CARLOS ALBERTO MEIER BASSO

UMA PROPOSTA PARA A EVOLUÇÃO DE ONTOLOGIAS A PARTIR DE PERSONOMIAS EM SISTEMAS BASEADOS EM *TAGGING*

MARINGÁ

2009

CARLOS ALBERTO MEIER BASSO

UMA PROPOSTA PARA A EVOLUÇÃO DE ONTOLOGIAS A PARTIR DE PERSONOMIAS EM SISTEMAS BASEADOS EM *TAGGING*

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual de Maringá, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Sérgio Roberto Pereira da Silva

MARINGÁ

2009

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

B322p

Basso, Carlos Alberto Meier
Uma proposta para a evolução de ontologias a partir de personomias em sistemas baseados em *tagging* / Carlos Alberto Meier Basso. -- Maringá, 2009.
103 p. : figs., tabs.

Orientador : Prof. Dr. Sérgio Roberto Pereira da Silva.
Dissertação (mestrado) - Universidade Estadual de Maringá, Programa de Pós-Graduação em Ciência da Computação, 2009.

1. Web - Recuperação da informação. 2. Tagging - Identificação de semântica. 3. Ontologia - Organização da informação. 4. Recuperação da informação - Web - Ontologia. I. Silva, Sérgio Roberto Pereira da, orient. II. Universidade Estadual de Maringá. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 21.ed. 006.333

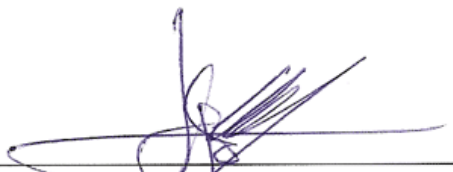
CARLOS ALBERTO MEIER BASSO

UMA PROPOSTA PARA A EVOLUÇÃO DE ONTOLOGIAS A PARTIR DE PERSONOMIAS EM SISTEMAS BASEADOS EM *TAGGING*

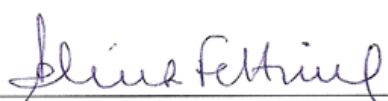
Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Estadual de Maringá, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Aprovado em 28/08/2009.

BANCA EXAMINADORA



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



Profa. Dra. Valéria Delisandra Feltrim
Universidade Estadual de Maringá – DIN/UEM

Prof. Dr. Cesar Augusto Tacla
Universidade Tecnológica Federal do Paraná – CPGEI/UTFPR

Agradecimentos

A Deus;

Aos meus pais Rui e Gerda por todo incentivo e apoio em mais essa etapa da minha vida;

A minha irmã Ângela, minha sobrinha Nattalie e minha namorada Michelli pelo carinho, companhia e risos sempre que foi possível;

Ao meu orientador, professor Sérgio Roberto P. da Silva, pela confiança e por ter me guiado nessa jornada;

Aos colegas do Grupo de Sistemas Interativos Inteligentes pela amizade, incentivos e sugestões, em especial: Josiane M. P. Ferreira, Valéria Delisandra Feltrim e Roberto Pereira;

Finalmente, agradeço ao apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Resumo

Devido à facilidade com que usuários criam e publicam conteúdo na *World Wide Web*, nunca antes houve tanta informação disponível nesse meio. Isso gera uma sobrecarga de informação sobre os usuários, fazendo com que o processo de identificação da informação desejada exija um esforço cognitivo grande por parte dos mesmos para discernir o que é relevante e o que não é. Por essa razão, tornam-se necessários meios para organizar esses dados de forma mais adequada para facilitar o processo de identificação da informação desejada. Uma forma de fazer isso é utilizando taxonomias ou ontologias. A utilização desse tipo de estrutura é interessante para a recuperação da informação, mas sua definição é muito custosa e incompatível com a quantidade de informações disponível na *web*, uma vez que é uma tarefa complexa para a maioria dos usuários. Dessa forma, a técnica de *tagging* tem se tornado muito popular, sendo sua grande vantagem a facilidade com que os próprios usuários categorizam a informação. No entanto, com um aumento expressivo no número de categorizações é gerada novamente uma sobrecarga de informação sobre os usuários no momento da recupração de informação. Por esta razão, propomos nessa dissertação utilizar os benefícios da técnica de *tagging* para a categorização da informação e os benefícios das ontologias como estrutura para a recuperação de informação. Nossa proposta utiliza uma metodologia não-supervisionada para a evolução de ontologias a partir dos dados de *tagging* de uma pessoa, ou seja, de sua personomia. Essa técnica visa representar o modelo mental do usuário referente ao seu conjunto de categorizações para contornar problemas inerentes a utilização de *tags* no processo de recuperação de informação, considerando alguns estudos de psicologia cognitiva.

Abstract

Creating and publishing content in World Wide Web has become much easier for the users. As a result, there has never been so much information available in that mean before. It generates an information overload on the users since the process of identifying the desired information requires a great cognitive effort from them in order to discern which is relevant and which is not. Therefore, means to organize data more adequately become necessary to facilitate the identification of the desired information. One way to carry this out is the use of taxonomies or ontologies. The definition of this kind of structure is interesting for the information retrieval process, but it is costly and incompatible with the huge amount of information available on the web, since this is a complex task for most users. For that reason, the tagging technique is becoming popular once its main advantage is the ease with which users categorize the information. However, with a significant increase in the amount of categorizations another information overload is generated on users at the moment of the information retrieval. For this reason, in this dissertation we propose the use of the benefits of tagging process for the information categorization and the benefits of ontologies as a structure for helping in the information retrieval process. Our proposal uses an unsupervised approach for the ontologies emergence from a person's tagging data — its personomy. This technique aims to shape the user's mental model relating to his/her tagspace to solve problems inherent to the tags use in the information retrieval process, considering some cognitive psychology studies.

Lista de Ilustrações

Lista de Abreviaturas e Siglas

AJAX	*Asynchronous JavaScript and XML*
API	*Application Programming Interface*
DHTML	*Dynamic HTML*
DOM	*Document Object Model*
HTML	*Hypertext Markup Language*
IC	*Information Content*
LCS	*Last Common Subsumer*
OWL	*Web Ontology Language*
RDF	*Resource Description Framework*
RSS	*Really Simple Syndication*
TM	*TagManager*
TOM	*TagOntologyManager*
URL	*Uniform Resource Locator*
W3C	*World Wide Web Consortium*
XML	*Extended Markup Language*

Sumário

Capítulo I

Introdução

Nos últimos anos temos acompanhado um crescimento de grandes proporções no volume de informação que chega até nós pelos meios de comunicação, dentre eles a *World Wide Web*. Esse crescimento do conteúdo da *web* vem ocorrendo devido à facilidade com que pessoas comuns podem publicar informações, juntamente com a ausência de mecanismos de controle de qualidade para as publicações. Essa quantidade excessiva de informações na *web* é conhecida como **sobrecarga de informação** (*information overload*) e chega a ser prejudicial aos usuários, pois ela exige um esforço cognitivo cada vez maior para se discernir o que é ou não relevante.

Nos primeiros sistemas de busca da *web* a informação era organizada utilizando-se de taxonomias, as quais eram mantidas por especialistas que analisavam o conteúdo das páginas *web* para colocá-las na categoria mais adequada. Hoje em dia, devido à popularização da *web*, não é plausível considerar a existência de especialistas verificando e controlando todos os conteúdos publicados. O custo e a complexidade dessa tarefa são inviáveis e incompatíveis

com a natureza aberta da *web*. Além disso, as taxonomias, mesmo fazendo sentido para a organização de alguns tipos de informações, tem um processo custoso de classificação da informação se analisado do ponto de vista do usuário, principalmente quando a informação que deve ser classificada não se encaixa em lugar algum ou pode ser encaixada em mais de uma categoria. Além disso, quando as taxonomias são muito especializadas, elas se tornam confusas para o usuário, tanto na classificação, quanto na recuperação da informação (BREITMAN, 2005).

Outra tentativa de reduzir a sobrecarga de informação sobre o usuário é a utilização de algoritmos que pré-processem a informação, reduzindo, assim, o tempo gasto com o processamento manual pelo usuário. A exemplo disso, temos sistemas de busca como o *Google*[1] que possuem algoritmos complexos para a indexação das páginas *web*, procurando melhorar os resultados obtidos nas buscas feitas pelos usuários. O problema, porém, é que mesmo trazendo resultados úteis, muitas vezes os melhores resultados não são os primeiros a serem mostrados aos usuários e, freqüentemente, nem se encontram na primeira página de resultados (IPROSPECT, 2007). Segundo dados levantados pela IPROSPECT (2007), cerca de 81,7% dos usuários de sistemas de busca não vão além da terceira página de resultados obtidos e cerca de 22,6% analisam apenas os primeiros resultados obtidos.

Uma forma de melhorar significativamente a organização da informação, e sua posterior recuperação, é pelo uso de ontologias, as quais representam conhecimento estruturado por meio de conceitos, instâncias, atributos e relações que são modelados na forma de um grafo ou rede (ECHARTE *et al*., 2007). Diferentemente das taxonomias, que permitem a definição apenas da relação pai-filho, as ontologias permitem a definição de vários tipos de relacionamentos, tais como: pai-filho, todo-parte, de associação, entre outros (BREITMAN, 2005). Essas relações semânticas mais estruturadas podem tornar o processo

[1] http://www.google.com

de recuperação da informação menos custoso para o usuário, pois ele terá outras formas de acessar a mesma informação. Além disso, segundo Breitman (2005), "por modelarem estritamente um domínio de informação, as ontologias servem como base para garantir uma comunicação livre de ambigüidades capturando e deixando explícito o vocabulário utilizado", o que também facilita o processo de recuperação da informação. O problema de utilizar ontologias para a classificação da informação é que essa é uma tarefa que também tem um alto custo cognitivo para o usuário. Isso ocorre, principalmente, porque as ontologias, assim como as taxonomias, são difíceis de se construir e manter (ECHARTE *et al.*, 2007), pois a coerência da ontologia deve ser mantida após a classificação de uma nova informação.

Se por um lado a tarefa de organizar a informação na forma de uma ontologia tem um alto custo cognitivo, a mesma tarefa utilizando rótulos de texto (*tags*) torna-se mais simples. Vários sistemas hoje em dia permitem que os próprios usuários façam uso de *tags* escolhidas por eles mesmos para organizar algum tipo de informação. Esse processo é normalmente denominado de ***tagging*** (SMITH, 2008). O conjunto de categorizações e *tags* de um usuário compõe sua **personomia** (HOTHO *et al.*, 2006) e um conjunto de personomias disponibilizados para uma comunidade de usuários caracteriza uma **folksonomia** (WAL, 2005) (MATHES, 2004). A vantagem desse processo de atribuição de *tags* é derivado do fato de que as pessoas usam seu próprio vocabulário, adicionando, assim, significado explícito ao recurso, o qual pode vir do entendimento delas sobre a informação ou sobre o objeto que está sendo categorizado (MATHES, 2004), tornando a classificação da informação uma tarefa de baixo custo.

O uso de *tagging* para organização de informação é muito interessante se analisarmos a facilidade com a qual os usuários categorizam um recurso. Por outro lado, o processo de recuperação é prejudicado por alguns motivos. O primeiro deles é que as personomias sofrem de problemas de organização e ambigüidade, que o desenvolvimento de vocabulários

controlados e esquemas hierárquicos podem melhorar (MATHES, 2004). Sobre os problemas de ambigüidade podemos citar, principalmente, o uso de acrônimos[2], homônimos[3], sinônimos[4] e o fato de que os sistemas atuais parecem não ser projetados para lidar com palavras compostas nas *tags*. O segundo motivo vem com o aumento no número de categorizações, pois a maioria dos sistemas baseados em *tagging* utiliza listas ou nuvens de *tags* para iniciar a recuperação da informação (as quais se tornam caóticas quando o número de *tags* utilizadas é grande), exigindo bastante esforço cognitivo por parte do usuário. Além disso, como a única relação entre as *tags* é a de co-ocorrência[5] (DAMME *et al.*, 2007), as opções de visualização e acesso à informação são bastante limitadas, pois, por ser semanticamente fraca, essa relação gera apenas uma estrutura plana entre as *tags*.

De modo geral, pode-se dizer que o problema de recuperação de informação dos sistemas de *tagging* está relacionado à dificuldade de lembrança do termo associado ao recurso, tarefa na qual os seres humanos têm dificuldades cognitivas (ANDERSON, 1995). Uma possibilidade para contornar esse problema seria utilizar os ganhos da ontologia como estrutura para recuperação da informação e do *tagging* para organização da informação. Alguns autores (GRUBER, 2005; WU 2006) propõe ontologias para a representação de dados de *tagging*, porém, não adianta transpor uma estrutura semanticamente mais fraca (do *tagging*) em uma mais forte (de uma ontologia) para suprir suas limitações. Outros autores (BEGELMAN, 2006; MIKA, 2005) propõem técnicas baseadas em estatísticas de co-ocorrência entre as *tags* para identificar conjuntos (*clusters*) de *tags* correlatas, cada um representando um conceito ou uma faceta. A técnica de *clustering* permite emergir alguma estrutura a partir das *tags* de um usuário, podendo, por exemplo, auxiliá-lo na navegação no

[2] Acrônimos são popularmente conhecidos como siglas. Ex.: "*NLP*" é o acrônimo de "*Natural Language Processing*".

[3] Homônimos são palavras com a mesma escrita, mas significado diferente. Ex.: "Jaguar" pode estar se referindo a um animal ou uma marca de carros.

[4] A é sinônimo de B se A e B possuem o mesmo significado. Ex.: "*airplane*" é sinônimo de "*plane*".

[5] Quando duas tags são utilizadas em conjunto em uma mesma categorização

conjunto de *tags*. Um problema dessa técnica, é que a relação entre as *tags* não é explícita, ou seja, sabe-se que um par de *tags* está relacionado, mas fica difícil determinar automaticamente que tipo de relação é essa (ex.: equivalência, generalização, etc.), deixando também em aberto os problemas de *tags* sinônimas e da falta de níveis hierárquicos explícitos entre as *tags*. Por esta razão, alguns estudos como o de Laniado *et al.* (2007), o de Specia e Motta (2007) e o de Angeletou *et al.* (2008) utilizam outras fontes de informação semântica para emergir uma estrutura a partir das *tags* e/ou obter termos relacionados às mesmas.

O trabalho de Laniado *et al.*(2007) trata exclusivamente da exibição das *tags* na forma de uma hierarquia de vários níveis para abstrair do usuário o grande número de *tags* de sua personomia no sistema *Delicious*, porém, não trata o problema da ambigüidade de uma *tag* ao recuperar a informação categorizada. Já os trabalhos de Specia e Motta (2007) e de Angeletou *et al.* (2008) focam em identificar o sentido das *tags* de um *cluster* e associá-las a entidades da *web* semântica em outras ontologias, para que, entre outras coisas, agentes de *software* possam "entender" o significado dos dados enriquecidos. Dessa forma, a semântica obtida por esses dois últimos estudos é mais voltada à interpretação dos dados de *tagging* por computadores, deixando de lado a possibilidade de ajuda na lembrança dos usuários dos termos utilizados e não se preocupando em gerar estruturas que possam auxiliar na navegação no espaço de *tags*. Além disso, essas propostas demandam um grande número de *tags*, que normalmente são obtidas a partir de um grupo de usuários em sistemas baseados em folksonomia, sem dar ênfase, por exemplo, na recuperação de informação em uma personomia.

Dado o espaço de problemas apresentado, propomos a elaboração de um processo para a evolução de ontologias com diversas relações (hierárquicas e não hierárquicas) com o intuito de facilitar a tarefa de recuperação da informação a partir de uma personomia. Ao contrário das metodologias de Specia e Motta (2007) e de Angeletou *et al.* (2008), cujo foco é

associar *tags* a conceitos da *web* semântica, nosso objetivo é emergir uma estrutura de vários níveis a partir das *tags* de uma personomia, visando contornar/minimizar vários dos problemas presentes nos sistemas baseados em *tagging*. Para tal, as principais ações propostas são as seguintes:

- Agrupar sinônimos e acrônimos para permitir buscas por conceitos, uma vez que usuários têm mais facilidade na lembrança de conceitos do que de rótulos de texto (ANDERSON, 1995);
- Permitir que sejam feitas buscas mais ou menos abrangentes para retornar um conjunto de conteúdos de interesse;
- Possibilitar buscar recursos minimizando o problema de ambigüidade das *tags*; e
- Permitir que as *tags* sejam mostradas para o usuário em formas alternativas as caóticas nuvens e listas de *tags*.

Existem vários problemas a serem contornados ao emergir ontologias a partir de *tags*. Um dos problemas é que as *tags* empregadas em uma categorização são relacionadas entre si apenas por co-ocorrência, tornando-se necessária outra fonte de informações para obter relações semanticamente mais fortes entre elas. Existem várias taxonomias, tesauros e ontologias cujas relações podem ser utilizadas para a emergência de uma estrutura de vários níveis a partir das *tags*. Para se obter relações de hierarquia, como as de generalização/especialização, todo/parte, além de termos alternativos, uma das possibilidades é a de se empregar relações lingüísticas. Por essa razão, optamos pelo uso da *WordNet*[6] (WORDNET, 2006), que é um grande banco de dados léxicos eletrônico com relações formais entre seus termos. Porém, como as *tags* são atribuídas livremente pelo usuário, muitas vezes elas possuem variações de escrita, devido a ocorrência de plurais, de termos compostos,

[6] http://wordnet.princeton.edu/

etc. Dessa forma, as *tags* devem ser normalizadas (para reduzir suas derivações, flexões e identificar os *tokens* de termos compostos) antes da obtenção das informações na *WordNet*.

Outro problema que deve ser tratado ao trabalhar com relações lingüísticas é o da polissemia, visto que uma mesma *tag* pode ter vários significados. Dessa forma, torna-se necessária também que seja efetuada uma desambiguação do sentido de algumas *tags* para evitar que termos ou conteúdos sem relação ao que o usuário tem interesse sejam sugeridos. A desambiguação pode ser feita de forma manual ou automática. Neste trabalho optou-se por um processo automático, uma vez que os usuários freqüentemente possuem centenas de categorizações e demandaria muito esforço cognitivo para a definição manual do sentido das *tags*. Isso pode ser feito comparando entre si os sentidos possíveis das *tags* co-ocorrentes de uma categorização e a escolha do sentido desejado é feita a partir de resultados de métricas (LIN, 1998; BARNERJEE e PEDERSEN, 2002) baseadas na *WordNet.* Como forma de melhorar o processo, outras informações das categorizações também podem ser utilizadas para ajudar na identificação do seu contexto de uso, como, o título e/ou a descrição do recurso categorizado, bem como, em alguns casos, o próprio recurso categorizado. Após a normalização, obtenção e desambiguação dos sentidos na *WordNet*, as relações entre as *tags*, bem como termos auxiliares, podem ser obtidos, completando o processo de evolução de uma ontologia a partir de uma personomia.

Esta dissertação está organizada da seguinte forma: no Capítulo II são apresentadas as principais bases teóricas para o entendimento deste estudo, incluindo algumas estruturas de classificação e categorização de informação, além de algumas pesquisas que estão sendo realizadas para obter melhorias nesses processos. No Capítulo III é apresentada a metodologia adotada para a evolução de ontologias a partir de personomias. No Capítulo IV são apresentados alguns estudos preliminares para a utilização das ontologias de personomias como: possibilidades alternativas para a exibição/navegação no espaço de *tags*, possibilidades

mais avançadas de recuperação de recursos categorizados e a sugestão de termos semânticos para a categorização. Finalmente, no Capítulo V são apresentadas as conclusões e considerações finais do presente estudo.

Capítulo II

Taxonomias, Ontologias e *Tagging*

Encontrar uma coleção de documentos e não conseguir encontrar a informação que sabemos que existe é um problema tão antigo quanto a existência das coleções de documentos (GARSHOL, 2004). Atualmente, a facilidade do acesso à *Internet* e a facilidade com que usuários criam e publicam conteúdos na *web* são fatores que contribuem para que encontremos uma quantidade nunca antes vista de informação disponível. Devido a essa sobrecarga de informação, torna-se indispensável a utilização de formas e técnicas para a categorização e organização da mesma a fim de facilitar o processo de recuperação pelos usuários. Em Rosch (1978) são apresentados os princípios da "**economia cognitiva**" e da "**percepção da estrutura do mundo**". O primeiro diz que a tarefa dos esquemas de categorização é o de prover o máximo de informação com o menor esforço cognitivo. O segundo argumenta que os objetos do mundo são percebidos com uma estrutura co-relacional alta e não apenas como um conjunto desestruturado de atributos co-ocorrentes. Dessa forma,

o máximo de informação com o mínimo de esforço cognitivo é obtido quando as estruturas mapeiam a nossa percepção do mundo da forma mais próxima possível da realidade.

Neste capítulo são apresentadas as bases teóricas para os principais tópicos abordados no decorrer deste trabalho. São explicadas algumas formas de organização da informação, destacando seus pontos fortes e fracos e como essas estruturas podem ajudar a atingir os objetivos citados no Capítulo I. Entre os tópicos abordados estão as taxonomias, as ontologias e o *tagging*, além de alguns exemplos dessas estruturas e algumas propostas para organizar os dados de sistemas baseados em *tagging* em ontologias.

2.1 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". Em uma definição mais contextualizada ná área da tecnologia de informação, Daconta *et al*. (2003) diz que uma taxonomia "é a classificação de entidades de informação no formato de uma hierarquia de acordo com relacionamentos que estabelecem com entidades do mundo real que representam". Resumindo, taxonomias servem para classificar informação de forma hierárquica, como uma árvore, utilizando relacionamentos de generalização/especialização (BREITMAN, 2005). Um exemplo de uma taxonomia simples sobre objetos básicos é mostrado na Figura 1.

Podemos observar que esses conceitos representam entidades do mundo real e que as entidades mais especializadas estão agrupadas em uma entidade mais generalizada (*i.e.*, "Mobília"). Quanto aos tipos de taxonomias, podemos ter as hierarquias "tipo-de" formais e as "tipo-de" informais. As hierarquias tipo-de formais incluem instâncias de um domínio bem definido, sendo que os relacionamentos de generalização são respeitados integralmente. Como exemplo, podemos citar uma taxonomia de seres vivos na qual cada elemento possui um lugar

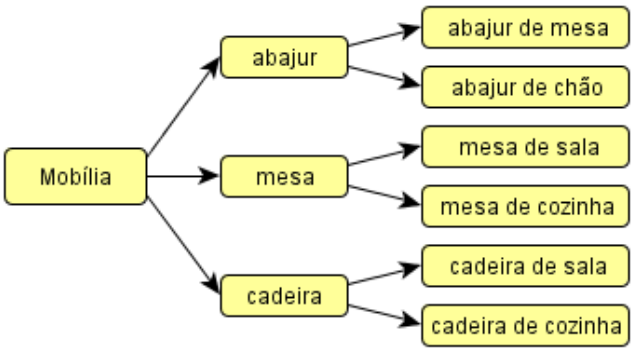


Figura 1: Exemplo de uma taxonomia usada em pesquisa de objetos básicos (adaptado de ROSCH, 1988).

certo e imutável na hierarquia. Já nas taxonomias que estão entre as hierarquias tipo-de informais, os conceitos podem ser agregados a uma categoria, mesmo que não respeitem integralmente o conceito de generalização. Um exemplo deste tipo de estrutura é a do *Open Directory Project* (NETSCAPE, 2009), que classifica páginas da *web* de acordo com sua estrutura hierárquica, porém, sem necessariamente respeitar de forma integral o conceito de generalização. Segundo Breitman (2005), a semântica desses relacionamentos é considerada fraca.

