Uma Proposta para a Evolução de Ontologias a partir de Personomias em Sistemas Baseados em *Tagging*

Carlos Alberto M. Basso, Josiane M. P. Ferreira, Sérgio Roberto P. da Silva

Departamento de Informática – Universidade Estadual de Maringá (UEM)
Av. Colombo, 5790, zona 07 – 87020-900 – Maringá – PR – Brazil

mail.carlosalberto@gmail.com,{jmpinhei,srsilva}@din.uem.br

Abstract. *Tagging based systems are becoming a widely used tool since they are simple and quick to categorize resources. However, due to the free vocabulary used for tagging and due to its plane structure, there are some drawbacks inherent to this kind of system, mainly, for the information retrieval by the users. This paper presents a proposal for the evolution of ontologies from the vocabulary of tagging based systems, seeking to reduce the information overload and the cognitive effort of the users in the information retrieval process.*

Resumo. *Os sistemas baseados em tagging estão se tornando ferramentas amplamente utilizadas principalmente pela sua facilidade de organizar informações. Por outro lado, devido ao vocabulário descontrolado utilizado nas suas categorizações e da sua estrutura plana, existem alguns problemas inerentes a estes sistemas, principalmente, na recuperação da informação pelo usuário. Este artigo apresenta uma proposta para a evolução de ontologias a partir dos vocabulários de sistemas baseados em tagging, visando reduzir a sobrecarga de informação e o esforço cognitivo do usuário no processo de recuperação da informação.*

1. Introdução

Nos últimos anos temos acompanhado um enorme crescimento no volume de informação que chega até nós pelos meios de comunicação, dentre eles a *World Wide Web*. Como conseqüência cada vez mais os usuários são sobrecarregados com informações e isso exige um esforço cognitivo cada vez maior para discernir o que é ou não relevante.

Nos primeiros sistemas de busca na *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. Do ponto de vista do usuário, as taxonomias muitas vezes tornam o processo de classificação da informação custoso, principalmente quando a informação que o usuário deseja classificar 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].

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*. 2004]. Diferentemente das taxonomias, que permitem a definição apenas da relação pai-filho, as ontologias permitem a definição de vários tipos de relacionamentos tais como: o pai-filho, o todo-parte, o causa-efeito, o de associação, entre outros [Breitman 2005]. Estas relações semânticas mais estruturadas podem tornar o processo 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 classificação da informação é que esta 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 construir e manter [Echarte *et al.* 2004], 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, 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 informação. Este 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 socialmente para uma comunidade de usuários caracteriza uma **folksonomia** [Mathes 2004]. O valor deste 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. Mas, por outro lado, o processo de recuperação da informação é prejudicado por alguns motivos. O primeiro deles é que as personomias sofrem de problemas de organização e ambigüidade, que o bom desenvolvimento de vocabulários controlados [da Silva 2009] e esquemas de hierarquia efetivamente podem melhorar [Mathes 2004]. Sobre os problemas de ambigüidade podemos citar, principalmente, o uso de acrônimos, homônimos, sinônimos e o fato de que os sistemas atuais parecem não ser projetados para lidar com palavras compostas nas *tags*. O segundo problema 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[1] [Damme *et al.* 2007] não existem muitas outras opções de visualização e acesso à informação, pois, por ser semanticamente fraca, esta relação gera apenas uma estrutura plana entre as *tags*.

Visto 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 tem dificuldades cognitivas [Anderson 1995], uma possibilidade para

[1] Relação de duas *tags* que são utilizadas em conjunto em uma mesma categorização.

contornar este problema seria utilizar os ganhos da ontologia como estrutura para recuperação da informação, e os ganhos do *tagging* para organização da informação. Deste modo, a idéia central de nossa proposta é definir um processo para gerar uma ontologia com diversas relações (hierárquicas e não hierárquicas) que possa facilitar a tarefa de recuperação da informação a partir das *tags* de um usuário. Esta ontologia poderá ser utilizada futuramente para: (i) melhorar o processo de recuperação de informação, utilizando-se das relações semânticas entre as *tags* para permitir buscas por conceito ao invés da forma exata com que a *tag* foi criada; (ii) sugerir termos para categorização, baseando-se no interesse do usuário, uma vez que a ontologia obtida pode representar os assuntos de interesse do usuário; e (iii) melhorar a capacidade de "navegação" do usuário entre as *tags* relacionadas, por meio de estruturas de hierarquias ou grafos (que são estruturas melhores para o acesso à informação do que as listas e nuvens de termos), as quais podem ser obtidas por meio das relações semânticas da ontologia. Todas estas vantagens têm como principal objetivo reduzir o esforço cognitivo do usuário para recuperar a informação desejada.

