Os Limites das Folksonomias como Conceitualizações Compartilhadas na Especificação de Modelos Conceituais

Josiane M. P. Ferreira, Cesar Augusto Tacla
Universidade Tecnológica Federal do Paraná (UTFPR)
Av. Sete de setembro 3165, CEP 80230-901
Curitiba – PR – Brasil
+55 44 3310-4685
josianempf@gmail.com, tacla@utfpr.edu.br

Sérgio Roberto P. da Silva
Universidade Estadual de Maringá
Av. Colombo, 5790, CEP 87020-900
Maringá – PR – Brasil
+55 44 3011-4076
sergio.r.dasilva@gmail.com

ABSTRACT
Looking to reduce the problem of knowledge acquisition bottleneck, this work takes on the hypothesis that the folksonomy induced from collaborative tagging data on the Web, based on parameters of authorship and motivation of categorization, can represent a shared conceptualization of a domain. Thus, it is expected that the use of such data can generate a reduction of divergences in the elicitation of terms that will be part of the conceptual model when compared with folksonomy induction algorithms that do not use these parameters.

Categories and Subject Descriptors
I.2.6 [**Artificial Intelligence**]: Learning – *Knowledge acquisition.* H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering.*

General Terms
Algorithms, Design, Experimentation.

Keywords
Conceptual Models, Knowledge Acquisition, Collaborative *Tag*ging Systems, Folksonomies.

1. CONTEXTO TEÓRICO
Guizzardi [6] adota o nome conceitualização para designar o conjunto de conceitos utilizados para articular abstrações do estado das coisas em um domínio. Modelo é uma abstração de uma porção da realidade articulada segundo uma conceitualização de um domínio. Ainda, para Guizzardi [6], tanto conceitualizações como modelos existem somente nas mentes das pessoas. O que há de concreto são **especificações de modelos conceituais** feitas em uma linguagem de modelagem que permitem expressar (representar) conceitualizações. Desta forma, a especificação do modelo conceitual – denominada de modelo conceitual, é um artefato concreto que permite aos atores envolvidos no processo de construção do modelo compreender o domínio, **atingir consenso** sobre o significado das entidades representadas e se comunicar.

Na passagem das conceitualizações e modelos abstratos para modelos concretos ocorre o problema descrito por Feigenbaum [4] denominado de **gargalo de aquisição de conhecimentos** que diz respeito à dificuldade que os engenheiros de conhecimentos têm em capturar e representar conhecimentos a partir de interações com especialistas. Entretanto, nos dias de hoje, além dos especialistas, existem outras fontes de informação que podem ser utilizadas no processo de especificação do modelo conceitual, como, por exemplo, a *Web*. O número de atores envolvidos (engenheiros de conhecimento, especialistas no domínio e usuários) no processo de especificação do modelo conceitual também pode ser maior [18].

Realizar aquisição de conhecimentos em larga escala é um processo demorado e custoso. Atingir consenso com um número elevado de atores torna-se difícil, pois aumentam as divergências, assim como o número de interações para resolvê-la. Há abordagens de aprendizado de ontologias que se utilizam de métodos e técnicas de processamento de linguagem natural, aprendizado de máquina e mineração de textos para extrair conceitos, relações e instâncias de fontes de informação processáveis (ex. esquemas de bancos de dados, textos, etc.) [11]. Algumas abordagens de aprendizado de ontologias atuais têm utilizado dados dos sistemas baseados em ***tagging* colaborativo** como fonte de informação para estes algoritmos.

Sistemas de *tagging* colaborativo são aplicações ditas sociais que permitem aos seus usuários atribuírem etiquetas (*tags*) a recursos da *Web*. Um recurso pode ser etiquetado por vários usuários com quantas e quais *tags* eles acharem convenientes. O fato interessante é que, apesar de não existir um vocabulário controlado, depois de certo tempo as *tags* utilizadas pelos usuários para etiquetar um recurso parecem se estabilizar [14]. Ao associarem as mesmas *tags* aos mesmos recursos, os usuários constroem um **vocabulário consensual** para um determinado conjunto de recursos que pode ser representativo em um domínio, como mencionado por vários autores [1, 8, 11, 13], e pode ser visto como uma forma simples de **conceitualização compartilhada** na forma de uma lista de termos (*tags*, neste caso). Por isso, várias abordagens [1, 2, 3, 6, 8, 11, 13, 14, 15, 18] utilizam estes dados como entrada para construir algum tipo de estrutura "consensual" (vocabulário compartilhado, taxonomia, ontologia) dos dados do *tagging*.