Apesar da utilização de taxonomias fazer sentido em muitos domínios, muitas vezes do ponto de vista do usuário que deseja classificar sua informação elas são confusas e rígidas. São confusas porque os usuários freqüentemente não entendem classificações muito especializadas, pois essas impõem um alto custo cognitivo de definição de um objeto dentro das classes pré-programadas da hierarquia. Isso pode forçar os usuários a verem o mundo de forma não familiar devido à terminologia utilizada na criação da taxonomia (SUNDELOF, 2005). Já a rigidez das taxonomias simples se deve ao fato de que um objeto pode possuir apenas uma classe — a classe mãe — e o único tipo de relacionamento existente é o de generalização (SMITH, 2008). Quando o objeto se enquadra em mais de uma classe, o usuário tende a classificá-lo em apenas uma delas ou duplicar a informação em mais classes (BREITMAN, 2005). Outra deficiência é que quando se deseja classificar um recurso em uma

classe que ainda não exista na estrutura, deve-se aguardar a inserção dessa classe por pessoas que fazem a manutenção da taxonomia, o que não é nada prático, além de ser bastante custoso, principalmente em uma época na qual a quantidade de informação cresce rapidamente (SUNDELOF, 2005).

Quando o assunto analisado é a recuperação da informação nas taxonomias, elas se tornam interessantes pelo fato de permitirem uma abstração da informação por meio de estruturas hierárquicas. Essas estruturas podem servir para ajudar os usuários a especializar a navegação até chegar à informação desejada, porque a informação de alto nível contribui para a interpretação dos níveis mais baixos, além de estimular a lembrança (ANDERSON, 1995). Porém, acreditamos que se a estrutura for muito especializada e possuir muitos níveis, o esforço cognitivo do usuário para encontrar uma informação tende a aumentar.

2.2 Ontologias

As ontologias são outra forma de organização de informação na qual pode-se atribuir características ou propriedades aos objetos, além de possibilitar a definição de vários tipos de relacionamentos, como todo-parte, causa-efeito, associação, entre outros (BREITMAN, 2005). Uma definição comumente encontrada no contexto de compartilhamento da informação é que as "ontologias são especificações formais e explícitas de conceitualizações compartilhadas" (GRUBER, 1993). Breitman (2005) discorre sobre a definição de Gruber afirmando o seguinte:

> Aqui uma "conceitualização" representa um modelo abstrato de algum fenômeno que identifica os conceitos relevantes para o mesmo; "explícita" significa que os elementos e suas restrições estão claramente definidos; "formal" significa que a ontologia deve ser passível de processamento automático; e "compartilhada" reflete a noção de que uma ontologia captura conhecimento consensual, ou seja, aceito por um grupo de pessoas.

Uma ontologia define um conjunto de primitivas representacionais que visam modelar um domínio de conhecimento. Essas primitivas representacionais são tipicamente classes (ou conjuntos), atributos (ou propriedades) e relacionamentos (ou relações entre os membros das

classes). As definições das primitivas representacionais incluem informação sobre seus significados e restrições em sua aplicação lógica e consistente (GRUBER, 2008). Dessa forma, como uma ontologia modela rigorosamente um domínio, ela se torna um conjunto de conceitos que pode ser usado por agentes de software para dialogar empregando uma linguagem em comum (ECHARTE *et al.*, 2007). Um exemplo de uma ontologia obtida a partir de dados da *ConceptNet*[7] (LIU e SINGH, 2004) pode ser visto na Figura 2. Nessa figura podem ser observados diversos tipos de relações entre os conceitos de uma ontologia para representar um domínio bem específico.

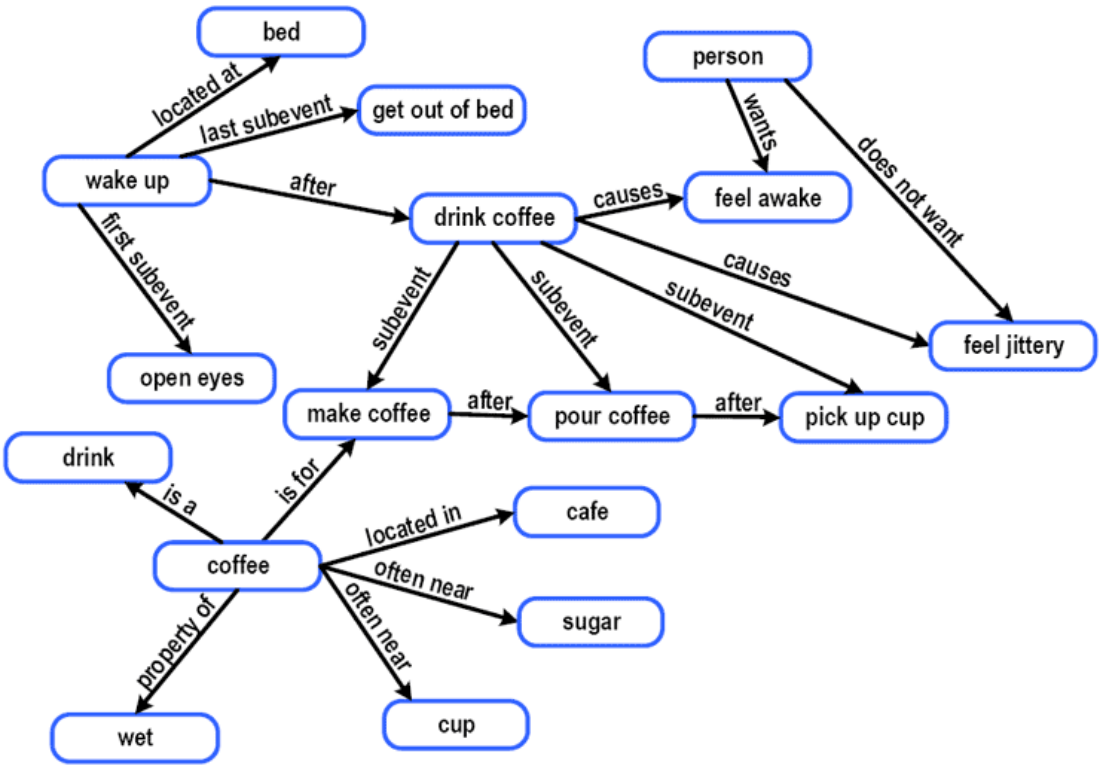


Figura 2: Um exemplo de uma ontologia com alguns conceitos e relações que mapeiam o "senso comum" dos seres humanos (LIU e SINGH, 2004).

Freqüentemente ontologias são tratadas erroneamente como sinônimos de *RDF*[8] (HERMAN *et al.*, 2004) e *OWL*[9] (MCGUINESS e HARMELEN, 2004), que são linguagens

[7] Ontologia que tenta mapear o "senso-comum" dos seres humanos para acesso por computadores.

[8] *Resource Description Framework* é uma linguagem para a definição de ontologias.

[9] *Web Ontology Language* é uma linguagem para a definição de ontologias.

para definir uma estrutura para documentos e explicitar relações semânticas entre diferentes recursos (ECHARTE *et al*., 2004). A definição de ontologias nessas linguagens é também a base para a *web* semântica (BREITMAN, 2005).

Como vantagem da utilização de ontologias, podemos citar que elas estão entre as mais eficientes estruturas para navegação, podendo auxiliar na busca baseada em palavras-chaves. Além disso, pelo fato de modelarem estritamente um domínio de informação, as ontologias servem como base para garantir uma comunicação livre de ambigüidades, capturando e deixando explícito o vocabulário utilizado nas aplicações semânticas (BREITMAN, 2005).

2.2.1 *WordNet*

Um exemplo interessante de uma ontologia representando conceitos relacionados semananticamente é a *WordNet*, a qual é bastante utilizada em aplicações de inteligência artificial e de análise automática de textos, além de ser facilmente utilizada por seres humanos como um dicionário ou tesauro (WORDNET, 2006). A *WordNet* foi desenvolvida inicialmente para a língua inglesa e suas características mais ambiciosas são a organização da informação léxica por meio dos significados das palavras (em vez das formas das palavras) e o fato dela ter sido desenvolvida com base em teorias psicolingüísticas, concernentes à organização do léxico na memória humana, chamado de léxico mental (MILLER *et al.*, 1993). Na *WordNet*, substantivos, verbos, adjetivos e advérbios são agrupados em conjuntos de sinônimos cognitivos chamados de *synsets*, cada um expressando um conceito distinto. Por exemplo, na *WordNet*, o conjunto *{world, human race, humanity, humankind, human beings, humans, mankind, man}* é um *synset* consistindo de palavras e colocações definidas como "todos os habitantes da Terra" *("All inhabitants of the earth")* (KIKAS e TREUMUTH, 2007). Sendo assim, palavras homônimas podem estar presentes em mais de um *synset*

dependendo de quantos significados elas possam ter. Até o início do ano de 2009, a *WordNet* continha mais de 155 mil palavras distintas, distribuídas em mais de 117 mil *synsets*, totalizando quase 207 mil pares de palavra/sentido, dos quais mais de 70% são de substantivos (WORDNET, 2009). Os *synsets*, por sua vez, são conectados de acordo com seu significado a outros *synsets* por meio de vários tipos de relações, dentre as quais podemos citar (FELLBAUM, 1998; KIKAS e TREUMUTH, 2007):

- **hiperonímia**: A é hiperonímia de B se B for do tipo de A. Ex: "*programming language*" é hiperônimo de "*Pascal*" e de "*Prolog*".
- **hiponímia**: B é hiponímia de A se B é do tipo de A. Ex: "*car*" e "*bike*" são hipônimos de "*vehicle*".
- **holonímia**: A é holonímia de B se B for parte de A. Ex: "*airplane*" é holônimo de "*wing*".
- **meronímia**: B é meronímia de A se B for parte de A. Ex: "*wheel*" e "*pedal*" são merônimos de "*bike*".
- **termo coordenado**: A é um termo coordenado de B se A e B compartilharem um hiperônimo. Ex: "*man*" e "*woman*" são termos coordenados, uma vez que compartilham "*human*" como hiperônimo.
- **troponímia**: o verbo A é um tropônimo de B se A é B de um modo específico/particular mas o contrário não é válido (B não é A de um modo específico/particular). Ex.: "*to trim*" e "*to slice*" são tropônimos de "*to cut*"
- etc.

Algumas dessas relações entre *synsets* podem ser observadas na Figura 3, na qual foi efetuada uma busca pela palavra "*bike*" e três *synsets* foram encontrados para a mesma. Os dois primeiros *synsets* são substantivos e representam o conceito de uma motocicleta e de uma

bicicleta respectivamente. O terceiro *synset* é um verbo e representa o ato de andar de bicicleta.

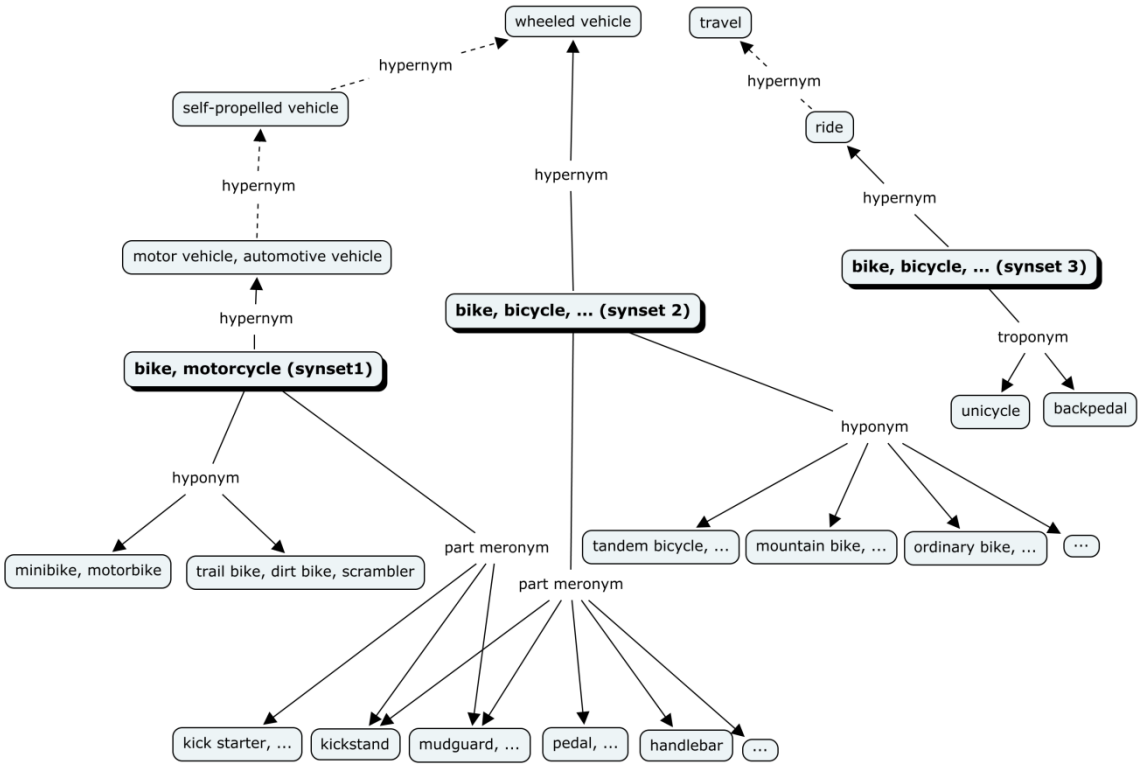


Figura 3: Exemplo das relações que podem ser obtidas na WordNet, resultantes da busca pela palavra "bike".

Outro detalhe importante sobre a *WordNet* é que as relações de hiperonímia/hiponímia entre *synsets* de substantivos podem ser interpretadas como relações de especialização entre categorias conceituais. Em outras palavras, os substantivos estão dispostos em uma estrutura semelhante a uma árvore, a qual pode ser reconstruída seguindo-se a trilha das relações de hiperonímia. Porém, para ser utilizada como uma representação de árvore por seres humanos, essa estrutura deve ser corrigida, uma vez que ela possui centenas de inconsistências, como (i) a existência de especializações em comum para categorias mutuamente exclusivas e (ii) redundâncias na hierarquia de especialização (MILLER *et al.*, 1993). Assim como os substantivos, os verbos também estão dispostos em hierarquias, porém, essa é uma hierarquia distinta da de substantivos (MILLER *et al.*, 1993). As hierarquias citadas podem ser observadas na Figura 3, na qual foram recuperadas recursivamente as relações de hiperonímia

dos dois primeiros *synsets*, os quais são substantivos, até obtemos uma hierarquia com uma raiz em comum (*i.e. "wheeled vehicle"*).

Uma detalhe da utilização da *WordNet* é que novos conceitos são adicionados apenas pelos especialistas que fazem a manutenção da mesma. Essa é uma desvantagem muito comum proveniente do uso de ontologias, bem como das taxonomias. Sua construção e manutenção é uma tarefa complexa para ser feita por usuários comuns, pois a consistência dos dados deve ser mantida quando novos conceitos são adicionados (ECHARTE *et al.*, 2007). Por esta e por outras razões uma forma de organização de informação conhecida como *Tagging* tem ganhado popularidade.

2.3 *Tagging*

Se por um lado a tarefa de organizar a informação na forma de uma ontologia tem um alto custo cognitivo, organizar a informação utilizando *tags* é uma tarefa bem simples. Vários sistemas hoje em dia permitem que os próprios usuários façam uso de *tags* escolhidas por eles mesmos para organizar algum tipo de objeto. Em outras palavras, em vez de especialistas fazerem a indexação dos objetos, os usuários a fazem.

O processo completo é essencialmente baseado em três pivôs (Figura 4): o **usuário** — que realiza a categorização; o **recurso** — que é categorizado; e as ***tags*** — que geram a categorização descrevendo o objeto (SMITH, 2008; RUSSELL, 2005).

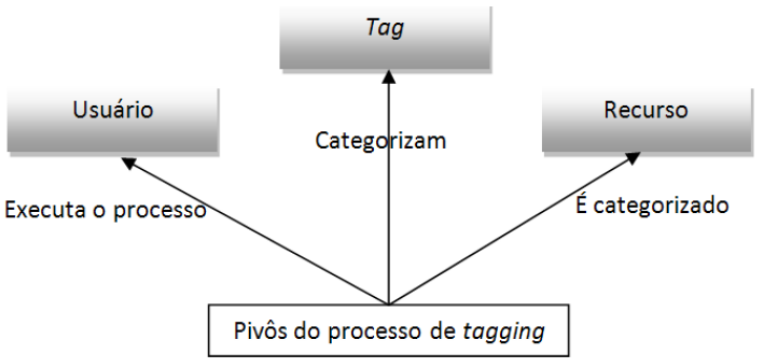


Figura 4: Os três pivôs do processo de tagging.

Esse processo de atribuição de *tags* a recursos por usuários pode ocorrer em um ambiente privado ou em um ambiente social (WAL, 2007; VOSS, 2007) e é, normalmente, denominado de ***tagging*** (SMITH, 2008). O conjunto de categorizações e *tags* de um usuário compõe sua **personomia** (HOTHO *et al.*, 2006) e um conjunto de personomias disponibilizados para uma comunidade de usuários caracteriza uma **folksonomia** (SMITH, 2008) (WAL, 2007). Um exemplo de usuários categorizando recursos com *tags* em um ambiente social pode ser visto na Figura 5.

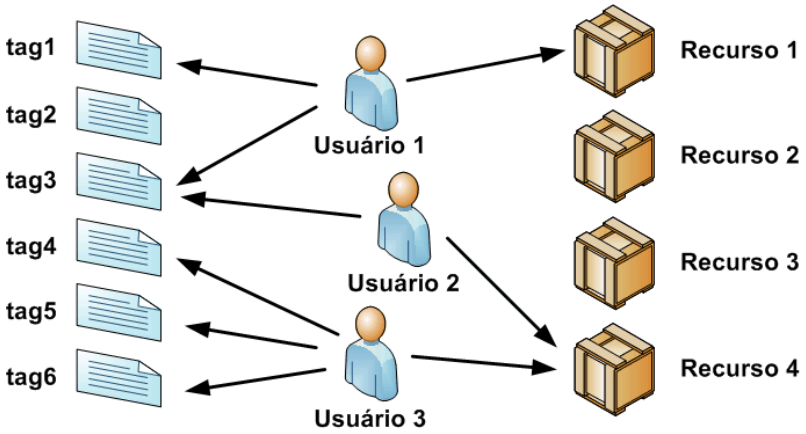


Figura 5: Exemplo de usuários utilizando tags para categorizar recursos.

Diferentemente das taxonomias, que "classificam" a informação, quando se fala em *tagging* há certa preferência em se utilizar o termo "categorização", que, segundo alguns especialistas, sugere um esquema menos rígido de organização (STURTZ, 2004; MATHES, 2004). Na categorização, os usuários podem adicionar várias *tags* a um objeto sem a necessidade de especialistas para adicionarem nova terminologia, o que é visto como uma grande vantagem em relação às taxonomias e às ontologias. A vantagem desse processo de atribuição de *tags* é derivada do fato de que as pessoas usam seu próprio vocabulário e adicionam significado explícito ao recurso, o qual pode vir do entendimento delas sobre a informação ou sobre o objeto que está sendo categorizado, sem a necessidade de aprenderem nomenclaturas taxonômicas complicadas. De fato, ao utilizar as *tags* para categorizar recursos, as pessoas estão criando meios de conectar objetos, provendo significado para seu

próprio entendimento e estabelecendo a forma como as informações serão organizadas, tudo isso com um baixo custo cognitivo (MATHES, 2004) (WU *et al.*, 2006).

Assim como a natureza não controlada do uso de *tags* em categorizações é responsável por vários fatores positivos, como, a fácil adaptação do vocabulário às necessidades dos usuários, ela também traz vários fatores negativos com seu uso. Segundo Al-Khalifa e Davis (2007), as *tags* comumente utilizadas em sistemas baseados em *tagging* podem ser classificadas em três tipos, de acordo com a intenção do usuário no momento da categorização:

- ***Tags* fatuais**: identificam fatos a respeito do recurso, como, "*blog*", "*java*" e "*programming*";
- ***Tags* subjetivas**: exprimem a opinião do usuário a respeito do recurso como, por exemplo, "interessante" e "divertido";
- ***Tags* pessoais**: relacionadas com uma necessidade pessoal de seu criador, muitas vezes, utilizadas para referências próprias ou para o gerenciamento de suas tarefas. Exemplos desse tipo de *tag* são "*toread*", "*todo*" e "*mysite*".

Devido ao fato de não haver controle sobre a forma de categorização e termos utilizados, há um risco e uma tendência à criação de informações "ruidosas" e inúteis e, conseqüentemente, à redução da utilidade dos sistemas de *tagging* (WU *et al.*, 2006; MATHES, 2004). As *tags* possuem de problemas de organização e ambigüidade que o bom desenvolvimento de vocabulários controlados e esquemas de hierarquia efetivamente melhoram (MATHES, 2004). Quando falamos em problemas de ambigüidade em sistemas baseados em *tagging* nos referimos, principalmente, ao problema dos acrônimos (siglas), dos homônimos, dos sinônimos e do fato de que os sistemas atuais parecem não ser projetados para lidar com *tags* compostas (*i.e. tags* formadas por mais de um termo separados por espaço). Tanto os acrônimos quanto os homônimos criam problemas por definirem *tags* de

escrita igual que possam se referir a recursos com significados distintos, fazendo com que duas idéias totalmente diferentes sejam categorizadas sob um mesmo termo. Como exemplo, podemos citar o acrônimo ANT, que pode ser utilizado para se referir a "Agência Nacional de Transportes", a "*Actor Network Theory*", ou mesmo a um *software* chamado "*Ant*", demonstrando vários domínios e idéias diferentes categorizados por uma mesma *tag*. Nos sistemas baseados em *tagging* também não há controle de sinônimos, conduzindo o usuário a utilizar várias palavras com o mesmo sentido para categorizar um mesmo recurso (MATHES, 2004). Outro fator que conduz a problemas é que, normalmente, os sistemas não permitem que sejam utilizados espaços como separadores de termos compostos, fazendo com que os usuários utilizem *CamelCase* ou caracteres especiais para diferenciar os termos que compõem uma *tag* (ex.: "*SemanticWeb*", "*artificial_intelligence*", etc.) (RIDDLE, 2005) (GUY e TONKIN, 2006). Além desses problemas, Guy e Tonkin (2006) afirmam que muitos usuários não dão muita atenção na forma em que utilizam *tags* para categorizar recursos. Isso contribui para que sejam comuns casos nos quais haja *tags* escritas erradas, *tags* que não seguem uma convenção (ex.: as vezes no singular e as vezes no plural) e *tags* que são utilizadas uma única vez (por não serem lembradas posteriormente). Esses problemas e inconsistências no espaço de *tags* prejudicam a recuperação de informações categorizadas.