Este artigo está organizado da seguinte forma: na Seção 2 são descritos alguns trabalhos correlatos e como se diferenciam do nosso. Na Seção 3 são detalhados todos os passos da nossa proposta para a evolução de ontologias a partir de personomias, incluindo uma proposta para a desambiguação do sentido de *tags*, a obtenção de relações semânticas entre as *tags* e alguns resultados. Para finalizar, na Seção 4 são apresentadas nossas conclusões, as limitações deste trabalho e algumas propostas de trabalhos futuros.

2. Propostas para emergir estrutura em sistemas baseados em *tagging*

Existem basicamente dois tipos de abordagens para extrair/emergir estrutura em sistemas baseados em *tagging*. Algumas propostas utilizam uma análise estatística baseando-se na co-ocorrência do conjunto de *tags* para identificar conjuntos (*clusters*) de *tags* relacionadas. Isto é interessante, uma vez que os sistemas com este recurso conseguem encontrar conjuntos de *tags* relacionadas identificando diferentes contextos. O problema é que, devido à falta de níveis hierárquicos, não há critério para 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 2008], os quais são recursos que podem orientar o usuário na navegação no conjunto de *tags* e ajudá-los na recuperação de informação. Como exemlo de abordagens de *clustering* podemos citar o trabalho de Begelman *et al.* (2006) e o sistema *Flickr*[2].

Outra abordagem é a de propostas que utilizam fontes externas de informação, além dos dados de sistemas de *tagging*, para obter relações semânticas entre as *tags*. Como exemplo desse tipo de abordagem podemos citar os trabalhos de Damme *et al.* (2007), de Laniado *et al.* (2007) e de Angeletou *et al.* (2008). Em Damme *et al.* (2007) são sugeridas possibilidades para mapear diversos tipos de relações entre *tags* em uma ontologia, porém, essas possibilidades não foram implementadas. Outro trabalho semelhante é o de Laniado *et al.* (2007), no qual 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*[3]. Para isso torna-se necessária a utilização de uma

[2] Sistema baseado em folksonomia para a categorização de fotos, disponível em: http://flickr.com/

[3] Sistema que permite a categorização de *bookmarks*, disponível em: http://del.icio.us/

fonte externa de informação (além da própria personomia) para a obtenção das relações hierárquicas entre os termos e, também, um processo para a desambiguação do sentido das *tags* (para que sejam apenas utilizadas relações que sejam de interesse do usuário). Nosso trabalho se diferencia do de Laniado *et al.* (2007) porque nosso foco está na evolução de uma estrutura com vários tipos de relações entre as *tags* de um usuário, enquanto eles estão preocupados somente com a obtenção de relações hierarquicas importantes para a navegação. Por último, há ainda o trabalho de Angeletou *et al.* (2008), o qual também obtém algumas relações semânticas em uma fonte externa de informações além das próprias folksonomias, mas este trabalho tem o objetivo de encontrar entidades da *web* semântica em outras ontologias disponíveis na *web* para associar com as *tags*, ao contrário do nosso cujo objetivo é emergir uma estrutura de vários níveis a partir das *tags*.

3. Uma proposta para transformar os dados de personomias em ontologias

Para evoluir uma ontologia a partir de uma personomia torna-se necessário, primeiramente, entender o processo de *tagging*. Nesse sentido, existem alguns trabalhos como o de Knerr (2007) e o de Echarte (2004) que definem modelos ontológicos para esta tarefa. Os modelos diferem-se um pouco uns dos outros, mas todos possuem no mínimo os três pivôs dos sistemas de *tagging* que são: o usuário, o recurso categorizado e as *tags* utilizadas. Em nosso trabalho se fez necessário estender estes modelos para que se torne possível colocar sentidos nas *tags* e relacioná-las com outras *tags* por meio de relações semânticas. Na Figura 1 as relações em cinza representam o modelo ontológico proposto por Knerr (2007) e as relações em preto expressam o nosso entendimento de como a tarefa de *tagging* seria modelada com algumas relações semânticas entre as *tags*.

