*, *folksonomy*, *personomy*, *iphone*, *mashup*, *weblogs*, etc.; (iii) nome de empresas, projetos ou sistemas como: *Microsoft*, *IBM*, *Ericsson*, *Ebay*, *Flickr*, *Wikipedia*, *Technorati*, *Facebook*, *Gmail*, *Photoshop*, *ACM*, *IEEE*, etc.

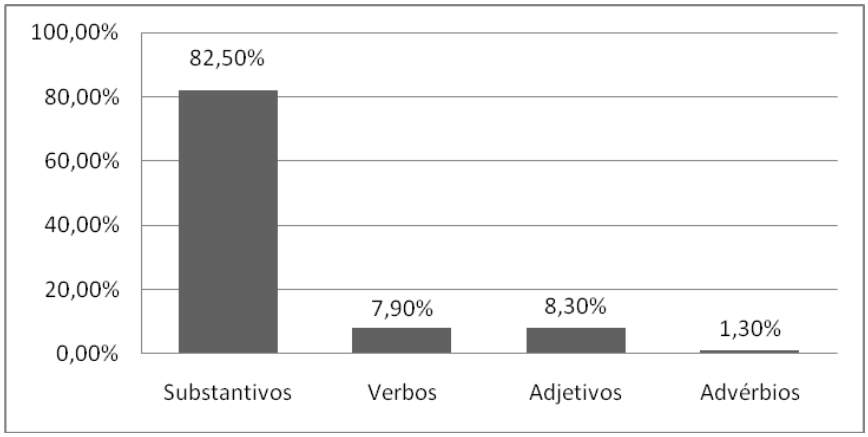

Figura 14. Identificação das classes gramaticais na WordNet para os termos das páginas Web.

Nesta proposta, o processo de identificação de termos das páginas *Web* na *WordNet* é importante, pois os que não são reconhecidos ficarão excluídos do processamento de emergência da ontologia para a página *Web*. Todavia, para aumentar a quantidade de termos reconhecidos, torna-se necessário realizar um *mashup* com outras base de dados mais dinâmicas do que a *WordNet*, como a *DBPedia* (http://dbpedia.org) que é um esforço da comunidade para extrair informações estruturadas a partir da *Wikipedia* (http://wikipedia.org) e a *ConceptNet* (http://conceptnet.media.mit.edu/) que é uma ontologia que tenta mapear o "senso-comum" dos seres humanos para permitir o acesso por computadores, dentre outras. Contudo, isso não foi explorado, pois não é o foco deste trabalho.

Com esse experimento foi identificado os motivos de vários termos das páginas *Web* não serem encontrados na *WordNet*. Além disso, percebemos que é interessante extrair termos das páginas para as quatro classes gramaticais a fim de melhor representar seu contexto, uma vez que não identificando os verbos, adjetivos e advérbios a característica da página *Web* pode ser prejudicada, dado que esses termos podem ser relevantes.

[32] Sigla de *Portable Document Format*
[33] Sigla de *Service-Oriented Architecture*
[34] Sigla de *Really Simple Syndication* (RSS 2.0)
[35] Sigla de *eXtensible Markup Language*

Nesse capítulo, foram mostrados os quatro possíveis cenários que um usuário pode se deparar em uma eventual categorização. Além disso, foram mostrados os principais benefícios de utilizar a personomia, a folksonomia e o recurso *Web* no processo de recomendação de *tags*. No próximo capítulo é apresentada a metodologia da recomendação de *tags* deste trabalho.

Capítulo

4

Uma Abordagem para Recomendação *de Tags* Semânticas

Conforme discutido no capítulo anterior, a abordagem deste trabalho faz recomendações de *tags* baseando-se na análise de três fontes de informação (a personomia, a folksonomia do sistema baseado em *tagging* e a página *Web*). Para cada fonte de informação é emergida uma ontologia específica, de tal modo que cada *tag* sugerida representará um conceito ancorado pela *WordNet* e não apenas rótulos de texto sem um significado formal relacionado. As estruturas semânticas adquiridas na *WordNet* possibilitam descrever e organizar as *tags* de uma categorização, tornando-as livres de ambiguidade, uma vez que cada termo recomendado estará relacionado a um único *synset* pelo algoritmo do processo de recomendação proposto (o processo de escolha de um sentido para cada *tag* é detalhado em Basso *et al*. (2009)), conforme ilustra a Figura 15. Por essa razão, a interface de categorização pode ser igual a dos sistemas convencionais, não sendo necessário que o usuário adquira novos conhecimentos para utilizar esta abordagem de recomendação.

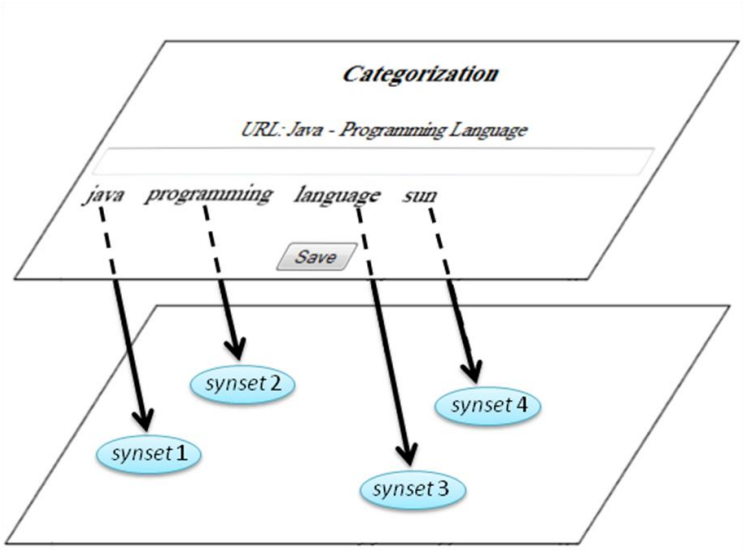

Figura 15. Exemplo de uma categorização da abordagem proposta neste trabalho.

Para a geração da recomendação não é solicitada nenhuma informação ou opinião do usuário, ou seja, estamos empregando uma abordagem não supervisionada. A Figura 16 mostra de modo geral os passos desta proposta de recomendação de *tags*. Para cada uma das três fontes de informação é extraído um conjunto de termos que representará as características referentes à página *Web* que estiver sendo categorizada (passo 1, passo 2 e passo 3). Em seguida, é gerada uma ontologia para cada fonte de informação a partir do conjunto de termos extraídos (passo 4), baseando-se em uma adaptação da proposta de Basso *et al*. (2009). Então, é feita uma comparação entre as três ontologias (passo 5) a fim de identificar os conceitos mais relevantes relacionados entre si, processo esse denominado de mapeamento entre as ontologias. Por fim, são recomendados os conceitos que estiverem mais relacionados entre as três ontologias. Os passos desse processo são detalhados a seguir.

4.1. Passo 1: Extração de Termos de Páginas *Web*

A extração de termos da página *Web* é uma tarefa fundamental no escopo deste trabalho, pois os termos que melhor identificam as características e propriedades do conteúdo de uma página serão utilizados para representar seu contexto. Para representar um documento a partir de termos há três abordagens principais: a estatística, a linguística e a híbrida. Para a primeira, cada documento é representado por um vetor de termos e cada termo possui um valor associado que indica a frequência do termo no documento (Termo-Frequência). Para a segunda, é necessário que os textos tenham anotações com alguma informação linguística (morfológica, sintática ou semântica), sendo que essas anotações são analisadas no processo

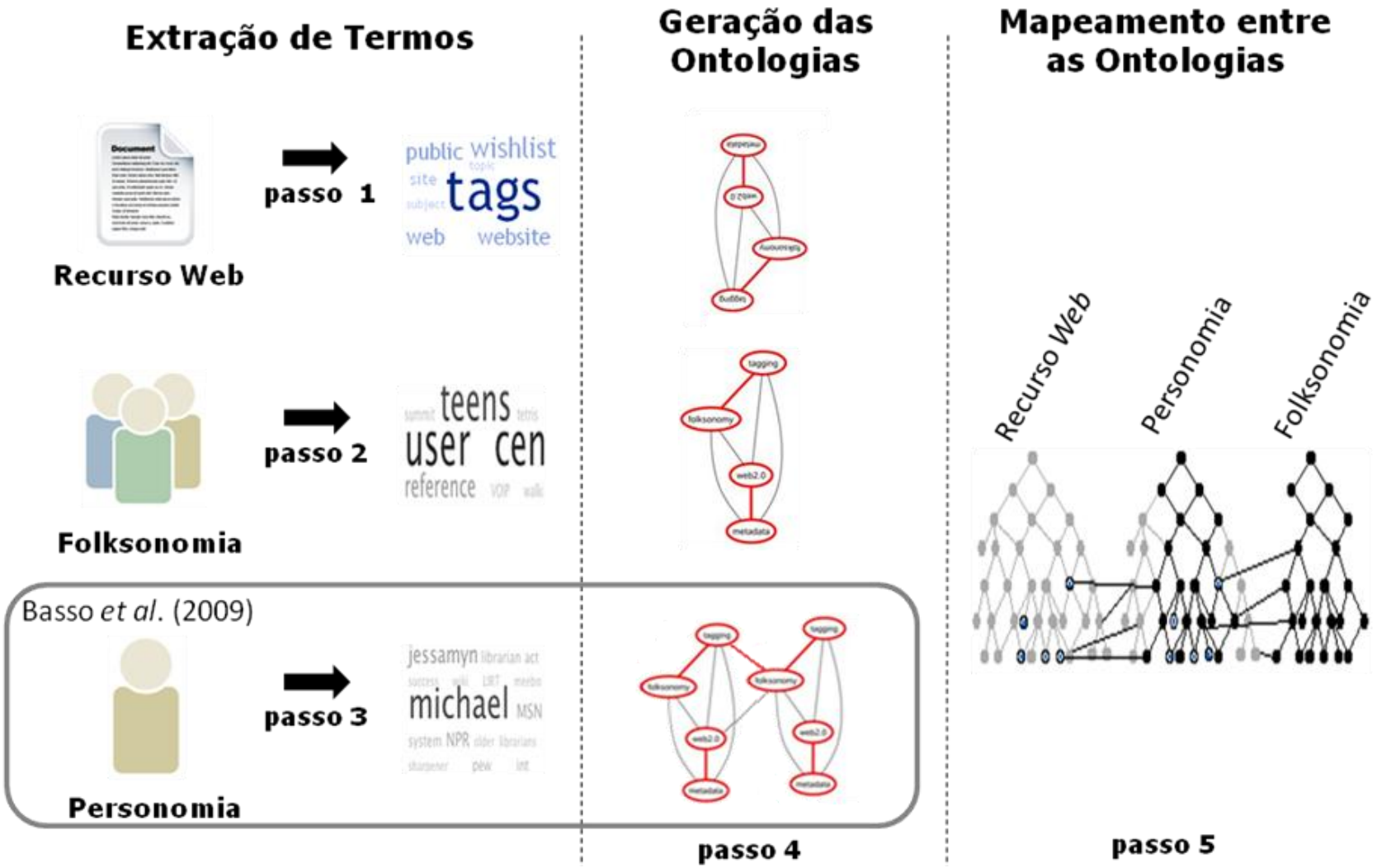

Figura 16. O processo desta abordagem de recomendação de tags (BORTH et al., 2010).

de extração. Na terceira, é utilizado a estatística juntamente com o conhecimento linguístico. Neste trabalho, é usada a abordagem estatística, uma vez que esse tipo de abordagem permite quantificar a importância entre os termos no documento (MANNING, 2008) (MICARELLI, 2007) (RIJSBERGEN, 1999). Além disso, esse é um modelo considerado clássico e, em geral, bem aceito e utilizado academicamente e comercialmente.

Na extração[36] de termos da página *Web*, dois conjuntos de termos são extraídos: (1) uma lista de termos presentes no texto com suas respectivas frequências e, (2) um conjunto de termos a partir do título da página. O uso do segundo conjunto é feito para aumentar a relevância (frequência) dos termos comuns entre o título e o texto, visto que o título de um documento geralmente é um resumo significativo do seu conteúdo. Em seguida, para que esses dois conjuntos de termos possam ser utilizados de forma eficaz, realizamos um "processo de limpeza", cujo objetivo é identificar e eliminar as palavras que não estão relacionadas ao conteúdo de um recurso *Web*. O "processo de limpeza" consiste em (i) desconsiderar sinais de pontuações e símbolos, tais como: "=", "-", "(", ")", "[", "]", "{", "}", "", "\"", "\\", "|", ";", ":", ".", ",", "!", "?" e (ii) remover as *stopwords*, que são palavras que aparecem inúmeras vezes no conteúdo de um texto e não são representativas para o documento (CROFT, 2009). Essas palavras podem ser preposições, artigos, pronomes e outras classes que servem apenas para auxiliar na construção de textos. Conforme afirma Silva (2007), entre 40 e 50% das palavras em um documento são insignificantes. Por essa razão, utilizar uma lista bem elaborada de *stopwords* permite que sejam eliminados muitos dos termos irrelevantes, aumentando, assim, a eficiência do resultado obtido pelo processo de extração dos termos para o documento. A lista das *stopwords* utilizada neste trabalho é a sugerida por *LexTek International*[37], sendo composta por 429 termos identificados por um grupo de pesquisadores e desenvolvedores de tecnologia em Processamento de Linguagem Natural. Os processos (i) e (ii) podem ser mais bem visualizados na Figura 17. Nesse exemplo, considere que a frase extraída da página "*Java (Programming Language)*" da *Wikipedia* está mostrada na cor cinza e, as respectivas remoções estão grifadas em cada fragmento do texto. Logo, ao final do primeiro processo, não haverá nenhuma pontuação e, ao final do segundo processo nenhuma *stopwords*.

[36] O processo de extração de informação das páginas *Web* é realizado apenas para termos simples, não identificando termos compostos.
[37] http://www.lextek.com/manuals/onix/stopwords1.html

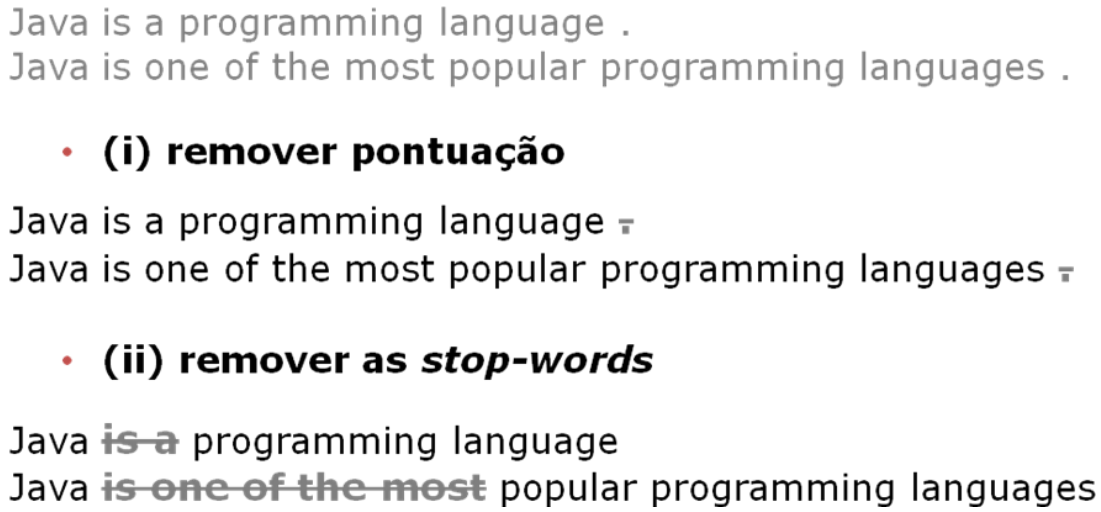

Figura 17. Exemplo do processo de remoção da pontuação e das stopwords.

Após o "processo de limpeza" os termos remanescentes passarão por outro processo, cujo objetivo é normalizar os conceitos que possuem formas léxicas de escrita distintas, porém, que sejam semanticamente iguais. Essa tarefa é realizada na tentativa de evitar que haja variações léxicas entre os termos que possuem o mesmo significado, pois as palavras contidas em um documento podem possuir variantes como, plural ou singular, palavras adicionadas de sufixo, etc. As várias formas de flexionar uma palavra não necessariamente alteram sua representação semântica, causando um problema no momento da geração da matriz dos termos para o documento, pois, em geral, não faz sentido a palavra "*car*" compor a matriz Termo-frequência de forma diferente da palavra "*cars*". Para minimizar esse problema, utilizamos uma técnica de conflação[38], cujo objetivo é unir as palavras que possuem variações léxicas em uma única palavra. Assim, o documento fica representado por termos únicos, de forma que cada par de termos variantes em sua forma léxica identificados com a mesma semântica sejam representados por apenas um deles e quando houver essa junção suas frequências são somadas.

Os processos mais comuns de conflação são o *stemming* (FRAKES *et al*., 1992) (KROVETZ, 1993) (ALLAN *et al*., 2003) e a lematização (ARAMPATZIS *et al*., 2000) (KORENIUS *et al*., 2004). O objetivo do *stemming* é reduzir a grande dimensionalidade do termo à sua provável raiz (*stem*) (parte fundamental da palavra), ou seja, as palavras que se diferenciam basicamente pela flexão e as que possuírem significados semelhantes são reunidas em uma mesma palavra. A lematização, por sua vez, reduz as palavras que são variantes à sua correspondente forma canônica. A principal diferença entre *stemming* e a

[38] Ato de fusão ou combinação para igualar as variantes morfológicas.

lematização é que, no primeiro, palavras de diferentes categorias morfológicas podem ser reduzidas a um mesmo *stem*, enquanto que para a lematização a categoria morfológica é mantida (GONZALEZ *et al*., 2006). A Tabela 4 ilustra a diferença nos resultados entre essas técnicas. Neste trabalho a técnica de conflação utilizada foi a lematização.

Tabela 4. Exemplo de stemming e lematização.

Stemming	Lematização
smoke **smok**ing **smok**er → **smok**	smoke smoking smoker → **smoke**

Para exemplificar o processo de lematização, considere o mesmo fragmento de texto ilustrado na Figura 17. A lematização é realizada apenas para o fragmento de texto resultante do processo (i) e (ii). Nesse exemplo, os termos no gerúndio e no plural serão lematizados. Por exemplo, o termo "*programming*" será lematizado para "*program*" e, o termo "*languages*" será lematizado para "*language*", conforme ilustra a Figura 18.

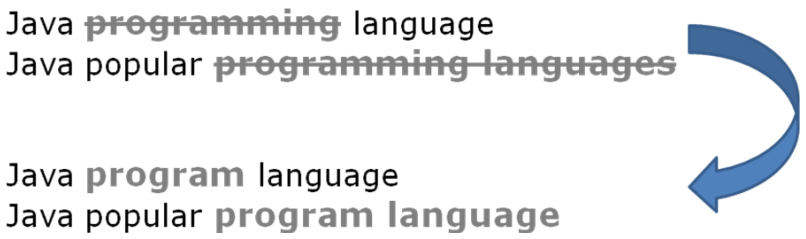

Figura 18. Exemplo de lematização de um fragmento de texto extraído da página Web "Java (Programming Language)" da Wikipedia.

Após o processo de lematização, é realizada a tarefa de detecção de similaridade entre os termos do conteúdo da página *Web* e seu título[39] (BASSO *et al*. (2009). Essa tarefa tem por objetivo verificar as palavras do texto que possuem significados similares com o título da página, aumentando sua relevância. Esse aumento de relevância dos termos com similaridade ao título poderá ajudar na identificação do conjunto de termos que melhor retratam às características de uma página *Web*. Essa tarefa utiliza a métrica de similaridade semântica chamada *Lesk* adaptada para a *WordNet* (BANERJEE, 2002). O objetivo dessa métrica é medir o quão forte estão interligados o conceito entre duas palavras.

Na comparação dos termos, usando a métrica *Lesk* (BANERJEE, 2002), cada par de termos (um da página *Web* e outro do título) produzirá um valor que representa o grau de similaridade entre ambos. Esse valor se encaixará em uma das três classes de similaridade

[39] Maiores detalhes sobre essa tarefa pode ser encontrada em BASSO (2009).

aceitas neste trabalho: (i) “alta similaridade”, (ii) “similar” e (iii) “baixa similaridade”. Após a classificação, apenas os termos que pertencerem às categorias (i) e (ii) terão suas frequências aumentadas, somando ao valor da frequência atual 10 e 5 pontos, respectivamente (obtivemos esses valores a partir de vários testes com páginas *Web* aleatórias), pois na abordagem estatística a frequência é o único atributo que pode refletir o quanto o termo está relacionado a um documento textual. Como resultado, haverá a redução dos termos entre os mais frequentes que não estão diretamente relacionados ao contexto da página *Web*. Por exemplo, termos relacionados a “*car*” ou “*bike*” podem estar presentes no conteúdo de uma página *Web*, no entanto, eles podem não estar suficientemente relacionados à semântica do recurso, pois o título da página pode ser “*Programming Language*” e, consequentemente, não terão suas frequências aumentas. Em uma tarefa como a recomendação de *tags*, esse processo pode reforçar a escolha de um conceito semanticamente mais adequado para representar a ontologia da página *Web*.

De fato, esse primeiro passo possui grande importância para esta proposta, pois extrair bons termos, que possam ser os mais representativos às características do recurso, é fundamental e garantirá que a ontologia da página *Web* seja criada com características e propriedades adequadas do recurso.