2.3.1 Recuperação de Informação em Sistemas Baseados em *Tagging*

A recuperação da informação é um dos pontos críticos em sistemas baseados em *tagging*. A maioria desses sistemas utiliza listas ou nuvens de *tags* (*tag clouds*) para iniciar a recuperação da informação. As nuvens de *tags* são um recurso de visualização e navegação no espaço de *tags* que se tornou popular juntamente com a ascensão dos sistemas baseados em *tagging*/folksonomia. Isso ocorreu, em parte, devido ao fato de que a única relação entre as *tags* nesses sistemas é a co-ocorrência, ou seja, quantas vezes um par de *tags* foi utilizado em

conjunto para categorizar recursos distintos de um mesmo usuário. Essas relações são consideradas semanticamente fracas e geram uma estrutura plana entre as *tags*, o que limita a capacidade de representação e de buscas nessas informações. Devido à falta de níveis hierárquicos entre as *tags*, as nuvens de *tags* se tornam uma possibilidade para mostrar as *tags*, dando ênfase às mais utilizadas em categorizações (Figura 6a). As listas de *tags*, por sua vez, também exibem as *tags* em uma lista extensa, normalmente ordenada de acordo com a quantidade de uso das *tags* em categorizações ou em ordem alfabética (Figura 6b). Um problema, porém, é que essas estruturas de navegação tendem a ficar caóticas com o aumento significativo do número de termos, requerendo, assim, um grande esforço cognitivo do usuário para recuperar o que ele deseja. Na Figura 6a, podemos observar que fica difícil encontrar uma *tag* desejada, principalmente se ela for utilizada poucas vezes e estiver com pouco destaque. Na Figura 6b, por sua vez, torna-se difícil encontrar os termos porque a lista fica muito extensa, necessitando que o usuário utilize a "rolagem" da página ou da lista.

Figura 6: Exemplos de estruturas para navegação no espaço de tags: a) nuvens de tags; b) lista de tags.

Por essas razões, existem estudos visando proporcionar estruturas mais eficientes para navegação entre as *tags*. Para evitar o esforço do usuário para identificar as *tags* que deseja

em um emaranhado de muitos termos, autores como, Begelman *et al.* (2006) e Mika (2005) propõem técnicas de identificação de conjuntos (*clusters*) de *tags* relacionadas. Essa técnica se baseia em estatísticas de co-ocorrência entre as *tags*, possibilitando que os sistemas com esse recurso consigam encontrar conjuntos de *tags* relacionadas, identificando diferentes contextos. Técnica semelhante a de Begelman *et al.* (2006) é utilizada pelo sistema *Flickr*[10] (YAHOO, 2009), no qual são exibidas apenas as *tags* mais relevantes ao usuário e, no momento que ele seleciona um desses termos são exibidos os temos relacionados. O problema é que, devido à falta de níveis hierárquicos, não há critério para se organizar as *tags* e apenas um pequeno conjunto delas pode ser mostrado ao usuário. Além disso, não há conexão explícita entre o sentido das *tags* ou relações semânticas entre elas (ANGELETOU *et al.*, 2008), os quais são recursos que podem orientar o usuário na navegação no conjunto de *tags* e ajudá-los na recuperação de informação. Em Laniado *et al.* (2007) é proposta uma ferramenta para organizar as *tags* de uma personomia em uma hierarquia para ser mostrada no lugar da lista de *tags* do sistema *Delicious*. Para tal, a ferramenta utiliza a *WordNet* para a obtenção das relações hierárquicas entre os termos. Essa é uma proposta interessante, pois permite que apenas termos abrangentes sejam mostrados, permitindo ao usuário navegar e especializar sua busca. Um problema, porém, é que essa ferramenta só está disponível para o sistema *Delicious* e é necessária a instalação de *plugins* no navegador *web* do usuário.

Além de utilizar estruturas de navegação, o usuário também pode iniciar a recuperação de algum recurso categorizado por meio da digitação de *tags* no campo de busca do sistema baseado em *tagging*, porém os usuários têm dificuldades para lembrar do termo exato que utilizaram em uma categorização (ANDERSON, 1996).

A recuperação de informação nos sistemas baseados em *tagging* tradicionais também se torna crítica devido aos problemas anteriormente citados da falta de qualidade das *tags*,

[10] Sistema baseado em *tagging* para a categorização de fotos, disponível em: http://flickr.com.

incluindo o problema de sinônimos e homônimos, uma vez que o mecanismo de busca desses sistemas não consegue identificar essas relações entre os termos. É importante destacar que vários problemas discutidos nesta seção poderiam ser resolvidos utilizando estruturas como as taxonomias e ontologias, o que vem ao encontro com os objetivos deste trabalho.

2.3.2 Sistemas Baseados em *Tagging* e sua Falta de Interoperabilidade

Devido ao crescimento constante na quantidade de informações da *web* e da facilidade de categorizar recursos usando *tags*, surgiram vários sistemas baseados em *tagging* para as mais diversas funcionalidades. O *Delicious* (YAHOO, 2008) e o *Flickr* (YAHOO, 2009) estão entre os mais citados no meio científico. O primeiro é um gerenciador de *bookmarks* social, o qual permite aos usuários utilizar *tags* para categorizar *URL*s[11], facilitando assim sua recuperação. O *Delicious* utiliza o que chamamos de folksonomia larga (*broad*) a qual permite ao usuário categorizar conteúdos que não são seus (qualquer *URL* disponível na *web*). Já o *Flickr* permite que os usuários categorizem fotografias e imagens pertencentes a eles mesmos, caracterizando uma folksonomia estreita (*narrow*). Na folksonomia estreita não é permitido aos usuários categorizar um recurso que não lhes pertença (RIDDLE, 2005; RUSSEL, 2005; SHEN, 2005; STURTZ, 2004).

Um problema nos sistemas baseados em folksonomia atuais é que não há interoperabilidade entre eles (KNERR, 2006). Todos eles exigem que seus usuários criem e mantenham um vocabulário de termos utilizados para a categorização de seus recursos (sua personomia). Assim, um usuário que utilize vários sistemas baseados em folksonomia deverá manter, obrigatoriamente, um vocabulário em cada sistema, independentemente desses termos serem os mesmos ou completamente diferentes (da SILVA, 2009). Visando criar formas

[11] Localizador Uniforme de Recursos, que são basicamente endereços de recursos na *web*.

padronizadas para a representação de dados de *tagging* em um ambiente "entre aplicações" (*cross applications*), alguns autores definiram ontologias para a representação de *tagging*, conforme apresentado a seguir.

2.3.3 Ontologias para a Representação de *Tagging*

Em Gruber (2005) e Wu *et al.* (2006) são apresentadas as premissas para representar os dados de sistemas baseados em *tagging* e folksonomias em ontologias. Existe pouca diferença entre as ontologias criadas pelos autores supracitados, uma vez que os dados mínimos necessários para essa representação são o **usuário**, o **recurso** categorizado e as ***tags*** utilizadas (os três pivôs do *tagging*). Wu *et al.* (2006) adiciona ainda a data em que a categorização ocorreu e Gruber (2005) adiciona dados sobre em qual sistema a categorização ocorreu (ex.: *Flickr*, *Delicious*, etc.) e permite que usuários votem coletivamente a favor ou contra uma *tag* utilizada na categorização de um recurso (no caso de folksonomias largas). Já Newman (2005), Knerr (2006) e Echarte (2007) focaram seus trabalhos em definir as premissas de Gruber (2005) e de Wu *et al.* (2006) em *RDF* e/ou *OWL*.

Esse tipo de ontologia para a representação de folksonomias traz diversas vantagens sobre as folksonomias e ontologias puras, como (i) a utilização de repositórios de *tags* para um conjunto de aplicações baseadas em folksonomias, (ii) a possibilidade de categorizar objetos utilizando *tags* em um ambiente entre aplicações e (iii) a interoperabilidade entre os sistemas baseados em folksonomia por meio de um formato padronizado e extensível. Os dados dessas ontologias podem ser utilizados diretamente pelo usuário e/ou disponibilizados socialmente (KNERR, 2006).

Na Figura 7 pode ser visto como os perfis de usuários e dados de categorizações ficam disponíveis atualmente nos sistemas baseados em folksonomia. As informações sobre as

categorizações podem ser obtidas no formato *RSS*[12], mas cada *website* tem sua própria estrutura para a representação no formato citado.

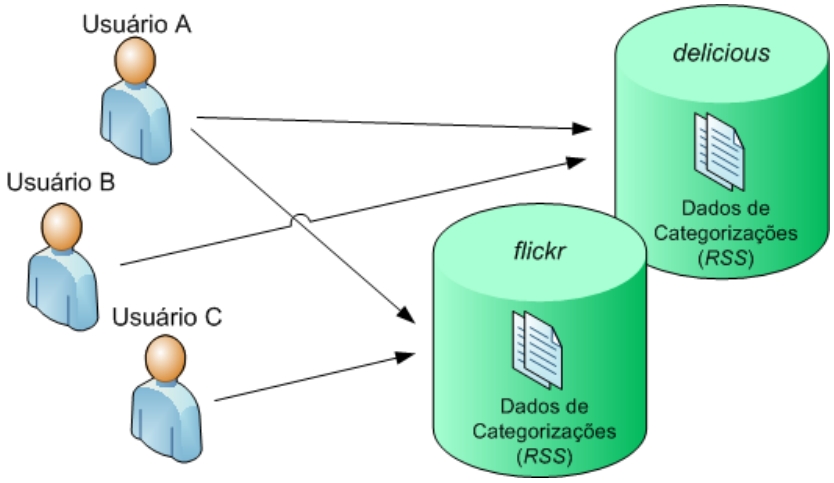


Figura 7: Sistemas baseados em folksonomia atuais, nos quais os dados de categorizações e tags são retidos no servidor (adaptado de KNERR, 2006).

Já na Figura 8 é representada a estrutura proposta por Knerr (2006), na qual os dados das categorizações (*tags.rdf*) ficam unificados e disponíveis para o usuário, e os sistemas

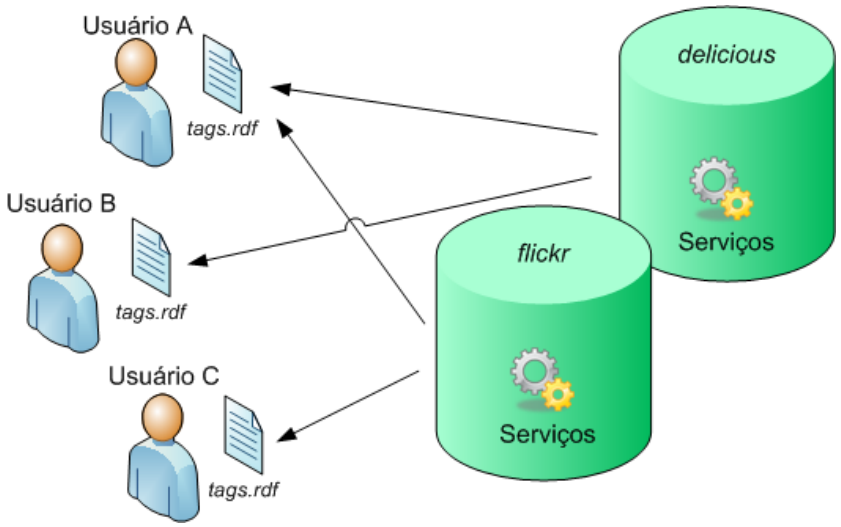


Figura 8: Com a utilização de ontologias para a representação de folksonomias, o usuário pode ficar em poder de seus dados sobre categorizações e tags criados nos sistemas baseados em folksonomia que utilizar (adaptado de KNERR, 2006).

[12] *Really Simple Syndication* permite que dados freqüentemente modificados possam ser "assistidos" por usuários que tenham interesse.

baseados em folksonomia são utilizados apenas para armazenar os recursos categorizados, por exemplo, o *Flickr* armazenaria apenas as fotos e o *Delicious* armazenaria os *bookmarks*.

Por enquanto, nenhuma das ontologias para a representação de *tagging* se tornou padrão, apesar de exitirem discussões sobre o assunto nas listas de discussão do *W3C*[13]. Além disso, para obter interoperabilidade entre os sistemas e permitir que os usuários possam gerenciar seus dados de *tagging* de forma unificada, é necessário que esses dados sejam utilizados por uma aplicação ou por um agente de *software*, conforme pode ser visto na Figura 9.

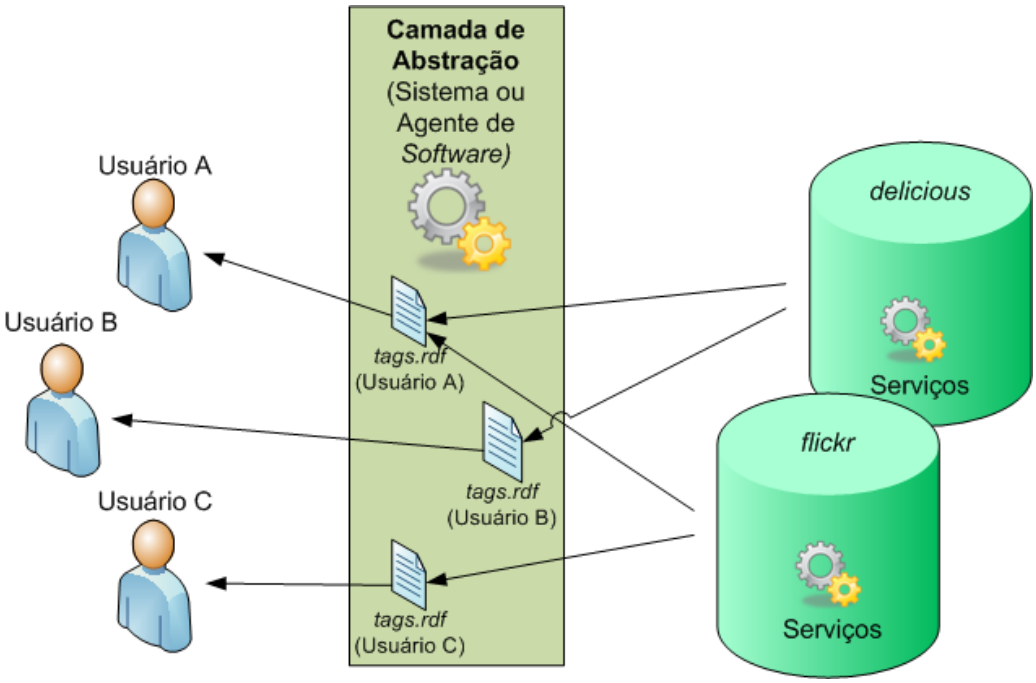


Figura 9: Exemplo de como as ontologias para a representação de tagging podem ser utilizadas em uma camada de abstração entre os sistemas de tagging e os usuários.

Enquanto não há uma ontologia padrão para a representação de *tagging*, cada sistema disponibiliza esses dados de forma diferente por meio de *APIs*[14] ou *RSS*. Uma forma de abstrair essa tarefa de obtenção dos dados de personomias nos diversos sistemas baseados em *tagging* é por meio da utilização de um *software* que funcione como um mecanismo de

[13] *World Wide Web Consortium* é um consórcio de empresas que define especificações e padrões da *web*.

[14] É uma interface que define os caminhos para uma aplicação requisitar serviços de bibliotecas de dados ou de outras aplicações.

abstração, obtendo os dados das personomias nos diferentes formatos disponibilizados e padronizando-os. Uma possibilidade para essa abstração é a utilização do sistema denominado *TagManager* (da Silva, 2009), o qual obtém e gerencia os dados de *tagging* de um conjunto de sistemas desse tipo, semelhante a camada de abstração mostrada na Figura 9.

2.3.4 O *TagManager*

Um sistema que gerencia as múltiplas personomias de um usuário é o ***TagManager***. Segundo da Silva (2009), o *TagManager* pode auxiliar o usuário monitorando sua personomia global, que é criada utilizando os dados oriundos dos diversos sistemas baseados em *tagging* que o mesmo utilize. Ainda segundo da Silva, esse monitoramento deverá trazer melhorias que serão refletidas nas personomias dos vários sistemas que um usuário utiliza e, como conseqüência, a recuperação da informação nesse sistema será melhorada tanto no contexto pessoal de busca quanto no contexto global para todos os usuários desse sistema. Atualmente o *TagManager* gerencia as personomias dos usuários nos sistemas *Delicious*, *Flickr*, *SlideShare*[15] e *YouTube*[16], permitindo que o usuário sincronize seu conjunto de *tags* e passe a monitorar seu vocabulário, permitindo ainda buscas com resultados unificados entre os sistemas citados. Outro recurso importante do *TagManager* é a possibilidade de "limpeza" no conjunto de *tags* do usuário. Esse recurso analisa vários aspectos de uma *tag* como, por exemplo, a data da última utilização e a quantidade de vezes que a mesma foi utilizada, a fim de sugerir melhorias ao usuário (CÔGO e DA SILVA, 2008).

O *TagManager* será utilizado no presente estudo para obter os dados das personomias de usuários nos diversos sistemas baseados em *tagging*. Futuramente esse sistema utilizará alguns dos benefícios das ontologias obtidas no presente estudo para a representação do

[15] Sistema baseado em *tagging* para a categorização e compartilhamento de apresentações de *slides*. Disponível em <http://slideshare.net>

[16] Sistema baseado em *tagging* para compartilhamento de vídeos, disponível em <http://youtube.com>.

espaço de *tags* e para permitir buscas semânticas no conteúdo categorizado. O *TagManager* pode auxiliar o usuário a melhorar seu vocabulário, porém não consegue identificar a semântica das *tags*. Para isso, existem algumas propostas que são discutidas na seção seguinte.

2.3.5 Alguns Estudos sobre a Emergência de Estruturas a partir de Dados de *Tagging*

Existem basicamente dois tipos de abordagens para extrair/emergir uma estrutura em sistemas baseados em *tagging*. Conforme discutido anteriormente, algumas propostas utilizam uma análise estatística baseando-se na co-ocorrência do conjunto de *tags* para identificar conjuntos (*clusters*) de *tags* relacionadas. Outra abordagem é a de propostas que utilizam outras fontes de informação, além dos dados de sistemas de *tagging*, para obter relações semânticas explícitas entre as *tags*. Como exemplo dessa abordagem, além do trabalho de Laniado *et al.* (2007) discutido na Seção 2.4.1, podemos citar os trabalhos de Damme *et al.* (2007), de Angeletou *et al.* (2008) e de Specia e Motta (2007). Em Damme *et al.* (2007) são sugeridas possibilidades para mapear diversos tipos de relações entre *tags* em uma ontologia, como, relações de sinonímia obtidas na *WordNet* e relações de equivalência entre termos em diferentes linguagens obtidas por meio do uso de dicionários eletrônicos, etc. Porém, essas possibilidades não foram implementadas. Há ainda os trabalhos de Angeletou *et al.* (2008) e o de Specia e Motta (2007), os quais também obtém relações semânticas em outras fontes de informações além das próprias folksonomias (inclusive na *WordNet*) e tem como objetivo encontrar entidades da *web* semântica em outras ontologias disponíveis na *web* para associar com as *tags*. Diferente dos trabalhos supracitados, o principal objetivo do presente trabalho é apresentar uma proposta para a evolução de ontologias a partir de dados de *tagging* visando melhorias na representação e recuperação de informação pelos usuários. Essa proposta é detalhada no capítulo seguinte.

Capítulo III

Evolução de Ontologias a partir de Personomias

Como pôde ser observado na seção anterior, existem diversas abordagens para emergir estruturas a partir de dados obtidos em sistemas baseados em *tagging*, entre elas, algumas considerando apenas o conjunto de dados pessoal (de uma personomia) e outras considerando um conjunto de dados social (de folksonomias); algumas baseadas em estatística e outras baseadas em outras fontes de informações; algumas com supervisão humana e outras não supervisionadas. Neste capítulo é descrita a metodologia da evolução de ontologias proposta por este trabalho. Esta metodologia se baseia na utilização de personomias em vez de folksonomias, uma vez que acreditamos que primeiramente devemos solucionar o problema pessoal para posteriormente atacar o problema coletivo da recuperação de informação. Da mesma forma, optamos pela utilização de uma fonte de informações semânticas para o enriquecimento da ontologia em vez de estatísticas baseadas na co-ocorrência, devido ao fato de que para obter uma semântica mais forte em técnicas como a de clusterização, seriam necessários muitos dados no espaço amostral de *tags*, requerendo assim

a utilização de dados de uma folksonomia com vários usuários, ao invés de uma personomia. Além disso, como discutido na seção anterior, com a técnica de clusterização são identificados diferentes interesses de um usuário e significados que as *tags* possuem, mas essa identificação não deixa explícita qual é a relação semântica entre os conceitos elas representam. Por esta razão, optamos pela utilização de fontes auxiliares de informação semântica para ajudar na identificação explícita das relações entre as *tags* de uma personomia. Quanto a uma abordagem ser ou não ser supervisionada, optamos pela forma não supervisionada, pois um fator que leva os usuários a utilização de *tags* para a categorização de recursos é que o custo cognitivo para o usuário é baixo se comparado com a classificação em esquemas hierárquicos e/ou ontológicos (PEREIRA e DA SILVA, 2008). Dessa forma, acreditamos que uma abordagem não supervisionada seja mais adequada, uma vez que não requer esforço cognitivo do usuário para a tomada de decisões no momento da evolução da ontologia. Certamente, uma abordagem supervisionada permitiria a obtenção de resultados com qualidade superior, permitindo que a posterior recuperação da informação fosse bastante melhorada, porém, reduziria a atratividade da utilização de *tags* para a categorização de recursos, uma vez que seria necessário um esforço maior do usuário durante o processo.

3.1 Uma Ontologia para a Representação de Semântica entre as *Tags*

Para evoluir uma ontologia a partir de uma personomia torna-se necessário, primeiramente, entender o processo de *tagging* e como pode ser expressa a semântica entre as *tags*. Para isso, analisamos algumas ontologias, como a de Echarte (2007), a de Newman (2005) e a de Knerr (2006). Como já observado na Seção 2.3.3, apesar do fato de as ontologias dos referidos autores possuírem detalhes diferentes entre elas na terminologia utilizada e na implementação, todos eles estão de acordo com as observações de Gruber (2005) sobre a representação da tarefa de *tagging* em ontologias, ou seja, que ela deve incluir

o **recurso** categorizado, as **tags** utilizadas, o **usuário** que efetuou a categorização e o **sistema** no qual a categorização ocorreu. No presente estudo se fez necessário estender essas ontologias para que se tornasse possível colocar sentidos nas *tags* e relacioná-las com outras por meio de relações semanticamente mais ricas do que a co-ocorrência. Na Figura 10 as relações em cinza representam a ontologia proposta por Knerr (2006) e as relações em preto expressam o nosso entendimento de como a tarefa de *tagging* deve ser modelada expressando relações semânticas entre as *tags*.