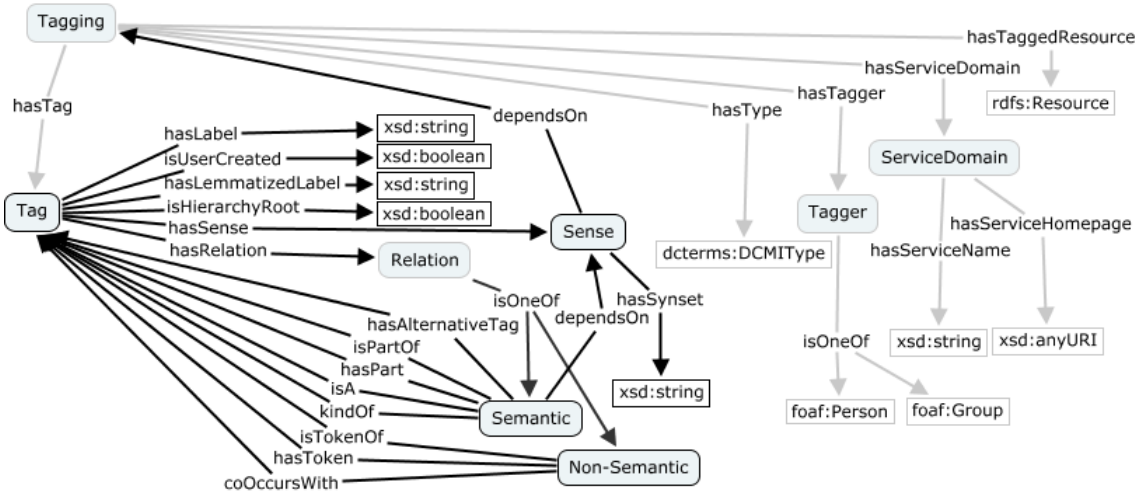

Figura 1: Modelo ontológico que representa um *tagging* estendido a partir do modelo de Knerr (2007) com relacionamentos semânticos entre as *tags*.

Segundo Anderson (1995), "depois de processar uma mensagem lingüística, as pessoas lembram apenas do sentido e não exatamente das palavras utilizadas". Em outras palavras, em geral, os usuários lembram o sentido das *tags* utilizadas na categorização, mas não de sua forma escrita exata, o que pode prejudicar em muito a recuperação da informação na personomia. Como pode ser observado na Figura 1, a relação *hasAlternative* é ideal para resolver este problema, uma vez que ela relaciona uma *tag* a

formas alternativas de escrita. As relações *hasPart* e *isPartOf* agrupam partes a um todo e vice-versa. Além dessas, as relações *isA* e *kindOf* associam semanticamente termos mais abrangentes com *tags* cujo significado é mais específico e vice-versa. Estas duas últimas relações exercem 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. Esta taxonomia pode ser utilizada no lugar das caóticas nuvens e listas de *tags*, permitindo ao usuário ir especializando uma busca até chegar ao resultado esperado. Ela também permite ao usuário fazer uma busca mais abrangente, retornando todos os objetos categorizados com *tags* mais específicas como, por exemplo, uma busca por *"vehicle"* poderia retornar objetos categorizados tanto com a *tag "bike"*, quanto com a *tag "car"*.

Fixado o modelo ontológico do processo de *tagging* a ser utilizado, podemos partir para o processo de construção da estrutura a partir da personomia. Para obter os dados das categorizações de um usuário em vários sistemas baseados em *tagging*, utilizamos o sistema denominado *TagManager* [da Silva 2009], o qual tem a função de gerenciar as informações das personomias de um usuário nos diversos sistemas baseados em *tagging* que ele utilize.