4.2. Passo 2: Recuperando Informação a partir da Folksonomia

Para recuperar as *tags* de vários usuários para um respectivo recurso *Web* utilizamos o sistema *TagManager* (DA SILVA, 2009), cujos objetivos são obter, agrupar e gerenciar informações das personomias de vários usuários para vários sistemas baseados em *tagging* (*Delicious*, *Flickr*, *YouTube*, *SlideShare*, etc.) que os respectivos usuários utilizam.

Outra alternativa para recuperar os termos de uma folksonomia é desenvolver um “*screen scrapper*” de uma *URL* contida em um sistema baseado em *tagging*, como o sistema *Delicious*, visando extrair as *tags* utilizadas nas categorizações por vários usuários a um determinado recurso. Contudo, para ambos os casos (usando o *TagManager* ou executando o “*screen scrapper*”) é possível que o recurso não tenha sido categorizado ainda pelos usuários e, consequentemente, não haverá informações sociais para o processamento da ontologia da folksonomia.

Após a recuperação/extração das *tags* da folksonomia para um determinado recurso, cada *tag* possuirá um valor que estará relacionado à sua frequência, o qual irá dizer o quanto

ela foi utilizada pelos usuários no sistema. Considerando que, nesse momento, os termos da página *Web*, juntamente com suas respectivas frequências também já foram extraídos, percebe-se que, em uma eventual comparação entre a frequência dos termos de cada fonte (a página *Web* e a folksonomia), poderá haver uma desproporcionalidade. Por exemplo, ao extrair os termos e as frequências da página *Web* e da folksonomia para a página "*Java (Programming Language)*"[40] da *Wikipedia*, percebe-se que a frequência do termo "*java*" do conteúdo da página é muito superior à frequência de uso da *tag* "*java*" na folksonomia, conforme mostra a Tabela 5. Essa desproporcionalidade acontece porque, em geral, uma página *Web* possui sua matriz de termo-frequência fixa, uma vez que a tendência do conteúdo de uma página é não se alterar após sua publicação, enquanto que a frequência de uso das *tags* da folksonomia tende a crescer periodicamente conforme aumenta o número de categorizações para o respectivo recurso.

Tabela 5. Relação Termo-Frequência extraídos da página "Java (Programming Language)" da Wikipedia e a frequência de uso das tags pelos usuários (folksonomia).

Página *Web*		**Folksonomia**	
java	117	*java*	69
sun	30	*programming*	30
retrieved	27	*wikipedia*	18
language	21	*language*	11
programming	21	*wiki*	9

Uma vez que este estudo trata dois tipos de fontes heterogêneos e, pelo fato de não priorizarmos os termos da página *Web* sobre os da folksonomia ou vice-versa, torna-se necessário equalizar a escala de frequência dos termos entre essas fontes. Não é necessário equalizar a frequência dos termos da personomia, pois é com base nessa fonte de informação que os termos resultantes da página *Web* e da folksonomia serão mapeados para a identificação final dos conceitos relevantes. Assim, o processo de equalização para os termos da página *Web* e as *tags* da folksonomia para um determinado recurso tem por objetivo uniformizar a frequência entre os termos de ambas as fontes. A fórmula para a equalização dos termos da folksonomia é dada pela Equação 1:

$$\text{Frequência } t = t * \frac{Frequência\ do\ termo\ mais\ frequente\ da\ página\ Web}{Frequência\ do\ termos\ mais\ frequente\ da\ folksonomia}$$

Equação 1. Cálculo da equalização para uma tag da folksonomia em relação aos termos da página Web.

[40] http://en.wikipedia.org/wiki/Java_(programming_language)

na qual *t* é a frequência da respectiva *tag* da folksonomia.

Como pode ser visto na Tabela 6, após o processo de equalização, a *tag* "*java*" da folksonomia também possuirá a frequência de 117. Logo, os demais termos dessa fonte de informação terão suas frequências atualizadas proporcionalmente.

Tabela 6. Resultado da equalização das tags da folksonomia em relação aos termos da página Web.

Termos da folksonomia	Frequência antes da equalização	Frequência após a equalização
java	69	117
programming	30	50
wikipedia	18	30
language	11	18
wiki	9	15

4.3. Passo 3: Recuperando Informação a partir da Personomia

Da mesma forma que foi realizada a recuperação dos termos da folksonomia é realizada a recuperação dos termos da personomia. Entretanto, a diferença é que nesse passo, as categorizações recuperadas serão apenas para um usuário, não sendo necessário o processo de equalização entre os termos.

Recuperado os dados para cada um dos três tipos de fontes, podem ser geradas as ontologias para cada fonte de informação.

4.4. Passo 4: Geração das Ontologias

Um grande problema ao gerar uma ontologia para uma fonte de informação com muitos dados é que, geralmente, a quantidade de termos é muito alto e, quanto maior a quantidade de termos para uma fonte maior será o tempo gasto para a geração da ontologia, pois para cada termo é necessário identificar um conceito e suas relações. Além disso, nesse processo é possível que seja encontrado mais de um conceito para um termo, tornando-se necessário realizar uma desambiguação de sentido[41] que, consequentemente, irá aumentar ainda mais o tempo da geração da ontologia. Dessa forma, para reduzir esse tempo é necessário encontrar

[41] O processo de desambiguação de sentido é detalhado em Basso (2009).

uma alternativa para diminuir o número de termos de baixa relevância extraídos de cada fonte de informação, pois quanto menor a quantidade de termos na geração de uma ontologia menor será a quantidade de conceitos e relações e, por consequência, menor o tempo para a geração da estrutura ontológica.

Para tanto, realizamos uma análise da distribuição de frequência dos termos extraídos de cada fonte de informação, como mostra a Figura 19. Essa imagem ilustra parte da distribuição de frequência dos termos extraídos da folksonomia para a página "*Java (Programming Language)*" da *Wikipedia*, na qual é possível notar um efeito conhecido como "cauda longa" (*long tail*) (ANDERSON, 2006) previsto pela lei de Zipf's (NEWMAN, 2006) (ZIPF, 1947). A lei de Zipf's está associada com o Princípio de Pareto (NEWMAN, 2006), também conhecido pela regra dos 80-20, o qual determina que 80% das causas são responsáveis por 20% dos efeitos, enquanto que os outros 20% das causas são responsáveis por 80% dos efeitos (BRYNJOLFSSON *et al*., 2007) (KOCH, 2000). Após a distribuição de frequência, percebemos que o comportamento para os três tipos de fontes de informação foi semelhante, tendo poucos termos sendo usados muitas vezes e muitos termos sendo usados poucas vezes.

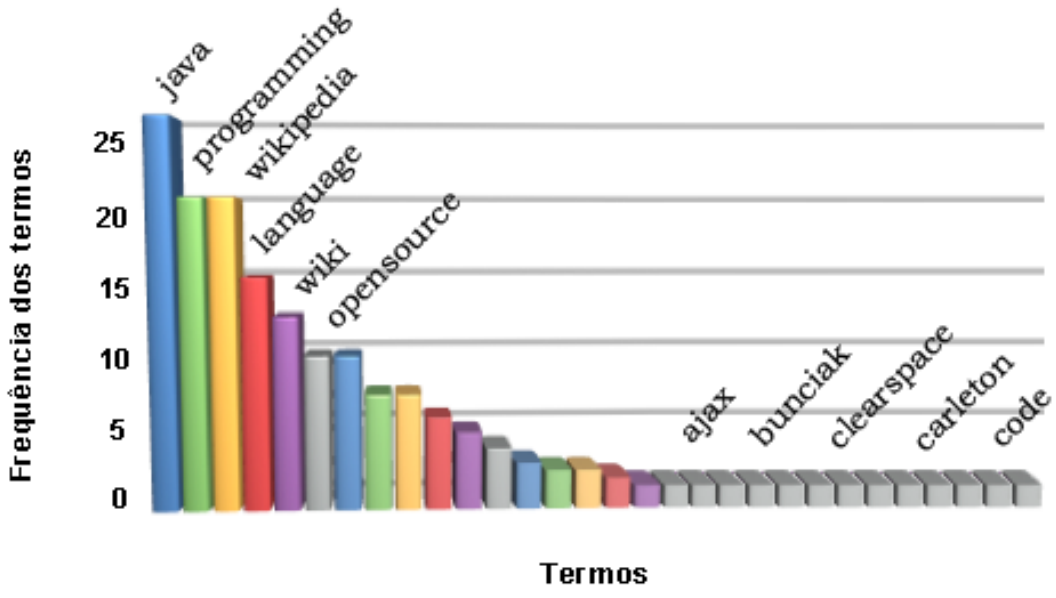

Figura 19. Distribuição de frequência dos termos para a página "Java (Programming Language)" da Wikipedia.

Desse modo, para reduzir a quantidade de termos de cada fonte de informação visando à emergência de ontologias, foi adotado neste trabalho o uso da regra dos 80-20. Assim, a ontologia de cada fonte de informação será gerada apenas com os 20% dos termos mais frequentes e, com a tendência de ser os 20% mais representativos das características e

propriedades do recurso *Web*. Com isso, foi reduzida um pouco a qualidade das ontologias, uma vez que a ontologia não será emergida com todos os termos, porém, acreditamos que o ganho em tempo computacional para emergir as ontologias compensa a perda de qualidade.

Uma vez que o conjunto de termos de cada fonte de informação está definido, uma ontologia é gerada para cada tipo de fonte (a personomia, a folksonomia e a página *Web*) baseando-se na adaptação feita na proposta de Basso *et al*. (2009). Atualmente, a ontologia gerada neste trabalho é uma ontologia leve (GIUNCHIGLIA, 2006), pois ela foi adaptada para emergir apenas as relações de "*is-a*" e "*part-of*". É importante salientar que, a abordagem de Basso *et al*. possui o processo de extração de conteúdo apenas para os dados da personomia que, consequentemente, também funciona para os termos da folksonomia. Assim, para essas duas fontes de informação podem ser identificados os termos de forma composta. Por exemplo, termos como, "*Semantic_Web*", "*artificial_intelligence*" e "*programming-language*", poderão ser identificados como um único conceito na ontologia. Por outro lado, o processo de extração de dados das páginas *Web* deste trabalho não faz a extração de termos compostos. Logo, a emergência das ontologias suporta conceitos compostos apenas para a personomia e a folksonomia.

Após a geração das ontologias é necessário identificar os conceitos semelhantes entre cada fonte de informação, processo esse denominado de mapeamento entre as ontologias.

4.5. Passo 5: Mapeamento entre as Ontologias

Atualmente, há várias expressões semelhantes utilizadas para o propósito de mapeamento de ontologias, tais como: fusão, integração, alinhamento, mapeamento, etc. Su (2002) define nosso entendimento sobre o mapeamento de ontologias como: "Dadas duas ontologias A e B, mapear uma ontologia com outra significa que para cada conceito (nó) da ontologia A, tentaremos encontrar um conceito (nó) correspondente, o qual possui a mesma semântica na ontologia B e vice-versa"[42], caso isso não ocorra não há um mapeamento exato. Além disso, o autor relata que são necessárias duas condições para possibilitar o mapeamento: (i) desenvolver um algoritmo que descubra os conceitos com significados semelhantes e (ii) definir as relações semânticas que existam entre dois tópicos relacionados. Logo, para o mapeamento de ontologias, criamos um processo que corresponde à definição e aos requisitos citados pelo autor. Para identificar os conceitos similares entre duas ontologias, o algoritmo

[42] Tradução livre para "*Given two ontologies A and B, mapping one ontology with another means that for each concept (node) in ontology A, we try to find a corresponding concept (node), which has the same or similar semantics, in ontology B and vice verse*".

deste trabalho compara a igualdade entre *synsets*, mapeando apenas os *synsets* similares. Esses *synsets*, por sua vez, serão equivalentes e sinônimos. Isso é uma consequência do uso da *WordNet*, pois quando dois *synsets* são iguais eles são sinônimos cognitivos, porém, não necessariamente seguem a mesma forma léxica de escrita. Por exemplo, os conceitos "*programming*" e "*computer programming*" possuem formas de escrita distintas, no entanto, são representados pelo mesmo *synset* na base de dados da *WordNet*.

Nesse momento, todos os *synsets* presentes na ontologia da página *Web* e na ontologia da folksonomia são comparados, um a um, com cada *synset* presente na ontologia da personomia. Os *synsets* comparados entre as ontologias são apenas aqueles que foram extraídos/recuperados de cada fonte de informação. Logo, respeitando o Princípio de Pareto, a comparação ocorrerá apenas entre os 20% dos termos mais frequentes de cada fonte. Quando dois *synsets* são iguais, eles são mapeados para ambas as ontologias, isto é, para cada conceito de uma das duas ontologias que se encontra na ontologia da personomia, cria-se uma relação de mapeamento entre ambos. O objetivo é descobrir quais os conceitos de cada ontologia que têm relações com os conceitos da ontologia da personomia. Assim, quando um conceito da ontologia da personomia é mapeado com um conceito de outra ontologia, esses são fortalecidos. Logo, um conceito contido na ontologia da personomia pode ser fortalecido até duas vezes, uma para o mapeamento da página *Web* e outra para o mapeamento da folksonomia.

Fica explícito que a ontologia central deste estudo é a personomia, pois não haverá mapeamentos entre os conceitos da ontologia da página *Web* e os da folksonomia. Em vista disso, no momento da recomendação, os conceitos presentes na ontologia da personomia que estiverem mapeados com conceitos da página *Web* e/ou da folksonomia serão mais relevantes que os outros não mapeados. Essa observação é importante porque se um mesmo conceito estiver presente nas três ontologias, esse terá duas relações de mapeamento (personomia – página *Web* e personomia – folksonomia) e, portanto, sua relevância será maior que os demais conceitos que possuem apenas uma ou nenhuma relação de mapeamento. O resultado desse processo será a identificação dos conceitos mais relacionados entre as ontologias e, possivelmente, com a tendência de serem os mais relevantes ao interesse do usuário.

Para exemplificar, vamos considerar que um usuário está categorizando o *website* "*Java (Programming Language)*" da *Wikipedia*. O processo de mapeamento entre as ontologias pode ser visualizado na Figura 20. Pelo fato da ontologia gerada ser muito grande para as três fontes de informação, a ilustração mostra apenas alguns conceitos e relações de cada ontologia. Os conceitos que estão com as caixas em negrito são os termos extraídos/recuperados de cada fonte. Esses são os conceitos que realmente importam no processo de mapeamento, pois são considerados relevantes para a respectiva fonte de informação. Logo, o mapeamento será realizado apenas entre os conceitos que estão com a caixa em negrito. Por exemplo, cada conceito com a caixa em negrito da personomia será comparado com cada conceito com a caixa em negrito para a página *Web* e a folksonomia. Nesse caso, o conceito "*Programming Language*" da ontologia da personomia será mapeado com o mesmo conceito presente na ontologia da folksonomia. A mesma situação acontece

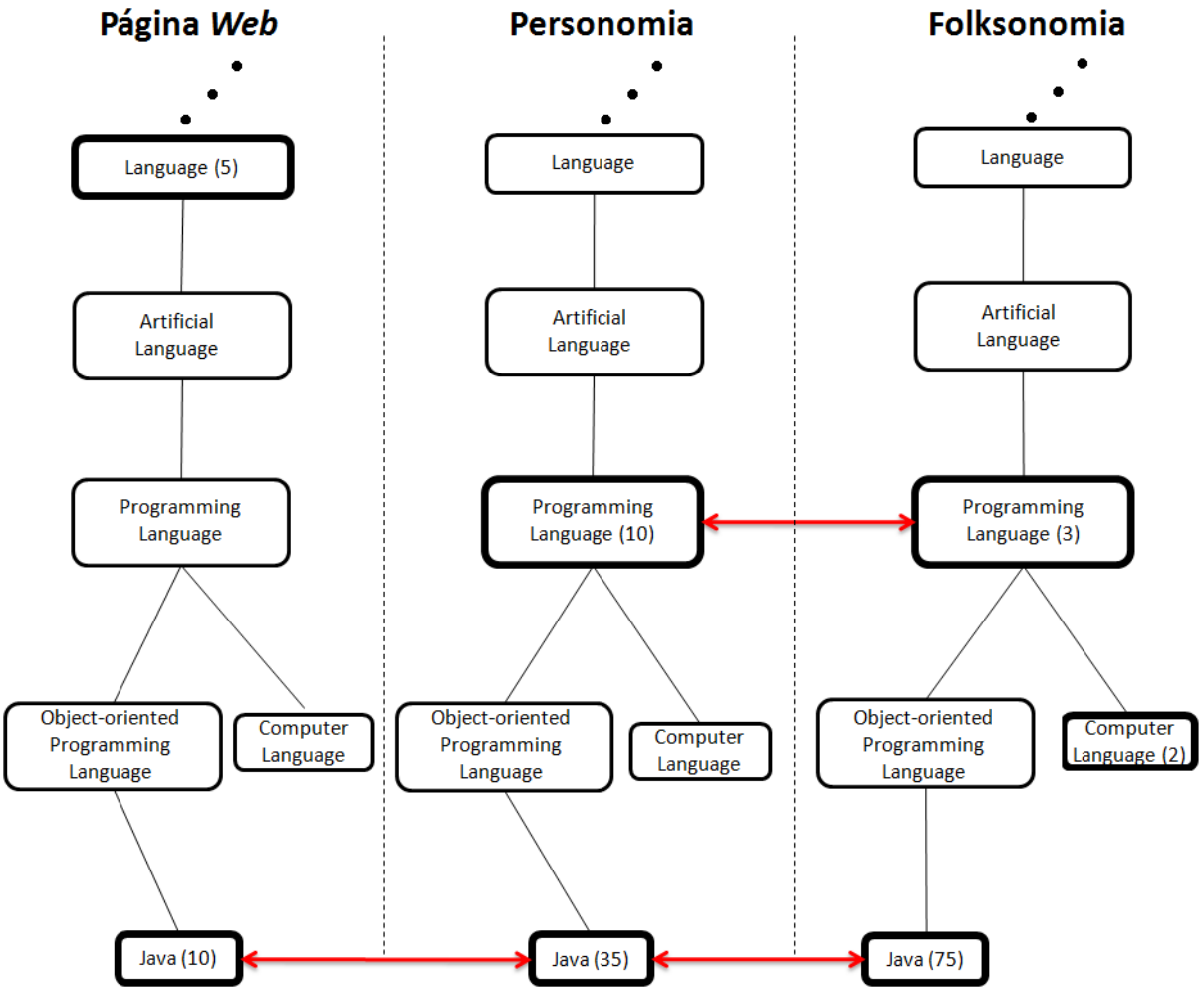

Figura 20. Exemplo do mapeamento entre as ontologias para o website "Java (Programming Language)" da Wikipedia.

com o conceito "*java*", embora para esse é realizado dois mapeamentos, um para a ontologia da página *Web* e outro para a da folksonomia. Os conceitos que não estão com a caixa em negrito não entram no processo de mapeamento, pois foram criados automaticamente pelo processo de emergência. Esses são conceitos de alto nível identificados pela relação de hiperonímia na *WordNet* e, assim, não serão analisados, pois foram criados apenas para sustentar a hierarquia entre os conceitos da *WordNet*. Dessa maneira, o conceito "*Language*" da ontologia da folksonomia e o conceito "*Computer Language*" da ontologia da folksonomia não serão mapeados com os mesmos conceitos presentes na ontologia da personomia, pois eles são apenas conceitos auxiliares criados na construção da estrutura hierarquia das ontologias.

Outra situação no processo de mapeamento pode acontecer quando, em uma categorização, houver dados na personomia do usuário e o conteúdo da página *Web* para ser analisado, porém, sem informações na folksonomia. Nesse caso, o mesmo processo de mapeamento será efetuado, porém, agora o mapeamento será realizado apenas entre a ontologia da personomia e da página *Web*, visto que sem informações na folksonomia, nenhuma ontologia será emergida. Assim, cada conceito da ontologia da personomia que estiver com a caixa em negrito poderá ser mapeado com um conceito correspondente da página *Web* que estiver com a caixa em negrito.

Uma última situação possível é quando o usuário não tiver dados em sua personomia, por exemplo, um novo usuário no sistema. Nesse caso, o algoritmo de mapeamento não realizará o mapeamento entre as ontologias, pois não haverá conceitos na ontologia central (personomia) para ser mapeado com as ontologias da página *Web* e da folksonomia. Dessa forma, a recomendação de *tags* é realizada apenas pela frequência (em ordem descendente) de uso dos conceitos extraídos de ambas as fontes de informação, a página *Web* e a folksonomia. Assim, um usuário novo no sistema poderá ter conceitos recomendados de ambas as fontes. Portanto, mesmo tentando personalizar a recomendação de *tags* ao vocabulário do usuário, esta proposta não se limita a recomendar apenas termos da personomia do usuário. Assim, entende-se que quando não existem dados na personomia, conceitos da página *Web* e da folksonomia são interessantes de serem adicionados ao espaço pessoal do usuário e, consequentemente, poderão ser recomendados. A ordem e a prioridade em que as *tags* são recomendadas são detalhadas na próxima seção.

4.6. Passo 6: Seleção dos Conceitos para a Recomendação de *Tags*

Ao final do processo, a recomendação exibirá os conceitos que possuírem as maiores relevâncias, baseando-se na quantidade de mapeamentos entre as ontologias e, em seguida, na frequência de uso de cada conceito na personomia. A prioridade para a recomendação é feita da seguinte maneira:

1. Primeiramente, são selecionados os conceitos da personomia que possuem duas relações de mapeamento, com as ontologias da folksonomia e da página *Web*. Caso haja mais de um conceito com duas relações, esses são ordenados a partir de sua frequência na personomia;
2. Em seguida, são selecionados os conceitos da personomia que possuem apenas uma relação de mapeamento para uma das ontologias, da folksonomia ou da página *Web*. Caso haja mais de um conceito com apenas uma relação, esses são ordenados a partir de sua frequência na personomia; e
3. Por fim, são selecionados os conceitos das ontologias da folksonomia e da página *Web* que fazem parte do conjunto de termos extraídos. A ordenação será de forma descendente à frequência de cada conceito para as duas fontes de informação, visto que o processo de equalização de frequência entre os termos foi realizado previamente e, por isso, os termos estarão distribuídos de maneira uniforme. Com essa regra, é possível que conceitos da folksonomia e da página *Web* sejam recomendados ao usuário, não se limitando à apenas conceitos presentes no vocabulário do usuário.