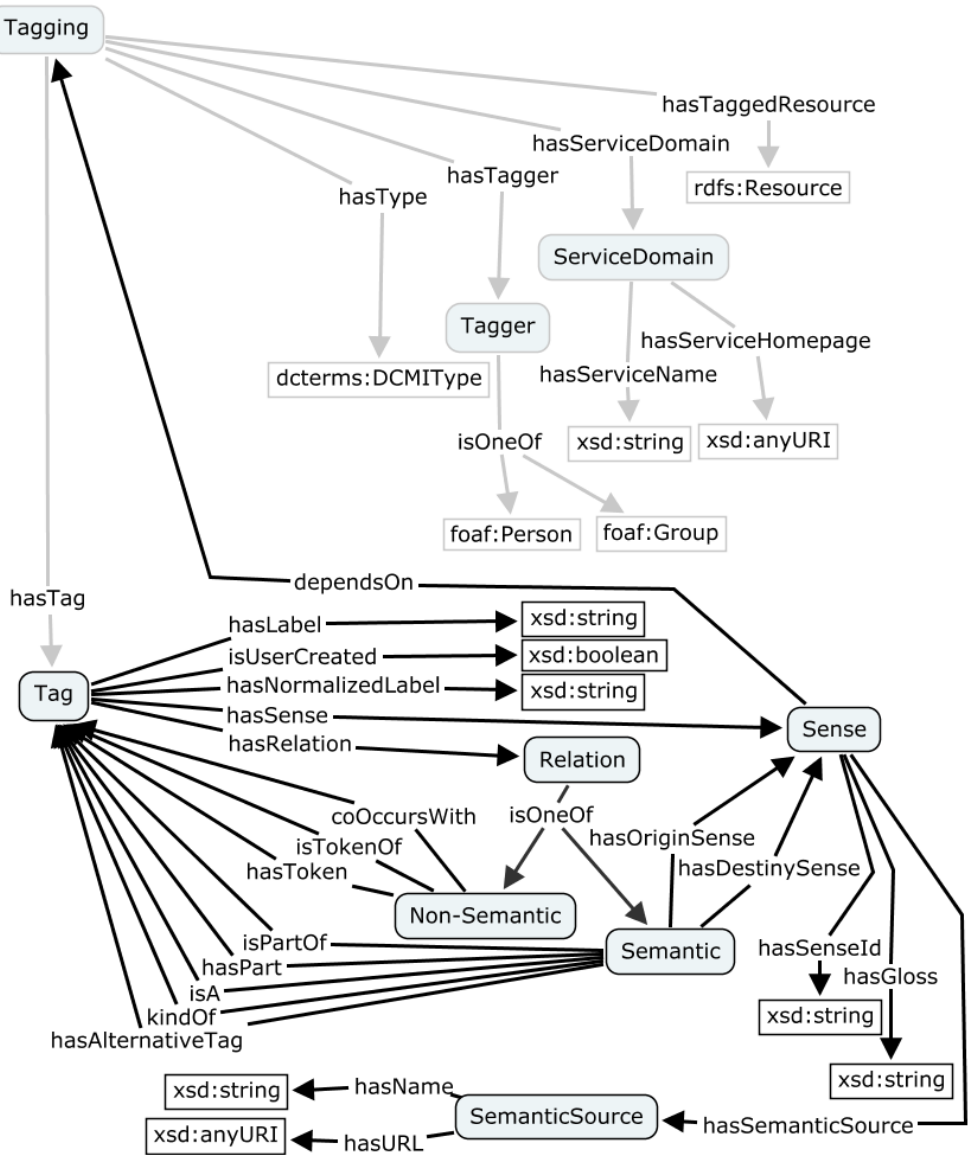


Figura 10: Ontologia que representa um tagging estendido a partir da ontologia de Knerr (2006) com relacionamentos semânticos entre as tags.

As relações da ontologia de Knerr (2006) são: *hasTag*, *hasType*, *hasTagger*, *hasServiceDomain* e *hasTaggedResource*, que representam, respectivamente: as *tags* utilizadas na categorização, o tipo do recurso categorizado (ex.: uma foto, um *bookmark*, etc.), o usuário, o sistema (domínio) no qual foi feita a categorização (ex.: o *Delicious*, o *Flickr*, etc.) e o próprio recurso. As classes da ontologia representada pela Figura 10 estão descritas na Tabela 1.

Tabela 1: Descrição das classes da ontologia.

Classe	Descrição
Tagging	Representa uma categorização.
Tagger	Representa o usuário que efetuou a categorização.
ServiceDomain	O serviço utilizado para efetuar a categorização (ex.: *Flickr*, *Delicious*)
Tag	Uma *tag* utilizada em uma categorização.
Sense	Um sentido possível de uma *tag*, obtido de uma fornte externa.
Relation	Uma relação entre duas *tags*, a qual pode ser semântica ou não semântica.
NonSemantic	Uma relação não-semântica (que não depende do sentido das *tags* relacionadas).
Semantic	Uma relação semântica (que depende do sentido das *tags* relacionadas).
SemanticSource	A fonte externa de informações semânticas utilizada para representar um sentido de uma *tag* (ex.: *WordNet*).

Os atributos adicionados para representar a semântica entre as *tags* — representados na Figura 10 — são descritos na Tabela 2.

Tabela 2: Atributos da ontologia adicionados para enriquecer as tags.

Atributo	Descrição
hasLabel Domínio: *Tag*, Tipo: *string*	O rótulo textual de uma *tag*, da forma em que foi escrita originalmente pelo usuário.
isUserCreated Domínio: *Tag*, Tipo: *boolean*	Se é uma *tag* que foi criada pelo usuário (e faz parte da sua personomia) ou se é uma *tag* de relação, obtida em uma fonte de informação semântica externa para ajudar a relacionar os termos da personomia.
hasNormalizedLabel Domínio: *Tag*, Tipo: *string*	O rótulo textual da *tag* após um processo de normalização da mesma, para que fique da mesma forma em que é representada em uma fonte externa de informações semânticas.
hasSense Domínio: *Tag*, Tipo: *Sense*	Relaciona uma *tag* com uma fonte externa de dados semânticos.
hasRelation Domínio: *Tag*, Tipo: *Relation*	Representa uma relação entre *tags*.
dependsOn Domínio: *Relation*, Tipo: *Tagging*	O sentido de uma *tag* depende do contexto de uma categorização.
isOneOf Domínio: *Relation*, Tipo: *Semantic*, *Non-Semantic*	Uma relação pode ser semântica (*ie*. dependente do sentido da *tag*) ou não semântica (ex.: estatística, lingüística, etc.).
hasOriginSense Domínio: *Semantic*, Tipo: *Sense*	Uma relação semântica depende do sentido da *tag* de origem.
hasDestinySense Domínio: *Semantic*, Tipo: *Sense*	Uma relação semântica depende do sentido da *tag* de destino.
hasSenseId Domínio: *Sense*, Tipo: *string*	O identificador de um sentido em uma fonte externa de dados semânticos.
hasSenseGloss Domínio: *Sense*, Tipo: *string*	A descrição de um sentido em uma fonte externa de dados semânticos.

hasSemanticSource Domínio: *Sense*, Tipo: *SemanticSource*	Identifica a fonte de dados semânticos a qual o sentido está descrito.
hasName Domínio: SemanticSource, Tipo: *anyURI*	Possui o nome da fonte de dados semânticos do sentido utilizado.
hasURL Domínio: SemanticSource, Tipo: *anyURI*	Possui a *URL* da fonte de dados semânticos do sentido utilizado.
coOccursWith Domínio: *Non-Semantic*, Tipo: *Tag*	Mapeia a co-ocorrência entre duas *tags*.
hasToken Domínio: *Non-Semantic*, Tipo: *Tag*	Relaciona uma *tag* agrupada com os seus *tokens*[17]. Ex.: "*SemanticWeb*" possui dois *tokens*: "*semantic*" e "*web*".
isTokenOf Domínio: *Non-Semantic*, Tipo: *Tag*	É a relação contrária de *hasToken*.
isPartOf Domínio: *Semantic*, Tipo: *Tag*	Relaciona duas *tags* das quais a *tag* de origem é parte da *tag* de destino. Ex: uma "roda" é parte de um "carro".
hasPart Domínio: *Semantic*, Tipo: *Tag*	É a relação contrária de *isPartOf*.
isA Domínio: *Semantic*, Tipo: *Tag*	Relaciona uma *tag* mais específica com uma mais abrangente. Ex.: um "carro" é um "veículo".
kindOf Domínio: *Semantic*, Tipo: *Tag*	É a relação contrária de *isA*.
hasAlternativeTag ***Domínio: Semantic, Tipo: Tag***	Relaciona duas tags que possuam o mesmo sentido. Ex.: "plane" é um termo alternativo para "airplane".

[17] Um *token*, neste caso, identifica um pedaço de uma *tag* composta, o qual muitas vezes possui um significado mais abrangente do que a *tag* de origem.

A importância destes novos atributos está no fato de que eles permitem mapear uma semântica mais forte entre as *tags* em relação a co-ocorrência (que é a única relação entre as tags nos sistemas baseados em *tagging* tradicionais).

As relações semânticas entre as *tags* são especialmente importantes, pois, segundo Anderson (1995), "depois de processar uma mensagem lingüística, as pessoas lembram apenas do sentido e não exatamente das palavras utilizadas". Em outras palavras, os usuários lembram com mais facilidade do sentido das *tags* utilizadas na categorização, mas não de sua forma escrita exata, o que pode prejudicar a recuperação da informação na personomia. Todas as relações semânticas presentes na ontologia da Figura 10 podem ajudar o usuário na tarefa de relembrar o conceito usado na categorização, uma vez que relacionam termos de acordo com seus sentidos, criando, assim, grupos de palavras relacionadas por seu conceito. A relação *hasAlternative* é ideal para resolver esse problema, uma vez que ela relaciona uma *tag* a formas alternativas de escrita que representam o mesmo conceito. As relações *hasPart* e *isPartOf* agrupam partes a um todo e vice-versa, recursos que também podem auxiliar o usuário a recuperar a informação, pois o ajudam no reconhecimento do conceito utilizado na categorização, tarefa em que os seres humanos tem mais facilidades cognitivas do que a lembrança (ANDERSON, 1995). O mesmo pode ser dito sobre as relações *isA* e *kindOf*, que podem estimular o reconhecimento do contexto em que um recurso foi categorizado e, por meio do uso da ontologia ajudar a recuperar a informação desejada. Estas relações possuem também um papel especial na evolução da ontologia proposta neste trabalho, pois podem formar estruturas hierárquicas, se forem obtidas recursivamente até um nó raiz da ontologia, as quais podem ser utilizadas no lugar das caóticas nuvens e listas de *tags*. Isso permite ao usuário ir especializando uma busca até chegar ao resultado esperado. Segundo Anderson (1995), quando o contexto ou o conhecimento geral do mundo guiam a percepção, os níveis mais altos de abstração do conhecimento contribuem para a interpretação das unidades

perceptuais de mais baixo nível. Uma estrutura hierárquica também permite ao usuário fazer uma busca mais abrangente, retornando todos os objetos categorizados com *tags* mais específicas. Por exemplo, uma busca por *"vehicle"* poderia retornar objetos categorizados tanto com a *tag "bike"* quanto com a *tag "car"*.

Já as relações não-semânticas, representadas por *coOccursWith*, *hasToken* e *isTokenOf*, são importantes principalmente para auxiliar na obtenção do contexto e das relações semânticas entre as *tags*. A relação *coOccursWith* pode ajudar na identificação do contexto de uma categorização, pois o usuário pode utilizar mais de um termo com o mesmo contexto em uma categorização. As relações *hasToken* e *isTokenOf* podem ajudar na identificação de relações semânticas entre as *tags*, uma vez que as *tags* compostas que dão origem aos *tokens* nem sempre são reconhecidas por fontes externas de informação semântica.

Um aspecto a ser destacado é que a ontologia mostrada na Figura 10 pode ser estendida com mais relações, na medida em que elas se tornem necessárias em aplicações futuras.

3.2 Processo de Obtenção e Enriquecimento das *Tags*

Fixada a ontologia de *tagging* a ser usada, podemos iniciar o processo de construção da estrutura a partir da personomia. Dessa forma, a entrada de dados para este processo consiste nas categorizações de um usuário em todos os sistemas baseados em *tagging* que o mesmo deseje utilizar. Apesar de alguns *scrappers* e *parsers* terem sido desenvolvidos para obter dados de personomias em sistemas baseados em *tagging*, o presente trabalho não prevê formas para essa tarefa. Virtualmente, qualquer conjunto de categorizações pode ser usado como entrada. Porém, recomendamos a utilização do *TagManager* (DA SILVA, 2009), o qual também funciona tendo como base os três pivôs da tarefa de *tagging* (usuário, recurso e *tags*), conforme as ontologias de Gruber (2005), Knerr (2006), Echarte (2007) e Newman (2005).

Após a obtenção dos dados da personomia de um usuário deve-se passar a enriquecer os relacionamentos entre as *tags* com relações semanticamente mais fortes do que a co-ocorrência, a qual é, até então, a única relação entre as *tags*. Para obter esses relacionamentos semanticamente mais ricos entre as *tags* é necessário que seja feito um *mashup*[18] com outra fonte de dados que permita a obtenção de tal informação. Como as *tags* são elementos textuais foram analisadas as seguintes fontes de informações semânticas: a *WordNet*, a *ConceptNet* (LIU e SINGH, 2004), a *DBpedia* (THIBODEAU, 2009), entre outras que não se mostraram tão interessantes para essa abordagem (METAWEB, 2009; NETSCAPE, 2009). A *WordNet* foi a escolhida inicialmente por possuir relações bastante formais entre as palavras, por estas relações serem baseadas em aspectos cognitivos e por existir bastante estudos e ferramentas sobre a mesma. A *ConceptNet* e a *DBpedia* também se mostraram interessantes e serão consideradas em trabalhos futuros.

Um detalhe que foi tratado diferentemente é que na *WordNet* as palavras são agrupadas por seus significados (em *synsets*) e na ontologia proposta neste trabalho os sentidos são agrupados nas palavras/*tags*, uma vez que estas são as unidades principais para a recuperação de recursos categorizados. Isto não é considerado como um problema, mas apenas como uma forma diferente de organizar os dados na qual o sentido de uma *tag* depende da categorização em que a mesma foi utilizada. Porém, pode correr um problema para se obter relações lingüísticas a partir das *tags* devido às diferentes formas de escrita das *tags*, como, os sinônimos e plurais. Além disso, uma palavra pode ter mais de um sentido (ambigüidade), o que pode prejudicar a obtenção das relações semânticas e, posteriormente, prejudicar a recuperação da informação desejada.

Nosso processo para a evolução de ontologias a partir de personomias é basicamente composto de três etapas. Primeiramente, temos o **processamento léxico**, no qual as *tags* são

[18] No contexto de manipulação de informação, um *mashup* é uma junção de *N* fontes de dados para gerar informações diferenciadas de suas fontes.

normalizadas para que possam ser reconhecidas como conceitos semânticos em fontes externas de informação e, além disso, são adicionadas as relações não-semânticas na ontologia. Já no processo de **atribuição de sentido** é feita uma escolha do sentido de *tags* ambíguas (*i.e.* que tenham mais de um sentido obtido na fonte externa de informações semânticas). Finalmente, no processo de **estruturação semântica** são obtidas relações semânticas para associar *tags* e termos auxiliares.

3.2.1 O Processamento Léxico

A primeira relação que é mapeada para a ontologia é a de co-ocorrência, a qual, conforme descrito anteriormente, é inerente aos sistemas baseados em *tagging*. Essa relação é obtida analisando-se quais termos são utilizados em conjunto nas categorizações e armazenada no atributo *coOccursWith* da ontologia.

Para identificar os conceitos da *WordNet* aos quais uma *tag* se relaciona, elas devem passar por um processo de normalização, o qual é diferente para *tags* comuns e para *tags* agrupadas. A maioria dos principais sistemas baseados em *tagging* não permite que uma *tag* contenha espaços, assim os usuários tendem a utilizar as *tags* de forma agrupada, como, em "*semanticWeb*", "*semantic_web*", etc. Segundo Guy e Tonkin (2006), as *tags* agrupadas constituem 10% do total de *tags* nos principais sistemas baseados em *tagging*. Por esta razão, primeiramente as *tags* agrupadas são "quebradas" em *tokens*. A identificação dos *tokens* é feita separando as *tags* em lugares onde ocorram caracteres especiais (ex.: "/", "_", "-", ".", etc.), caracteres maiúsculos indicando o uso de "*camelCase*" e números. Os *tokens* encontrados são então associados às *tags* por meio de relações *hasToken* e *isTokenOf*. Um exemplo do processo de identificação dos *tokens* de *tags* compostas pode ser visto na Figura 11, na qual os termos em negrito representam as *tags* compostas e os demais termos representam seus *tokens*.

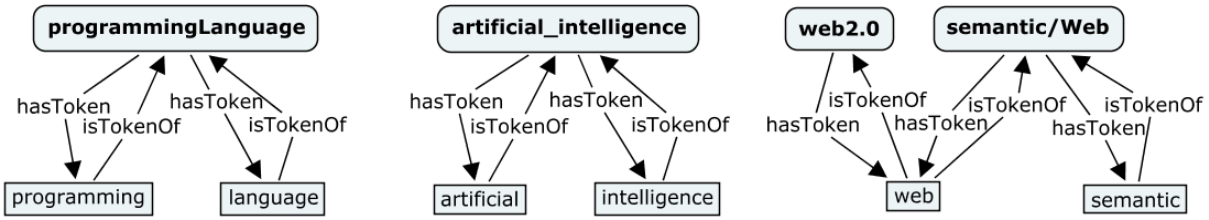


Figura 11: Tags (em negrito) após processo de identificação dos tokens.

Existe também a possibilidade de identificar os *tokens* comparando pedaços de *tags* agrupadas, isto é, que não possuam *CamelCase* ou símbolos separando os termos, à palavras existentes na *WordNet* ou outra fonte de informações semânticas. Porém, esse processo não foi utilizado, pois se torna bastante custoso buscar por todas as possibilidades de escrita de todas as *tags* do usuário e pode não representar a vontade de utilização do usuário, que pode ter deixado as *tags* daquela forma propositalmente. De qualquer forma, com a utilização de um mecanismo para a "limpeza" das *tags*, como o módulo com essa finalidade presente no *TagManager* (CÔGO e DA SILVA, 2008), torna-se possível oferecer sugestões que possam vir a ajudar o usuário na identificação e na correção de termos agrupados.

Após o processo de identificação dos *tokens*, estes, bem como as *tags*, passam por um processo de lematização (*lemmatization process*), o qual serve para reduzir uma palavra a sua forma canônica (normalizada), ou seja, sem derivações ou flexões. A etapa de lematização é um pré-processamento importante para aplicações de mineração textual, processamento de linguagem natural e outros campos que trabalhem com lingüística no geral (PLISSON *et al.*, 2004). Por exemplo, se o algoritmo receber o termo no plural "*women*", ele retornará "*woman"*, no singular, o qual está contido na base da *WordNet*. O processo de lematização é bastante parecido, e confundido, com o processo de *stemming*, o qual serve para se obter o radical de uma palavra. O processo de lematização não precisa encontrar o radical de uma palavra, ele deve apenas substituir o sufixo da palavra (PLISSON *et al.*, 2004). Por exemplo, as palavras "*computes*", "*computing*" e "*computed*" após o processo de *stemming* se tornariam "comput", e, após o processo de lematização se tornariam a forma infinitiva do verbo:

“compute”. Como o nosso objetivo é identificar na *WordNet* o sentido das *tags* e as relações entre elas, o processo de lematização se mostrou o mais adequado. O resultado da lematização das *tags* é armazenado no atributo *hasNormalizedLabel* da ontologia.

Quando as *tags* agrupadas correspondem a apenas um conceito na *WordNet*, elas devem ser mantidas agrupadas, porém, ainda assim, os *tokens* devem ser identificados. A diferença é que na *WordNet* esses termos normalmente estão separados por espaço, *underscore* ou hífen. Desta forma, deve ser verificado na *WordNet* as variações dos *tokens* separados pelos símbolos citados, para identificar os conceitos agrupados a que se referem. Quando uma *tag* composta é encontrada na *WordNet* utilizando-se todos seus *tokens*, sua forma escrita presente nessa fonte de informações é associada ao atributo *hasNormalizedLabel*. Como exemplo desse processo, podemos citar a *tag* “*programmingLanguage*”, que está presente na *WordNet* como “*programming language*” (separada por espaço), forma que é então utilizada no atributo *hasNormalizedLabel*. Os termos separados “*programming*” e “*language*”, também presentes na *WordNet*, são associados com a *tag* “*programmingLanguage*” como relações *isTokenOf* e *hasToken*, respectivamente. Um algoritmo em alto nível de abstração descrevendo o processamento léxico pode ser visto na Figura 12.

```
//PROCESSAMENTO LÉXICO E IDENTIFICAÇÃO DE POSSÍVEIS SENTIDOS
for each Tagging {
    //obter co-ocorrência
    for each pair of Tags {
        relationList.addCoOccurrenceRelations(tag1, tag2);
    }
    //obter tokens e efetuar a lematização
    for each Tag {
        if Tag has camelCase or specialCharacter {
            tokens[] = breakTagInTokens(Tag);
        }
        //se a tag não possuir tokens, tentar normalizá-la
        if (tag has not tokens) {
            tag.normalizedLabel = lemmatize(tag);
        }
        if (tag has tokens) {
            //tentar identificar se os tokens agrupados por espaço ou por hífen
            //conseguem ser lematizados com base na WordNet
            for each variation of tokens grouped with ' ' or '-' {
                tag.normalizedLabel = lemmatize (token1 + ' ' + token2 + ...);
            }
            //relacionar tags agrupadas com seus tokens
            for each token {
                tag.addTokenRelation(token);
            }
        }
    }
    //obter possíveis synsets na WordNet
    for each tag or token {
        tag.possibleSenses = getSynsets(tag.normalizedLabel);
    }
}
```

Figura 12: Algoritmo em alto nível de abstração descrevendo o processamento léxico.

3.2.1.1 Avaliação do Processamento Léxico

Ao final do processamento léxico, os termos processados (*tags* e *tokens*) são então utilizados para buscar os conceitos na base da *WordNet*. Em testes efetuados sobre uma amostra casual simples de *tags* (1.730.056 *tags* obtidas em aproximadamente 2.100 personomias do sistema *Delicious*), observamos que 45% delas foram reconhecidas na *WordNet* sem o processo de lematização e da identificação dos *tokens*. Com o processo de lematização, este número sobe para 52%. Se considerarmos as *tags* compostas que tiveram pelo menos um dos *tokens* identificados na *WordNet*, este valor chega a 62%. Sabendo que os *synsets* da *WordNet* são divididos em substantivos, verbos, adjetivos e advérbios, também foi testada a quantidade de *tags* identificada em cada uma dessas categorias. No trabalho de Laniado *et al.* (2007) foram considerados apenas os substantivos, uma vez que, segundo os

autores, apenas os termos dessa categoria estão dispostos em hierarquia, a qual é imprescindível para a proposta deles. De fato, segundo Fellbaum (1990) os verbos também estão dispostos em hierarquias na *WordNet*, porém não foi possível fazer com que estas hierarquias possuam apenas um nó raiz como no caso dos substantivos. Observamos que 94% das *tags* reconhecidas na *WordNet* são identificadas como substantivos, valor semelhante aos 95% obtidos por Laniado *et al.* (2007). Por esta razão, os substantivos tornam-se imprescindíveis para a evolução de ontologias a partir de *tags*. Quanto ao uso de verbos, observamos que 24% das *tags* podem ser identificadas dessa forma na *WordNet*, mas consideramos a sua utilização optativa, uma vez que 94% dos termos reconhecidos como verbos também são reconhecidos como substantivos. Os advérbios e adjetivos não são considerados, pois após o processo de lematização os advérbios podem retornar a sua forma normalizada, o que comumente constitui um verbo; e os adjetivos não trazem muitos benefícios para os objetivos propostos por este trabalho, uma vez que não estão dispostos hierarquicamente, e a maioria das *tags* reconhecidas nestes grupos (adjetivos e advérbios) também são reconhecidas como substantivos. Na Figura 13 pode ser observado que a maioria das *tags* reconhecidas como verbos, adjetivos e advérbios também podem ser reconhecidos como substantivos.

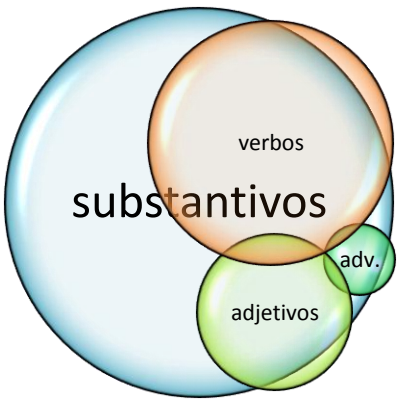


Figura 13: Disposição das tags em substantivos, verbos, adjetivos e advérbios e sua intersecção.

Uma questão que também foi levada em consideração é a das *tags* que não são reconhecidas na *WordNet*. Essas *tags* continuarão relacionadas com outras apenas por co-ocorrência (caso não possuam nenhum *token* reconhecido). Dessa forma, a *tag* não obterá diretamente os benefícios proporcionados pela semântica. Outra questão considerada é o caso de haver mais de um conceito na *WordNet* para uma mesma *tag*, tornando-se necessária uma desambiguação de sentidos para identificar a melhor opção de conceito.