Depois da 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 a única relação entre os termos até então. Para obter estes relacionamentos semanticamente mais ricos é necessário que seja feito um *mashup*[4] com outras fontes de dados externa que permitam a obtenção de tal informação. Como as *tags* são elementos textuais uma possibilidade para obter este tipo de informação é utilizar a *WordNet*, a qual é um grande banco de dados léxico eletrônico desenvolvido com base em teorias psicolingüísticas concernentes à organização do léxico na memória humana [Wordnet 2006]. A *WordNet* não é um dicionário comum, pois nela os substantivos, os verbos, os adjetivos e os advérbios são agrupados em conjuntos de sinônimos cognitivos, chamados de *synsets*, cada um expressando um conceito distinto. Os *synsets* por sua vez, são associados de acordo com o significado por meio de relações semânticas, formando uma rede de palavras. Assim sendo, a *WordNet* demonstra-se uma excelente fonte de informações para obter relações entre os termos que compõe as *tags* e termos associados de acordo com o significado dos mesmos. Porém, ao obtermos relações lingüísticas a partir das *tags* chegamos a alguns problemas devido às diferentes formas de escrita das *tags* como, por exemplo, 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.

Para identificar os conceitos da *WordNet* aos quais uma *tag* se relaciona, primeiramente quebramos as *tags* agrupadas (ex.: *CamelCase*, programming_language, etc.) em *tokens*. Após isso as *tags* e *tokens* passam por um processo de *stemming*, o qual serve para identificar a raiz de uma palavra. Por exemplo, se receber o termo no plural "*women*", o algoritmo retornará "*woman"*, no singular (o qual está na base da *WordNet*). Esses termos processados são então utilizados para buscar os conceitos na base da

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

WordNet. Se nenhuma ocorrência do termo for encontrada, a única relação desse termo com outros continuará sendo a co-ocorrência. Se a partir da *tag* mais de um conceito for encontrado na *WordNet*, torna-se necessária uma desambiguação de sentidos (a qual será descrita na Seção 3.1) para identificar a melhor opção. Por exemplo, a palavra "jaguar" pode estar se referindo a marca de carros ou ao animal, uma vez que nenhuma semântica explícita é associada pelo usuário às *tags* utilizadas na categorização, dificultando a definição das relações lingüísticas automaticamente.

Após a identificação dos conceitos que representam as *tags* do usuário na *WordNet* as relações semânticas entre elas pode ser obtida. Para isso, a partir das *tags* obtemos as relações de sinonímia, hiperonímia[5], hiponímia[6], meronímia[7] e holonímia[8], as quais são mapeadas no modelo ontológico da Figura 1 respectivamente para *hasAlternative*, *isA*, *kindOf*, *hasPart* e *isPartOf*, porque consideramos que estes termos seriam mais facilmente entendidos e por serem mais apropriados para a utilização de outras fontes de informação além da *WordNet*. É obtido somente um nível de relações de sinônimos e de termos mais específicos (holonímia e hiponímia), que servem para ajudar o usuário a lembrar das *tags* utilizadas, e *N* níveis para relações mais abrangentes (até o nó raiz da *WordNet*), uma vez que estas servem para organizar os dados na forma de uma hierarquia, podendo ser mostrados ao invés das nuvens e listas de tags. Na Figura 2 pode ser observado um exemplo simplificado de entrada e da respectiva ontologia obtida. A partir do conjunto de *tags* da personomia (*programmingLanguage*, *Java* e *Prolog*), cuja única relação é a de co-ocorrência, foi evoluída uma ontologia com diversas relações semânticas entre as *tags*. Além disso, termos que não constituem *tags* foram adicionados para ajudar, por exemplo, na busca por conceitos, ao invés de palavras-chaves, e para mostrar o conjunto de *tags* em forma de uma hierarquia. Pode-se observar que mais de um sentido foi encontrado para a *tag* "java", gerando um ramo fora de contexto no grafo obtido. Por esta razão, torna-se necessário um processo de desambiguação do sentido das *tags* para evitar esse tipo de informação indesejada.