Alguns autores chamam os dados do *tagging* colaborativo de folksonomia. Neste trabalho, o termo folksonomia será utilizado para designar a estrutura coletiva consensual (vocabulário compartilhado, taxonomia, ontologia) que emerge do *tagging* colaborativo por meio de um algoritmo de indução de folksonomias [17].

2. IDENTIFICAÇÃO DO PROBLEMA
Várias das abordagens citadas que utilizam os dados do *tagging* para derivar uma estrutura consensual de um conjunto de recursos pressupõem que estes dados ajudam no desenvolvimento de modelos conceituais de consenso pelo simples fato de resultarem

de um processo humano e coletivo sem, no entanto, verificar com profundidade a natureza do conhecimento existente no *tagging*. A maioria delas procura avaliar a abordagem, ou o algoritmo utilizado que induz a estrutura consensual dos dados de *tagging*, sem, no entanto, avaliar a utilidade da estrutura derivada – se ela realmente representa um consenso; ou as características dos dados de entrada – se eles são mais interessantes do que outros conjuntos de termos para determinada tarefa. A questão principal aqui é se as folksonomias que emergem dos dados do *tagging* colaborativo são realmente consensuais, já que as abordagens que as utilizam não verificam isso na prática. Será que os dados do *tagging*, por possuírem a dimensão social, são mais úteis do que um conhecimento extraído automaticamente de textos por meio de algoritmos de extração de termos, por exemplo? Será que a forma na qual a folksonomia é induzida dos dados de *tagging* (os parâmetros utilizados) determina sua utilidade – grau de consenso – na construção de modelos conceituais?

3. OBJETIVO

O objetivo deste trabalho é buscar evidências, por meio de experimentos de modelagem conceitual, de que estruturas que emergem da dimensão social do *tagging*, de acordo com alguns parâmetros, atenuam o gargalo de aquisição que ocorre na especificação de modelos conceituais de domínios, por representarem uma conceitualização compartilhada em uma comunidade de usuários.

Especificamente, pretende-se construir um algoritmo que leve em conta informações de autoria das *tags* (o conhecimento de quem a fez) e de motivação de etiquetagem (para quê a fez) e avaliar se as folksonomias produzidas com base nestes parâmetros realmente auxiliam a atenuar o gargalo da aquisição de conhecimento na construção de modelos conceituais. Espera-se que a utilização destas folksonomias provoque a diminuição de divergências na elicitação de termos que farão parte de um modelo conceitual em comparação com algoritmos de indução de folksonomias que não utilizam estes parâmetros e, também, com algoritmos baseados em técnicas tradicionais de recuperação de informações (ex. *TF-IDF*).

4. CONTRIBUIÇÕES ESPERADAS

As contribuições desta proposta interessam aos pesquisadores que lidam com modelagem conceitual, em particular, com a atenuação do gargalo de aquisição de conhecimentos na modelagem conceitual, bem como no entendimento da utilização e dos limites de uso das folksonomias como fonte de informação na modelagem conceitual. Particularmente, propõe-se melhorar os algoritmos de indução de folksonomias pelo uso de autoria (autoridade cognitiva) e motivação das etique*tag*ens.

5. METODOLOGIA ADOTADA

Para determinar se as folksonomias induzidas podem realmente ser consideradas como conceitualizações compartilhadas em um domínio, pretende-se realizar uma série de experimentos para a construção colaborativa de modelos conceituais. Parte-se do princípio de que o grupo que utilizar como entrada a folksonomia induzida que realmente represente consenso sobre determinado domínio irá se deparar com um número menor de divergências durante um experimento de modelagem do que os grupos que utilizam outros dados de entrada.