3.2.2 A Atribuição de Sentido às *Tags*

A ambigüidade é inerente a linguagem humana. Em particular a ambigüidade no sentido das palavras existe em todas as linguagens naturais, com um grande número de palavras tendo mais de um sentido possível (MIHALCEA, 2007). A ambigüidade é um assunto bastante discutido por ser um dos problemas conhecidos do uso de *tagging* (WU *et al.*, 2006), o qual se tornou mais evidente nos trabalhos que tentam obter a semântica explícita a partir das *tags* (LANIADO *et al*., 2007) (BASSO e DA SILVA, 2008) (ANGELETOU *et al.*, 2008) (BASSO *et al*., 2009).

Para as palavras que constituem *tags*, dadas sem um contexto, existe a ambigüidade de como elas devem ser interpretadas (MANNING e SCHÜTZE, 1999). Por exemplo, a palavra "jaguar" pode estar se referindo a marca de carros ou ao animal, e a palavra "banco" pode se referir ao objeto no qual as pessoas se sentam ou ao local onde elas vão fazer transações monetárias. Isso constitui um problema para a geração da ontologia proposta neste trabalho, uma vez que nenhuma semântica explícita é associada pelo usuário às *tags* utilizadas nas categorizações. Por esse motivo, torna-se difícil a definição das relações lingüísticas automaticamente, já que a máquina não consegue identificar com a mesma facilidade dos seres humanos em qual sentido uma *tag* foi utilizada. Uma possibilidade para contornar o

problema de ambiguidade é a utilização de uma técnica de **desambiaguação do sentido de palavras** (*word sense disambiguation*).

A tarefa da desambiguação é determinar qual dos sentidos de uma palavra ambígua é invocado em uma utilização particular desta palavra (MANNING e SCHÜTZE, 1999), neste caso, da *tag*. Uma palavra deve possuir um número finito de sentidos (freqüentemente fornecidos por um dicionário, tesauro, ou outra fonte de referência) e a tarefa do processo de desambiguação é fazer uma escolha forçada entre esses sentidos para cada uso de uma palavra ambígua baseado no seu contexto de uso (MANNING e SCHÜTZE, 1999). No problema aqui abordado, os vários sentidos de uma palavra estarão dispostos na base da *WordNet*, representados pelos *synsets*. Como possibilidades para a identificação do contexto das *tags*, levantamos as seguintes possibilidades:

i) Utilizar as ***tags* co-ocorrentes**: que são úteis para a desambiguação de sentido, porque é comum usuários utilizarem mais de uma *tag* representando conceitos semelhantes em uma categorização. A ajuda provinda das *tags* co-ocorrentes só é válida caso estejam contidas na *WordNet*.

ii) Utilizar o **título da categorização**: também pode ser utilizado no processo de desambiguação, uma vez que normalmente são utilizadas palavras presentes na *WordNet* e que são semanticamente relacionadas ao recurso categorizado.

iii) Utilizar a **descrição do recurso**: esse elemento, assim como o título, pode ajudar na desambiguação por descrever o recurso, mas muitas vezes ele pode não estar disponível ou pode conter muitas palavras, tornando o processo de desambiguação muito custoso computacionalmente.

iv) Utilizar o próprio **recurso categorizado**: em alguns casos o conteúdo do recurso categorizado pode ser utilizado para ajudar na identificação do contexto das *tags* da categorização. Por exemplo, se o recurso for uma página

web ela pode ser minerada para identificar termos de maior relevância. Já se o recurso categorizado for, por exemplo, um vídeo, ele não ajudaria na recuperação do contexto de uma *tag* pela máquina — a não ser que estivesse associado a algum tipo de metadado textual.

A primeira possibilidade que testamos para a resolução do problema da desambiguação de sentido das *tags* foi aplicar um algoritmo para analisar as *tags* co-ocorrentes em cada categorização do usuário e compará-las com os sinônimos, hiperônimos e hipônimos dos diversos *synsets* encontrados na *WordNet*. Assim, poderiam ser obtidos resultados próximos ao que o usuário quis dizer, de acordo com a suposição de que, palavras com sentidos semelhantes são utilizadas em conjunto para categorizar um mesmo recurso (BASSO e DA SILVA, 2008). Esta primeira tentativa não retornou bons resultados e poucas *tags* foram desambiguadas.

Sendo assim, foram feitos testes utilizando a métrica de **similaridade semântica** (*semantic similarity*) chamada ***Lin*** (1998) e da métrica de **relação semântica** (*semantic relatedness*) chamada ***Lesk*** adaptada para a *WordNet* (BARNERJEE e PEDERSEN, 2002). Segundo Warin (2004), a similaridade semântica é um caso especial de relação semântica. A relação semântica leva em consideração o quão dois conceitos são relacionados usando qualquer tipo de relação, enquanto a similaridade semântica considera apenas as relações de hiperonímia e hiponímia (generalização e especialização). Por exemplo, as palavras "*car*" e "*gasoline*" podem estar bastante relacionadas uma com a outra, uma vez que a gasolina é o combustível mais utilizado pelos carros. Já as palavras "*car*" e "*bicycle*" são semanticamente similares, não porque ambos os objetos possuem rodas e meios de propulsão, mas porque eles são instâncias de "*vehicle*" na hierarquia da *WordNet*. A relação entre semanticamente similar e semanticamente relacionado é assimétrica: se dois conceitos são similares, eles também são relacionados, mas eles não são necessariamente similares apenas porque eles são relacionados

(WARIN, 2004). Por esta razão, ambas as métricas foram testadas separadamente para identificar a mais adequada para a desambiguação de termos utilizados em categorizações.

Considerando que algumas das relações da *WordNet* formam uma hierarquia, a medida proposta por Lin (1998) soma o valor do *Information Content*[19] (*IC*) do último *subsumer*[20] que dois conceitos tem em comum (*LCS*) com a soma do *IC* de cada um deles individualmente, conforme pode ser visto na fórmula:

$$Sim_{Lin} = \frac{2 * IC\ (LCS)}{IC\ (synset1) + IC\ (synset2)}$$

O resultado do cálculo utilizando a fórmula da medida *Lin* retorna um valor entre zero e um (inclusive), no qual quanto maior o valor, maior a similaridade entre os *synsets*.

Já o princípio da medida *Lesk* adaptada (BARNERJEE e PEDERSEN, 2002) é que quanto mais palavras houver em comum entre as descrições (*glosses*) de dois *synsets*, mais relacionados eles serão. Sua implementação não apenas usa as descrições dos *synsets*, mas também as relações entre eles na *WordNet* para comparar com as descrições de *synsets* próximos. O resultado da medida *Lesk* adaptada é dado pela soma dos quadrados do tamanho da sobreposição (*overlap lengths*), conforme pode ser visto na sua fórmula:

$$Sim_{Lesk} = \sum_{i}^{\#overlaps} length^{\,2}(overlap_i)$$

Por exemplo, a sobreposição de uma palavra traz o resultado 1; a sobreposição de duas palavras não consecutivas traz o resultado 2; a sobreposição de duas palavras consecutivas traz o resultado 4; a sobreposição de 3 palavras consecutivas traz o resultado 9; e assim por diante. Esses exemplos são mais facilmente entendidos com a comparação das descrições de dois *synsets*, o primeiro representando uma "motocicleta" e o segundo um "automóvel":

[19] *Information Content*, que mede a especificidade de um dado conceito. Essa medida é baseada em valores pré-determinados obtidos por testes em *corpus*.

- "*A motor vehicle with two wheels and a strong frame*"; e
- "*A motor vehicle with four wheels; usually propelled by an internal combustion engine*".

Como pode ser observado no exemplo anterior, se ignorarmos as *stop words*, ocorrem três sobreposições (as quais estão sublinhadas), sendo duas delas consecutivas. O resultado da comparação dessas duas frases produz como resultado o valor cinco ($2^2 + 1^2 = 5$).

A desambiguação do sentido das *tags* é feita comparando entre si todos os pares de *synsets* correspondentes as *tags* de uma categorização que estejam contidas na *WordNet*, na qual os *synsets* mais fortemente relacionados são os utilizados para a obtenção das relações semânticas que servem para a evolução da ontologia a partir das *tags* do usuário. Um exemplo disso é representado na Figura 14.

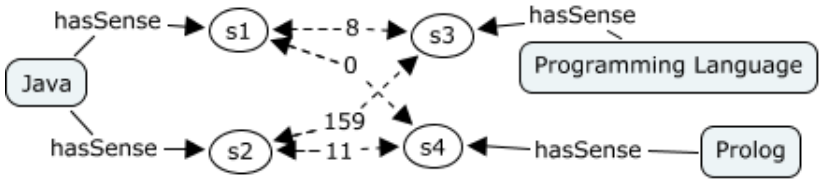


Figura 14: Processo de comparação dos dois sentidos da tag "Java" com os sentidos das tags co-ocorrentes utilizando a métrica lesk.

No exemplo da Figura 14, foram comparados entre si os *synsets* das *tags* "*Java*", "*Programming Language*" e "*Prolog*" — as quais co-ocorrem em uma categorização — utilizando a métrica *Lesk* adaptada. A única *tag* ambígua do exemplo é "*Java*", cujos sentidos estão representados por "s1" — "*a beverage consisting of an infusion of ground coffee beans*" — e por "s2" — "*a platform-independent object-oriented programming language*". O sentido da *tag* "*Programming Language*" ("s3"), é descrito como "*a language designed for programming computers*" e o sentido da *tag* "*Prolog*" ("s4") é descrito como "*a computer language designed in Europe to support natural language processing*". Ainda conforme a

[20] Último conceito mais abrangente que dois conceitos em uma hierarquia compartilham entre si.

Figura 14, o sentido da *tag* ambígua "*Java*" mais fortemente relacionado semanticamente com os sentidos das *tags* co-ocorrentes é o "s2" (com uma nota 159), que é então o sentido utilizado para a obtenção das relações semânticas.

Como passo opcional para o processo de desambiguação, podemos também considerar o título da categorização para ajudar na identificação do contexto da *tags*, principalmente em categorizações que possuam apenas uma *tag*, ou mais de uma *tag* pouco relacionadas semanticamente. Para isso, após a comparação das *tags* de uma categorização entre si, as palavras significativas que fazem parte do título do objeto também serão comparadas com as *tags* para reforçar a escolha do *synset* semanticamente mais adequado. Para que o título seja utilizado de forma mais eficiente nesse processo ele deve passar por uma "limpeza" no qual deve ser (i) removida a pontuação; as palavras devem (ii) passar por um processo de lematização; e deve-se (iii) remover as "*stop words*" (as quais são palavras que não trazem benefícios ao processo como, por exemplo, "*a*", "*but*", "*for*", entre outras). Um exemplo com o resultado destas etapas pode ser visto na Figura 15, a qual representa o título de uma categorização relacionada a férias na ilha de Java.

```
Input    ) Vacations in Java!
Step i   ) Vacations in Java!
Step ii  ) Vacations in Java!
step iii) Vacations in Java!
Output   ) {Vacation,Java}
```

Figura 15: Exemplo das etapas de limpeza do título de uma categorização para auxiliar no processo de identificação do contexto das tags.

Após esse processo, a lista de palavras remanescentes do título pode ser comparada com as *tags* da categorização para verificar a maior similaridade semântica entre os *synsets* das *tags* e, assim, definir o conceito mais adequado a ser utilizado. O *synset* (de uma *tag* ambígua) a ser utilizado é o que possui a maior relação, ou similaridade semântica, com o *synset* de uma *tag* co-ocorrente ou de uma palavra retirada do título.

A descrição do recurso, bem como o próprio recurso categorizado não foram utilizados na desambiguação. A descrição, como anteriormente relatado, muitas vezes não está disponível e torna o processo de desambiguação computacionalmente mais custoso. Além disso, a descrição pode não estar relacionada com o recurso categorizado, porque é adicionada pelo próprio usuário e pode conter anotações fora de contexto. Já o próprio recurso categorizado não foi considerado porque no trabalho aqui proposto serão considerados todos os tipos de recursos que possam ser categorizados por *tags*, inclusive fotos e vídeos, os quais nem sempre ajudam na identificação do contexto das mesmas. Um algoritmo descrevendo em alto nível o processo de desambiguação pode ser observado na Figura 16.

```
//DESAMBIGUAÇÃO DO SENTIDO DAS TAGS
for each pair of tags from a tagging {
    titleRelevantWords[] = processTitleForDisambiguation(tagging.getTitle());
    //os termos aqui podem ser tags ou palavras provindas do título
    for each pair of terms from a tagging {
        for each pair of synsets from two terms {
            //a similaridade é calculada com alguma métrica baseada na WordNet
            //como possibilidades podemos utilizar a Lin ou a Lesk
            score = verifySimilarityScoreBetweenSynsetsWithLesk(synset1, synset2);
            //adicionar o score obtido pelo synset para uma tag em uma categorização.
            //Internamente, o método "addSimilarityScore" vai verificar se o score
            //obtido para um dos synsets nessa interação é maior do que os obtidos
            //na interação anterior (da mesma categorização).
            disambiguationResults.addSimilarityScore(synset1, score);
            disambiguationResults.addSimilarityScore(synset2, score);
        }
    }
    //Associar as tags da categorização com seus respectivos synsets que obtiveram
    //os maiores scores.
    tag.addSense(disambiguationResults.getStrongerSynset(tag));
}
```

Figura 16: Algoritmo descrevendo o processo de desambiguação do sentido das tags.

Nos trabalhos de Laniado *et al.* (2007) e de Angeletou *et al.* (2007) também são propostas formas de desambiguação do sentido de *tags.* Ambos os trabalhos utilizam a similaridade semântica entre os *synsets* de um conjunto de *tags* para determinar automaticamente o sentido mais adequado. No trabalho de Laniado *et al.* (2008) os autores consideram apenas as *tags* mais populares de cada categorização, não apenas as *tags* utilizadas pelo usuário na categorização do recurso, mas também as *tags* mais populares de

todos os usuários da folksonomia utilizadas para aquele mesmo recurso. Isso se torna possível em um sistema que utilize a folksonomia larga, como o *Delicious*, porém, não pode ser feito em sistemas que utilizem a folksonomia estreita, como a maioria dos sistemas gerenciados pelo *TagManager* (*i.e. YouTube*, *Flickr* e *SlideShare*) uma vez que nesses sistemas apenas o dono do recurso pode aplicar *tags* ao mesmo. Por esta razão foram consideradas apenas as *tags* co-ocorrentes da personomia no presente estudo. Já no trabalho de Angeletou *et al.* (2008) todos os sentidos do conjunto de *tags* de entrada são comparados entre si. Esse conjunto de *tags* pode ser, como na proposta de desambiguação do presente trabalho, as *tags* co-ocorrentes de uma categorização. Nem o trabalho de Laniado *et al.* (2007) e nem o de Angeletou *et al.* (2008) consideram a utilização do título da categorização para a desambiguação e nenhum deles apresenta os resultados dessa etapa.

3.2.2.1 Avaliação do processo de Atribuição e Desambiguação de Sentidos

Em experimentos realizados sobre 1.730.056 *tags* de 2.100 personomias obtidas aleatóriamente no sistema *Delicious*, 64% das *tags* encontradas na *WordNet* possuíam mais de um sentido possível e necessitavam desambiguação. Após o processo de desambiguação de 180.000 categorizações (obtidas aleatoriamente a partir de 2.100 personomias), utilizando-se apenas das *tags* co-ocorrentes para a identificação do contexto, 79% das *tags* ambíguas tiveram um nível de similaridade semântica positivo entre seus sentidos utilizando a medida *Lin* e 92% tiveram um nível de relação semântica positivo utilizando a medida *Lesk*. Com a utilização do título do objeto categorizado para auxiliar na identificação do contexto os valores melhoraram passando para 90% com a medida *Lin* e 97% com a medida *Lesk*. Porém, identificar automaticamente algum grau de similaridade ou relação semântica não indica que as escolhas dos sentidos foram corretas.

Para medir a qualidade da desambiguação, foi obtida uma amostra aleatória de 40 categorizações de diferentes usuários e foram verificados manualmente os sentidos escolhidos

para cada *tag* de acordo com as descrições obtidas na *WordNet (glosses)* e com o contexto da categorização (verificado por meio do título e das *tags* co-ocorrentes). As categorizações da amostra foram provenientes de diversos assuntos, tais como: economia, psicologia, política, culinária, aprendizagem, fotografia, tecnologia e história. Das *tags* que passaram pelo processo de desambiguação, 64% das escolhas de sentido foram corretas utilizando-se da medida *Lin* e 68% utilizando-se da medida *Lesk*, considerando apenas as *tags* co-ocorrentes para identificar o contexto das *tags* ambíguas. Estes valores mudaram para 64% e 71% respectivamente quando foram utilizadas as *tags* co-ocorrentes e o título da categorização para a identificação do contexto.

Como pode ser observado na Figura 17, o percentual de resultados corretos não é significativamente melhor utilizando o título para ajudar na identificação do contexto, mas a quantidade de *tags* que possuam algum grau de similaridade é melhorado. Dessa forma, podemos afirmar que mais *tags* foram desambiguadas. Devido aos resultados obtidos, optamos pela utilização da medida *Lesk*, considerando tanto as *tags* co-ocorrentes quanto o título da categorização na identificação do contexto, porque melhores resultados são obtidos tanto na quantidade quanto na qualidade da desambiguação do sentido das *tags*.

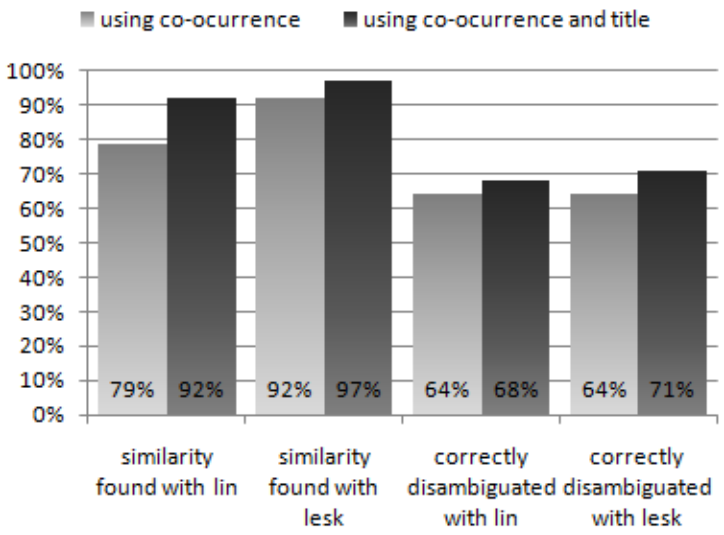


Figura 17: Resultados obtidos no processo de desambiguação das tags.

Analisando os problemas na identificação dos sentidos, pudemos observar que as categorizações relacionadas a tecnologia foram as que mais obtiveram erros nas escolhas de sentido. Quando ocorre um erro no processo de desambiguação e um sentido errado é associado a uma *tag*, pode ocorrer o problema de conceitos serem relacionados de forma errada no momento da expansão semântica (discutida na seção seguinte). Para evitar esse transtorno, o ideal seria que os sentidos fossem confirmados manualmente pelo usuário, em vez de ser um processo automático. Isso, porém, acarretaria em um esforço cognitivo grande por parte do usuário, principalmente quando o mesmo já possui muitas categorizações e *tags*, o que torna este processo manual inviável. Para melhorar os resultados obtidos com a desambiguação uma opção é a de o usuário utilizar um mecanismo para detectar e solucionar problemas no conjunto de *tags* como, por exemplo, o discutido por Côgo e Da Silva (2008).

Outra possibilidade que poderia ajudar bastante na identificação dos sentidos de *tags* ambíguas é a da utilização pelo usuário de métodos para melhorar a qualidade das *tags*, como os descritos por Guy e Tonkin (2006), os quais, entre outras coisas, sugerem que usuários utilizem termos específicos, sinônimos e termos mais abrangentes para delimitar bem o contexto da categorização. Uma possibilidade que também poderia ser considerada é a da criação por especialistas de um *corpus* com anotações semânticas de *tags* e seus significados, de acordo com a freqüência de sua co-ocorrência nos diversos sistemas baseados em folksonomia. Essa é uma possibilidade que, mesmo não sendo trivial, poderia ser de grande ajuda no processo de desambiguação. Já uma técnica que poderia eliminar o processo de desambiguação seria a utilização de conceitos semânticos (*i.e. tags* associadas a um sentido) no momento da categorização, em vez da utilização de rótulos de texto atribuídos livremente. Desta forma, o sentido dos conceitos poderia ser armazenado e utilizado posteriormente no processo de evolução de ontologias. Porém, isso ainda não é permitido na maioria dos sistemas baseados em *tagging* e, além disso, existem muitas categorizações de usuários que

foram feitas sem nenhuma associação explícita de semântica. Assim, algum tipo de desambiguação de sentido automático das *tags* se torna imprescindível.

Outro problema que foi observado na desambiguação foi na utilização de termos mais abstratos como, por exemplo, "*design*", os quais possuem vários significados na *WordNet* e também não obtiveram resultados satisfatórios na identificação dos sentidos mais adequados. Observamos ainda que muitos *synsets* dessas palavras são bastante similares entre eles e, futuramente, será testada a atribuição de mais de um sentido para uma *tag* no contexto de uma categorização, desde que esses sentidos possuam um nível alto de similaridade ou relação semântica entre eles.

Identificados os conceitos na *WordNet* a partir das *tags*, pode-se passar a obter as relações semânticas entre elas, processo que denominamos de Estruturação Semântica.

3.2.3 A Estruturação Semântica das Relações entre as *Tags*

Após a identificação das relações não-semânticas e a identificação e desambiguação dos conceitos que representam as *tags* do usuário na *WordNet*, pode se passar a obter as relações semânticas entre elas. Para isso, obtemos a partir das *tags* as relações de sinonímia, hiperonímia, hiponímia, meronímia e holonímia, as quais são mapeadas na ontologia da Figura 10 respectivamente para *hasAlternative*, *isA*, *kindOf*, *hasPart* e *isPartOf*, porque consideramos que estes termos são mais apropriados para a utilização de outras fontes de informação além da *WordNet*.