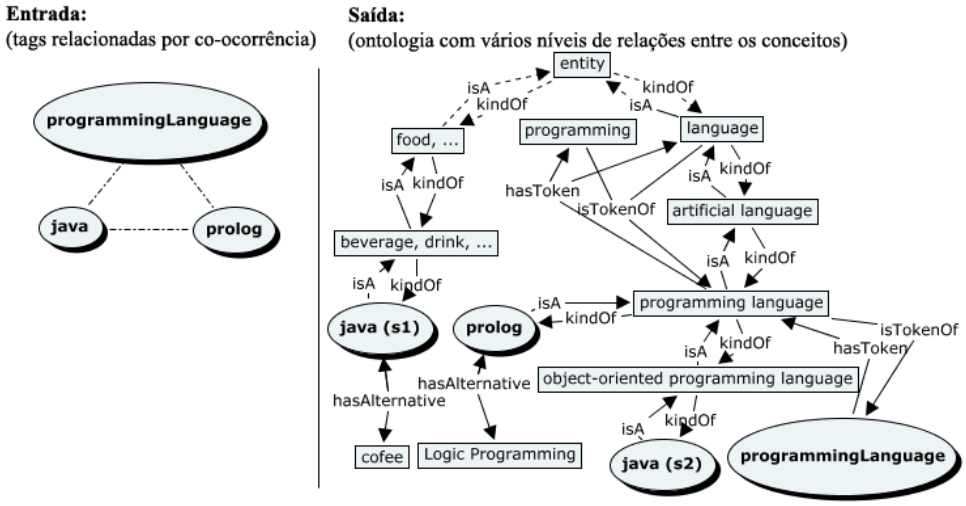

Figura 2: Exemplo simplificado de entrada e da ontologia obtida.

[5] A é hiperônimo de B se B for do tipo de A. Ex: *programming language* é hiperônimo de *Pascal* e de *Prolog*.

[6] B é hipônimo de A se B é do tipo de A. Ex: *car* e *bike* são hipônimos de *vehicle.*

[7] B é merônimo de A se B for parte de A. Ex: *wheel* e *pedal* são merônimos de *bike*.

[8] A é holônimo de B se B for parte de A. Ex: *airplane* é holônimo de *wing*.

3.1. O processo de desambiguação do sentido das *tags*

Uma palavra possui um número finito de sentidos (freqüentemente fornecidos por um dicionário ou outra fonte de referência) e a tarefa da **desambiguação** é fazer uma escolha forçada entre estes sentidos para cada uso de uma palavra ambígua baseado no seu contexto de uso [Manning e Shütze 1999]. No problema aqui abordado, os vários sentidos de uma palavra estarão dispostos na base da *WordNet*, representados pelos *synsets*. Para identificar o contexto de uma *tag* em uma categorização, pode-se utilizar (i) das *tags* co-ocorrentes, (ii) do título dado ao objeto categorizado, (iii) de uma possível descrição e, em alguns casos (iv) do próprio recurso categorizado.

Em Basso e da Silva (2008), propusemos que a desambiguação fosse feita considerando que as *tags* co-ocorrentes de uma categorização fossem comparadas entre si verificando a proximidade entre seus *synsets*. Como os *synsets* da *WordNet* estão relacionados na forma de um grafo, para a desambiguação utilizar-se-ia os que estivessem mais próximos um do outro. Após a implementação do algoritmo proposto, os resultados obtidos não foram satisfatórios. Além disso, quando havia uma única *tag* na categorização e esta fosse ambígua não era possível obter seu contexto.

Sendo assim, implementamos e testamos a utilização da métrica de **similaridade semântica** chamada ***lin*** [Lin 1998] e da métrica de **relação semântica** chamada ***lesk*** adaptada para a *WordNet* [Banerjee e Pedersen 2002]. Ambas as métricas possuem o mesmo objetivo: descrever o quão forte os sentidos de palavras estão interconectados. Por esta razão, elas 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 *IC*[9] do último *subsumer*[10] de dois conceitos com a soma do *IC* de cada um deles individualmente. Já o princípio da medida *lesk* adaptada [Banerjee 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 [Banerjee e Pedersen 2002]. A desambiguação é feita comparando entre si todos os pares de *synsets* correspondentes as *tags* de uma categorização. 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. Isso eliminaria, por exemplo, o ramo que está fora de contexto na Figura 2, o qual representa conceitos do "café java", e não da linguagem de programação com este nome.

Como passo opcional para o processo de desambiguação, podemos também considerar o título da categorização para a desambiguação. 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" na qual deve ser (i) removida a

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

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

pontuação; as palavras devem (ii) passar por um processo de *stemming*; e deve-se (iii) remover as "*stopwords*" (as quais são palavras que não trazem benefícios ao processo como, por exemplo, "*a*", "*but*", "*for*", entre outras). 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 o conceito mais adequado será utilizado.