Com base no exemplo da Figura 20, a ordenação dos conceitos para a recomendação de *tags* seria da seguinte maneira: primeiramente, seriam recuperados os conceitos da personomia que possuem duas relações de mapeamento. Nesse caso, o conceito selecionado seria "*Java*", visto que esse possui uma relação de mapeamento com a ontologia da página *Web* e outra com a da folksonomia. Em seguida, seriam recuperados os conceitos da personomia que possuem apenas uma relação de mapeamento, independente se é com a ontologia da página *Web* ou da folksonomia. Para essa situação, o conceito "*Programming Language*" da personomia satisfaz a regra, uma vez que ele possui apenas a relação de mapeamento com a ontologia da folksonomia. Por fim, após recuperar todos os conceitos que possuem relações de mapeamentos, seriam recuperados os conceitos da página *Web* e da folksonomia que não possuem relação de mapeamento. Nesse caso, os conceitos "*Language*"

e "*Computer Language*" seriam selecionados baseados em sua frequência (ordem descendente). A ordem pela qual as *tags* seriam recomendadas é dada abaixo:

- *java*: o conceito possui duas relações de mapeamento;
- *programming language*: o conceito possui uma relação de mapeamento;
- *language*: o conceito da página *Web* não possui relação de mapeamento, entretanto, após o processo de mapeamento, ele foi identificado como o mais relevante dentre os que não contêm relação de mapeamento, visto que sua frequência é maior que o conceito *computer language*; e
- *computer language*: o conceito da folksonomia não possui relação de mapeamento, porém, após o processo de mapeamento, ele obteve a segunda maior frequência dentre os termos que não possuem relação de mapeamento.

Nesse capítulo, foi apresentada a metodologia da recomendação de *tags* semânticas desta dissertação, a qual se baseia em três fontes de informação distintas. No próximo capítulo são mostrados os aspectos de implementação deste trabalho.

Capítulo

5

Aspectos de Implementação

Para instrumentar a abordagem proposta do Capítulo 4 foi desenvolvido um sistema denominado ***TOM Tag Recommender*** (***T****ag****O****ntology****M****anager Tag Recommender*). A aplicação desenvolvida faz a recomendação de *tags* semânticas a partir de três fontes de informação distintas (a personomia, a folksonomia e a página *Web*) com o objetivo de evitar os problemas discutidos nos capítulos anteriores. Esse capítulo aborda os principais aspectos de codificação da aplicação desenvolvida que reflete esta abordagem.

5.1. Tecnologias Utilizadas para o Desenvolvimento da Aplicação

A presente aplicação de recomendação de *tags* foi desenvolvida utilizando as tecnologias da plataforma Java. Essa linguagem de programação foi escolhida por ser bem difundida e por disponibilizar diversas *APIs* e *frameworks* que auxiliam no desenvolvimento, os quais facilitam e minimizam o tempo de desenvolvimento de várias tarefas essenciais como,

extração de texto, requisições via *HTTP*, dentre outros. O servidor de aplicação *Tomcat*[43] foi usado para o gerenciamento do ciclo de vida. Abaixo estão listadas as principais *APIs* utilizadas e suas respectivas funcionalidades.

API do *Delicious*[44] é uma biblioteca desenvolvida em Java, em sua versão atual 1.14, que permite recuperar e manipular *tags* das categorizações realizadas por um usuário no sistema. Embora essa *API* tenha sido empregada no *TOM Tag Recommender* para acessar a base de dados do *Delicious* e recuperar as *tags* utilizadas nas categorizações de um usuário, ela não fornece métodos atualizados para recuperar as *tags* utilizadas por vários usuários a um respectivo recurso – dados da folksonomia. Para essa tarefa, foi preciso utilizar uma *API* cujo objetivo é extrair informações das páginas *Web* do *Delicious*, por exemplo, a *Jericho HTML Parser*.

O *Jericho HTML Parser*[45] é uma biblioteca desenvolvida em Java que permite analisar e manipular documentos *HTML* (*HyperText Mark-up Language*), possibilitando extrair informações de forma facilitada a partir de uma *URL*. Neste trabalho, é utilizada essa biblioteca para a tarefa de extração de informações dos recursos *Web* e, para a extração de informações das páginas do *Delicious*, visto que esse sistema nem sempre disponibiliza uma forma simplificada de acesso aos seus dados a partir de sua *API*. A tarefa de extração de textos das páginas *Web* é denominada de "*screen scrapping*". Nessa tarefa é necessário identificar os conteúdos relevantes que estão presentes entre as *tags HTML* e criar algoritmos para extrair esses dados. Por exemplo, a Figura 21 ilustra parte do código-fonte de uma página do *Delicious*. Para esse caso foi necessário localizar as informações pertinentes como, as *tags* que estão sendo utilizadas em uma categorização e desenvolver um *parser* que seja capaz de interpretar o código-fonte para recuperar as respectivas informações. Uma desvantagem dessa técnica é que por se tratar de uma abordagem que trabalha com recursos dinâmicos, as páginas *Web*, há a necessidade de atualizar o algoritmo de extração de informação sempre que a estrutura da página for modificada.

A *API* da *WordNet*, chamada de *JWI* – "*MIT Java WordNet Interface*", é uma biblioteca desenvolvida em Java que permite acessar a base de dados da *WordNet* (FINLAYSON, 2009). Essa biblioteca permite que o *TOM Tag Recommender* faça a lematização dos termos das páginas *Web* e os deixem identificados de forma única. Conforme descrito na Seção 4.1 (seção referente à extração de termos de páginas *Web*), o acesso aos

[43] http://tomcat.apache.org/
[44] http://sourceforge.net/projects/delicious-java/
[45] http://jericho.htmlparser.net/

```
<h5 class="tag-chain-label">TAGS</h5>
<div class="tagdisplay">
   <ul class="tag-chain">
      <li class="off first">
         <a class="noplay" rel="tag" href="/marceloborth/ontology; ylv=3">
            <span class="tagItem">ontology</span>
         </a>
      </li>
      <li class="off">
         <a class="noplay" rel="tag" href="/marceloborth/mapping; ylv=3">
            <span class="tagItem">mapping</span>
         </a>
      </li>
      <li class="off last">
         <a class="noplay" rel="tag" href="/marceloborth/semantic; ylv=3">
            <span class="tagItem">semantic</span>
         </a>
      </li>
   </ul>
</div>
```

Figura 21. Código-fonte de uma página do sistema Delicious.

dados da *WordNet* é de grande importância para o processo de recomendação de *tags* deste trabalho e, também, para a emergência de ontologias. A sequência de execução dessas tecnologias pode ser vista na Figura 22.

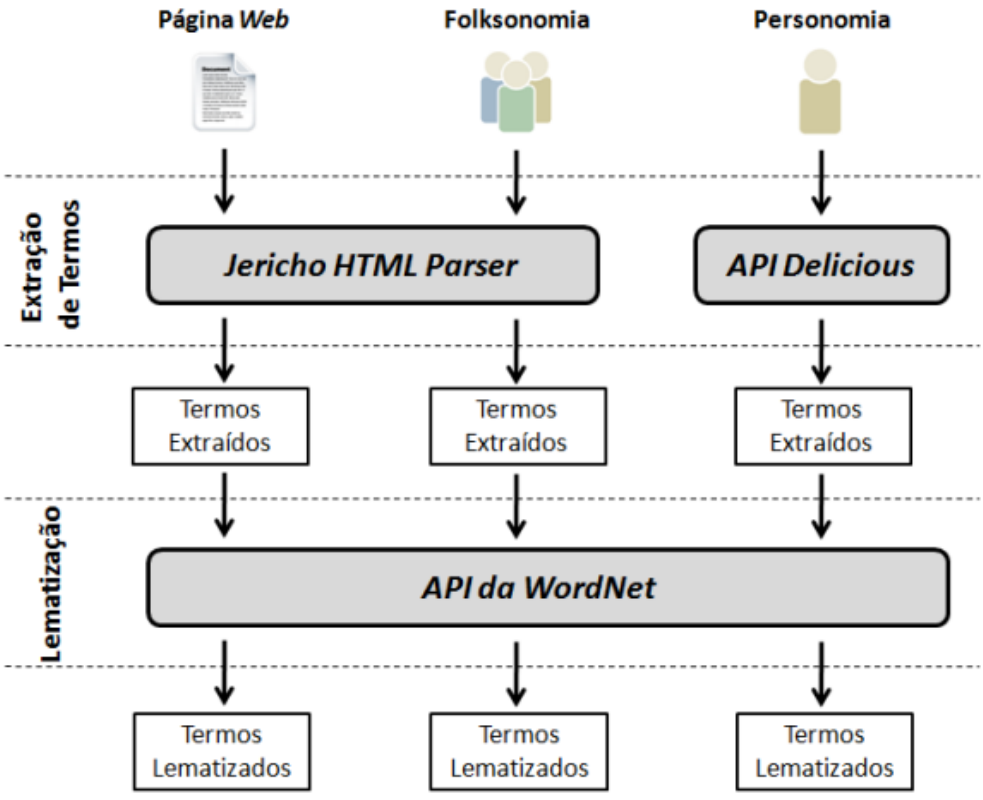

Figura 22. Sequência da aplicação para as principais APIs utilizadas neste trabalho.

Dada uma visão geral sobre as tecnologias utilizadas no desenvolvimento dessa aplicação, torna-se possível compreender de forma mais eficaz os aspectos gerais de implementação.

5.2. Visão Geral da Implementação

A abordagem apresentada no Capítulo 4 foi usada como um guia para a construção de cada funcionalidade do *TOM Tag Recommender*. Essas funcionalidades foram agrupadas em três subsistemas que compõem a arquitetura da aplicação da recomendação de *tags*, conforme ilustra a Figura 23.

A seguir, os subsistemas que compõem a aplicação desenvolvida são brevemente descritos:

- ***TagOntologyManager* (*TOM*)**: Sua responsabilidade é recuperar (ou extrair) um conjunto de *tagging* de um usuário ou um conjunto de termos de um recurso e, construir a estrutura semântica de uma ontologia com conceitos e relações. O *TOM* é responsável pela criação e gerenciamento da ontologia de um usuário. Esse subsistema teve sua implementação criada por Basso (2010) e adaptada pelo autor deste trabalho;
- ***Scrapping***: Esse subsistema é responsável pela extração e manipulação dos textos das páginas *Web*. Esse subsistema foi desenvolvido pelo autor deste trabalho; e
- ***TOM Tag Recommender***: Esse subsistema é responsável por gerenciar todo o ciclo de vida de uma recomendação de *tags*. Esse subsistema realiza todas as

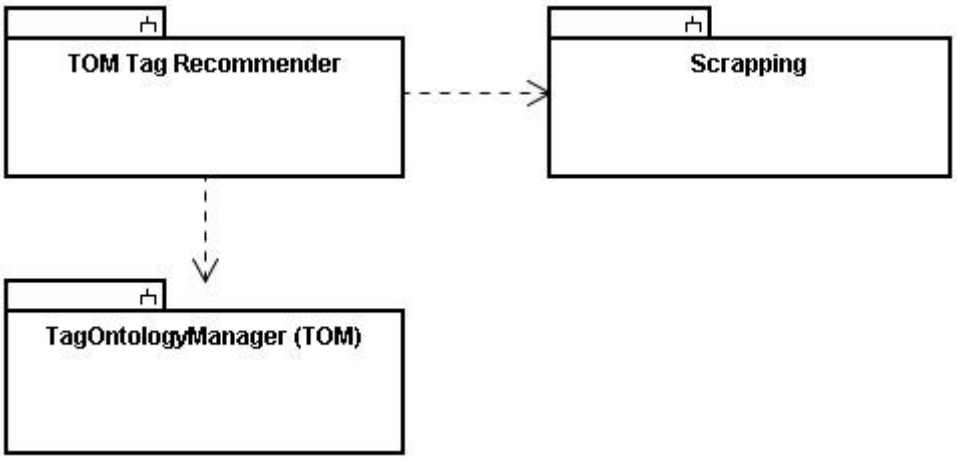

Figura 23. Subsistemas do projeto do TOM Tag Recommender.

chamadas aos métodos de extração e manipulação de textos das páginas *Web* (subsistema *Scrapping*) e para a geração das ontologias para as três fontes de informação (subsistema *TOM*). Esse subsistema foi desenvolvido pelo autor deste trabalho e depende dos subsistemas *Scrapping* e *TOM*.

A fim de proporcionar um baixo acoplamento entre as entidades do mecanismo de recomendação desenvolvido (*TOM Tag Recommender*) com as entidades e processos já existentes no *TOM*, a integração dos subsistemas é realizada por meio da classe `TagRecommendation`, conforme ilustrado no modelo conceitual da Figura 24. Essa classe é a responsável por gerenciar todo o processamento da recomendação de *tags*, bem como instanciar e gerenciar o ciclo de vida da ontologia para cada fonte de informação (personomia, folksonomia e página *Web*). Para gerar as ontologias é utilizada a classe `RelationProcessor` do subsistema *TOM* (representado em cinza no modelo conceitual).

O modelo conceitual exibe as principais entidades envolvidas no desenvolvimento do projeto da recomendação de *tags*, sendo que as representadas na cor branca modelam a aplicação desenvolvida neste estudo e a representada na cor cinza mostra a classe que fornece o acesso para a geração das ontologias do sistema *TOM*. Além disso, é possível perceber no modelo que a classe `TagRecommendation` é a entidade central e a mais importante de todo o modelo conceitual. A entidade que possibilita a extração das informações contidas nos

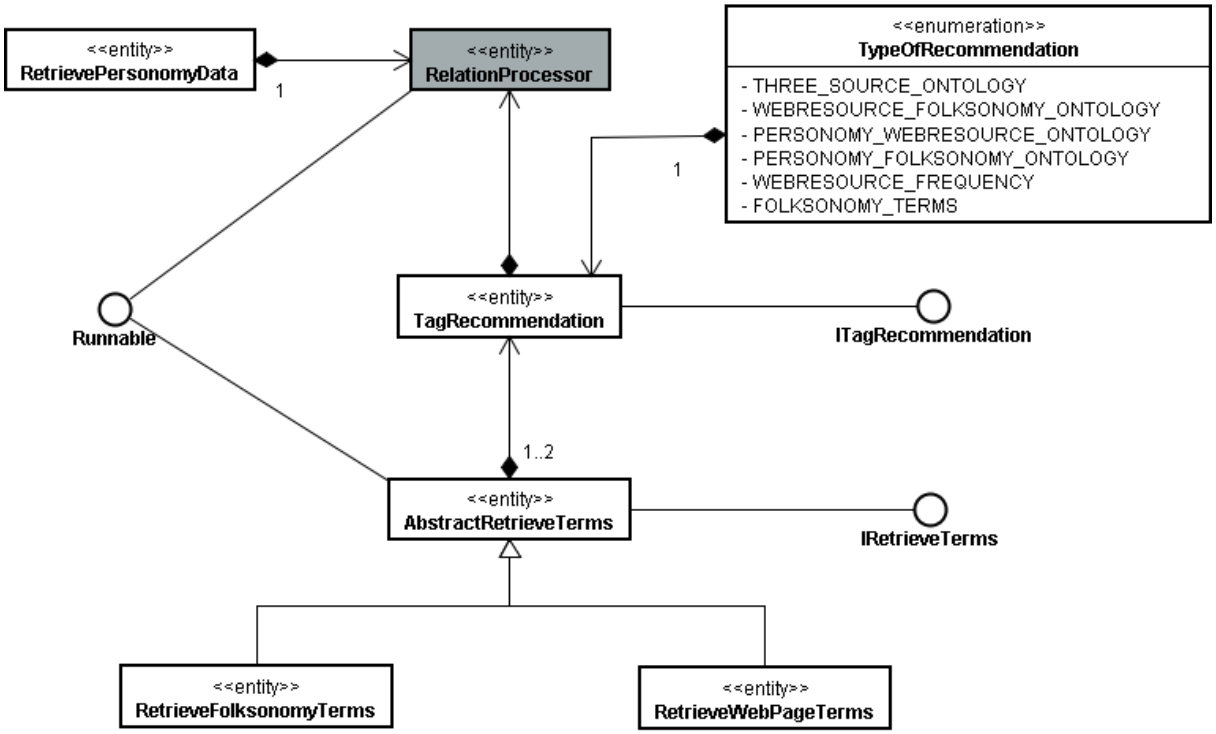

Figura 24. Modelo conceitual da recomendação de tags

recursos é a `AbstractRetrieveTerms`, podendo ser derivada para `RetrieveWebPageTerms`, `RetrieveFolksonomyTerms` ou `RetrievePersonomyTerms`, conforme o tipo de fonte de informação utilizada na respectiva instância do objeto. A interface `ITagRecommendation` disponibiliza os métodos necessários para que uma aplicação externa possa usufruir do processo de recomendação.

Uma vez que o foco deste trabalho é recomendar *tags* a partir da intersecção entre três tipos de fontes de informação, é possível gerar recomendações diferenciadas mediante uma combinação das fontes. Para tanto, a entidade `TypeOfRecommendation` fornece os tipos de recomendações possíveis. Esses diferentes tipos de recomendações são detalhados a seguir.

- *THREE_SOURCE_ONTOLOGY*: a recomendação de *tags* é realizada baseando-se nos três tipos de fontes de informação, conforme a proposta detalhada no Capítulo 4;
- *WEBRESOURCE_FOLKSONOMY_ONTOLOGY*: a recomendação de *tags* é realizada baseando-se na proposta deste trabalho, no entanto, são utilizados apenas dois tipos de fontes, a página *Web* e a folksonomia;
- *PERSONOMY_FOLKSONOMY_ONTOLOGY*: a recomendação de *tags* é realizada baseando-se na proposta deste trabalho, no entanto, são utilizados apenas dois tipos de fontes, a personomia e a folksonomia;
- *PERSONOMY_WEBRESOURCE_ONTOLOGY*: a recomendação de *tags* é realizada baseando-se na proposta deste trabalho, no entanto, são utilizados apenas dois tipos de fontes, a personomia e a página *Web*;
- *WEBRESOURCE_FREQUENCY*: a recomendação de *tags* é realizada a partir dos termos mais frequentes de uma página *Web*; e
- *FOLKSONOMY_TERMS*: a recomendação de *tags* é realizada a partir das *tags* mais utilizadas por outros usuários a um determinado recurso em um sistema baseado em *tagging* – folksonomia.

Para um melhor entendimento da aplicação que realiza a recomendação de *tags*, criamos um algoritmo que ilustra de forma abstrata as principais etapas da abordagem desenvolvida (Figura 25). A partir dos dados de entrada (uma lista com as categorizações realizadas pelo usuário (*tagging*) e a *URL* da página que será categorizada (*url*)) é recomendado um conjunto *tags*.

Input data:
- User categorizations – **tagging**
- URL of the Web page – **url**

Result: A set of tags: **RecommendedTags**

begin

/* Step 1 – Extract terms from the Web Page */
$WebPage_{Terms}$ ← extractWebPageTerms(url)
foreach Term ∈ $WebPage_{Terms}$ **do**
Term ← LexicalProcess (Term)
Term ← Normalization (Term)
Term ← TitleSimilarityDetection (Term)

/* Step 2 – Retrieve terms from the Folksonomy */
$Folksonomy_{Terms}$ ← retrieveFolksonomyTerms (url)
foreach Term ∈ $Folksonomy_{Terms}$ **do**
Term ← LexicalProcess (Term)

/* Step 3 – Retrieve terms from the User Personomy */
$Personomy_{Terms}$ ← tagging

/* Step 4 – Ontology Generation */
$WebPage_{Terms}$ ← ParetoPrinciple ($WebPage_{Terms}$)
$Folksonomy_{Terms}$ ← ParetoPrinciple ($Folksonomy_{Terms}$)
$Personomy_{Terms}$ ← ParetoPrinciple ($Personomy_{Terms}$)
$Ontology_{WebPage}$ ← generateOntology ($WebPage_{Terms}$)
$Ontology_{Folksonomy}$ ← generateOntology ($Folksonomy_{Terms}$)
$Ontology_{Personomy}$ ← generateOntology ($Personomy_{Terms}$)

/* Step 5 – Mapping the ontologies */
foreach $Tag_{Personomy}$ ∈ $Ontology_{Personomy}$ **do**
$Synset_{Personomy}$ ← getSynset ($Tag_{Personomy}$)
foreach $Tag_{WebPage}$ ∈ $Ontology_{WebPage}$ **do**
$Synset_{WebPage}$ ← getSynset ($Tag_{WebPage}$)
if $Synset_{Personomy}$ = $Synset_{WebPage}$ **then**
mapSynsets ($Tag_{Personomy}$, $Tag_{WebPage}$)
foreach $Tag_{Folksonomy}$ ∈ $Ontology_{Folksonomy}$ **do**
$Synset_{Folksonomy}$ ← getSynset ($Tag_{Folksonomy}$)
if $Synset_{Personomy}$ = $Synset_{Folksonomy}$ **then**
mapSynsets ($Tag_{Personomy}$, $Tag_{Folksonomy}$)

/* Step 6 – Recommend tags with highest priority */
RecommendedTags ← getStrongerTags()

end

Figura 25. Algoritmo descrevendo o processo de recomendação de tags.