As relações de mais alto nível, como as de hiperonímia e de holonímia, devem ser obtidas até o nó raiz da *WordNet*, que nos substantivos é representado por "*entity*" e nos verbos pode haver mais de um. Por esta razão, se os verbos forem considerados no processo de evolução de ontologias, deve ser adicionado um "falso" nó raiz para agrupar os nós de raiz obtidos na *WordNet*. Essa abordagem é bastante utilizada por alguns autores como, por

exemplo, Pedersen *et al.*, (2004), que utiliza os nós "falsos" para agrupar todos os nós raiz dos verbos, permitindo, assim, a obtenção de similaridade semântica entre conceitos que não estejam na mesma hierarquia. Este nó não é necessariamente exibido para o usuário, ele serve apenas como um ponto de partida para os algoritmos que utilizem a ontologia. Um exemplo

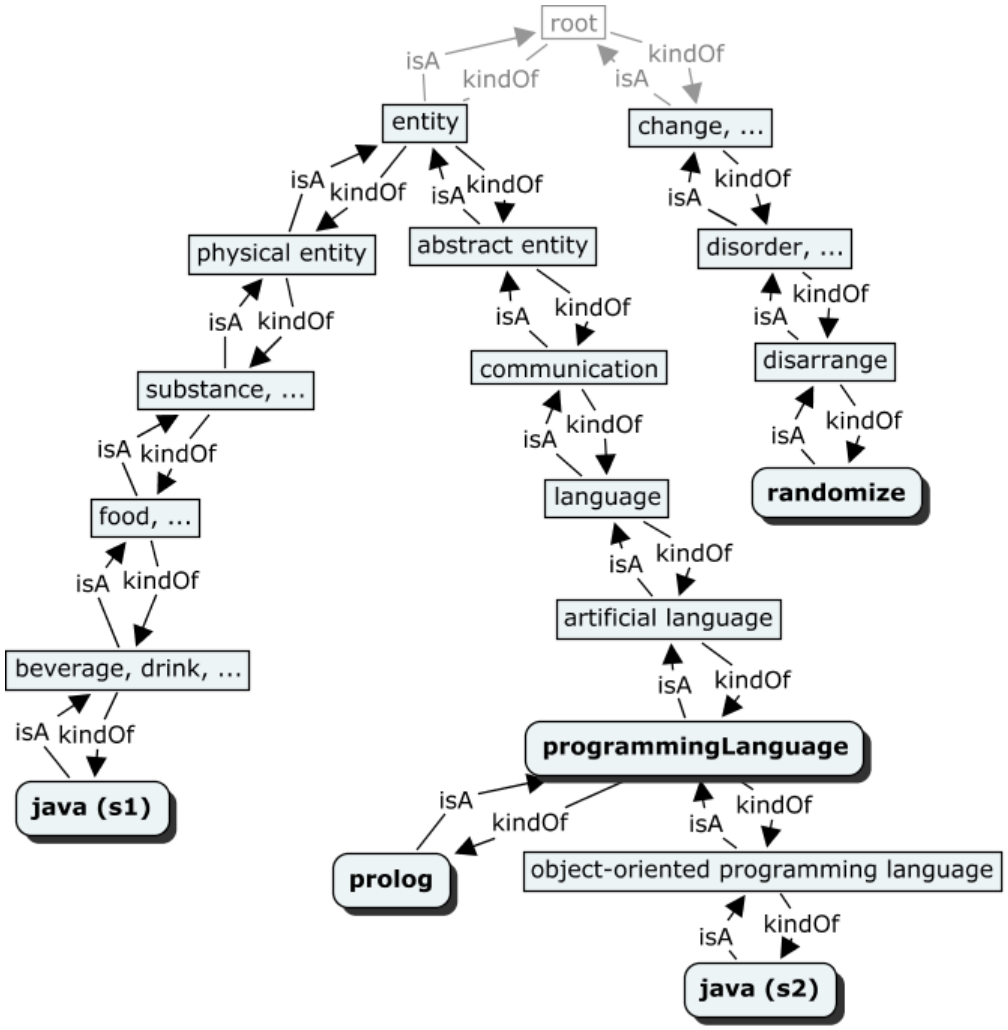


Figura 18: Hierarquia obtida por meio de relações de hiperonímia a partir das tags de uma personomia, as quais estão em negrito.

da obtenção recursiva de relacionamentos com termos mais abrangentes, bem como de um nó de raiz discutido é mostrado na Figura 18.

Apesar das relações terem sido obtidas a partir das relações de hiperonímia na *WordNet*, na Figura 18 elas são mostradas como *isA* e o seu contrário *kindOf* por serem as relações equivalentes da ontologia proposta neste trabalho. Um aspecto a ser observado é que

termos que não fazem parte do conjunto de *tags* do usuário (*i.e.* de sua personomia) também são adicionados na ontologia constituindo o que chamamos de "*tags* auxiliares". Estas *tags* são adicionadas para auxiliar na estruturação das *tags* da personomia na ontologia, permitindo, por exemplo, buscas por conceitos mais abrangentes e mostrar as *tags* em forma de hierarquias. Se fossem utilizadas apenas as *tags* da personomia, haveria o risco de não haver a possibilidade de organizá-las corretamente de forma hierárquica, porque muitas vezes não há *tags* suficientes para isto na personomia. A possibilidade de manter apenas as *tags* da personomia na ontologia, sem termos auxiliares obtidos na *WordNet*, talvez pudesse ser uma realidade se considerássemos todas as *tags* de uma folksonomia e não apenas as *tags* de uma personomia. Outro detalhe que pode ser observado na Figura 18 é que se no processo de obtenção de níveis mais abrangentes um termo que já constitui uma *tag* da personomia do usuário for encontrado, essa *tag* continua diferenciada das outras pelo atributo *isUserCreated* da ontologia proposta neste trabalho. Um exemplo disso, que pode ser observado na Figura 18, é o termo "*programmingLanguage*" (que pode ser reconhecido na *WordNet* pela sua forma normalizada de escrita "*programming language*"), o qual constitui uma *tag* da personomia e é um termo intermediário na hierarquia de "*Java*" e "*Prolog*" até "*entity*".

Existem algumas relações semânticas que não são obtidas recursivamente como, por exemplo, a de sinonímia, a de hiponímia e a de meronímia. A relação de sinonímia possui apenas um nível. Já as relações de hiponímia e de meronímia podem ser utilizadas recursivamente. Porém, em testes efetuados, observamos que a obtenção de mais de um nível mais específico destas relações aumenta, desnecessariamente, o tamanho da ontologia gerada, uma vez que é difícil identificar a qual conceito mais específico uma categorização está relacionada. Dessa forma, quanto mais níveis de especialização são buscados para uma *tag*, mais esses conceitos acabam ficando "distantes" do significado da *tag* da personomia a que estão relacionados. Por essa razão, optamos inicialmente pela obtenção de apenas um nível de

relações mais específicas, as quais podem ajudar na recuperação de informação sem provocar um aumento expressivo no tamanho da ontologia gerada. A Figura 19 mostra um exemplo da obtenção de um nível das relações de (a) sinonímia, (b) meronímia e (c) hiponímia.

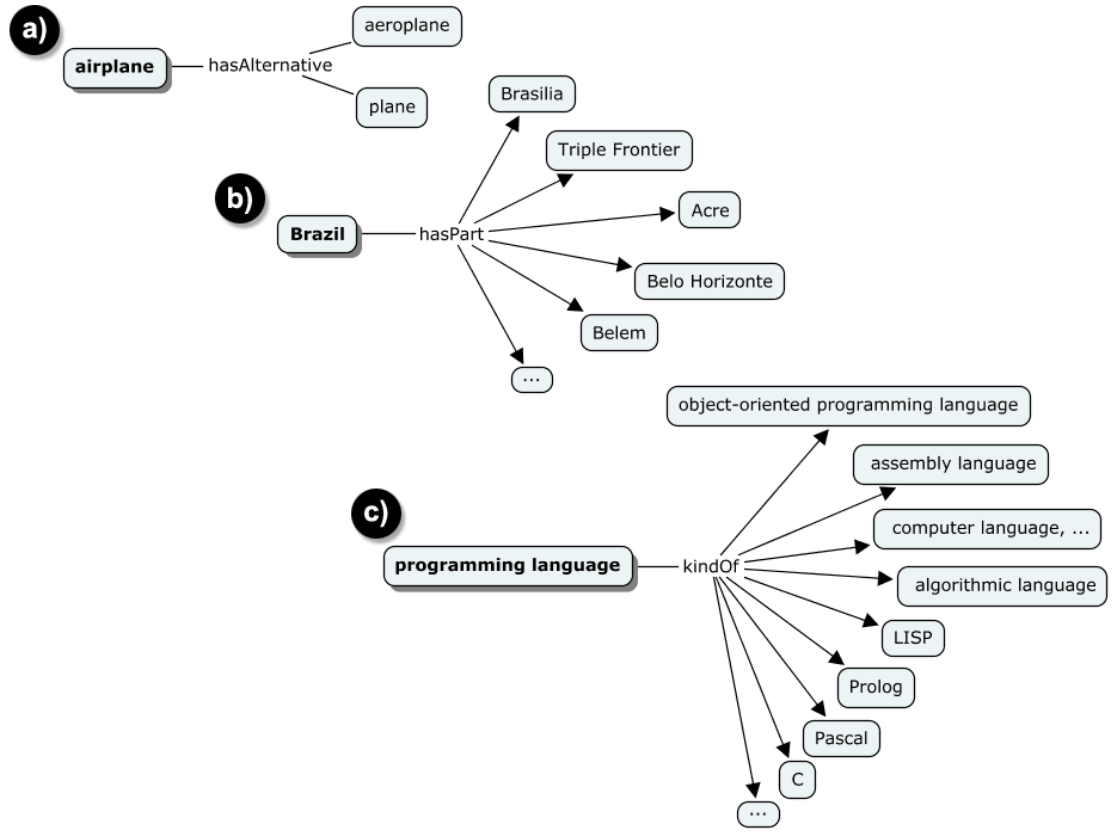


Figura 19: Exemplo de relações de sinonímia, meronímia e hiponímia obtidos respectivamente a partir dos termos "airplane", "Brazil" e "programming language".

A Figura 20 representa um algoritmo de alto nível descrevendo o processo de estruturação semântica.

```
//ESTRUTURAÇÃO SEMÂNTICA
for each tag or token {
    for each synset obtained in disambiguation {
        relationList.add( obtainIsARelations(tag, synset) );
        relationList.add( obtainKindOfRelations(tag, synset) );
        relationList.add( obtainHasPartRelations(tag, synset) );
        relationList.add( obtainIsPartOfRelations(tag, synset) );
        relationList.add( obtainHasAlternativeRelations(tag, synset) );
    }
}

//MÉTODO PARA A OBTENÇÃO DE RELAÇÕES "IS A"
method obtainIsARelations(originTag, originSynset):relationList {
    //pega o synset do hiperônimo do synset de origem na WordNet
    relationSynset = getWordnetHypernym(originSynset);
    //identificar atributos da tag de destino do relacionamento
    //tentar inicialmente verificar se o synset encontrado faz parte da personomia
    relationTag = personomyTagList.findTagBySynset(relationSynset);
    if (relationTag is null) //se o relationSynset não está presente no conj. de tags
        //criar uma "tag de relação", obtendo dados como lemma e gloss do synset
        relationTag = populateNewTag(relationSynset);
        relationTagList.add(relationTag);
    }
    //criar uma relação entre as duas tags (de origem e de destino)
    relationList.add(originTag, originSynset, relationTag, relationSynset, IsA);
    //se a hiperonímia obtida não for um nó raiz na WordNet, continuar a
    //busca em profundidade
    if (hypernym is not null) {
        obtainIsARelations(relationSynset);
    }
}

//MÉTODO PARA A OBTENÇÃO DE RELAÇÕES "KIND OF"
method obtainKindOfRelations(originTag, originSynset):relationList {
    //pega o synset do hipônimo do synset de origem na WordNet
    relationSynset = JWI.getWordnetHyponym(originSynset);
    //identificar atributos da tag de destino do relacionamento
    //tentar inicialmente verificar se o synset encontrado faz parte da personomia
    relationTag = personomyTagList.findTagBySynset(relationSynset);
    if (relationTag is null) //se o relationSynset não está presente no conj. de tags
        //criar uma "tag de relação", obtendo dados como lemma e gloss do synset
        relationTag = populateNewTag(relationSynset);
        relationTagList.add(relationTag);
    }
    //criar uma relação entre as duas tags (de origem e de destino)
    relationList.add(originTag, originSynset, relationTag, relationSynset, KindOf);
}

//MÉTODO PARA A OBTENÇÃO DE RELAÇÕES "IS PART OF"
method obtainIsPartOfRelations(originTag, originSynset):relationList {
    //esse método é idêntico ao "obtainIsARelations",
    //com exceção do tipo da relação obtida.
}

//MÉTODO PARA A OBTENÇÃO DE RELAÇÕES "HAS PART"
method obtainHasPartRelations(originTag, originSynset):relationList {
    //esse método é idêntico ao "obtainIsARelations",
    //com exceção do tipo da relação obtida.
}

//MÉTODO PARA A OBTENÇÃO DE RELAÇÕES "HAS ALTERNATIVE"
method obtainHasAlternativeRelations(originTag, originSynset):relationList {
    //esse método é idêntico ao "obtainIsARelations",
    //com exceção do tipo da relação obtida.
}
```

Figura 20: Algoritmo descrevendo o processo de estruturação semântica.

Para exemplificar o resultado do processo de evolução de ontologias a partir de personomias descrito nesta seção, na Figura 21 é apresentado um esquema simplificado de todas as etapas sobre um conjunto de dados. A partir dos dados de entrada, os quais consistem em duas categorizações, cuja única estrutura entre as *tags* é a co-ocorrência, é evoluída uma estrutura ontológica com vários níveis de relações, bem como termos auxiliares. Um aspecto a ser destacado é a estrutura hierárquica obtida por meio das relações de generalização/especialização. As relações de todo/parte também podem gerar estruturas semelhantes, podendo, porém, ter mais de um nó raiz. Outro aspecto a ser destacado na figura é que, na parte referente ao processo de desambiguação, dois sentidos foram utilizados para a *tag* "Java", cada um referente ao contexto de uma das duas categorizações do exemplo.

Quanto as tecnologias empregadas para o desenvolvimento da metodologia descrita neste capítulo, os algoritmos foram desenvolvidos com a linguagem de programação "Java" (SUN, 2009b), com o auxílio de uma *API* de acesso a base de dados da *WordNet* chamada *"JWI"* — "*MIT Java WordNet Interface*" (FINLAYSON, 2008). A representação das ontologias geradas é feita por meio de estruturas de dados, também em "Java", baseadas na ontologia da Figura 10.

Neste capítulo foi apresentada a proposta de uma metodologia para a evolução de ontologias a partir de personomias. Esta proposta pode ser adaptada dependendo da finalidade em que será aplicada. No próximo capítulo são descritos o sistema de gerenciamento das ontologias geradas, denominado *TagOntologyManager*, e algumas possibilidades de utilização das mesmas.

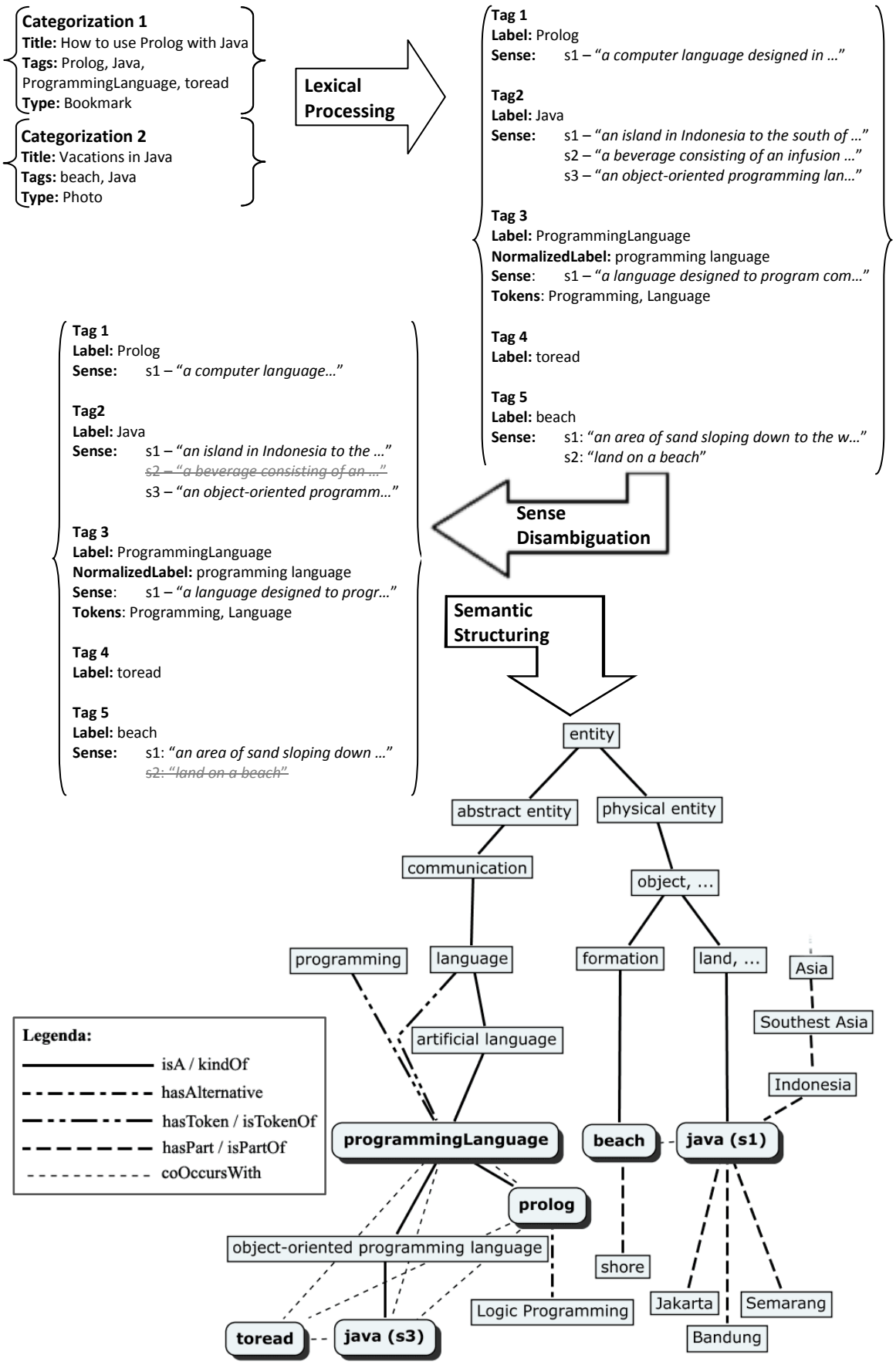


Figura 21: Exemplo simplificado do processo de evolução de ontologias a partir de personomias.

Capítulo IV

TagOntologyManager

O *TagOntologyManager (TOM)* é um sistema — desenvolvido pelo autor deste trabalho — que visa gerenciar uma ontologia obtida a partir de uma personomia cujo processo foi descrito no capítulo anterior, a qual iremos nos referir como "ontologia de personomia". O *TOM* funciona como uma "caixa preta" e deixa implícita a forma de obtenção dos dados de *tagging.* Este esquema foi escolhido para que ele possa ser utilizado em conjunto com qualquer sistema baseado em *tagging*. Além disso, o *TOM* disponibiliza uma *API* para que se possa implementar novas funcionalidades aos sistemas baseados em *tagging* utilizando as ontologias por ele definidas.

A Figura 22 representa a arquitetura do sistema e como ele se relaciona com outros recursos, inclusive para a persistência de dados, e com fontes de dados semânticos como a *WordNet*. Também pode ser observado na Figura 22 que o *TOM* pode usar diretamente alguns sistemas baseados em *tagging*, porém, para isso deve ser implementado nesses sistemas, ou em uma camada intermediária, *parsers* para a conversão dos seus dados para o padrão do

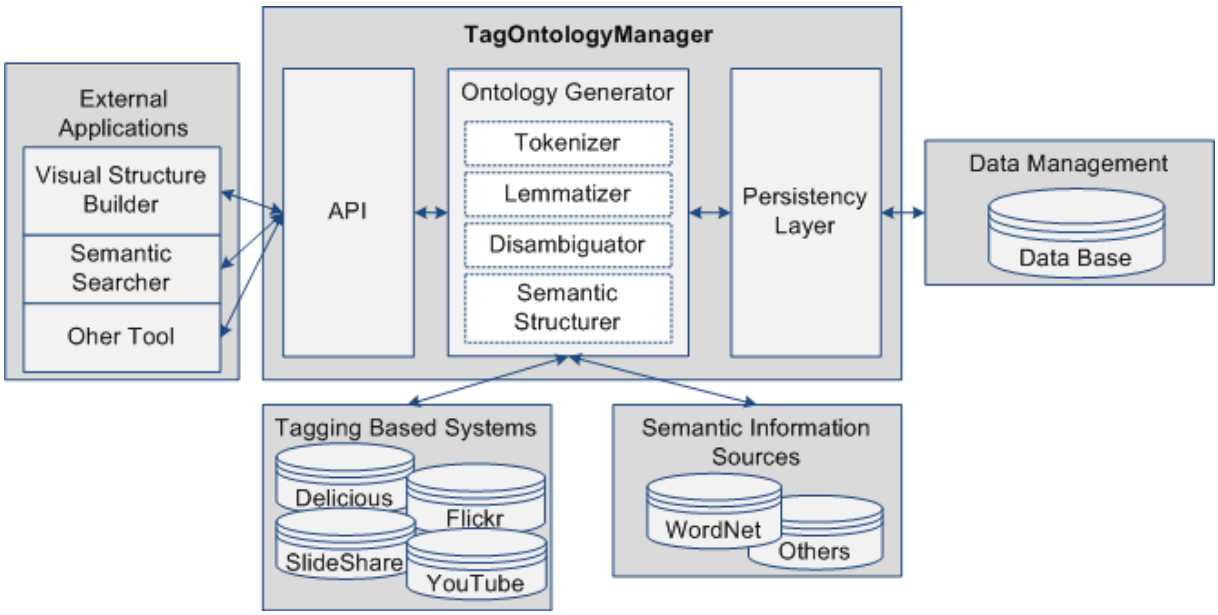


Figura 22: A Arquitetura do TagOntologyManager e como ele se relaciona com outros recursos.

TOM. Outra opção para a obtenção dos dados de categorização de vários sistemas baseados em *tagging* é por meio da utilização do sistema *TagManager* (DA SILVA, 2009). Por esta razão, na seção seguinte é descrita a possibilidade de integração desse sistema com o *TOM*.

4.1 A Possibilidade de Integração do *TagOntologyManager* com o *TagManager*

O *TagManager (TM)* (DA SILVA, 2009) serve para auxiliar o usuário no monitoramento de sua personomia global, a qual é criada utilizando os dados oriundos dos diversos sistemas baseados em *tagging* do usuário. Segundo Da Silva (2009), esse monitoramento deverá trazer melhorias que serão refletidas nas personomias dos vários sistemas que um usuário utiliza e, como conseqüência, a recuperação da informação poderá ser melhorada tanto no contexto pessoal de busca quanto no contexto global para todos os usuários desse sistema (*i.e.* da folksonomia).

O *TOM* pode tanto obter benefícios a partir do *TM* quanto proporcionar melhorias ao mesmo. Os benefícios que o *TOM* pode conseguir são oriundos da obtenção dos dados de categorizações em vários sistemas baseados em *tagging*, além da possibilidade de obter um

conjunto mais limpo e organizado de *tags* (CÔGO e DA SILVA, 2008). Já os benefícios que o *TM* pode obter com o *TOM* provêem de aplicações que utilizem as ontologias de personomias para melhorar o processo de recuperação da informação categorizada e, também no processo de visualização e navegação no espaço de *tags*. Um esquema mostrando a integração desses dois sistemas pode ser observada na Figura 23.