A descrição do objeto, assim como o próprio objeto, não foram utilizados para desambiguação, pois são informações que muitas vezes não estão disponíveis ou, dependendo do recurso, não ajudam na recuperação automática do contexto de uma *tag*.

Nos experimentos realizados sobre as *tags* de um usuário, das *tags* encontradas na *WordNet* 64% possuíam mais de um sentido possível (e necessitavam desambiguação). Após o processo de desambiguação utilizando-se apenas das *tags* co-ocorrentes para a identificação do contexto 79% das *tags* ambíguas tiveram sentidos identificados com a medida *lin* e 92% com 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*.

No entanto, a identificação do sentido das *tags* não implica que as escolhas de sentido foram corretas. Para medir a qualidade da desambiguação, foram analisadas várias categorizações de diferentes usuários e verificados manualmente os sentidos escolhidos para cada *tag* de acordo com o contexto da categorização. Foram utilizadas categorizações de diversos assuntos 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*. Notou-se que as categorizações relacionadas com tecnologia foram as que mais obtiveram erros na escolha dos sentidos de *tags* ambíguas. Além disso, termos mais abstratos como, por exemplo, "*design*", que possuem muitos significados na *WordNet* também não obtiveram resultados satisfatórios.

Devido aos resultados obtidos, decidiu-se pela utilização da medida *lesk*, uma vez que, além de identificar o sentido de um número maior de *tags* ambíguas, suas escolhas foram em maior quantidade corretas em relação a medida *lin*.

Para testar a metodologia proposta, executamos o algoritmo para a evolução de ontologias sobre os dados de 2.100 personomias de usuários escolhidos ao acaso, totalizando 1.730.056 *tags*. Após a etapa de enriquecimento das *tags*, 53% delas foram reconhecidas na *WordNet*. Segundo Laniado *et al.* as *tags* mais populares possuem uma probabilidade bem maior de pertencerem a *WordNet*. Além disso, acreditamos que uma quantidade maior de *tags* possa ser reconhecida na *WordNet* se o usuário utilizar a ferramenta *TagTydier* [da Silva 2009], que trabalha em conjunto com o *TagManager* e tem como objetivo a detecção de inconsistências no conjunto de *tags* e uma personomia, deixando-o mais limpo para a evolução de uma ontologia. As *tags* não encontradas no dicionário continuarão relacionadas com outras *tags* apenas por co-ocorrência, apenas não obtendo as vantagens proporcionadas pela semântica.

Para verificar a utilidade da ontologia foram feitos alguns testes sobre as seguintes possibilidades: (i) melhorar o processo de recuperação da informação em personomias, (ii) sugerir termos para a categorização baseando-se no interesse do

usuário e (iii) possibilitar que as *tags* sejam apresentadas na forma de uma hierarquia, alternativamente a nuvens e listas de *tags*. Os resultados preliminares destes testes foram satisfatórios e serão apresentados em Basso e da Silva (2009).

4. Conclusão

Acreditamos que a geração de uma ontologia a partir de uma personomia seria melhor se, no momento da categorização, já fossem utilizados conceitos semânticos (descritos por outras ontologias), porém isso não é possível na maioria dos sistemas de *tagging* atuais. Por essa razão, neste artigo propusemos uma abordagem de obtenção automática dos conceitos e relações semânticas, uma vez que muitos dos usuários já possuem grandes quantidades de informações categorizadas. Nesses casos, acreditamos que nossa proposta torna-se muito satisfatória por auxiliar esses usuários, reduzindo o esforço cognitivo dos mesmos para recuperar a informação desejada.

Quanto ao idioma das *tags*, a princípio estarão sendo apenas consideradas as escritas na língua inglesa, por ser o idioma básico utilizado pela *WordNet*. Como trabalho futuro pretendemos utilizar outras fontes de informação para reconhecer mais *tags* (além das que são reconhecidas na *WordNet*), e obter outros tipos de relações semânticas além dos disponíveis na *WordNet*. Dentre as fontes de informação que estão sendo investigadas para uso futuro podemos citar a *ConceptNet*[11], a qual utiliza-se do "senso comum" para relacionar conceitos e a *DBpedia*[12], a qual é um esforço comunitário para extrair informação estruturada a partir da *Wikipedia*[13] e tornar essa informação disponível na *Web* no formato de uma ontologia.