De fato, esta aplicação foi desenvolvida com o objetivo de ser acoplada ao sistema *TagManager* (DA SILVA, 2009). Por essa razão, a próxima seção mostra como esta aplicação está integrada ao *TagManager*.

5.3. A Integração do *TOM Tag Recommender* ao *TagManager*

O *TagManager* (DA SILVA, 2009) tem por objetivo auxiliar o usuário no gerenciamento de sua personomia para vários sistemas baseados em *tagging*, a qual é criada a partir da utilização desses vários sistemas. Além disso, o *TagManager* fornece uma funcionalidade que possibilita a limpeza e organização das *tags* presentes na personomia, denominado de *TagTydier* (CÔGO e DA SILVA, 2008). Da Silva (2009) afirma que esse gerenciamento traz melhorias que podem refletir nas personomias para os vários sistemas que o usuário utiliza. Portanto, os benefícios que o *TagManager* terá com o sistema desenvolvido são provenientes da melhor estruturação que a recomendação de *tags* semânticas pode fornecer ao utilizar um sentido associado a cada *tag* no momento da categorização, possibilitando uma melhor organização do vocabulário do usuário. Por consequência disso, a tarefa de recuperação de informação poderá se tornar mais eficaz, tanto para quesitos individuais quanto para sociais (folksonomia), uma vez que o *TagManager* também disponibiliza um espaço para pesquisar objetos de outros usuários.

As entidades utilizadas para integrar o *TOM Tag Recommender* ao *TagManager* são a `TagRecommendation` e a `ITagRecommendation` – conforme ilustrado na Figura 24. É por meio dessa interface que o *TagManager* tem o conhecimento dos métodos e os tipos de recomendações disponíveis para o uso do *TOM Tag Recommender*. Assim, pelo fato da desta aplicação fornecer um serviço, o *TagManager* integra-se ao *Tom Tag Recommender* conforme ilustra a Figura 26. Nessa ilustração, o elemento ao lado esquerdo representa o *TagManager* e os subsistemas relacionados internamente são outras pesquisas desenvolvidas pelo Grupo de Sistemas Interativos e Inteligentes[46] (GSII) durante outros dois trabalhos de mestrado (DA SILVA, 2009) (BASSO, 2009) e um trabalho de graduação (CÔGO, 2009), respectivamente, *TagManager*, *TagOntologyManager* e *TagTydier*. Os elementos ao lado direito representam as aplicações externas que são acessadas por nossos projetos como, a *WordNet*, o *Delicious*, o *Youtube*, etc.

[46] O laboratório do grupo GSII está fisicamente presente no Departamento de Informática da Universidade Estadual de Maringá (UEM).

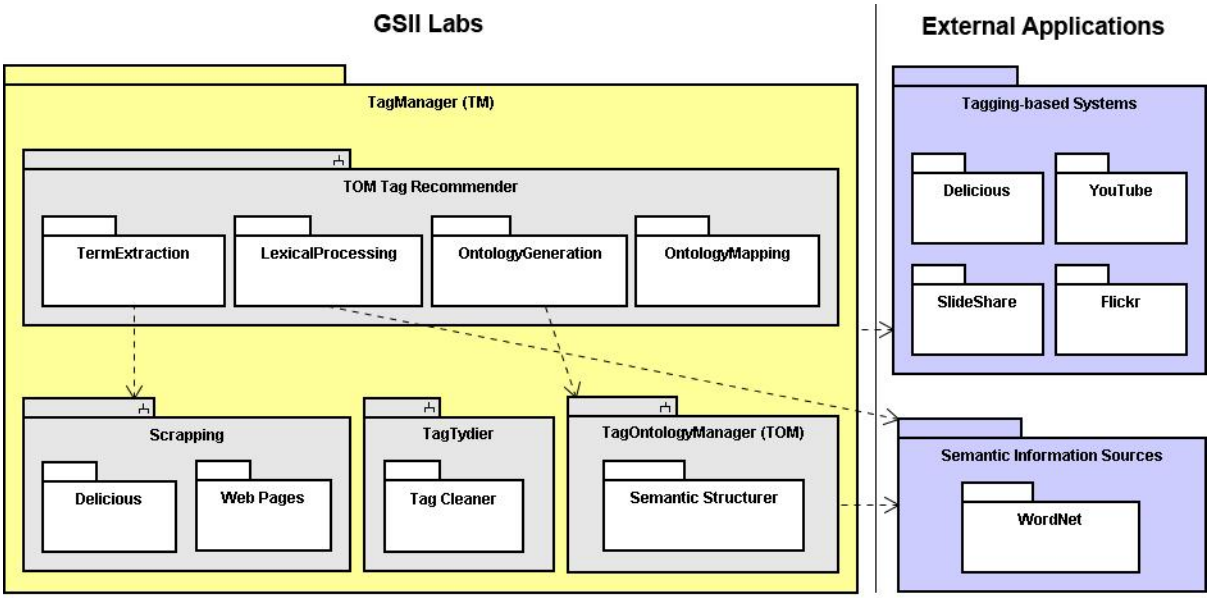

Figura 26. Exemplo da integração do TagManager e o TOM Tag Recommender.

5.4. Tempo de Processamento da Recomendação de *Tags* perante o algoritmo desenvolvido

Para gerar a recomendação de *tags* o algoritmo possui alguns processos, sendo eles:

- extração e processamento léxico dos termos de uma página *Web*;
- recuperação dos termos da folksonomia;
- recuperação dos termos da personomia;
- geração das três ontologias;
- mapeamento entre as ontologias; e
- seleção dos termos que serão recomendados.

Para realizar todas as etapas e obter o resultado da recomendação de *tags* é necessário aguardar pelo tempo de processamento que, em média, é de 25 segundos. Embora o tempo de processamento seja alto, é necessário ressaltar que não foi realizada nenhuma otimização no algoritmo de recomendação de *tags* e de geração das ontologias. Acreditamos que isto seja o fato responsável pelo baixo desempenho do algoritmo, uma vez que se tivéssemos uma equipe especializada para o desenvolvimento dos algoritmos, o tempo de resposta da recomendação possivelmente seria próximo dos sistemas atuais.

Por outro lado, ao avaliar o tempo de processamento da recomendação de *tags* do sistema *Delicious*, foi percebido que sua recomendação é gerada, em média, em 1,5 segundos. Acreditamos que o sistema *Delicious* usa técnicas de *cache* para obter esse baixo tempo de

resposta na sua recomendação, realizando, assim, a indexação das *tags* para cada página *Web* categorizada. A principal vantagem desse processo de *cache* consiste em evitar o acesso ao banco de dados - que pode ser demorado -, armazenando as informações em meios de acesso rápidos.

Nesse capítulo foi apresentada uma visão geral sobre os aspectos de implementação do projeto desenvolvido que visa uma abordagem de recomendação de *tags* semânticas. No próximo capítulo são apresentados os experimentos realizados com alguns usuários e os resultados obtidos a partir da execução do sistema desenvolvido.

Capítulo

6

Experimentos e Resultados

Esse capítulo apresenta três experimentos cujo objetivo foi verificar o percentual de aceitação das *tags* recomendadas aos usuários pela abordagem de recomendação de *tags* proposta, O primeiro experimento avalia, de forma automática, se as *tags* recomendadas por esta abordagem foram previamente utilizadas por outros usuários no sistema *Delicious*, isto é, as *tags* contidas na folksonomia para cada respectivo recurso *Web* analisado. O segundo experimento avalia a recomendação de *tags* com usuários, identificando o percentual de aceitação dos conceitos mais relevantes. O terceiro experimento avalia o percentual de aceitação das *tags* pelos usuários a partir dos conceitos mais relevantes (identificados por este algoritmo de recomendação de *tags*) e seu respectivo conceito hiperônimo. A aplicação desenvolvida – apresentada no Capítulo 5 – foi utilizada como instrumento para a execução desses experimentos.

6.1. Experimento 1: Avaliação da Recomendação de *Tags* em Comparação com a recomendação de *Tags* do Sistema *Delicious*

Esse experimento analisou aproximadamente 400 *bookmarks*, de forma automática, com o objetivo de medir a relevância das recomendações desta abordagem em comparação com as 30 *tags* mais utilizadas no sistema *Delicious* para os respectivos *bookmarks*. Esse sistema possui uma seção que exibe um histórico das 30 *tags* mais usadas previamente por outros usuários para um determinado *bookmark*.

O experimento abordou dois tipos de cenários relatados no Capítulo 3, sendo eles: a situação em que temos um novo usuário no sistema que não possui *tags* em sua personomia; e a situação em que temos um usuário experiente que possui um grande número de *tags* em sua personomia. Para os dois cenários, existiam informações na folksonomia, caso contrário, não haveria parâmetros para se realizar a comparação de forma automática. Dessa forma, todas as *tags* recomendadas por esta proposta foram comparadas com as *tags* presentes no espaço da folksonomia do sistema *Delicious*. Por isso, não foi possível considerar os outros dois cenários citados na Seção 3.1, nos quais os *bookmarks* não foram previamente categorizados no sistema *Delicious*.

Deste modo, vamos considerar que esta aplicação está recomendando *tags* para uma *URL* da *Wikipedia*. Nesse caso, o algoritmo desenvolvido neste trabalho irá gerar 10 *tags* de recomendação, as quais serão comparadas uma a uma com as 30 *tags* mais utilizadas pelos usuários do sistema *Delicious* para a mesma *URL*. Para cada *tag* recomendada é verificado se ela está presente entre as 30 *tags* mais utilizadas pelos usuários. Ao final das comparações para todas as *URLs*, é obtido o percentual das *tags* encontradas no sistema *Delicious* para cada uma das 10 *tags* recomendadas por este algoritmo.

No gráfico ilustrado na Figura 27, o eixo horizontal exibe a quantidade de *tags* recomendadas, enquanto que o eixo vertical exibe o percentual de acerto da recomendação de *tags* em comparação com o sistema *Delicious*. Para as 10 *tags* recomendadas por esta proposta, o percentual de identificação no *Delicious* foi alto, uma vez que para um usuário novato a taxa de identificação da *tag* menos relevante para a *tag* mais relevante variou de 83% a 94%. Apesar da diferença entre os dois tipos de usuários analisados na recomendação, pode ser observado que os resultados foram semelhantes para ambos.

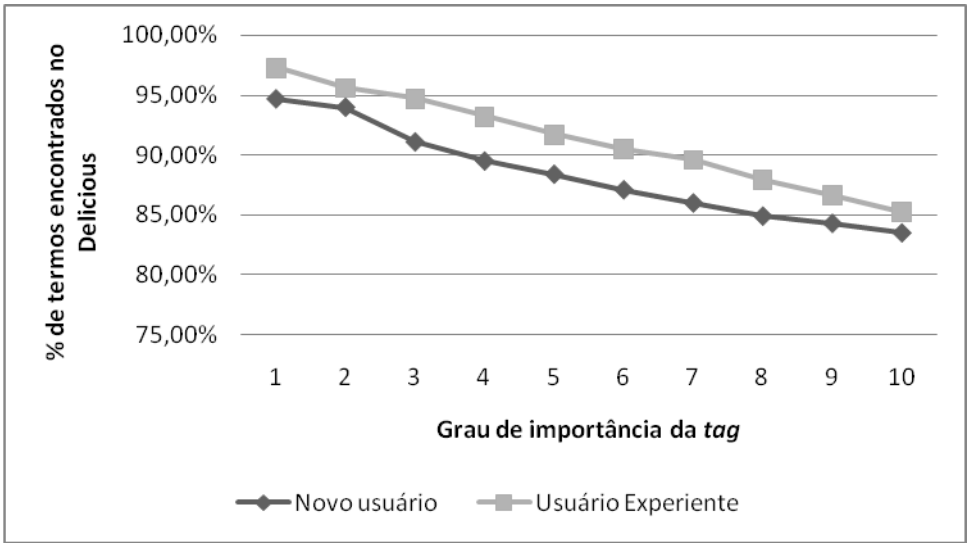

Figura 27. Avaliação da recomendação de tags em comparação com as tags presentes no sistema Delicious (BORTH et al., 2010).

Esse resultado mostrou que esta abordagem pode gerar recomendações relevantes para um usuário no momento da categorização de uma página *Web*, pois os dados contidos no sistema *Delicious* fornecem uma visão bastante representativa por se tratar de informações geradas por uma comunidade de usuários. Contudo, pelo fato desse experimento ser automatizado, ele não é suficiente para assegurar que esta proposta recomenda *tags* que sejam bem aceitas pelos usuários, uma vez que, para evidenciar isso, seria necessário realizar experimentos com usuários reais. A fim de inferir resultados mais confiáveis, as próximas seções apresentam outros dois experimentos com usuários reais.

6.2. Experimento 2: Avaliação da Recomendação de *Tags* com Usuários

A metodologia explicada nessa seção está diretamente relacionada a um processo de experimentação, como o estabelecido por Juristo e Moreno (2000). Esses autores discutem que um experimento deve ser dividido nas seguintes atividades (ver Figura 28):

1. Definição dos objetivos (é gerado um conjunto de objetos);
2. Planejamento do experimento (é produzido o projeto do experimento);
3. Execução do experimento (irá produzir os dados a serem avaliados); e
4. Análise dos resultados (validará ou não a hipótese definida na primeira atividade).

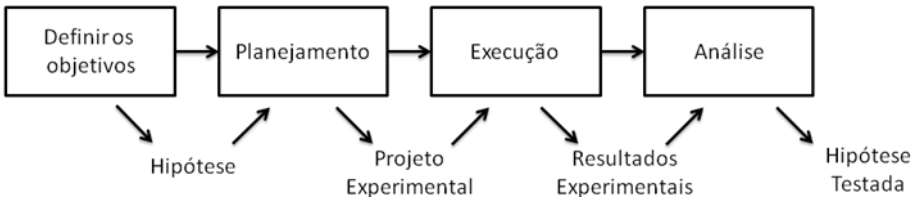

Figura 28. Processo de Experimentação (JURISTO E MORENO, 2000).

Por essa razão, para guiar esse experimento foi utilizado o *framework* DECIDE[47] (PREECE, 2005). Esse *framework* identifica as principais questões que precisam ser consideradas quando se tem por objetivo desenvolver uma avaliação. O *framework* DECIDE especifica algumas tarefas importantes, tais como: determinar as metas e objetivos da avaliação, explorando questões gerais e específicas; e identificar as questões práticas abordadas como a seleção dos participantes, as questões éticas, os modos de avaliação, a interpretação e a apresentação dos resultados. Para executar a recomendação de *tags* foi desenvolvida uma aplicação *Web*, a qual foi implantada e executada na própria estrutura do laboratório do Grupo de Sistemas Interativos Inteligentes. As pessoas que realizaram o experimento foram os integrantes do GSII e alguns alunos do curso de mestrado. Logo, o perfil desses usuários foi exclusivamente da área de exatas.

O objetivo desse experimento foi medir o percentual de aceitação da recomendação de *tags* pelos usuários. Nesse experimento, houve três fontes de informação disponíveis para a análise: (1) os termos do recurso *Web*; (2) as *tags* da folksonomia do sistema *Delicious*; e (3) os termos da personomia. A partir de uma combinação dentre as fontes, foi possível gerar várias abordagens. Assim, foi possível comparar o percentual de aceitação das *tags* recomendadas desta proposta com cada abordagem gerada. As abordagens avaliadas foram:

- **Abordagem 1**: Recomendação a partir da combinação entre (1), (2) e (3). Essa abordagem é representada pela descrição "*THREE_SOURCE_ONTOLOGY*". Essa abordagem é baseada neste estudo;
- **Abordagem 2**: Recomendação a partir da combinação entre (1) e (2). Essa abordagem é representada pela descrição "*WEBRESOURCE_FOLKSONOMY_ONTOLOGY*";
- **Abordagem 3**: Recomendação a partir da combinação entre (2) e (3). Essa abordagem é representada pela descrição "*PERSONOMY_FOLKSONOMY_ONTOLOGY*";

47 Em inglês, o acrônimo DECIDE é formado pelas iniciais das palavras *determine*, *explore*, *choose*, *identify*, *decide* e *evaluate* (respectivamente: determine, explore, escolha, identifique, decida e avalie).

- **Abordagem 4**: Recomendação a partir das *tags* mais frequentes da folksonomia (2). Essa abordagem utiliza as *tags* do sistema *Delicious* e é representada pela descrição "*FOLKSONOMY_TERMS*". Para esse experimento, essa foi considerada uma abordagem de referência;
- **Abordagem 5**: Recomendação a partir dos termos mais frequentes de uma página *Web* (1). Essa abordagem é representada pela descrição "*WEBRESOURCE_FREQUENCY*"; e

Nesses experimentos cada abordagem foi avaliada por duas pessoas, sendo que, cada usuário realizou as categorizações para apenas uma abordagem. Não houve nenhuma possibilidade de identificar qual abordagem o usuário estava avaliando. Cada participante realizou 30 categorizações a partir de *URLs* predefinidas pelo autor deste trabalho, sendo que 15 eram da área da computação e 15 de conhecimentos gerais (as *URLs* selecionadas encontram-se no Apêndice A). Dentre as 15 *URLs* de cada categoria, 4 eram páginas da *Wikipedia*. Dessa forma, foi obtido um grupo de 8 páginas *Web* da *Wikipedia* que abrangeram tanto páginas de computação quanto de conhecimentos gerais.

Durante a execução, nenhum *log* de usabilidade ou de comportamento foi registrado, visto que eram as abordagens de recomendação que estavam sendo avaliadas e não os participantes ou a interface. Para cada uma das páginas *Web* categorizadas foram recomendadas 10 *tags*. O participante teve a liberdade de selecionar as *tags* recomendadas e/ou informar *tags* adicionais às sugeridas. Entretanto, o algoritmo do experimento consegue distinguir quais foram recomendadas e selecionadas pelo usuário e, quais as que foram adicionadas que não estavam na recomendação de *tags*.

Entendemos que cada *tag* possui um grau de relevância, dessa forma elas foram distribuídas na janela de categorização com base na sua relevância para o recurso *Web* de forma decrescente, *i.e.* da esquerda para a direita. Dessa forma, a *tag* mais relevante é a primeira e a menos relevante é a décima. A Figura 29 ilustra uma recomendação de 10 *tags*

Figura 29. Exemplo do grau de relevância de uma tag em uma recomendação de tags.

para uma página *Web*. Nessa ilustração, a *tag* "*ubiquitous*" (primeira *tag* da esquerda para a direita) será a mais relevante da recomendação, seguida pela *tag* "*computing*" e, assim, sucessivamente para as demais *tags*.

Ao término de todas as categorizações, cada participante preencheu dois questionários: Q1, para coleta de dados referente à sua experiência – necessário para analisar a influência dos resultados; e Q2, para identificar as dificuldades e as ajudas solicitadas na utilização do experimento. Ambos os questionários podem ser vistos no Apêndice B.

6.2.1. Condução do Experimento

O experimento de avaliação da proposta de recomendação de *tags* deste trabalho foi composto de três passos principais. A Figura 30 mostra um diagrama de atividades que representa as interações entre o usuário e o sistema.

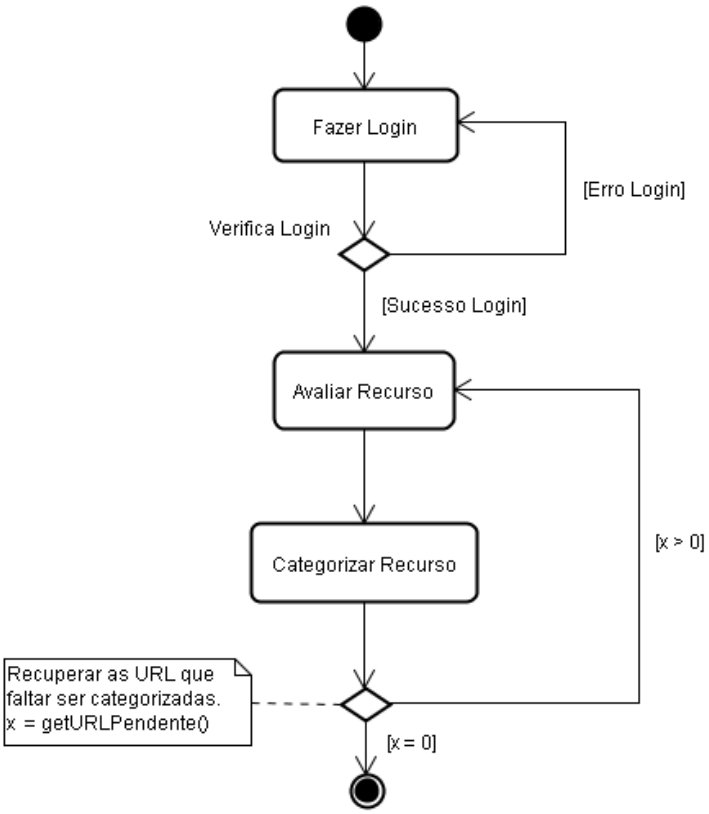

Figura 30. Diagrama de atividade do experimento.

Os passos são descritos a seguir:

1. O participante deverá fazer o *login* no sistema informando seu nome de usuário;

2. O participante abrirá a *URL* que aparecerá na janela de recomendação, analisará seu conteúdo e identificará as palavras-chave interessantes para a representação de seu conteúdo dentro do contexto de seu interesse; e
3. O participante selecionará/informará as *tags* que melhor representam seu interesse na janela da categorização, finalizando, assim, a categorização no botão salvar.