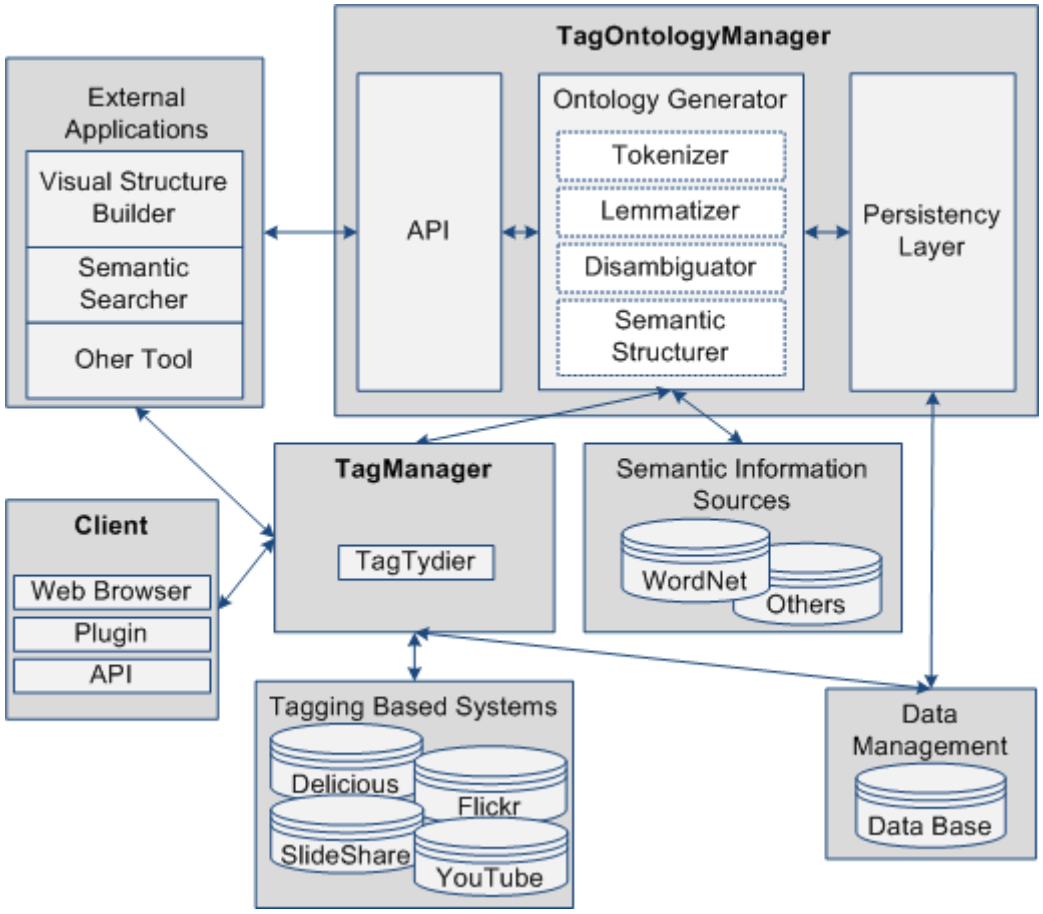


Figura 23: Esquema de como o TagManager e o TagOntologyManager são integrados.

Outro ponto que ilustra a integração do *TOM* ao *TM* é uma ontologia unificada para a organização dos dados de ambos os sistemas, o qual pode ser observado na Figura 24. Nessa ontologia as relações em cinza correspondem à ontologia de categorização do TM, que visa principalmente ter um controle maior sobre o vocabulário do usuário. Já as relações destacadas em preto correspondem à ontologia do *TOM*, que visam representar a semântica das *tags* e entre as *tags*. Um aspecto importante nesta nova ontologia é que as relações

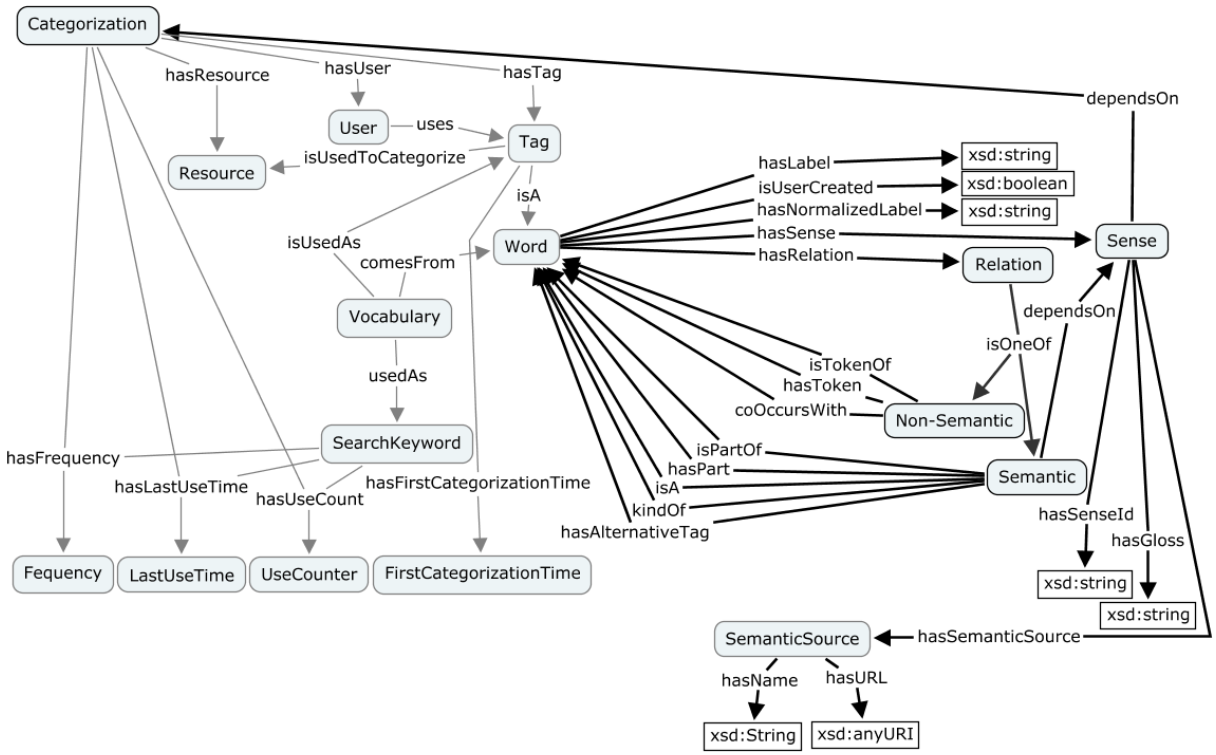


Figura 24: Ontologia unificada do TagManager com o TagOntologyManager.

semânticas estão relacionadas entre palavras (*Word*) e não necessariamente entre *tags*. Isso ocorre porque na ontologia do *TM* uma *Tag* corresponde exclusivamente a um termo criado pelo usuário para a sua personomia. Já uma palavra pode ser uma *tag* da personomia bem como um termo auxiliar, como os termos auxiliares obtidos na *WordNet* no processo de evolução de ontologias descrito no capítulo anterior.

Definidos a arquitetura e a ontologia unificada entre os dois sistemas, podemos passar a descrição de algumas aplicações que testamos que tem como base as ontologias de personomias geradas pelo *TOM*. Nas próximas seções trataremos da utilização das ontologias de personomias para possibilitar que a recuperação da informação em sistemas baseados em *tagging* seja mais eficiente, permitindo buscas por conceitos em vez de rótulos de texto puros e para gerar estruturas visuais alternativas as tradicionais e, muitas vezes, caóticas listas e nuvens de *tags*.

4.2 Buscas por Recursos Categorizados Utilizando a Ontologia

Um dos pontos críticos dos sistemas baseados em *tagging* é a recuperação da informação categorizada. Essa dificuldade é proveniente, principalmente, do fato de que na maioria desses sistemas não há controle sobre o vocabulário empregado na criação das *tags*, as quais constituem a base para a recuperação da informação categorizada. Isso gera alguns inconvenientes como, por exemplo, *tags* que constituam sinônimos ou homônimos, *tags* compostas ou com erros de digitação, variações de singular e plural, além do fato de que as *tags* criam uma estrutura plana (SMITH, 2008) (GUY e TONKIN, 2006).

4.2.1 O Problema das *Tags* que constituem sinônimos

Tags que constituam sinônimos prejudicam a recuperação da informação, pois os usuários tem dificuldade em lembrar da palavra exata que utilizaram em uma categorização (ANDERSON, 1998). Dessa forma, ao categorizar recursos similares os usuários podem utilizar variações léxicas de um mesmo conceito. No momento da recuperação da informação, na maioria dos sistemas baseados em *tagging* atuais, o usuário entra com um termo e apenas os recursos categorizados com *tags* que possuam exatamente a mesma forma de escrita são retornados. Um cenário que exemplifica isso é a recuperação de recursos categorizados com as *tags* "*plane*" e "*airplane*", as quais são sinônimos cognitivos. Suponhamos que o usuário tenha categorizado algumas fotos de aviões com a *tag* "*plane*" e outras fotos e vídeos com a *tag* "*airplane*", conforme o exemplo da Figura 25.

Se o usuário buscar por "*airplane*" apenas as fotos 1 e 2 serão retornadas. A foto 3, categorizada apenas com a *tag* "*plane*", mesmo que representando o mesmo conceito não é retornada. Para retornar todos os recursos categorizados com ambas as *tags* citadas é necessária uma busca por todas as variações de escrita do conceito com operações "*OR*" entre elas, o que não é nada prático e requer um esforço cognitivo considerável por parte do

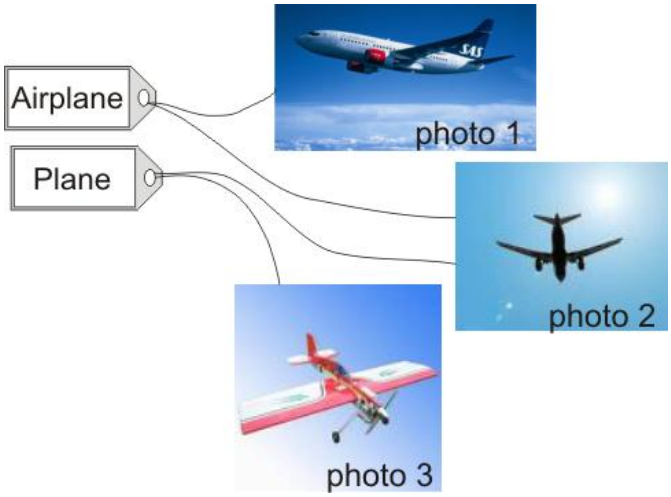


Figura 25: Fotos associadas à tags sinônimas.

usuário, já que o mesmo deve lembrar de todos os sinônimos que utilizou para um mesmo conceito.

Com a utilização da ontologia de personomia podemos facilmente contornar o problema do uso de *tags* sinônimas, uma vez que elas estão associadas por meio de relações *hasAlternative* de acordo com seus conceitos. Dessa forma, se uma busca for feita por "*airplane*", recursos categorizados com "*plane*" e "*aeroplane*" também podem ser retornados de forma transparente ao usuário.

4.2.2 O Problema das *Tags* que constituem homônimos

Os homônimos prejudicam a recuperação da informação porque podem retornar recursos categorizados com a mesma *tag*, mas com significados que fujam ao interesse do usuário naquele momento. Isso gera uma sobrecarga de informação e pode exigir um esforço cognitivo considerável do usuário para discernir quais resultados da busca são relevantes.

Com o uso da ontologia de personomia podemos fazer buscas por conceitos, retornando apenas resultados relevantes ao interesse do usuário naquele momento ou ordenando os resultados e posicionando os mais relevantes como os primeiros da lista. Para escolher o sentido de um termo ambíguo que deseja buscar, o usuário pode selecioná-lo dentro de um contexto em um grafo de termos (semelhante ao grafo de saída da Figura 21) ou

escolher por opções que remetam ao sentido do termo como, por exemplo, na Figura 26 (a) pelas descrições dos sentidos do termo ou (b) pelo termo abrangente mais próximo.

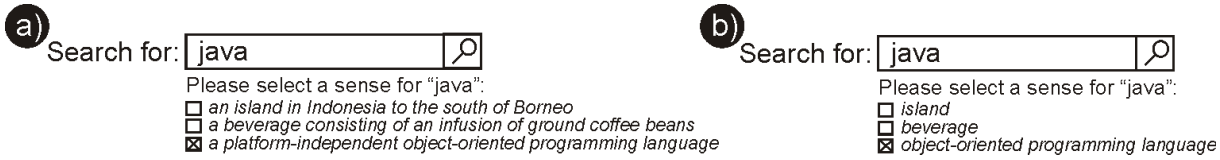


Figura 26: Duas possibilidades de escolha do sentido do termo de busca. a) Opções baseadas na descrição do conceito. b) Opções baseadas no termo abrangente mais próximo.

Selecionados um ou mais sentidos, a busca por conceitos relacionados torna-se possível, pois a desambiguação do sentido de uma *tag* é feito com base no contexto de cada categorização na qual ela é utilizada. Porém, há um possível problema quando ocorrem erros na desambiguação de sentidos de algumas *tags* durante o processo de evolução das ontologias. Dessa forma, uma busca por um termo ambíguo pode retornar recursos categorizados fora de contexto ou, o que é pior, ignorar resultados que deveriam ser retornados. Por esta razão, até que seja encontrada uma solução melhor para a desambiguação de sentido das *tags* (que possua maior percentual de acerto na escolha dos sentidos), é mais prudente apenas reordenar os resultados das buscas, mostrando antes os que se acredita estarem mais relacionados com o conceito da busca, sem cortar os que se acredita estarem fora de contexto.

4.2.3 O Problema da Falta de Níveis Hierárquicos entre as *Tags*

Outro problema em buscas por recursos em sistemas baseados em *tagging* é que devido à falta de níveis hierárquicos entre elas, não se torna possível fazer buscas mais abrangentes. Um cenário que exemplifica isso é o de um usuário tentando recuperar recursos relacionados a "pneus" de qualquer tipo de veículo. Para isso ele entraria, por exemplo, com a *tag* "*tire*" e com *tags* constituindo todos os tipos de veículos que o mesmo categorizou como, por exemplo, "*car*" e "*bus*". Semelhante ao problema dos sinônimos, isso necessita que o

usuário se lembre de todas as *tags* da sua personomia que constituam veículos, a qual é uma tarefa em que os seres humanos têm dificuldades cognitivas (ANDERSON, 1998). Utilizando a ontologia de personomia isso pode ser facilmente melhorado, bastando que o usuário busque pelos termos "*tire*" e "*vehicle*", mesmo que este último não constitua uma *tag* de sua personomia. O mecanismo de busca pode utilizar as relações de especialização da ontologia para verificar as *tags* mais específicas que o usuário utilizou em suas categorizações, como pode ser observado na Figura 27.

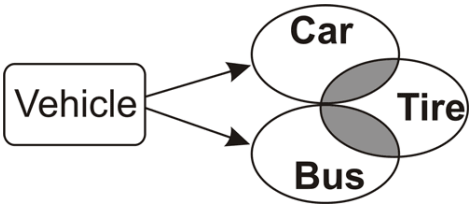


Figura 27: Uma busca hipotética por "tire" (que constitui uma tag da personomia) e "vehicle" (que não está contida na personomia).

Como pode ser observado a partir do termo auxiliar "*vehicle*" o mecanismo de busca pode encontrar termos mais específicos ("*car*" e "*bus*"), os quais constituem *tags* da personomia, e achar a interseção entre eles e o outro termo de busca ("*tire*"). Dessa forma o usuário não precisa lembrar e entrar com todos os tipos de veículos específicos que utilizou como *tags* nas categorizações.

Outra possibilidade interessante que podemos obter com a ontologia de personomia ocorre quando o usuário faz uma busca com um termo mais específico que não constitua uma *tag* da personomia. O agente de busca pode sugerir um termo mais abrangente que consista em uma *tag* da personomia para que resultados relacionados sejam recuperados. Isso é possível desde que o termo mais específico esteja relacionado com uma *tag* da personomia do usuário, conforme descrito na seção 3.2.3 sobre o processo de Estruturação Semântica.

O mesmo que foi dito sobre as relações de generalização/especialização também pode ser aplicado sobre as relações todo/parte, tanto da sugestão de *tags* mais específicas quanto de

tags mais abrangentes, uma vez que essas relações também podem gerar estruturas em forma hierárquica.

4.2.4 O Problema de Termos no Plural

Existe também o problema de que as *tags* no plural nem sempre são reconhecidas em fontes externas de informação semântica, o que também é uma realidade nas buscas, uma vez que os algoritmos normalmente comparam a igualdade da *string* de busca com a escrita das *tags*. Segundo Guy e Tonkin (2006), 8% das *tags* no *Flickr* e 11% das *tags* no *Delicious* consistem em plurais. A recomendação desses autores é que as *tags* sejam sempre criadas no singular para manter um padrão e facilitar a recuperação de recursos, uma vez que nos resultados de buscas são retornados apenas os recursos categorizados com *tags* escritas de forma exatamente igual ao termo de busca.

Utilizando as ontologias deste trabalho, torna-se possível ao mecanismo de busca recuperar os recursos categorizados com *tags* no plural, mesmo que o termo de busca esteja no singular (e vice-versa), já que no processo de lematização as *tags* no plural são convertidas para o singular e armazenadas no atributo *hasNormalizedLabel*.

4.2.5 Outros Problemas no Conjunto de *Tags* que Prejudicam as Buscas

As *tags* compostas também geram dificuldades no momento da recuperação de informação, uma vez que os usuários devem lembrar da forma exata que o termo foi agrupado, bem como o tipo de separador que foi utilizado. De certa forma, se mecanismos de "auto-completar" estiverem presentes no campo de busca, eles podem ajudar o usuário na identificação da forma de escrita que foi utilizada. É comum, porém, que usuários variem a escrita de um termo composto e, dessa forma, os mecanismos de "auto-completar" não podem ajudar de forma eficiente na recuperação de recursos categorizados, pois esse recurso atua

apenas na interface e não no mecanismo de busca. Conforme descrito anteriormente, na evolução de ontologias a partir de personomias as *tags* compostas são processadas para tentar identificar os conceitos que as compõem. Esses conceitos podem ser identificados na *WordNet* e passam a ser integrados pela ontologia por meio de relações *hasToken* e *isTokenOf*, as quais podem ser consideradas por um agente de busca como relações de especialização e generalização respectivamente. Dessa forma, essas *tags* e *tokens* também podem obter os benefícios descritos sobre buscas utilizando as relações de generalização/especialização das ontologias de personomias. Além disso, com a identificação de todos os *tokens* que compõem uma *tag* na *WordNet*, esse termo passa a ter as vantagens da identificação de sinônimos e outros tipos de relações.

Por fim, um problema complexo é o de *tags* digitadas erradas pelo usuário, não permitindo que recursos categorizados com elas sejam recuperados quando o termo de busca está digitado corretamente. As ontologias de personomias não ajudam a resolver esse tipo de problema, porém, com a utilização de um sistema que faça uma limpeza no conjunto de *tags* como, por exemplo, o proposto por Côgo e da Silva (2008), essas *tags* erradas podem ser identificadas e corrigidas pelo usuário para minimizar os problemas nas buscas.

Além de permitir buscas por palavras-chave, muitos sistemas baseados em *tagging* permitem a recuperação de informação por meio de estruturas de navegação nas *tags*, porém, como discutido anteriormente, algumas dessas estruturas exigem bastante esforço cognitivo por parte do usuário para identificar o termo desejado. Por essa razão também testamos algumas aplicações nesse sentido.

4.3 Geração de Estruturas Visuais Alternativas para Navegação no Espaço de *Tags*

A navegação no espaço de *tags* é diferente da navegação em outras estruturas como, por exemplo, taxonomias e ontologias. Na maioria dos sistemas baseados em *tagging* não há

facetas ou categorias para separar os conjuntos de *tags.* Isso ocorre principalmente devido a estrutura plana do conjunto de *tags*. Por esta razão a navegação entre as *tags* está mais para peneirar do que se mover deliberadamente a um destino específico (SMITH, 2008). A estrutura de navegação entre as *tags* mais conhecida é a nuvem de *tags*. Segundo Smith (2008), "*with the growing popularity of tagging, tag clouds have become a fashionable way of displaying tags. But like any fashion, what's hot today can look like an embarrassing fad tomorrow.*"

De fato, são bastante limitadas as possibilidades de visualização e navegação no espaço de *tags* e acreditamos que a popularidade das nuvens de *tags* esteja mais ligada a falta de alternativas do que do seu real benefício. Elas realmente podem ser utilizadas como ponto de entrada para navegar pelos recursos categorizados, porém, com o aumento significativo no número de *tags* da personomia, as nuvens, bem como as listas de *tags* se tornam caóticas, mostrando um emaranhado de termos que exigem um esforço cognitivo considerável do usuário para encontrar o termo desejado. Por essa razão, existem alguns estudos para melhorar e/ou substituir estas formas de visualização e navegação. Um deles é por meio do uso da já citada técnica de clusterização das *tags* (BEGELMAN *et al.*, 2006) (YAHOO, 2004). Essa técnica encontra grupos de *tags* relacionadas estatisticamente e, para ser utilizada como forma de navegação, normalmente agrupa o *cluster* sob um rótulo, o qual normalmente constitui a *tag* mais popular do conjunto. Essa técnica é interessante, mas limitada, pois exige muito da lembrança do usuário e desfavorece o uso das *tags* que foram pouco utilizadas em categorizações. Em Laniado *et al.* (2007) é apresentada uma alternativa que consiste na exibição do conjunto de *tags* na forma de uma hierarquia, a qual é herdada da hierarquia de substantivos da *WordNet*. Uma limitação da aplicação é o funcionamento apenas para o sistema *Delicious* (YAHOO, 2003) mediante a instalação de uma extensão no navegador do usuário.

Uma alternativa às estruturas navegacionais citadas é a da utilização das ontologias de personomias deste trabalho para a geração de estruturas mostrando as relações reais entre as *tags*. Uma forma interessante para representar esses dados é por meio de grafos mostrando termos mais gerais e, a partir deles, o usuário pode especializar sua busca até obter o resultado desejado. Esse modelo de navegação ajudaria o usuário no reconhecimento de conceitos utilizados em categorizações, tarefa na qual os seres humanos têm maior facilidade cognitiva do que a lembrança (ANDERSON, 1998), a qual normalmente é bastante exigida na recuperação de informação em sistemas baseados em *tagging*. O primeiro modelo navegacional que testamos foi o uso de estruturas hierárquicas para organizar as *tags*.

4.3.1 Geração de Hierarquias para a Navegação no espaço de *Tags*

Para criar uma estrutura para a exibição do conjunto de *tags* na forma de uma hierarquia a partir de uma ontologia de personomia, basta partir do nó raiz das relações de generalização da ontologia e pegar todos os termos (da personomia e auxiliares obtidos na *WordNet*) mais especializados por meio de um algoritmo de busca em profundidade.

Um problema ao gerar uma estrutura em forma de árvore a partir de uma ontologia é que muitas vezes um nó pode ter mais de um super-nó, como pode ser visto na Figura 28a. Para resolver isso, esses nós com mais de um "pai" devem ser duplicados, conforme pode ser observado na Figura 28b.

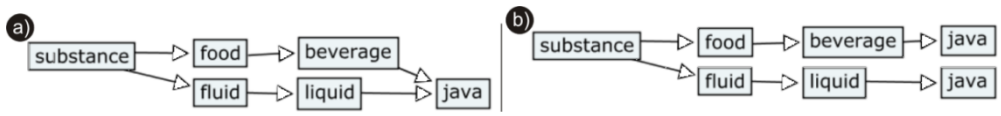


Figura 28: Para exibir o grafo em forma de hierarquia, nós com mais de um super-nó (a) devem ser duplicados (b).

Outro problema é que as relações obtidas na *WordNet* possuem granularidade muito fina, provendo uma hierarquia com muitos níveis, o que pode prejudicar sua navegação pelo usuário. Dessa forma, torna-se necessário um processo de limpeza para a remoção de alguns

nós desnecessários para a visualização de hierarquias. De forma semelhante ao processo de Laniado *et al.* (2007) nossa abordagem consiste em fazer uma busca em profundidade removendo todos os nós que tenham menos de dois sub-nós e que não represente uma *tag* da personomia do usuário (*i.e.* um nó só é removido caso seja um termo auxiliar obtido da *WordNet*). Após a remoção de um nó, seus sub-nós são adicionados ao seu super-nó. Além disso, os nós de mais alto nível como, por exemplo, "*entity*", "*physical entity*", "*abstraction*", etc. também podem ser removidos da exibição, uma vez que são muito abstratos. Uma hierarquia obtida a partir de uma ontologia de personomia sem o processo de limpeza, e mostrando nós que devem ser cortados, é mostrada na Figura 29. Os termos em negrito correspondem a *tags* da personomia e os termos auxiliares estão em cinza.