A proposta para a evolução de ontologias discutida nesse trabalho será futuramente utilizada pelo sistema *TagManager* [da Silva 2009] para a exibição do conjunto de *tags* de forma alternativa a listas e nuvens de *tags* e para melhorar o processo de recuperação de informação, permitindo buscas por conceitos ao invés da utilização apenas da forma escrita das *tags*. Acreditamos que as ontologias geradas também possam vir a ser utilizadas por outros sistemas de recuperação de informação e por sistemas de recomendação de conteúdo, uma vez que a ontologia pode possuir informações sobre os interesses do usuário.

5. Agradecimentos

Agradecemos a CAPES pelo apoio ao desenvolvimento do presente trabalho.

Referências

Anderson, John R. (1995) "Cognitive Psychology and its Implications". New York: W. H. Freeman and Company, 4 ed.

Angeletou, S.; Sabou, M.; Motta, E. (2008) "Semantically enriching folksonomies with FLOR". In Workshop Collective Intelligence & the Semantic Web, 2008.

[11] http://conceptnet.media.mit.edu/

[12] http://dbpedia.org/

[13] A *Wikipedia* é uma enciclopédia eletrônica livre que está sendo construída por milhares de colaboradores.

Banerjee, S. e Pedersen, T. (2002) “An adapted Lesk algorithm for word sense disambiguation using WordNet”. In CICLING’02, pages 136–145, Mexico City.

Basso, C. A. M. e da Silva, S. R. P. (2008) “Uma Proposta para a Evolução de Ontologias a partir de Folksonomias”. In: WebMedia/WTD 2008 - Workshop de Teses e Dissertações. Vila Velha - ES : SBC, 2008. v. 1. p. 197-200.

Basso, C. A. M. e da Silva, S. R. P. (2009) “Uma Proposta para a Evolução de Ontologias a partir de Personomias em Sistemas Baseados em Tagging”. Dissertação de mestrado, a ser publicada.

Begelman, G.; Keller, P. e Smadja, F. (2006) “Automated Tag Clustering: Improving search and exploration in the tag space” In WWW2006, May 22–26, Edinburgh, UK.

Breitman, K. (2005) “Web Semântica: a Internet do Futuro”. LTC, Rio de Janeiro.

da Silva, José V. (2009) “Gerenciamento do vocabulário do usuário em sistemas baseados em tagging”. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Maringá, Maringá-PR.

Damme, C. V.; Hepp, M.; Siorpaes, K. (2007) “FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies”, http://www.kde.cs.uni-kassel.de/ws/eswc2007/proc/FolksOntology.pdf.

Echarte, F.; Estrain, J. J.; Córdoba, A.; Villadangos, J. (2004) “Ontology of Folksonomy: A New Modeling Method”, In: Conference’04, Month 1–2, 2004.

Hotho, A.; Jäschke, R.; Schmitz, C.; Stumme, G. (2006) “Information retrieval in folksonomies: Search and ranking”. In: York Sure and John Domingue, editors, The Semantic Web: Research and Applications, volume 4011 of LNCS, pages 411-426. Springer, June 2006.

Knerr, T. (2007) “Tagging Ontology – Towards a Common Ontology for Folksonomies”, http://tagont.googlecode.com/files/TagOntPaper.pdf

Laniado, D.; Eynard, D. and Colombetti, M. (2007) “A Semantic Tool to Support Navigation in a Folksonomy”, In: HT’07, 10-12/09/2007, Manchester, UK.

Lin, D. (1998) “An information-theoretic definition of similarity”. In Proceedings of the International Conference on Machine Learning, Madison, August.

Mathes, A. (2004) “Folksonomies - Cooperative Classification and Communication Through Shared Metadata”, http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

Manning, C. D. e Schütze, H. (1999) “Foundations of Statistical Natural Language Processing”, The MIT Press, Cambridge.

Smith, G. (2008) “Tagging: People-Powered Metadata for the Social Web”, New Riders, Berkeley, CA.

WordNet (2006) “About Wordnet”, Cognitive Science Laboratory, Princeton University, http://wordnet.princeton.edu/