Ao total, houve 30 iterações entre o participante e o sistema. Em cada iteração foi realizada a categorização de uma *URL* diferente. Para não tornar o processo cansativo, cada usuário teve um período de quatro semanas para realizar todas as categorizações, não sendo necessário iniciar e terminar o experimento no mesmo momento. O sistema salvava as categorizações efetuadas em um banco de dados, possibilitando, assim, que o usuário continuasse da categorização de onde havia parado. Além disso, não era possível que o participante alterasse as categorizações previamente efetuadas. Detalhes sobre a modelagem do banco de dados são apresentados no Apêndice C. O protótipo da janela de recomendação é mostrado na Figura 31.

Figura 31. Exemplo da janela da recomendação de tags para a categorização de uma página Web.

Por fim, antes da execução do experimento, os usuários foram informados sobre alguns assuntos relevantes, tais como: objetivos do experimento, o tempo de duração aproximado e os tipos de dados que seriam coletados.

Com o objetivo de não beneficiar nenhuma abordagem e pelo fato de todos os usuários terem pelo menos alguma informação em sua personomia, eles foram separados em dois grupos: um grupo com várias informações em sua personomia e outro grupo com poucas

informações. Assim, cada abordagem foi avaliada por dois usuários, um de cada grupo, selecionados aleatoriamente.

Além das abordagens analisadas, esse experimento contemplou uma abordagem extra, na qual não houve recomendações de *tags* nas categorizações. Dessa forma, cada *tag* associada ao recurso *Web* na categorização foi identificada e informada pelo próprio usuário sem qualquer auxílio. Essa abordagem é referenciada por "*WITHOUT_RECOMMENDATION*".

Após a coleta dos dados, calculamos a quantidade média de *tags* utilizadas em cada categorização. A Figura 32 mostra os resultados para cada abordagem. Nela pode ser observado que o uso conjunto da folksonomia, da personomia e/ou da página *Web* melhora a quantidade média de *tags* utilizadas em uma categorização. Em particular, a abordagem proposta neste trabalho, a "*THREE_SOURCE_ONTOLOGY*", foi a que obteve melhor resultado, recebendo em média de 4,51 *tags* por categorização, o que representa um aumento de 34% em relação à abordagem "*FOLKSONOMY_TERMS*", a qual é tomada como padrão de comparação neste trabalho. Para as demais abordagens, a média de *tags* utilizadas por categorização decresce gradativamente, com exceção da abordagem "*WITHOUT_RECOMMENDATION*", que praticamente se iguala a quantidade de *tags* utilizadas pelos usuários da abordagem "*FOLKSONOMY_TERMS*". Além disso, pode-se ver que o uso do recurso *Web* somente não traz bons resultados. Isto mostra claramente o valor da recomendação social pelo uso da folksonomia e do valor da personalização da recomendação pelo uso da personomia.

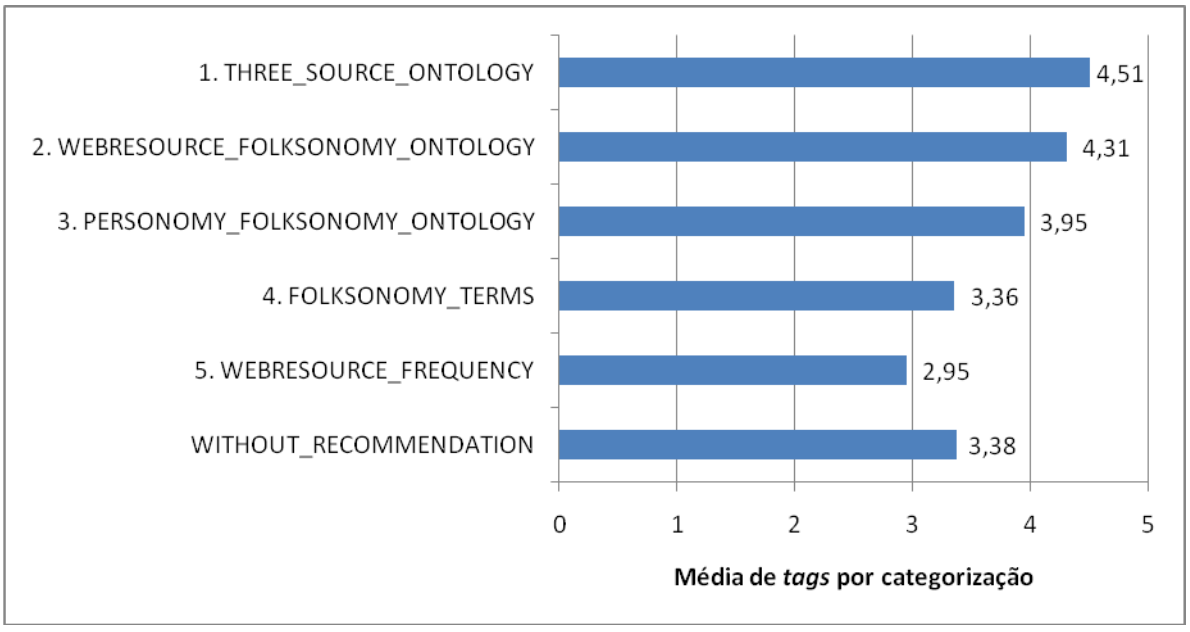

Figura 32. Média de tags utilizadas em cada categorização.

A diferença da quantidade de *tags* utilizada por cada abordagem com a abordagem de referência (o sistema *Delicious*, representado pela abordagem "*FOLKSONOMY_TERMS*") pode ser vista na Tabela 7. Por exemplo, a abordagem "*THREE_SOURCE_ONTOLOGY*" possui 34% a mais de *tags* em cada categorização que o sistema *Delicious*, enquanto que a abordagem "*WEBRESOURCE_FREQUENCY*" possui 12% a menos.

Tabela 7. Diferença entre cada abordagem e o sistema Delicious perante a quantidade média de tags utilizadas por categorização.

Abordagem	**Diferença**
THREE_SOURCE_ONTOLOGY	+ 34%
WEBRESOURCE_FOLKSONOMY_ONTOLOGY	+ 28%
PERSONOMY_FOLKSONOMY_ONTOLOGY	+ 17%
WEBRESOURCE_FREQUENCY	– 12%
FOLKSONOMY_TERMS	–
WITHOUT_RECOMMENDATION	+ 0,5%

Uma vez que o usuário possui a liberdade de associar qualquer *tag* em uma categorização, é possível que ele selecione as *tags* recomendadas ou informe *tags* adicionais às que foram recomendadas. Quando o usuário informa *tags* adicionais em uma categorização, significa que a recomendação não foi suficientemente interessante para suprir todas as necessidades momentâneas de interesse do usuário. Por essa razão, avaliamos a quantidade média de categorizações nas quais os usuários informaram *tags* adicionais.

A Figura 33 mostra o percentual de categorizações que utilizaram *tags* adicionais às recomendadas. Analisando o gráfico, percebe-se que a abordagem que utiliza os dados da personomia, folksonomia e da página *Web* ("*THREE_SOURCE_ONTOLOGY*") oferece *tags* mais interessantes aos usuários que as demais abordagens, pois somente em 51% das categorizações foram adicionadas *tags* complementares às recomendadas, o que corresponde a uma diferença de 12% em relação à abordagem "*FOLKSONOMY_TERMS*". Além disso, pode ser observado que para 80% das categorizações da abordagem que utiliza apenas a frequência dos termos de uma página *Web* ("*WEBRESOURCE_FREQUENCY*") houve a necessidade de informar *tags* adicionais, um percentual bastante alto comparado ao gerado pela abordagem "*THREE_SOURCE_ONTOLOGY*", o que confirma nossa observação na figura anterior da má qualidade desta abordagem. O percentual de *tags* adicionais de 76% para a abordagem "*WEBRESOURCE_FOLKSONOMY_ONTOLOGY*" mostra novamente o valor da personalização na recomendação, pois a falta do recurso personomia levou a uma maior necessidade de *tags* adicionais.

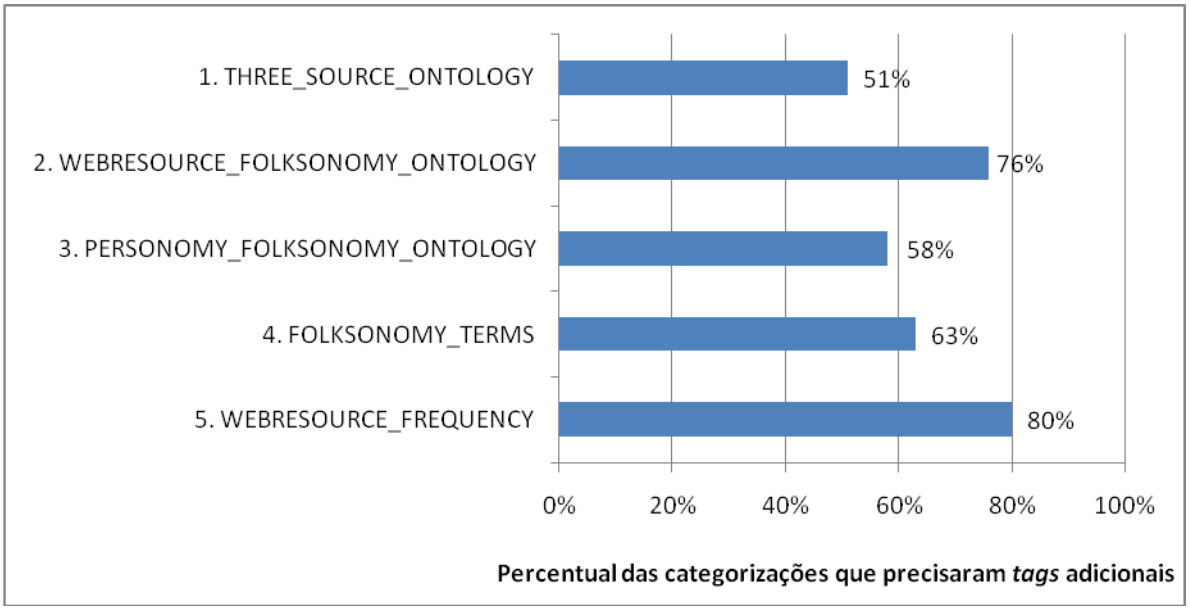

Figura 33. Percentual de categorizações que foram utilizadas tags adicionais às recomendadas pelos usuários.

Para identificar o percentual de aceitação das *tags* recomendadas de cada abordagem, foi necessário analisar as *tags* recomendadas que foram utilizadas pelos usuários nas recomendações. Para obter esse percentual, foi utilizado a Equação 2, a qual divide a quantidade de *tags* que o usuário utilizou em uma categorização (apenas entre as *tags* recomendadas) pelo total de *tags* recomendadas. Para exemplificar essa situação, considere que um usuário está categorizando uma página *Web* na qual foi recomendada 10 *tags*. Após o usuário analisar a relevância das *tags* recomendadas, ele decide categorizar a página *Web* com apenas quatro dentre as dez *tags* que foram sugeridas. Nesse caso, o percentual de aceitação da recomendação é de 40%.

$$\text{Percentual de aceitação} = \frac{\text{Quantidade de tags utilizadas da recomendação}}{\text{Quantidade de tags recomendadas na categorização}}$$

Equação 2. Cálculo do percentual de aceitação de uma recomendação de tags.

A Figura 34 mostra, de forma geral, o percentual de aceitação que cada abordagem obteve para 10 *tags* recomendadas aos usuários. Pode ser observado que a abordagem com o maior percentual é a "*THREE_SOURCE_ONTOLOGY*", obtendo uma aceitação de 40% dentre as *tags* recomendadas. Percebe-se, também, que a taxa de aceitação vai decrescendo entre cada abordagem até atingir 27% para a abordagem "*FOLKSONOMY_TERMS*" e 21% para a abordagem "*WEBRESOURCE_FREQUENCY*".

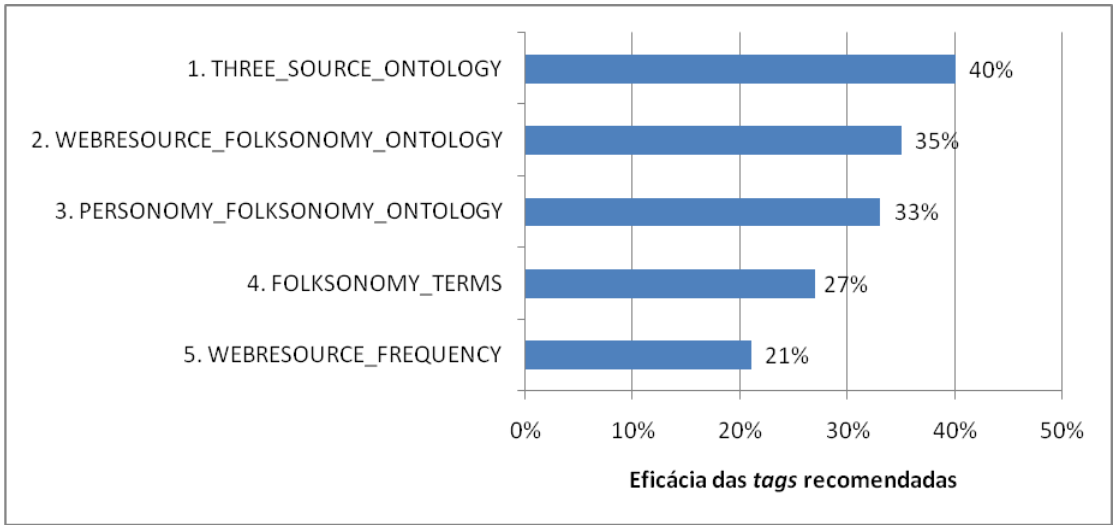

Figura 34. Percentual de aceitação das abordagens de recomendação de tags.

Com o objetivo de detalhar os resultados do experimento foi analisado separadamente o percentual de aceitação de cada *tag* com base no seu grau de relevância. Assim, em primeira instância, a Figura 35 mostra as *tags* aceitas para as recomendações com base na sua relevância. O grau de relevância varia de 1 a 10, sendo que 1 é o mais relevante e 10 o menos relevante. Nesse gráfico, o eixo "Abordagem" mostra as abordagens avaliadas (listadas na seção 6.2); o eixo "Relevância da *Tag*" mostra o grau de relevância da *tag* com base na recomendação realizada aos usuários; e o eixo "% de aceitação" mostra o percentual de *tags* que foram utilizadas pelos usuários com base no seu grau de relevância.

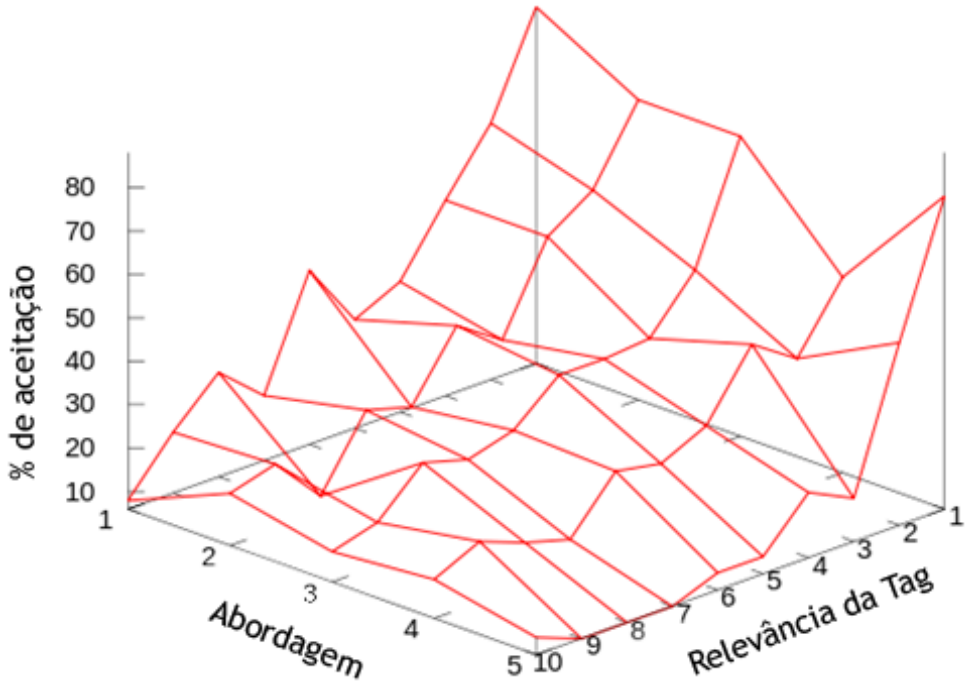

Figura 35. Tags aceitas na recomendação com base na importância da tag.

Ao analisar este gráfico, pode ser observado que, como era de se esperar, conforme aumenta o grau de relevância da *tag* recomendada, há uma tendência maior dessa *tag* ser mais interessante ao usuário. Esse comportamento não necessariamente segue um padrão, entretanto, há uma convergência para que isso ocorra. Por exemplo, na abordagem "*THREE_SOURCE_ONTOLOGY*", as *tags* de menor relevância possuem baixo percentual de aceitação, porém, conforme o grau de relevância aumenta a aceitação do usuário também aumenta, obtendo 88% de aceitação para a *tag* mais relevante nessa abordagem de recomendação. Para as demais abordagens o percentual de aceitação é gradativamente menor conforme seu grau de relevância.

A Figura 36 mostra o percentual de aceitação que cada abordagem obteve nas categorizações para os *websites* da computação, de conhecimentos gerais e da *Wikipedia*. Para a categoria de *websites* da computação é possível notar que a abordagem "*WEBRESOURCE_FREQUENCY*" obteve menos da metade da aceitação comparada com a abordagem "*THREE_SOURCE_ONTOLOGY*", a qual obteve o melhor percentual de

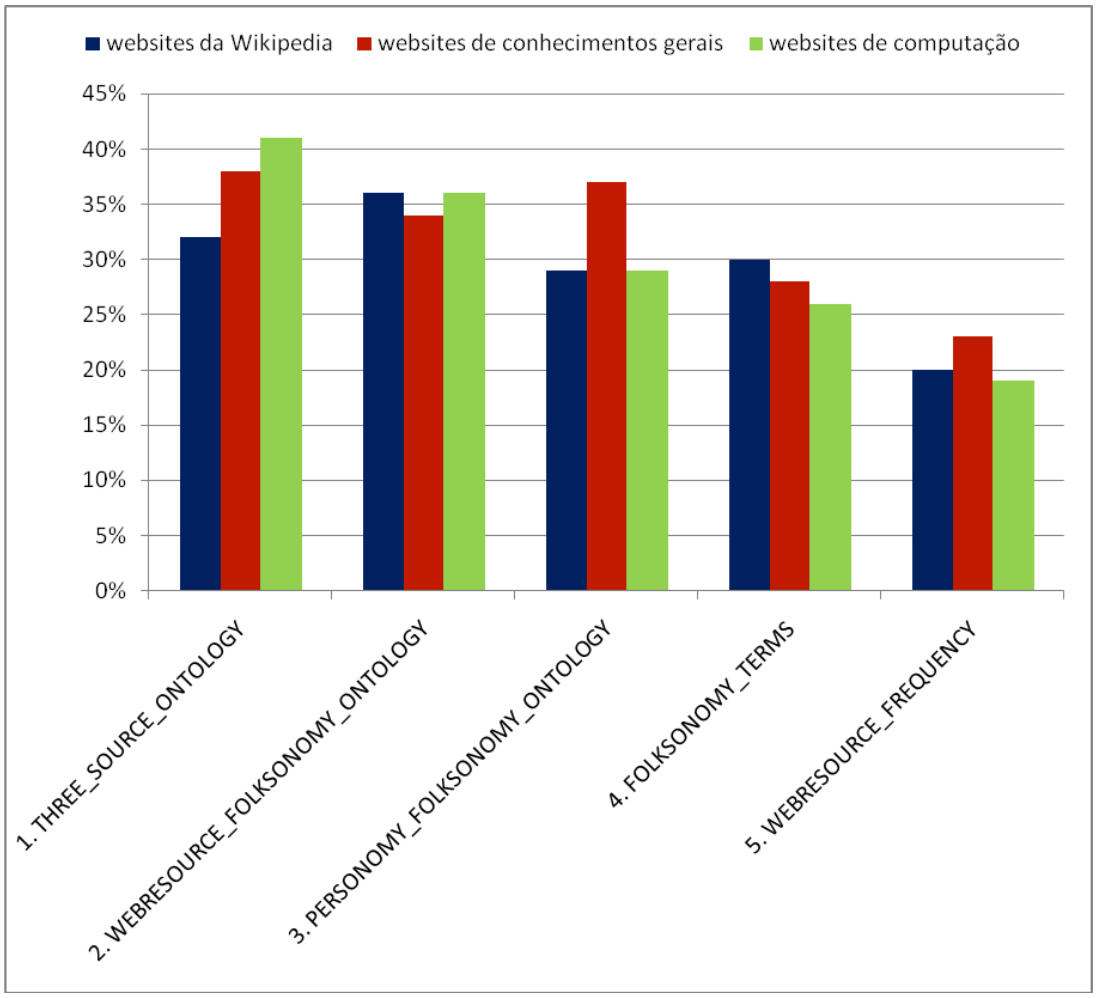

Figura 36. Detalhamento do percentual de aceitação para as abordagens avaliadas.

aceitação para essa categoria, obtendo 41% de aceitação pelos usuários, enquanto que a abordagem "*FOLKSONOMY_TERMS*" obteve 26% de aceitação. Para os *websites* de conhecimentos gerais, percebe-se que as abordagens "*THREE_SOURCE_ONTOLOGY*" e "*PERSONOMY_FOLKSONOMY_ONTOLOGY*" obtiveram praticamente o mesmo percentual de aceitação. Na abordagem "*FOLKSONOMY_TERMS*" o percentual de aceitação foi de 28%. Para os *websites* da *Wikipedia*, a abordagem que obteve a maior aceitação pelos usuários foi a "*WEBRESOURCE_FOLKSONOMY_ONTOLOGY*", obtendo o percentual de 36% de aceitação, enquanto que as abordagens "*THREE_SOURCE_ONTOLOGY*" e "*FOLKSONOMY_TERMS*" obtiveram 32% e 30%, respectivamente. Em comparação com as outras categorias de *websites* criadas, percebe-se que a abordagem "*THREE_SOURCE_ONTOLOGY*" não obteve a maior aceitação apenas para as páginas da *Wikipedia*, sendo que, nesse caso, ela obteve 4% a menos que a abordagem "*WEBRESOURCE_FOLKSONOMY_ONTOLOGY*". Logo, a partir da comparação entre as abordagens "*FOLKSONOMY_TERMS*" e "*THREE_SOURCE_ONTOLOGY*", percebe-se que a utilização da página *Web*, juntamente com a personomia e a folksonomia utilizada na abordagem "*THREE_SOURCE_ONTOLOGY*" pode gerar *tags* que tendem a ser mais aceitas pelos usuários.