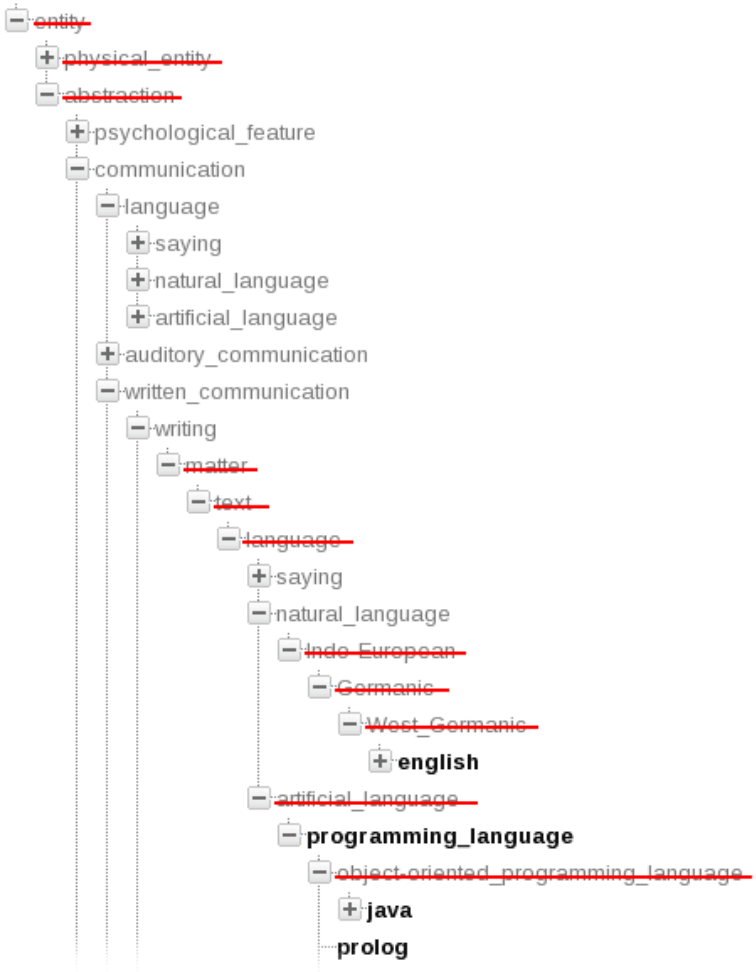


Figura 29: Estrutura hierárquica gerada a partir de uma ontologia de personomia com indicações de nós que podem ser automaticamente eliminados (tachados).

É importante observar na figura que a razão de alguns termos que estão sozinhos não terem sido cortados é que eles possuem nós filhos que estão omitidos ("*collapsed*"), o que pode ser identificado nos termos que possuam um símbolo "+" a sua esquerda. Outro aspecto a ser observado no exemplo da Figura 29 é que os termos auxiliares, os quais são obtidos na *WordNet* no momento da evolução da ontologia, são mantidos na estrutura hierárquica da ontologia formada. Outra possibilidade seria manter apenas as *tags* da personomia associadas hierarquicamente, porém, isso pode gerar um resultado não muito satisfatório, uma vez que os usuários nem sempre utilizam termos mais abrangentes e mais específicos relacionados a um mesmo conceito na categorização de recursos. Mantendo os termos auxiliares podem-se separar as *tags* da personomia de uma forma mais natural entre as categorias geradas. Um problema da utilização dos termos auxiliares entre as *tags* é que, em um primeiro momento, os usuários da hierarquia podem não estar familiarizados com a terminologia utilizada, a qual é, em parte, obtida na *WordNet*. Por outro lado, quando esses usuários se habituam a terminologia, a hierarquia lhes ajuda abstraindo a grande quantidade de dados que aparece nas listas e nuvens de *tags* tradicionais, o que requer bastante esforço cognitivo por parte do usuário na identificação dos termos desejados para a recuperação de recursos categorizados. Além disso, freqüentemente essas listas de *tags* não podem ser exibidas inteiramente na tela do computador do usuário, devido ao seu tamanho. Por meio de níveis de abstração, as hierarquias podem ser mostradas inteiramente na tela de um navegador web e ajudar o usuário a reconhecer os termos usados em categorizações. Um exemplo de uma estrutura hierárquica para navegação gerada conforme descrito nesta seção pode ser vista na Figura 30.

A recuperação da informação por meio de uma estrutura hierárquica, como a descrita nesta seção, é efetuada da mesma forma que uma das possibilidades de buscas descritas na Seção 4.2, uma vez que os termos selecionados já estão associados a um sentido.

Figura 30: Exemplo de uma estrutura hierárquica gerada a partir de uma ontologia de personomia pronta.

A representação do espaço de *tags* na forma de uma hierarquia apresenta vários aspectos positivos. Analisando do ponto de vista de implementação e da compatibilidade com navegadores *web*, é uma abordagem muito boa por não necessitar de *plugins* externos. A hierarquia pode ser gerada com *DHTML* e comunicação assíncrona com o servidor (*AJAX*), recursos suportados pela maioria dos navegadores *web* modernos (ZAKAS *et al.*, 2006). A princípio havíamos planejado a exibição das *tags* na forma de um grafo, sem as limitações de uma estrutura na forma de árvore, porém, isso necessitaria de *plugins*, como o *Flash Player* (ADOBE, 2009) ou o *Java Plugin* (SUN, 2009a). De qualquer forma, pretendemos futuramente testar essa forma de visualização das ontologias, as quais acreditamos ser mais naturais para a navegação por seres humanos e podem representar com mais facilidade os diferentes tipos de relação que obtemos entre as *tags*.

Neste capítulo foi descrito como o *TOM* pode ser integrado ao *TM*, bem como apresentadas algumas possíveis aplicações utilizando as ontologias de personomias, cujos benefícios iniciais podem ser resumidos como possibilidades de melhoria na recuperação da informação e na navegação do usuário no espaço de *tags*. Existem ainda algumas possibilidades futuras e limitações da evolução de ontologias a partir de personomias, as quais serão discutidas no próximo capítulo.

Capítulo V

Conclusões e Considerações Finais

Não conseguir encontrar a informação que sabemos que existe em uma coleção de documentos é um problema que os seres humanos procuram resolver a séculos (GARSHOL, 2004). Como citado anteriormente, a facilidade do acesso a *Internet* e a facilidade com que usuários criam e publicam conteúdo na *web* atualmente são fatores que contribuem para que encontremos uma quantidade nunca antes vista de informação disponível. Apesar de haver uma evolução nas técnicas de organização e de recuperação de informação, ainda temos dificuldade em obter os resultados desejados.

Por essa razão propusemos uma metodologia não-supervisionada para a evolução de ontologias a partir de personomias, permitindo, assim, que seja gerada uma ontologia que represente uma estrutura semelhante ao modelo mental do usuário referente às suas categorizações. Após a implementação dos algoritmos e de testes efetuados pudemos constatar que é possível a geração desse tipo de estrutura a partir dos dados de *tagging* de um usuário. Outro aspecto que constatamos é que a qualidade de uma ontologia gerada a partir de

uma personomia também depende da qualidade das *tags* que um usuário utiliza na categorização dos seus recursos. Em média, obtivemos 62% de *tags* que tiveram pelo menos um *token* reconhecido na *WordNet*, porém, em testes observamos que alguns usuários com um conjunto grande de dados de *tagging* possuíam 100% das *tags* reconhecidas nessa fonte de informação sem nenhuma forma de pré-processamento. De qualquer forma, as *tags* não reconhecidas na *WordNet* não prejudicam na estruturação da ontologia, apenas deixam de obter os benefícios da semântica e, assim, continuam sendo utilizadas na recuperação de recursos categorizados como simples rótulos de texto.

Um ponto a ser analisado é quanto a validação das ontologias geradas. Uma forma comum de fazer isso seria comparar com outras ontologias de domínio científico, porém, o objetivo do presente estudo não é criar uma ontologia para representar um domínio de forma extremamente formal, mas tentar estruturar o modelo mental do usuário referente às suas categorizações. Esse tipo de ontologia é considerada mais fraca (WIKIPEDIA, 2008) do que ontologias de domínio, mas, devido ao seu baixo custo, são ideais para colocar as *tags* utilizadas por um usuário em uma perspectiva frente a objetos do mundo real e suas relações. Apesar disso, por construção as ontologias geradas pela metodologia descrita por este estudo são conceitualmente corretas, pois elas estão ancoradas na estrutura da *WordNet*, a qual é bem estruturada e formal.

Quanto a possibilidade de usos das ontologias de personomias, apresentamos inicialmente duas aplicações: a possibilidade de buscas por conceitos ao invés de palavras-chave desprovidas de semântica e a elaboração de estruturas alternativas a listas e “nuvens” para a visualização e navegação no espaço de *tags*. Conseguimos obter com sucesso hierarquias para a organização, visualização e navegação nas *tags* de um usuário. Consideramos esse recurso importante, uma vez que essas estruturas podem abstrair uma grande quantidade de *tags* e ajudar o usuário no reconhecimento dos conceitos ao recuperar

um recurso categorizado. Isso se torna possível porque estruturas em forma de hierarquia ou grafo, por exemplo, podem descrever com mais detalhes um conceito e representar um conhecimento de forma mais semelhante a entidades do mundo real a que estão relacionadas (MINSKY, 1985). Além disso, as hierarquias de *tags* aliadas a um mecanismo de buscas semânticas, como o discutido na Seção 4.2, o qual será implementado futuramente, podem trazer diversos benefícios na recuperação de informação em sistemas baseados em *tagging*, contornando os problemas de sinônimos e homônimos, permitindo buscas mais abrangentes, entre outras funcionalidades.

Nossa proposta se diferencia dos trabalhos de Begelman *et al.* (2006) e de Mika (2005) por relacionar os conceitos de forma explícita, diferentemente das abordagens de clusterização. Ela também se diferencia das propostas de Angeletou *et al*. (2008) e de Specia e Motta (2007), pois nosso objetivo é relacionar as *tags* de acordo com entidades do mundo real que elas representam, visando reduzir alguns problemas inerentes aos sistemas baseados em *tagging*, e não o de obter entidades da *web* semântica. Da mesma forma, nossa proposta se diferencia da de Laniado *et al.* (2007), pois nosso estudo visa gerar uma estrutura ontológica, a qual pode, entre outras finalidades, ser utilizada para gerar estruturas visuais para a navegação no espaço de *tags*, como hierarquias. Outro ponto a ser destacado é que em todos os trabalhos correlatos que tivemos acesso os autores (LANIADO *et al.*, 2007; ANGELETOU *et al.*, 2008) consideram apenas as *tags* co-ocorrentes no processo de desambiguação de sentido das mesmas, nenhum deles considerou o título da categorização, o que trouxe bons resultados para o processo descrito por este trabalho.

Uma limitação do presente trabalho é que foi utilizada apenas a *WordNet* em inglês, no entanto, se for utilizada uma *WordNet* em outro idioma, o processo é o mesmo. Outro fato que pode ser considerado como uma limitação, é que a ontologia está representada em estruturas de dados da linguagem Java, porém, planejamos futuramente representá-la na

linguagem *OWL* (MCGUINNESS *et al.*, 2004), que é uma linguagem mais adequada para a representação de ontologias.

Como trabalhos futuros consideraremos também a utilização de outras fontes de informação além da *WordNet* como, por exemplo, a *DBpedia* e a *ConceptNet*. Também trataremos como trabalhos futuros a implementação do interfaceamento do sistema *TOM*, analisando as melhores formas de interação entre o usuário e a informação categorizada utilizando os benefícios de ontologias de personomia. O trabalho futuro prioritário, porém, será a implementação do mecanismo de busca semântica anteriormente comentado, o qual é um passo importante para a identificação de novos rumos em nossa pesquisa.

Por fim, a metotologia de evolução de ontologias a partir de personomias será utilizada futuramente em conjunto com o sistema *TagManager*, porém, acreditamos que possa vir a ser utilizada também por outros sistemas baseados em *tagging* que queiram prover mecanismos para facilitar a vida de seus usuários no processo de recuperação de informação, atribuindo semântica ao conjunto de *tags* utilizadas em categorizações. Além disso, acreditamos que as ontologias de personomias possam ser utilizadas em sistemas de recomendação de alguns tipos de conteúdo, uma vez que podem conter informações sobre áreas de interesse de um usuário.

Capítulo VI

Referências

ADOBE. **Adobe Flash Player**. 2009. Disponível em: <http://www.adobe.com/products/flashplayer/>. Acesso em: 20/06/09.

AL-KHALIFA, H. S.; DAVIS, H. C. **Towards better understanding of folksonomic patterns**. In: *Proceedings of the eighteenth conference on Hypertext and hypermedia*, pp. 163-166, 2007.

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

ANGELETOU, S.; SABOU, M.; MOTTA, E. **Semantically enriching folksonomies with FLOR**. In: *Workshop Collective Intelligence & the Semantic Web*, pp. 65 - 79, 2008.

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.; DA SILVA, S. R. P. **Uma Proposta para a Evolução de Ontologias a partir de Folksonomias**. In: *XIV Brazilian Symposium on Multimedia and the Web / VII Workshop on Thesis and Dissertations*. Vila Velha - ES : SBC, 2008. v. 2. pp. 197-200.

BASSO, C. A. M.; FERREIRA, J. M. P; DA SILVA, S. R. P. **Uma Proposta para a Evolução de Ontologias a partir de Personomias em Sistemas Baseados em Tagging**. *XXIX Congresso da Sociedade Brasileira de Computação / VII Encontro Nacional de Inteligência Artificial*. Bento Gonçalves – RS: SBC, 1 CD ROM, 2009.

BEGELMAN, G.; KELLER, P.; SMADJA, F. **Automated Tag Clustering: Improving search and exploration in the tag space**. In: *XV International World Wide Web Conference*, May 22–26, Edinburgh, Scotland. 2006 Também disponível em <http://www.pui.ch/phred/automated_tag_clustering/automated_tag_clustering.pdf>. Acesso em: 10/02/2009.

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

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.

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.

DACONTA, M.; OBRST, L.; SMITH, K. **The Semantic Web**. Wiley Publishing Inc, 2003.

DAMME, C. V.; HEPP, M.; SIORPAES, K. **FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies**. In: *4th European Semantic Web Conference*. Innsbruck, AU. pp. 57-70. 2007.

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

EDMONDS, P.; AGIRRE, E. **Word Sense Disambiguation**. *Scholarpedia*, 3(7):4358, 2008. Disponível em <http://www.scholarpedia.org/article/Word_sense_disambiguation>. Acesso em: 10/12/08.

FELLBAUM, C. **English Verbs as a Semantic Net**. *International Journal of Lexicography* 1990 3(4):278-301; doi:10.1093/ijl/3.4.278.

FELLBAUM, C. **WordNet: An Electronic Lexical Database**. Cambridge: The MIT Press, 1998.

FINLAYSON, M. A. **MIT Java WordNet Interface**. 2008. Disponível em <http://projects.csail.mit.edu/jwi/>. Acesso em: 20/03/2008.

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, DOI: 10.1177/0165551504045856.

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: 22/11/07.

GRUBER, T. **Ontology of Folksonomy: A Mash-up of Apples and Oranges**. In: *First On-Line conference on Metadata and Semantics Research (MTSR2005)*. Disponível em <http://tomgruber.org/writing/mtsr05-ontology-of-folksonomy.htm>. Acesso em 15/11/07.

GRUBER, T. **Ontology**. *Encyclopedia of Database Systems*, Ling Liu e M. Tamer Özsu (Eds.), Springer-Verlag, pp78-89, 2008. Também disponível em: <http://tomgruber.org/writing/ontology-definition-2007.htm>. Acesso em 20/03/09.

GUY, M.; TONKIN, E. **Folksonomies: Tidying up tags?** In: *D-Lib Magazine*, Volume 12, Number 1, ISSN 1082-9873, January-2006 pp

HERMAN, I.; SWICK, R.; BRICKLEY, D. **Resource Description Framework (RDF)**. 2004. Disponível em <http://www.w3.org/RDF/>. Acesso em 28/11/07.

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, pp.411-426. Springer, June 2006.

IPROSPECT. **iProspect Search Engine User Attitudes.** iProspect Search Engine Marketing Firm. 2007. Disponível em <http://www.iprospect.com/premiumPDFs/iProspectSurvey Complete.pdf>. Acesso em 15/10/07.

KIKAS, T.; TREUMUTH, M. **Word Sense Disambiguation: WordNet SenseRelate AllWords**. 2007. Disponível em <math.ut.ee/~treumuth/NLP/semantics2.pdf>. Acesso em: 05/01/09.

KNERR, T. **Tagging Ontology – Towards a Common Ontology for Folksonomies**. 2006 Disponível em <http://tagont.googlecode.com/files/TagOntPaper.pdf>. Acesso em 21/11/07.

LANIADO, D.; EYNARD, D.; COLOMBETTI, M. **A Semantic Tool to Support Navigation in a Folksonomy**, In: *IV Italian Workshop on Semantic Web Applications and Perspectives*, Bari, IT. pp 192-201, 2007.

LIN, D. **An information-theoretic definition of similarity**. In: *Proceedings of the International Conference on Machine Learning*, Madison. pp. 296-304.1998.

LIU, H.; SINGH, P. **Commonsense Reasoning in and over Natural Language** In: *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES'2004)*. Wellington, New Zealand. September 22-24. Lecture Notes in Artificial Intelligence, Springer, v. 3215/2004, DOI 10.1007/b100916, pp. 293-306, 2004

MCGUINNESS, D. L.; HARMELEN, F. V. **OWL Web Ontology Language**. 2004. Disponível em <http://www.w3.org/TR/owl-features/>. Acesso em: 28/11/07.

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 14/09/07.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge: The MIT Press, 1999.

METAWEB. **Freebase – A Welth of Free Data**. Disponível em <http://www.freebase.com/>. Acesso em: 05/01/2009.

MIHALCEA, R. **Using Wikipedia for Automatic Word Sense Disambiguation**. In: *Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Rochester, New York, USA. pp. 196-203. 2007.

MIKA, P. **Ontologies are us: A unified model of social networks and semantics**. In *Proceedings of the 4th International Semantic Web Conference, volume 3729 of Lecture Notes in Computer Science, pages 522–536*. Springer, 2005.

MILLER, G. A.; BECKWITH, R.; FELLBAUM, C.; GROSS, D. e MILLER, K.J. **Introduction to WordNet: An On-line Lexical Database**. In: *International Journal of Lexicography 1990 3(4):235-244*; *doi:10.1093/ijl/3.4.235*. *Revised August 1993*.

MINSKY, M. **The Society of Mind**. New York-NY: Simon e Shuster, 1985.

NETSCAPE. **Open Directory Project**. Disponível em <http://www.dmoz.org/>. Acesso em: 05/03/2009.

NEWMAN, R. **Tag Ontology Writeup**. 2005. Disponível em <http://www.holygoat.co.uk/projects/tags/>. Acesso em: 21/11/07.

PEDERSEN, T.; PATWARDHAN, S.; MICHELIZZI, J. **WordNet::Similarity - Measuring the Relatedness of Concepts**. In: *Proc. 19th National Conference On Artificial Intelligence*, San Jose, CA, USA. pp. 1024-1025. 2004.

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.

PLISSON, J.; LAVRAC, N.; MLADENIC, D. **A Rule based Approach to Word Lemmatization**. In: *7th International Multiconference on Information Society*, Ljubljana, Slovenia. 2004.

RIDDLE, P. **Tags: What are They Good For?**. *School of Information. University of Texas.* USA. 2005. Disponível em: <http://www.ischool.utexas.edu/~i385q/archive/riddle_p/ riddle-2005-tags.pdf>. Acesso em: 14/05/07.

ROSCH, E. **Principles of Categorization**. University of California, Berkeley, 1988.

RUSSELL, T. **Contextual Contextual Authority Tagging: Cognitive Authority Through Folksonomy**. *Unpublished manuscript. School of Information and Library Science.* University North Carolina. 2005. <http://www.terrellrussell.com/projects/contextualauthoritytagging/conauthtag200505.pdf>. Acesso em: 14/09/2007.

SHEN, K; WU, L. **Folksonomy as a Complex Network**. *Online Information*. 2005. Disponível em <http://arxiv.org/abs/cs/0509072>. Acesso em: 11/11/08.

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

SPECIA, L.; MOTTA, E. **Integrating Folksonomies with the Semantic Web**. *Proceedings of the European Semantic Web Conference (ESWC 2007)*, Innsbruck, AU: Springer, pp 624-639. 2007.

STURTZ, D. N. **Communal Categorization: The Folksonomy**. *INFO622: Content Representation*, December, 2004.

SUN. **Java Plugin Technology**. 2009a. Disponível em <http://java.sun.com/products/plugin/>. Acesso em 29/07/09.

SUN. **Java SE Overview – at a glance**. 2009b. Disponível em: <http://java.sun.com/javase/>. Acesso em: 29/07/09.

SUNDELOF, E. **Taxonomy - The Advantage of Tagging and Folksonomies for Communities**. *Online Information*. 2005. Disponível em <http://inthefieldonline.net/blog/2005/09/21/taxonomy-the-advantage-of-tagging-and-folksonomies-for-communities/>. Acesso em: 24/11/07.

THIBODEAU, T. **wiki.dbpedia.org : About**. 2009. Disponível em <http://dbpedia.org/About>. Acesso em 20/04/09.

VOSS, J. **Tagging, Folksonomy & Co: Renaissance of Manual Indexing?**. In: *X international Symposium for Information Science*, pp. 243-254. 2007.

WAL, T. V. **Folksonomy.** *Online Information. vanderwal.net*. 2005. Disponível em <http://www.vanderwal.net/random/entrysel.php?blog=1622>. Acesso em: 18/01/08.

WARIN, M. **Using WordNet and Semantic Similarity to Disambiguate an Ontology**.2004. Disponível em <ling16.ling.su.se:8080/PubDB/doc_repository/warin2004usingwordnet.pdf>. Acesso em: 20/03/2009.

WIKIPEDIA. **Weak Ontology.** 2008. Disponível em <http://en.wikipedia.org/wiki/Weak_ontology>. Acesso em 20/07/09.

WIKIPEDIA. **Wikipedia: About**. *Online Information. wikipedia.org*. 2009. Disponível em <http://en.wikipedia.org/wiki/Wikipedia:About>. Acesso em 08/07/09.

WORDNET. **About Wordnet**. Cognitive Science Laboratory, Princeton University, 2006. Disponível em <http://wordnet.princeton.edu/>. Acesso em: 30/05/09.

WORDNET. **WNStats – WordNet 3.0 Database Statistics**. Cognitive Science Laboratory, Princeton University, 2009. Disponível em <http://wordnet.princeton.edu/man/wnstats.7WN>. Acesso em: 24/01/09.

WU, X.; ZHANG, L.; YU, Y. **Exploring Social Annotations for the Semantic Web**. *XV International World Wide Web Conference'2006*, May 23–26, 2006, Edinburgh, Scotland. ACM 1595933239/06/0005.

YAHOO. **Delicious**. 2003. Disponível em <http://del.icio.us>. Acesso em 23/03/2008.

YAHOO. **Flickr**. 2004. Disponível em <http://flickr.com>. Acesso em 23/03/2008.

ZAKAS, N. C.; McPEAK, J.; FAWCETT, J. **Professional AJAX**. Indianapolis: Wiley Publishing, 2006.