A partir desse experimento, é possível sugerir que as *tags* fornecidas por esta proposta podem obter bons percentuais de aceitação pelos usuários na tarefa de categorização, principalmente quando a *tag* recomendada está entre as quatro primeiras (conforme ilustra a Figura 35), pois essas foram as *tags* mais aceitas pelos usuários. Além disso, a abordagem ("*THREE_SOURCE_ONTOLOGY*") obteve (i) o maior percentual de aceitação pelos usuários; (ii) foi a abordagem que, em média, mais agregou *tags* nas suas categorizações; (iii) obteve a menor probabilidade de ter uma *tag* adicional às recomendadas; (iv) obteve o maior percentual de aceitação pelos usuários para duas das três categorias criadas das *URLs* (os *websites* de computação e os *websites* de conhecimentos gerais). Acreditamos que esses percentuais de aceitação obtidos se justificam pelo fato de que esta abordagem realiza a análise das três possíveis fontes de informação no momento de uma categorização (a personomia, a folksonomia e a página *Web*), enquanto que as outras abordagens analisam apenas duas fontes ou, até mesmo, uma única fonte de informação, como é o caso do sistema *Delicious*. Por essa razão, é possível que esta abordagem tenha vantagens sobre as demais na tarefa de identificação das *tags* relevantes para a recomendação a uma determinada *URL*, pois cada fonte de informação fornece uma característica distinta da *URL*. Se compararmos nossa abordagem com o sistema *Delicious*, percebe-se que esta abordagem possui informações

adicionais para o processamento, como a personomia e a página *Web*. Além disso, nossa hipótese é a de que as informações complementares da página *Web*, juntamente com a folksonomia e a personomia influenciaram a escolha das *tags* pelos usuários, uma vez que, nesse caso há um grupo maior de informações (página *Web* e folksonomia) para direcionar à *tags* do próprio vocabulário do usuário.

Embora esta abordagem tenha sido a melhor dentre as avaliadas, ela ainda não tem um desempenho, em tempo computacional, aceitável para execução em uma aplicação *online*, visto que os internautas estão acostumados a um tempo de resposta rápido dos sistemas *Web*. Como visto na Seção 5.4, o tempo de processamento do algoritmo de recomendação de *tags* desenvolvido é em média 25 segundos por *URL* (sem otimização do algoritmo). Esse tempo inviabiliza sua utilização em sistemas baseados em *tagging* como o *Delicious*. Por essa razão, seria interessante realizar estudos de otimização do algoritmo cujo objetivo seria obter um tempo de execução próximo ao dos sistemas atuais, viabilizando sua implantação em sistemas *online*, pois dessa forma se aproximaria dos percentuais de aceitação obtidos nesse experimento.

Na próxima seção é apresentado um experimento realizado para avaliar a recomendação de *tags* com base em conceitos presentes na estrutura hierárquica da ontologia.

6.3. Experimento 3: Avaliação da Recomendação de *Tags* Hiperônimas com Usuários

Essa seção apresenta um terceiro experimento, o qual segue a mesma metodologia do experimento anterior, ou seja, para realizar as categorizações foi requerido o mesmo número de usuários por abordagem, as mesmas *URLs*, as mesmas métricas de avaliação, etc. A diferença desse experimento para o anterior é que nesse o foco foi avaliar a recomendação de *tags* com base na estrutura hierárquica da ontologia. Por essa razão, nesse experimento foi possível avaliar apenas as abordagens que realizam a recomendação de *tags* baseadas em uma ontologia, sendo elas:

- "*THREE_SOURCE_ONTOLOGY*";
- "*WEBRESOURCE_FOLKSONOMY_ONTOLOGY*"; e
- "*PERSONOMY_FOLKSONOMY_ONTOLOGY*".

Como mostrado na seção anterior (Figura 35), para as cinco abordagens avaliadas, em média, há uma aceitação mais expressiva pelos usuários para as quatro primeiras *tags* de uma

recomendação, ou seja, as quatro *tags* de maior relevância. Neste trabalho essas *tags* são denominadas de *tag* principais. Assim, esse experimento avalia apenas as quatro primeiras *tags* principais das três abordagens citadas acima, pois as demais tendem a ser menos aceitas pelos usuários. Nesse experimento, cada *tag* principal é recomendada juntamente com seu conceito hiperônimo[48]. Essa *tag* hiperônima representa o primeiro conceito mais genérico da *tag* principal recomendada. Por exemplo, se a *tag* "*java*" for recomendada, a *tag* "*object-oriented programming language*" também será recomendada, pois é o conceito hiperônimo de "*java*", conforme pode ser visualizado na Figura 37.

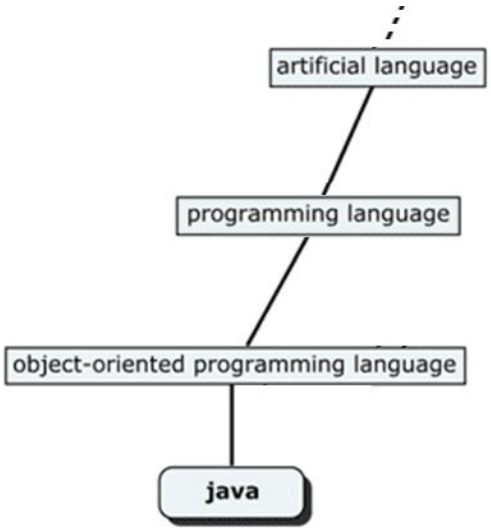

Figura 37. Estrutura hierárquica (parcial) da tag Java com base nas relações da WordNet.

Além do conceito hiperônimo recomendado, recomendamos o próximo conceito acima dessa *tag* hiperônima, com a restrição de que ele estivesse contido na personomia do usuário. Acreditávamos que um conceito mais genérico à *tag* principal poderia ser útil para o usuário na medida em que ele fosse encontrado em sua personomia. Por exemplo, se a *tag* "*artificial language*" estivesse na personomia do usuário, ela poderia ser recomendada, pois está hierarquicamente acima da *tag* principal "*java*", conforme ilustra a . Entretanto, ao executar esse experimento, foi encontrado um único conceito que satisfez essa condição. Logo, o conceito foi recomendado para o usuário, porém ele não foi aceito. Acreditamos que a dificuldade de encontrar um conceito que esteja na personomia do usuário e, que seja de um nível hierárquico mais alto do conceito principal, pode se justificar em vista da alta formalidade de nomenclatura dos conceitos na *WordNet*, não refletindo o vocabulário dos usuários. Essa formalidade também reflete os conceitos hiperônimos recomendados para cada

[48] Uma *tag* hiperônima é resultante da conexão entre a própria *tag* e sua superclasse. Esta conexão é representada pela relação de hiperonímia na ontologia.

tag principal, reduzindo, assim, o percentual de aceitação das *tags* hiperônimas. Por exemplo, o conceito hiperônimo de "*Java*" é "*object-oriented programming language*". Logo, pelo fato desse conceito ter uma descrição bastante formal, dificilmente um usuário irá categorizar um recurso relacionado a "*Java*" com o seu conceito hiperônimo.

O formulário desenvolvido para a categorização desse experimento é ilustrada na Figura 38, no qual é possível visualizar primeiramente as *tags* principais e, logo abaixo a sua respectiva *tag* hiperônima.

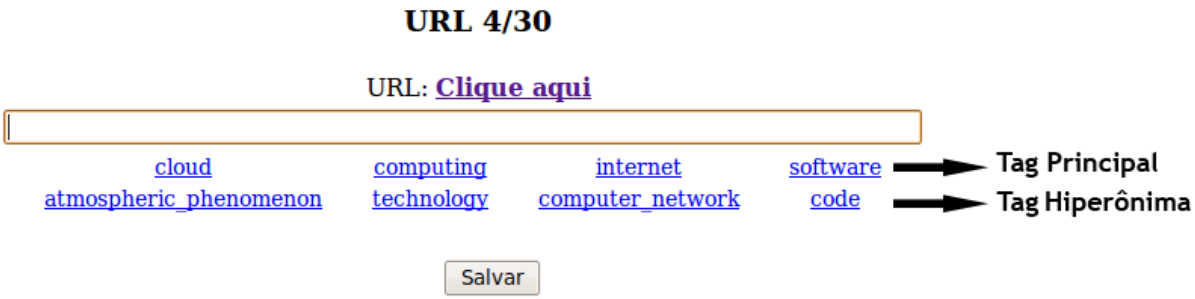

Figura 38. Formulário da categorização do experimento da recomendação de tags hiperônimas.

A Figura 39 mostra a quantidade média de *tags* utilizadas em cada categorização para cada abordagem. Considerando esses números, pode ser observado que a abordagem "*WEBRESOURCE_FOLKSONOMY_ONTOLOGY*" obteve o maior número de *tags* por categorização, totalizando, em média 4,98 *tags*, enquanto que as demais obtiveram 4,0 e 2,65 *tags* para as abordagens "*THREE_SOURCE_ONTOLOGY*" e "*PERSONOMY_FOLKSONOMY_ONTOLOGY*", respectivamente.

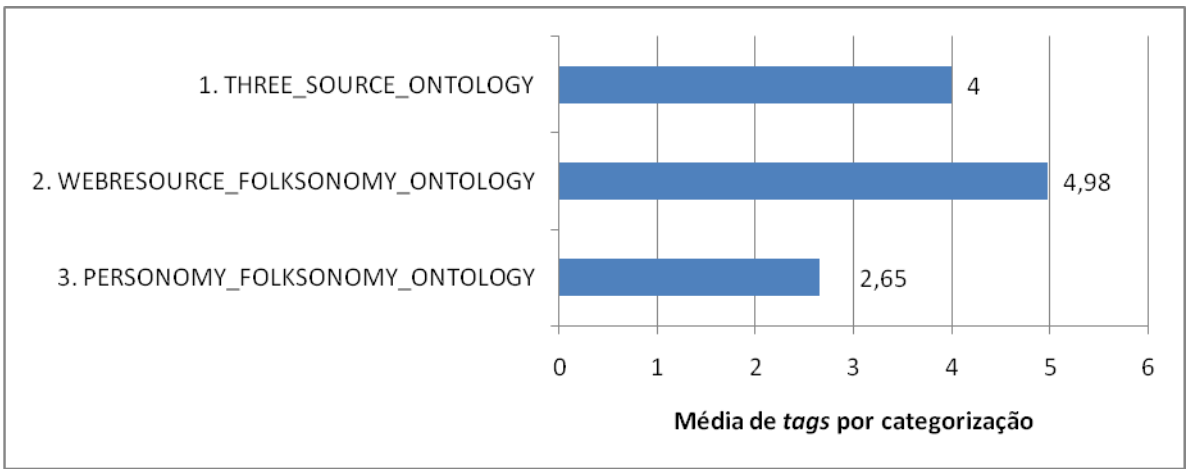

Figura 39. Média de tags utilizadas em cada categorização nas recomendações semânticas.

Com o objetivo de identificar o percentual de aceitação de cada abordagem, foram analisadas as *tags* utilizadas pelos usuários de cada recomendação para cada abordagem, conforme apresentado na Equação 2 (métrica idêntica à utilizada no experimento anterior). Dessa forma, a Figura 40 mostra o percentual de aceitação que cada abordagem obteve para todas as 8 *tags* recomendadas aos usuários (4 *tags* principais e 4 *tags* hiperônimas). Nesse gráfico pode ser observado que o percentual de aceitação mais alto é da abordagem "*THREE_SOURCE_ONTOLOGY*", obtendo 65% de aceitação para as *tags* principais e 17% para as *tags* hiperônimas. Para as demais abordagens, o percentual de aceitação se reduz gradativamente tanto para as *tags* principais quanto para as *tags* hiperônimas. Percebe-se que para as duas abordagens ("*THREE_SOURCE_ONTOLOGY*" e "*WEBRESOURCE_FOLKSONOMY_ONTOLOGY*"), que utilizam a página *Web* como fonte de informação para realizar a recomendação de *tags*, o percentual de aceitação dos usuários foi maior. Isso mostra que o uso dos termos da página *Web* possibilitam recomendar *tags* com maior aceitação por parte dos usuários quando se usa recomendação de *tags* hiperônimas.

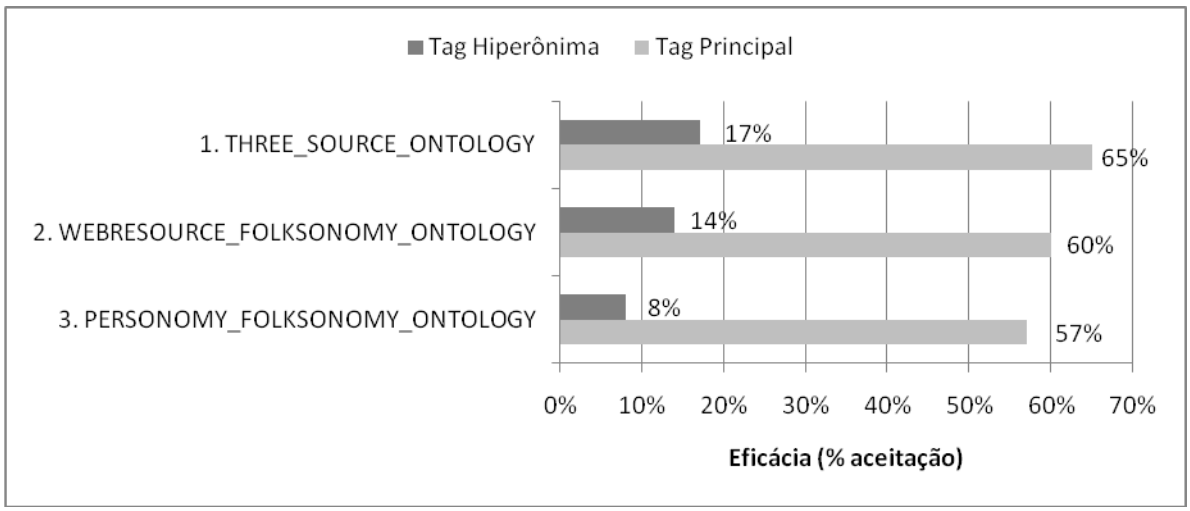

Figura 40. Percentual de aceitação das abordagens nas recomendações de tags hiperônimas.

Na Figura 41 é mostrado o percentual de aceitação que cada abordagem obteve nas categorizações para os *websites* da área da computação. Nesse gráfico é fácil visualizar que a aceitação das *tags* recomendadas pela abordagem "*THREE_SOURCE_ONTOLOGY*" é superior para as *tags* principais, obtendo 65% de aceitação, porém, para as *tags* hiperônimas essa abordagem se assemelha à abordagem "*WEBRESOURCE_FOLKSONOMY_ONTOLOGY*" com aproximadamente 15%.

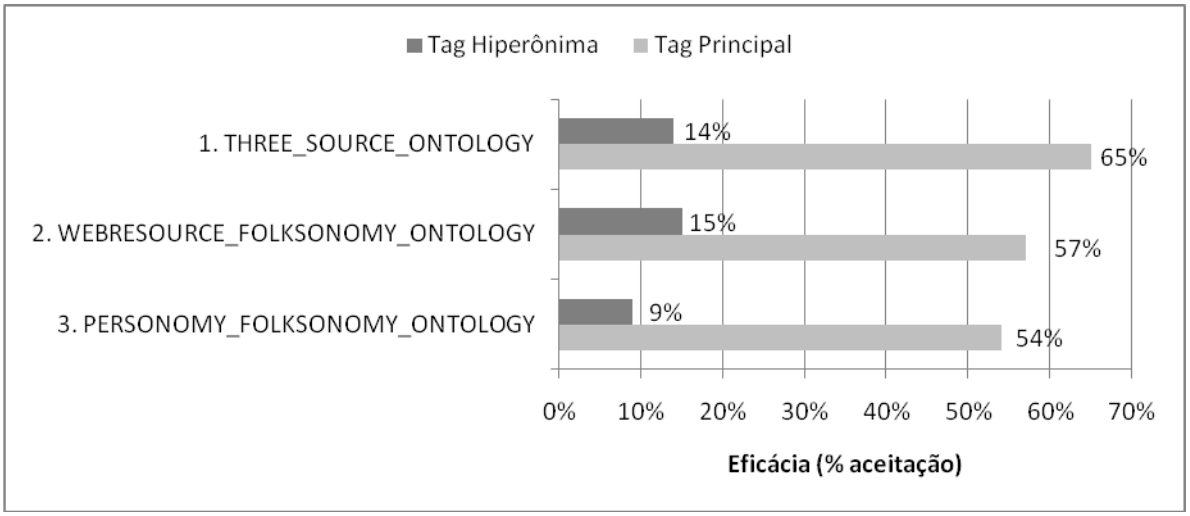

Figura 41. Percentual de aceitação para websites de computação para a recomendação de tags hiperônimas.

A categoria dos *websites* de conhecimentos gerais é ilustrada na Figura 42. Para esses *websites*, a abordagem "*THREE_SOURCE_ONTOLOGY*" é superior às demais, tanto para as *tags* principais quanto para as *tags* hiperônimas. Além disso, essa abordagem obteve 21% de aceitação para as *tags* hiperônimas, enquanto que para a abordagem "*PERSONOMY_FOLKSONOMY_ONTOLOGY*" obteve apenas 6%.

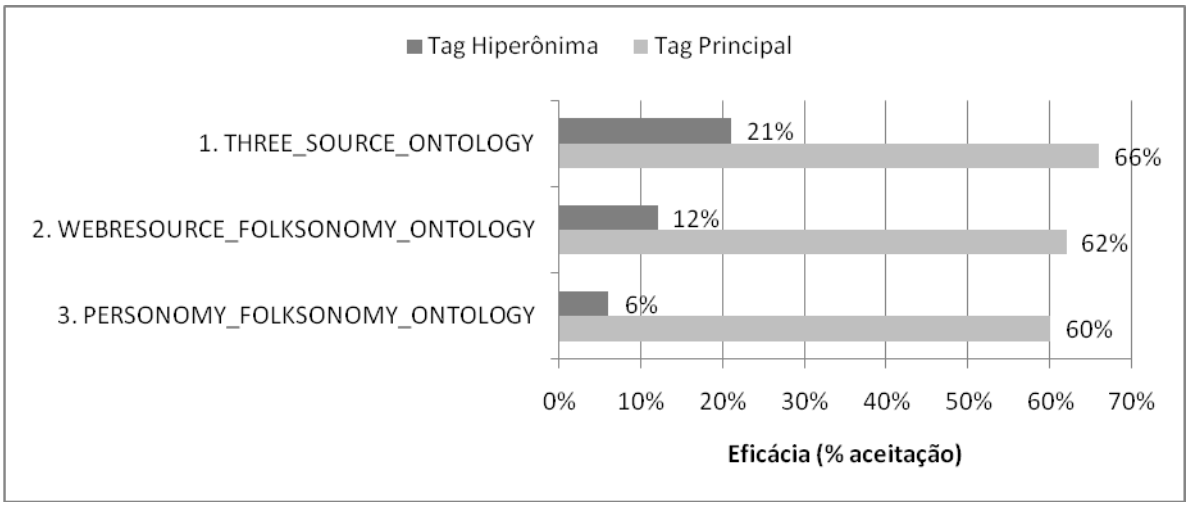

Figura 42. Percentual de aceitação das abordagens para websites de conhecimentos gerais para a recomendação de tags hiperônimas.

A terceira e última categoria criada nesse experimento é um grupo de *websites* da *Wikipedia*, a qual envolveu páginas da área de computação e páginas de conhecimentos gerais. O percentual de aceitação de cada abordagem é mostrado na Figura 43. A abordagem que obteve a maior aceitação pelos usuários para essa categoria foi a "*THREE_SOURCE_ONTOLOGY*". Essa abordagem obteve os maiores percentuais para as

tags principais e para as *tags* hiperônimas. Em comparação com as demais abordagens, percebe-se que a aceitação das *tags* hiperônimas obteve um percentual bastante significativo, alcançando 33% de aceitação, enquanto que as demais atingiram no máximo 12%. Percebe-se também que a abordagem "*THREE_SOURCE_ONTOLOGY*" obteve um percentual de aceitação superior para as *tags* hiperônimas dessa categoria comparado com as categorias de *websites* de computação e de conhecimentos gerais. Dessa forma, os gráficos sugerem que a abordagem "*THREE_SOURCE_ONTOLOGY*" pode fornecer uma maior contribuição aos usuários para *websites* da *Wikipedia*, seguido dos *websites* de conhecimentos gerais e *websites* de computação.

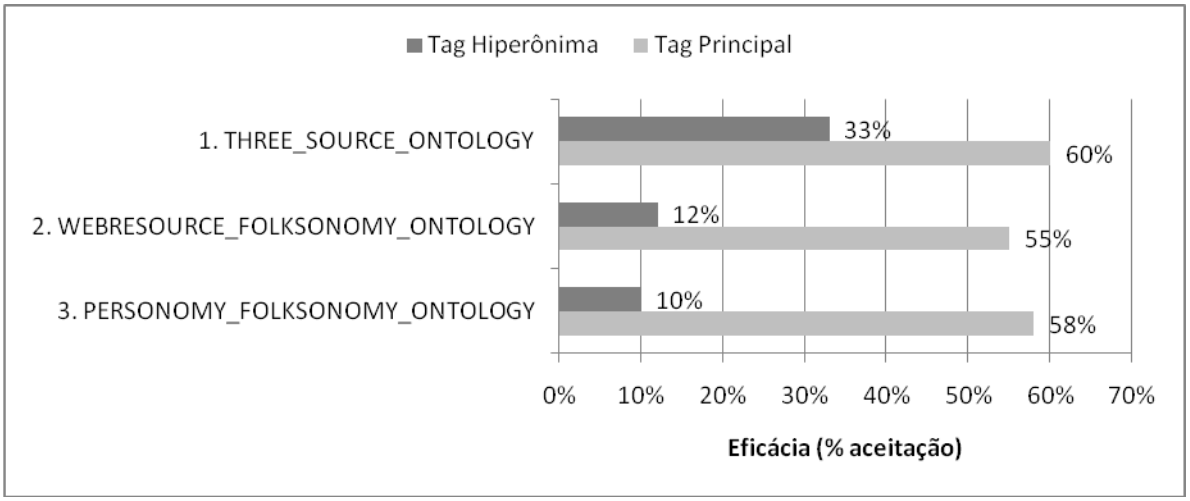

Figura 43. Percentual de aceitação para websites da Wikipedia para a recomendação de tags hiperônimas.

É possível concluir sobre esse experimento que as *tags* hiperônimas recomendadas foram relativamente bem aceitas pelos usuários para as abordagens "*THREE_SOURCE_ONTOLOGY*" e "*WEBRESOURCE_FOLKSONOMY_ONTOLOGY*", na qual obtiveram, respectivamente, 17% e 14% de aceitação. Além disso, a abordagem proposta neste trabalho obteve os maiores percentuais para as categorias de *websites* de conhecimentos gerais e *websites* da *Wikipedia*. Em comparação com os resultados do experimento anterior, percebemos que ocorreram duas semelhanças: (i) o percentual de aceitação foi maior para a abordagem "*THREE_SOURCE_ONTOLOGY*" e, (ii) dentre as categorias criadas, essa mesma abordagem obteve a maior aceitação para os *websites* de conhecimentos gerais. Embora nesse experimento a aceitação das *tags* pelos usuários tenha sido maior para cada abordagem, pelo fato da recomendação ter sido apenas com as quatro *tags* mais relevantes, os percentuais foram regularmente proporcionais ao do experimento anterior, o que pode alavancar a nossa hipótese dos resultados serem válido em situações reais.

Na próxima seção é mostrado o resultado dos questionários respondidos pelos usuários após a realização dos experimentos.

6.4. Experiência dos Usuários e Dificuldades Enfrentadas na Realização dos Experimentos

Essa seção caracteriza-se por uma pesquisa de opinião realizada após a conclusão dos experimentos pelos usuários. No total foram 12 (doze) pessoas distintas. A experiência dos participantes deve ser conhecida para entender a influência perante a validade dos resultados do experimento. Foram aplicados dois questionários aos usuários (ver Apêndice B), primeiramente o Q1 para coleta de dados referente à experiência de cada participante nas áreas de conhecimento envolvidas no experimento e, posteriormente o Q2 para coleta de dados referente às dificuldades na execução do experimento.

O experimento contou com a participação de 12 (doze) pessoas do meio acadêmico (área da computação), 1 (um) doutor, 2 (dois) mestres, 4 (quatro) mestrandos e 5 (cinco) graduandos. A Tabela 8 mostra a quantidade de participantes relacionando-os com suas experiências com os sistemas baseados em *tagging*. As colunas da tabela indicam quanto o participante conhece ou desconhece essas áreas, medidos pelas palavras **Nunca** que representa o total desconhecimento do assunto, **Raramente** que indica pouco conhecimento, **Frequentemente** que indica bastante conhecimento sobre o assunto, e **Constantemente** que indica total domínio sobre o assunto e comumente utilizado em suas atividades. Nessa tabela percebe-se que a maioria dos participantes já conhecia e praticavam atividades nos sistemas baseados em *tagging*. Os dados podem ser mais bem visualizados na Figura 44, pois representam a quantidade, em porcentagens, de pessoas com experiência nos sistemas baseados em *tagging* e na tarefa de categorização.

Tabela 8. Dados de experiência dos participantes.

Área/Conhecimento	**Nunca**	**Raramente**	**Frequentemente**	**Constantemente**
Sistemas baseados em *tagging*	0	4	5	3
Categorização	2	2	7	1

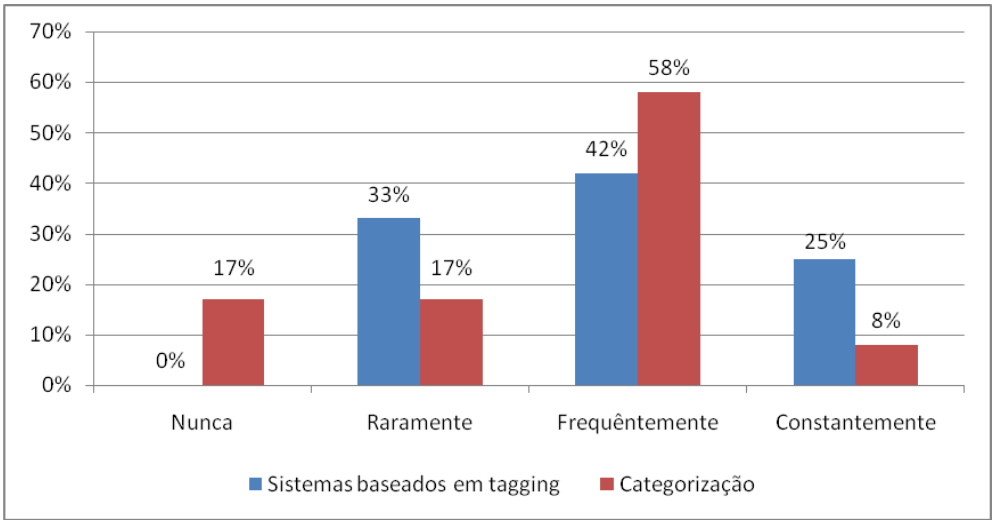

Figura 44. Experiência dos participantes.

Ao observar a Figura 44, identificamos que 17% dos participantes nunca realizaram uma categorização (associação de uma *tag* a um objeto), entretanto, todos já conheciam previamente o que é um sistema baseado em *tagging*. Embora houvesse 17% dos participantes que nunca realizaram uma categorização, a maioria das pessoas já tinha algum conhecimento sobre o processo.

Os dados coletados no questionário **Q2** representam as respostas para as dificuldades encontradas na realização do experimento. Nesse questionário foram obtidas respostas dos participantes sobre: (i) o nível de dificuldade enfrentado no experimento de modo geral e pela utilização da interface; (ii) a solicitação de ajuda durante o experimento; e (iii) os problemas enfrentados durante a execução do experimento. Dessa forma, analisamos todas as respostas a essas perguntas e geramos uma média dentre os usuários. Assim, em uma escala de 0 (zero) a 10 (dez), o nível de dificuldade enfrentado pelos participantes no experimento foi de 2,41 e, o nível de dificuldade da navegação foi 2. Além disso, apenas um participante precisou de ajuda para realizar as categorizações do experimento, adquirindo um dicionário para traduzir algumas palavras presentes nos textos dos *websites*, porém, acreditamos que isso não tenha interferido no resultado geral do experimento, pois é comum que alguns usuários tenham dificuldade em alguns tipos de textos do idioma inglês.

Nesse capítulo foi apresentado um experimento automatizado da recomendação de *tags* em comparação às categorizações realizadas no sistema *Delicious* e, dois experimentos realizados com usuários reais. Por fim, foram mostradas as opiniões, experiências e dificuldades enfrentadas pelos participantes nos experimentos. Existem ainda alguns trabalhos

futuros e limitações desta proposta de recomendação de *tags*, as quais são discutidas no próximo capítulo.

Capítulo

7

Conclusões e Considerações Finais

Encontrar dificuldades ou, até mesmo, não encontrar a informação desejada em uma coleção de documentos é um problema bastante antigo (GARSHOL, 2004). Nos últimos anos, a quantidade de informação tem aumentado de forma jamais vista devido à facilidade de acesso à *Internet* e a facilidade com que os usuários criam e publicam seus conteúdos na *Web*. Para os sistemas baseados em *tagging* a situação não é diferente, ocasionando vários problemas como a sinonímia e a polissemia. Esses problemas são gerados a partir da falta de critério na escolha e utilização das *tags* ou, até mesmo, pela liberdade proporcionada aos usuários em uma categorização. Por essa razão, a recuperação de um objeto categorizado torna-se prejudicada, acarretando em um alto esforço dos usuários na tarefa de filtragem de informação (HARDTKE, 2009) (BRUSILOVSKY, 2009) (MICARELLI, 2007) (PANT, 2003).

Incentivados pelos problemas da recuperação de informação nos sistemas baseados em *tagging*, propusemos uma abordagem de recomendação de *tags* para tentar minimizar os problemas gerados no processo de *tagging* como *tags* polissêmicas, *tags* sinônimas, singular e plural, etc. Os objetivos desta proposta foram recomendar *tags* que possam estar relacionadas

ao vocabulário do usuário e que possam estar relacionadas às características e interesses do usuário em relação a um recurso *Web*, pois acreditamos que isso pode facilitar a recuperação dos recursos categorizados. Para tal propomos uma metodologia de recomendação de *tags* semânticas que utiliza o conteúdo de três fontes de informação: as *tags* da personomia do usuário, os termos da folksonomia do sistema baseado em *tagging* e o conteúdo da página *Web*. Para promover os termos de cada fonte a conceitos formais utilizamos a estrutura criada por Basso *et al*. (2009). Essa estrutura permite gerar uma ontologia para cada fonte de informação e se beneficiar da formalidade de uma ontologia como a *WordNet*. Este trabalho faz uma extensão do trabalho de Basso *et al*. (2009). Assim, foi possível identificar quatro classes gramaticais (substantivos, verbos, adjetivos e advérbios) para os termos extraídos das fontes de informação, e não apenas substantivos como acontecia anteriormente. Além disso, adaptamos o algoritmo para emergir ontologias leves que utilizam apenas as relações de "*is-a*" e "*part-of*".

Analisando o conteúdo das páginas *Web*, percebemos que ao utilizar a *WordNet* para emergir a ontologia, 76% dos termos são identificados enquanto que 24% não são identificados. Neste trabalho, a identificação dos termos das páginas *Web* na *WordNet* é fundamental, pois os termos não reconhecidos não farão parte da ontologia dessa fonte de informação e, consequentemente, podem não ser recomendados. Assim, é interessante adicionar na ontologia os 24% dos termos não identificados, os quais são compostos por abreviações, termos novos adotados nas diversas áreas existentes, nomes de empresas, projetos e sistemas existentes. No entanto, para agregar esses termos na recomendação seria necessário a utilização de outras bases de dados, como a *DBPedia* e a *ConceptNet*, o que deverá ser realizado em trabalhos futuros.

Como resultado da realização de experimentos com os usuários, observou-se que, em geral, é possível gerar recomendações de *tags* com elevados percentuais de aceitação pelos usuários analisando as três fontes de informação adotadas neste trabalho. Os resultados dos experimentos realizados com esta proposta de recomendação de *tags* possibilitam sugerir que: (i) em comparação as 30 *tags* mais utilizadas pelo sistema *Delicious* para um respectivo recurso *Web*, esta abordagem gerou *tags* representativas em comparação com as *tags* contidas na folksonomia do sistema *Delicious*; (ii) esta abordagem produziu um aumento na quantidade média de *tags* utilizadas em uma categorização em comparação à média utilizada pelo sistema *Delicious*; (iii) esta abordagem reduziu a probabilidade de um usuário informar *tags* adicionais às recomendadas em uma categorização; (iv) esta abordagem recomendou *tags* com o maior percentual de aceitação dentre as cinco avaliadas, obtendo 40% de aceitação

geral pelos usuários, superando a recomendação gerada pelo sistema *Delicious* que obteve 27% de aceitação; e (v) esta abordagem obteve um percentual de aceitação superior às demais abordagens avaliadas em relação ao uso de *tags* hiperônimas, mostrando-se ser a mais relevante para recomendar *tags* hiperônimas a partir de uma ontologia baseada na *WordNet*.

Outro aspecto percebido é que a aceitação das *tags* recomendadas depende da qualidade das *tags* que estão em sua personomia, ou seja, as *tags* utilizadas pelo usuário nas categorizações de seus recursos. Dessa forma, a partir do momento em que um sistema estiver usando esta abordagem de recomendação de *tags*, ele estará melhorando os dados contidos nas personomias dos usuários e, consequentemente, a emergência da ontologia da personomia também será melhorada. Assim, uma vez que a emergência da ontologia da personomia depende das informações nela contidas e a recomendação pode melhorar as informações que serão agregadas na personomia (*tags* aceitas de uma recomendação que não estão na personomia), um ciclo se completa. Portanto, esta proposta de recomendação beneficiará a emergência da ontologia da personomia e vice-versa.

Uma limitação deste trabalho é que foi utilizada apenas a recomendação de *tags* para recursos *Web* do idioma inglês. No entanto, se for utilizada uma *WordNet* de outro idioma, o processo seria o mesmo. Outra limitação está em relação à interpretação dos conteúdos dos recursos *Web*, pois é possível interpretar apenas as páginas que estão em formato texto como *HTML* e *XHTML*, não tendo sido interpretado documentos nos formatos *pdf*, *doc*, etc.

Como trabalhos futuros serão considerados alguns estudos como:

- melhorar a identificação dos termos relevantes extraídos de uma página *Web*, uma vez que atualmente essa identificação é realizada baseando-se na frequência do termo no documento. Uma alternativa seria considerar o peso da *tag HTML* em que o termo da página está inserido, por exemplo, as *tags HTML* “<title>“, “<h1>“, “<h2>“, “<h3>“, etc., teriam pesos diferenciados;
- a possibilidade de extrair termos compostos das páginas *Web*. Dessa forma, *tags* compostas como “*artificial intelligence*” e “*Web semantic*” poderiam ser recomendadas como uma única *tag*, uma vez que isso poderia beneficiar ainda mais os usuários no momento de recuperar um recurso categorizado;
- avaliar como pode ser realizada a recomendação de *tags* em que os conceitos estejam presentes na hierarquia (profundidade) da ontologia, como um conceito genérico ou específico;

- avaliar como o processo de mapeamento de ontologias pode ser melhorado para considerar conceitos que não foram extraídos da ontologia da página *Web* ou da folksonomia para compor a lista de *tags* na recomendação;
- avaliar a eficácia das *tags* aceitas pelos usuários as quais foram recomendadas por esta proposta de recomendação de *tags* na tarefa de recuperação de informação; e
- avaliar os termos utilizados pelo usuário na tarefa de recuperação de informação de um recurso *Web*. O objetivo desse estudo será utilizar esses termos para compor a lista das *tags* recomendadas, pois acreditamos que os termos utilizados na recuperação de algum recurso possam ser úteis, ajudando a encontrar o recurso mais rapidamente.

Finalmente, esta abordagem de recomendação de *tags* será futuramente utilizada pelo sistema *TagManager*. Entretanto, acreditamos que ela possa ser utilizada também por outros sistemas baseados em *tagging* que precisam de formalidade nas categorizações.

Referências

ADOMAVICIUS, G.; TUZHILIN, A. **Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions**. IEEE Transactions on knowledge and data engineering, pp.734-749, 2005.

ADRIAN, B.; SAUERMANN, L. & ROTH-BERGHOFER, T. **ConTag: A semantic Tag Recommendation System**. In: Proceedings of I-Semantics' 07: pp. 297-304, 2007.

AL-KHALIFA, H. S.; DAVIS, H. C. **Towards better Understanding of Folksonomic Patterns**. In: HT '07: Proceedings of the 18th conference on Hypertext and hypermedia, New York, NY, USA: 163-166, 2007.

ALLAN, J.; KUMARAN, G. **Stemming in the Language Modeling Framework**. 26th Annual International ACM SIGIR Conference. Proceedings, p.455-456, 2003.

ANDERSON, C. **The Long Tail**. New York: Hyperion Books - Vol. I, 2006.

ANDERSON, J. R. **Cognitive Psychology and its Implications**. New York: W. H. Freeman and Company, 4 ed, 1995.

ARAMPATZIS, A. T.; WEIDE, T. P.; KOSTER, C. H. A.; BOMMEL, P. **Linguistically-motivated Information Retrieval**. Encyclopedia of Library and Information Science, V.69, pp. 201-222, 2000.

BANERJEE, S.; PEDERSEN, T. **An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet**. In: Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, pp. 136–145, 2002.

BASSO, C. A. M. **Uma Proposta para a Evolução de Ontologias a partir de Personomias em Sistemas Baseados em Tagging**. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Maringá, Maringá-PR, 2009. 103 p.

BASSO, C. A. M., FERREIRA, J. M. P., SILVA, S. R. P.: **An Unsupervised Approach for the Emergence of Ontologies from Personomies in Tagging-Based Systems**. In Proceedings of Latin American Web Congress, Merida, Yucatan, Mexico, pp. 193-200, 2009.

BEARMAN, D.; TRANT, J. **Unifying our Cultural Memory: Could Electronic Environments Bridge the Historical Accidents that Fragment Cultural Collections**. In: Information Landscapes for a Learning Society, Networking and the Future of Libraries, 3, 1998.

BORTH, M. R.; SILVA, S. R.; FERREIRA, J. M. P.; FELTRIM, V. D. **An Approach to Enrich Users' Personomy Using Semantic Recommendation of Tags**. In: III International Workshop on Web and Text Intelligence, São Bernardo do Campo - SP, pp. 876-885, 2010.

BREITMAN, K. **Web Semântica: a Internet do Futuro**. Rio de Janeiro: LTC, 2005.

BRUSILOVSKY, P.; AHN, J. **Adaptive Visualization of Search Results: Bringing User Models to Visual Analytics**. Information Visualization 8 (3), pp. 167-179, 2009.

BRYNJOLFSSON, E.; HU, Y.; SIMESTE, D. **Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales**. SSRN. 2007. Disponível em <http://ssrn.com/abstract=953587>. Acesso em 07 mai. 2010.

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. **What are Ontologies, and Why do We Need Them?** IEEE Intelligent Systems - Special Issue on Ontologies, v. 14, n. 1, pp. 20-26, 1999.

CHOY, S.; LUI, A. K. **Web Information Retrieval in Collaborative Tagging Systems**. Web Intelligence, 2006. WI 2006. IEEE/WIC/ACM International Conference on: pp. 352-355, 2006.

CHUN, S.; JENKINS, M. **Cataloguing by Crowd; a Proposal for the Development of a Community Cataloguing Tool to Capture Subject Information for Images**. Museums and the Web 2005, Vancouver Retrieved. Disponível em: <http://www.archimuse.com/mw2005/abstracts/prg_280000899.html>. Acesso em 31 mai. 2009.

CÔGO, F. R. **Uma Proposta de Organização do Vocabulário de Tags dos Usuários de Sistemas Baseados em Folksonomia**. Trabalho de Conclusão de Curso (Ciência da Computação) – Universidade Estadual de Maringá, Maringá-PR, 2009. 53 p.

CÔGO, F. R.; DA SILVA, S. R. P. **Uma Proposta de Organização do Vocabulário de Tags de Usuários de Sistemas Baseados em Folksonomia**. In: XIII Simpósio Brasileiro Sobre Fatores Humanos em Sistemas Computacionais. Porto Alegre - RS: ACM, v.1, pp. 288-291, 2008.

CROFT, W. B.; METZLER, D.; STROHMAN, T. **Search Engines: Information Retrieval in Practice**. Ed. Pearson Higher Education, 2009.

DA SILVA, J. V. **Gerenciamento do Vocabulário do Usuário em Sistemas Baseados em Tagging**. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Maringá, Maringá-PR, 2009. 124 p.

DIAS-DA-SILVA, B. C.; FELIPPO, A. D.; NUNES, M. G. V. **The Automatic Mapping of Princeton WordNet Lexical Conceptual Relations onto the Brazilian Portuguese**

WordNet Database. In Proceedings of the 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 2008.

DILL S., EIRON N., GIBSON D., GRUHL D., GUHA R., JHINGRAN A., KANUNGO T., RAJAGOPALAN S., TOMKINS A., TOMLIN J. A., ZIEN, J. Y. **SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation**. Proceedings of the 12th International Conference on World Wide Web (WWW'03). Budapest, Hungary. 2003.

ECHARTE, F.; ASTRAIN, J. J.; CÓRDOBA, A.; VILLADANGOS, J. **Ontology of Folksonomy: A New Modeling Method**. In: Conference'04, Month 1–2, 2004. Disponível em: <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-289/p08.pdf>. Acesso em 12 abr. 2009.

ECHARTE, F.; ESTRAIN, J. J.; CÓRDOBA, A.; VILLADANGOS, J. **Ontology of Folksonomy: A New Modeling Method**. In Proceedings of Semantic Authoring, Annotation and Knowledge Markup Workshop (SAAKM). Whistler, British Columbia, Canada. Outubro 28-31. 2007.

FELLBAUM, C. **WordNet: An Eletronic Lexical Database**. MIT Press, Cambridge, Massachusetts, 1998.

FENSEL, D. **Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce**. Springer, Heidelberg, Alemanha, 2ª edição, 2004.

FILHO, F. M. F.; ALBUQUERQUE, J. P; GEUS, P. L. **Sistemas de Recomendação e Interação na Web Social**. In: I Workshop de Aspectos da Interação Humano-Computador na Web Social, 2008, Porto Alegre. VIII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais, 2008. p. 24-27. Disponível em: <http://www.inf.pucrs.br/ihc2008/pt-br/assets/files/Sistemas_de_Recomendacao_e_Interacao_na_Web_Social.pdf>. Acesso em 11 abr. 2009.

FINLAYSON, M. A. **MIT Java WordNet Interface**. 2009. Disponível em <http://projects.csail.mit.edu/jwi/>. Acesso em 20 set. 2009.

FOUNTOPOULOS, G.I. (2007). **RichTags: A Social Semantic Tagging System**. Master's Thesis, University of Southampton, Southampton, UK. p. 45.

FRAKES, W. B.; BAEZA-YATES. **Information Retrieval: Data Structures and Algorithms**. Prentice-Hall, New York, 1992.

GARSHOL, L. M. **Metadata? Thesauri? Taxonomies? Topic Maps! Making Sense of it all.** In: Journal of Information Science, Vol. 30, No. 4, pp. 378-391, 2004.

GIUNCHIGLIA, F.; MARCHESE, M.; ZAIHRAYEU, I. **Encoding Classifications into Lightweight Ontologies**. In Proceedings of the 3rd European Semantic Web Conference (ESWC 2006), Budva, Montenegro, pp. 80-94, 2006.

GOLDER, S.A.; HUBERMAN, B.A. **The Structure of Collaborative Tagging Systems**.

HP Labs Technicals Report, 2006.

GONZALEZ, M.; de LIMA, V.L.S.; de LIMA, J.V.; **Termos, Relacionamentos e Representatividade na Indexação de Texto para Recuperação de Informação**. Letras de Hoje. Porto Alegre. v. 41, nº 2, junho, pp. 65-87, 2006.

GOOGLE. **Google Wonder Wheel – A Wonder Tool**. Disponível em <http://www.googlewonderwheel.com>. Acesso em 25 set. 2009.

GRUBER, T. **A Translation Approach to Portable Ontology Specifications**. Knowledge Acquisition, 5(2):199-220, 1993. Disponível em: <http://tomgruber.org/writing/ontolingua-kaj-1993.htm>. Acesso em 13 nov. 08.

GUY, M.; TONKIN, E. **Folksonomies: Tidying up Tags?** In: D-Lib Magazine, Volume 12, Número 1, ISSN 1082-9873, Janeiro, 2006.

HALPIN, H.; ROBU, V.; SHEPHERD, H. **The Complex Dynamics of Collaborative Tagging**. In: Proceedings of International World Wide Web Conference. ACM Press, New York, pp. 211-220, 2007.

HARDTKE, D.; WERTHEIM, M; CRAMER, M. **Demonstration of Improved Search Result Relevancy Using Real-Time Implicit Relevance Feedback**. SIGIR 2009 Workshop on Understanding the User – Logging and interpreting user interactions in information search and retrieval. Disponível em: <http://www.dfki.uni-kl.de/~elst/papers/Belkin2009.pdf>. Acesso em 20 jul. 2010.

HEPP, M. **Ontologies: State of the Art, Business Potential, and Grand Challenges**. In Martin Hepp, Pieter De Leenheer, Aldo de Moor, and York Sure, editors, Ontology Management: Semantic Web, Semantic Web Services, and Business Applications. Springer Verlag, pp. 3–22, 2007.

HEYMANN, P.; GARCIA-MOLINA, H. **Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems**. Report, InfoLab, Standford, 2006. Disponível em: <http://heymann.stanford.edu/taghierarchy.html>. Acesso em 09 dez. 2009.

HIMMA, K. E. **The Concept of Information Overload: A Preliminary Step in Understanding the Nature of a Harmful Information-Related Condition**. Ethics and Information Technology. Seattle. Springer. 2007.

HORNBY, A. S. **Oxford Advanced Learners Disctionary**. Oxford University Press, 6ª ed., 2005.

HOTHO, A.; JÄSCHKE, R.; SCHMITZ, C.; STUMME, G. **Information Retrieval in Folksonomies: Search and Ranking**. In: York Sure and John Domingue. The Semantic Web: Research and Applications, volume 4011 of LNCS. Springer, June, pp. 411-426, 2006.

IZQUIERDO, I. **A Arte de Esquecer: Cérebro, Memória e Esquecimento**. Vieira & Lent, 2004.

JÄSCHKE, R.; MARINHO, L.; HOTHO, A.; SCHMIDT-THIEME, L. & STUMME, G. **Tag**

Recommendations in Folksonomies. In Proceedings 11th Europ. Conf. on Principles and Practice of. Knowledge Discovery in Databases (PKDD), pp. 506-514, 2007.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Prentice-Hall, Upper Saddle River, EUA, 2000.

JURISTO, N.; MORENO, A. M. **Basics of Software Engineering Experimentation**. Universidad Politécnica de Madrid, 2000.

KDNUGGETS News. **Poll Results: The Most Popular Social Bookmarking Sites**. KDnuggets. 2007. Disponível em: <http://www.kdnuggets.com/news/2007/n08/1i.html>. Acesso em 25 jun. 2009.

KNERR, T. **Tagging Ontology – Towards a Common Ontology for Folksonomies**. 2006. Disponível em <http://tagont.googlecode.com/files/TagOntPaper.pdf>. Acesso em 13 out. 2009.

KOCH, N. P. **Software Engineering for Adaptive Hypermedia Systems Reference Model, Modeling Techniques and Development Process**. Tese de Doutorado em Engenharia de Software. Ludwig – Maximilian – Universitat Munchen. Munique, 2000.

KORENIUS, T.; LAURIKKALA, J.; JÄRVELIN, K.; JUHOLA, M. **Stemming and Lemmatization in the Clustering of Finnish Text Documents**. 13th ACM Conference on Information and Knowledge Management (CIKM). Proceedings, pp. 625-634, 2004.

KROVETZ, R. **Viewing Morphology as an Inference Process**. 16th Annual International ACM SIGIR Conference, pp.191-202, 1993.

LEVY, D. M. **To Grow in Wisdom: Vannevar Bush, Information Overload, and the Life of Leisure**. In Proceedings JCDL'05, ACM/IEEE-CS, 281-286, 2008.

LYMAN, P. **How Much Information?**. University of California. USA. 2000. Disponível em <http://www.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf>. Acesso em 10 jan. 2011.

MALTZ, D.; EHRLICH, K. **Pointing the Way: Active Collaborative Filtering**. In Proceedings of CHI-95, Denver, CO, pp. 202-209, 1995.

MANNING, D. C.; RAGHAVAN, P.; SCHÜTZE, H. **An Introduction to Information Retrieval**. Cambridge University Press, 1 ed. 2008.

MARCHETTI, A.; TESCONI, M.; RONZANO, F.; ROSELLA,M.; MINUTOLI,S.; **SemKey: A Semantic Collaborative Tagging System**. In: Proceedings of WWW 2007 Workshop on Tagging and Metadatafor Social Information Organization, 2007.

MATHES, A. **Folksonomies - Cooperative Classification and Communication Through Shared Metadata**. 2004. Disponível em: <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>. Acesso em 20 nov. 2008.

METEREN R. V., SOMEREN, M. V. **Using Content-Based Filtering for Recommendation**. MLnet/ECML2000 Workshop, May 2000, Barcelona, Spain. Disponível em: <http://www.ics.forth.gr/~potamias/mlnia/paper_6.pdf>. Acesso em 10 mar. 2009.

MICARELLI, A; GASPARETTI, F.; SCIARRONE, F.; GAUCH, S. **Personalized Search on the World Wide Web**. In The Adaptive Web. Ed. Springer-Verlag. Alemanha, Maio, pp.195-230. 2007.

MILLER, G. A. **WordNet: A Lexical Database for English**. Communications of ACM. Vol. 38, nº 11, Novembro, pp. 39-41, 1995.

MILLER, P. **Web 2.0: Building the New Library**. Ariadne, v. 45: 30, 2005. Disponível em: < http://www.ariadne.ac.uk/issue45/miller>. Acesso em 15 set. 2009.

MIZOGUCHI, R. **Knowledge Acquisition and Ontology**. Proceedings of KB&KS: Building Large-Scale Knowledge Bases and Knowledge Sharing, Tokyo, pp. 121-128, 1993.

NEWMAN, M. E. J. **Power Laws, Pareto Distributions and Zipf's Law**. Statistical Mechanics. Contemporary Physics, 46, pp. 323-351, 2006.

PANT, Gautam; BRADSHAW, Shannon; MENCZER, Filippo. **Search Engine-Crawler Symbiosis: Adapting to Community Interests**. 7th European Conference on Research and Advanced Technology for Digital Libraries. 2003. Disponível em: <http://dollar.biz.uiowa.edu/~pant/Papers/se-crawler.pdf>. Acesso em 19 out. 2009.

PEREIRA, R.; DA SILVA, S. R. P. **Folksonomias: Uma Análise Crítica Focada na Interação e na Natureza da Técnica**. In: XIII, Simpósio Brasileiro Sobre Fatores Humanos em Sistemas Computacionais. Porto Alegre - RS : ACM, v. 1, pp. 126-135, 2008.

PREECE, J. **Design de Interação: Além da Interação Homem-Computador**. Porto Alegre: Bookman, 2005.

RASHID, A.; ALBERT, I.; COSLEY, D.; LAM, S.; MCNEE, S.; KONSTAN, J.; RIEDL, J. **Getting to Know You: Learning New User Preferences in Recommender Systems**. In: Proceedings of Conference on Intelligent User Interfaces, pp.127-134, 2002.

RIDDLE, P. **Tags: What are They Good For?** School of Information. Technical Report. Disponível em: <http://www.ischool.utexas.edu/~i385q/archive/riddle_p/riddle-2005-tags.pdf>. 2005. Acesso em 30 out. 2008.

RIJSBERGEN, C. J. **Information Retrieval**. Department of Computing Science, University of Glasgow. Livro online, 1999. Disponível em: <http://www.dcs.gla.ac.uk/Keith/Preface.html>. Acesso em 13 set. 2009.

RUSSELL, T. **Contextual Contextual Authority Tagging: Cognitive Authority Through Folksonomy**. School of Information and Library Science. University North Carolina. 2005. Disponível em: <http://www.terrellrussell.com/projects/contextualauthoritytagging/conauthtag200505.pdf>. Acesso em 14 jul. 2009.

SHAW, B. **Building a Better Folksonomy: Web-based Aggregation of Metadata**. Technical Report, 2005. Disponível em: <http://www.metablake.com/webfolk/web-paper.pdf>. Acesso em 22 out. 2009.

SHEN, K; WU, L. **Folksonomy as a Complex Network**. Article. Departament of Computer Science. Shangai: Fudan University, 2005.

SHEPTSEN, A.; GEMMEL, J.; MOBASHER, B.; BURKE, R. **Personalized Recommendation in Social Tagging Systems Using Hierarchical Clustering**. In: proceedings of ACM Recommender Systems, pp.259-266, 2008.

SILVA, M. J.; CARDOSO, N. **Query Expansion through Geographical Feature Types**. In 4th Workshop on Geographic Information Retrieval, GIR 07 (held at CIKM'07), Lisbon, Portugal, 9th November, 2007.

SMITH, G. **Tagging: People-Powered Metadata for the Social Web**. New Riders, Berkley, 2008.

SKOS. **Simples Knowledge Organization System**. 2010. Disponível em: <http://www.w3.org/2004/02/skos/>. Acesso em 15 dez. 2009.

SOOD, S.; OWSLEY, S.; HAMMOND, K.; BIRNBAUM, L. **TagAssist: Automatic Tag Suggestion for Blog Posts**. In: Proceedings of the International Conference on Weblogs and Social Media, 2007.

STURTZ, D. N. **Communal Categorization: The Folksonomy**. Content Representation, 2004.

SU, X. **A Text Categorization Perspective for Ontology Mapping**. Technical report, Norwegian University of Science and Technology, Norway, 2002. Disponível em: < http://www.scs.carleton.ca/~armyunis/knowledge-managment/papers/Text-Categorization.pdf> Acesso em 10 out. 2009.

SYMEONIDIS, P.; NANOPOULOS, A.; MANOLOPOULOS, Y. **Tag Recommendations Based on Tensor Dimensionality Reduction**. In: RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems, New York, NY, USA, pp. 43-50, 2008.

VOSSEN, P. **Introduction to EuroWordNet**. Computers and the Humanities, Flushing, v.32, n.2-3, pp. 73-89, 1998.

WAL, T. V. **Folksonomy**. Online Information. vanderwal.net. 2005. Disponível em: <http://www.vanderwal.net/random/entrysel.php?blog=1622>. Acesso em 15 set. 2008.

WETZKER, R.; SAID, A.; ZIMMERMANN, C. **Understanding the User: Personomy Translation for Tag Recommendation**. In: Proceedings of European Conference on Machine Learning (ECML), 2009.

WIKIPEDIA. **Idiosyncrasy**. 2010. Disponível em: <http://en.wikipedia.org/wiki/Idiosyncrasy>. Acesso em 08 nov. 2010.

WORDNET. **About Wordnet**. Cognitive Science Laboratory, Princeton University, 2006. Disponível em: <http://wordnet.princeton.edu/>. Acesso em 30 ago. 2009.

WORDNET. **WNStats – WordNet 3.0 Database Statistics**. Cognitive Science Laboratory, Princeton University, 2010. Disponível em: <http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html >. Acesso em 30 jan. 2010.

WU, X.; ZHANG, L.; YU, Y. **Exploring Social Annotations for the Semantic Web**. In: Proceedings of the 15th international conference on World Wide Web. ACM Press, New York, pp.417-426, 2006.

YAHOO. **Delicious**. 2003. Disponível em: <http://del.icio.us>. Acesso em 03 abr. 2008.

ZHAO, S.; DU, N.; NAUERZ, A.; ZHANG, X.; YUAN, Q.; FU, R. **Improved Recommendation based on Collaborative Tagging Behaviors**. In: Proceedings of the International conference on intelligent user interfaces, pp.413-416, 2008.

ZIPF, G.K. **Human Behavior and the Principle of Least Effort**. Addison-Wesley, Cambridge, Massachusets, 1949.

Apêndice A

Esse apêndice apresenta as *URLs* utilizadas nos experimentos realizados com os usuários.

***URLs* de conhecimentos gerais**:

http://www.ehow.com
http://www.wordle.net
http://www.meebo.com
http://www.nytimes.com
http://www.pponline.co.uk
http://www.teslamotors.com
http://www.tattoojohnny.com
http://www.look4design.co.uk
http://www.readwritethink.org
http://effortlessenglishclub.com
http://wikitravel.org/en/Main_Page
http://the99percent.com/tips/6585/10-laws-of-productivity
http://en.wikipedia.org/wiki/Automobile
http://en.wikipedia.org/wiki/Photography
http://en.wikipedia.org/wiki/Nutrition

***URLs* de computação**:

http://asterisq.com
http://ontologyonline.org
http://www.javaworld.com
http://www.rietta.com/firefox/index.html
http://www.universalusability.com/index.html
http://www.dlib.org/dlib/january06/guy/01guy.html
http://tomgruber.org/writing/ontology-of-folksonomy.htm
http://www2003.org/cdrom/papers/refereed/p779/ess.html
http://www.vanderwal.net/random/entrysel.php?blog=1635
http://www.smashingmagazine.com/2009/10/07/minimizing-complexity-in-user-interfaces
http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html
http://en.wikipedia.org/wiki/Ontology
http://en.wikipedia.org/wiki/Semantic_Web
http://en.wikipedia.org/wiki/Cloud_computing
http://en.wikipedia.org/wiki/Programming_language

Apêndice B

Esse apêndice contém o questionário de experiência e o questionário de dificuldades elaborado para os participantes responderem após o término do experimento.

Q1. Questionário de experiência sobre sistemas baseados em *tagging*. Nas perguntas abaixo, marcar a alternativa que melhor se aplica ao seu caso.

1 Qual seu nível de formação?

( ) Graduando ( ) Graduado

( ) Especializando ( ) Especialista

( ) Mestrando ( ) Mestre

( ) Doutorando ( ) Doutor

2 Determine seu conhecimento/grau de experiência em sistemas baseados em *tagging*.

( ) Eu nunca usei um sistema baseado em *tagging.*

( ) Já utilizei previamente algum sistema baseado em *tagging* como requisito de um trabalho.

( ) Utilizo frequentemente os sistemas baseados em tagging.

( ) Os sistemas baseados em *tagging* fazem parte de meu trabalho (dia a dia).

3 Determine sua experiência para realizar categorizações em sistemas baseados em *tagging*.

( ) Nunca realizei nenhuma categorização.

( ) Já realizei algumas categorizações como parte de alguns trabalhos.

( ) Realizo categorizações periodicamente na universidade/empresa/casa.

( ) Realizo categorizações constantemente, isso faz parte da minha rotina para organizar as informações.

Q2. Questionário de dificuldades na utilização da recomendação de *tags*.

4 Você enfrentou problemas ao realizar as categorizações?

( ) Sim

( ) Não

Qual?

5 Você precisou de ajuda para completar o experimento?

( ) Sim

( ) Não

Qual?

6 Qual o nível de dificuldade enfrentado para realizar o experimento completo? (Entre 0 a 10)

( )

7 O experimento realizado foi de difícil navegação? (Entre 0 a 10)

( )

8 Comentários gerais:

Apêndice C

O banco de dados para a realização do experimento de recomendação de *tags* foi modelado apenas para suprir as necessidades. A seguir uma breve descrição sobre cada tabela do modelo de entidade e relacionamento (ver Figura 45):

- ***usersystem***: Responsável por armazenar as informações do participante: nome e usuário;
- ***system***: Responsável por armazenar as abordagens avaliadas;
- ***categorization***: Responsável por armazenar as informações de cada recurso *Web* categorizado;
- ***userChoices***: Responsável por armazenar as *tags* adicionais às recomendadas para o usuário em uma categorização; e
- ***recommendation***: Responsável por armazenar cada *tag* recomendada e se elas foram utilizadas pelo usuário na categorização. Essa tabela possui três campos-chaves que podem ser vistos na Tabela 9.

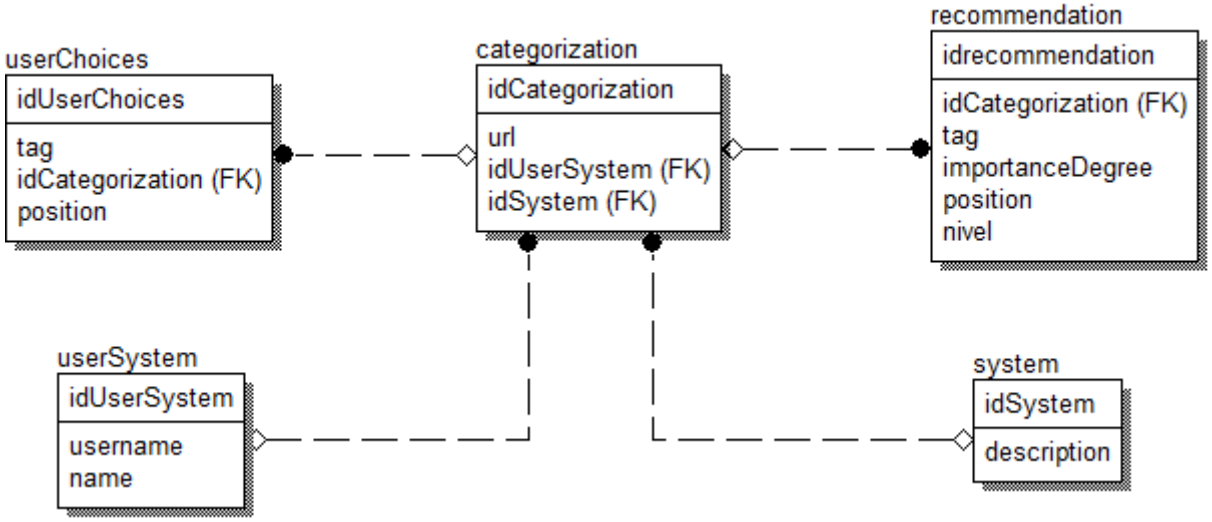

Figura 45. Modelo de entidades e relacionamentos do experimento.

Tabela 9. Descrição dos campos-chave da tabela recommendation.

Campo	**Descrição**
position	Indica a posição que a *tag* foi usada na categorização do recurso *Web*.
importanceDegree	Indica a importância ou o grau de relevância que a *tag* pertenceu na geração da recomendação.
nivel	Indica o nível da *tag* recomendada, isto é, mostra se a *tag* é de nível principal ou de nível hiperônima.

Apêndice D

Esse apêndice apresenta um artigo publicado referente esta dissertação até o momento.

Título	*An Approach to Enrich Users' Personomy Using Semantic Recommendation of Tags*
Autores	Marcelo R. Borth, Sérgio R. P. da Silva, Josiane M. P. Ferreira e Valéria D. Feltrim
Evento	*III International Workshop on Web and Text Intelligence*
Local	São Bernardo do Campo, SP, Brasil
Data	Outubro, 2010