Para avaliar o grau de divergência na modelagem conceitual, será utilizado o método *CoFolkconcept* [8]. O processo de modelagem no *CoFolkoncept* é colaborativo e se desenvolve da seguinte maneira: i) cada usuário constrói um modelo conceitual individualmente utilizando-se de um conjunto de *tags*/termos produzindo, desta forma, um modelo conceitual particular; ii) os diferentes modelos conceituais de cada usuário são comparados a fim de se detectar divergências nas *tags*/termos escolhidos por cada usuário quanto ao tipo (conceito, instância ou relação) e à posição taxonômica (quando forem conceito ou instância); iii) resolvem-se as divergências por meio de discussões estruturadas de acordo com a metodologia DILIGENT [18]; iv) gera-se uma nova versão do modelo conceitual que é consensual e repete-se o processo modificando-se individualmente o modelo consensual.

Para selecionar as *tags* sobre o domínio de interesse dos dados de *tagging* propõem-se um algoritmo que leve em conta informações de autoria e de motivação de etiquetagem, como mostra a Figura 1.

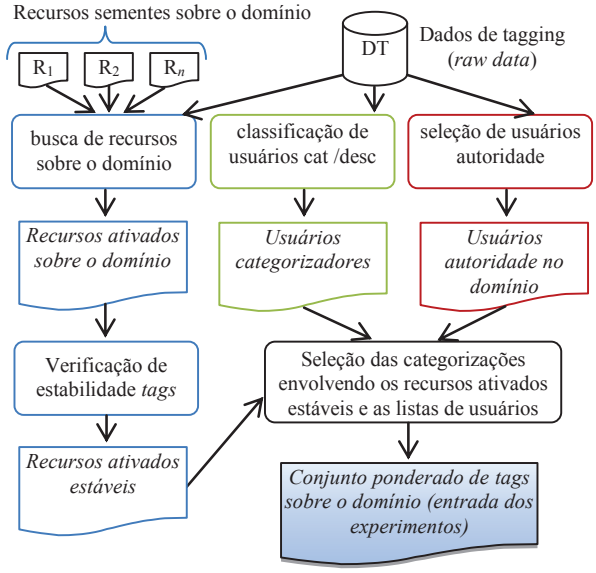

Figura 1: Algoritmo proposto para seleção do conjunto de *tags* que será utilizado como entrada dos experimentos.

Para selecionar os recursos pertinentes ao domínio de interesse em um sistema de *tagging* que permite a categorização de recursos sobre os mais diversos assuntos será utilizada a abordagem de García-Silva *et al.* [5]. Os autores desenvolveram um algoritmo que busca recursos sobre um domínio nos dados de *tagging* partindo de um conjunto de recursos sementes. As categorizações são modeladas na forma de um grafo não-direcionado no qual os vértices são os recursos e as arestas são *tags* em comum entre os dois recursos. Desta forma, dois recursos são adjacentes somente se eles contem ao menos uma *tag* em comum (independentemente de qual usuário fez a categorização). Partindo-se dos recursos sementes sobre o domínio a busca é feita em largura no grafo, com uma abordagem de *spreading activation* na qual a ativação de um recurso depende de quantas *tags* ele tem em comum com o recurso atual e do grau de ativação do recurso atual. Somente recursos com grau de ativação acima de um limiar serão visitados e considerados ativados. Outro ponto importante é verificar se o conjunto de *tags* utilizadas para categorizar estes recursos é estável, ou seja, é consensual. A abordagem descrita por Robu *et al.* [14] será utilizada para esta verificação.

Quanto à **motivação** do usuário ao realizar uma etiquetagem, pressupõe-se que ela pode ser reveladora do significado

pretendido para a *tag*, o que é importante no momento de se construir um modelo conceitual. Defende-se a ideia de que a motivação para criar uma *tag* tem influência no seu uso (ou não) durante a criação de um modelo conceitual. Körner *et al*. [9] abordam a motivação dos usuários durante a etiquetagem e tentam identificá-la automaticamente separando-as em dois grandes grupos: usuários categorizadores e usuários descritores de recursos. No conjunto de *tags* dos usuários categorizadores de recursos, há pouco uso de sinônimos (o que deve facilitar o consenso entre os atores envolvidos na especificação do modelo conceitual) e a estrutura induzida dos dados de *tagging* se aproxima de uma taxonomia. Por outro lado, no conjunto de *tags* dos usuários descritores de recursos, há uso mais proeminente de sinônimos e o vocabulário é, portanto, frequentemente maior, dificultando o consenso na especificação do modelo conceitual. Por isso, com base nas conclusões do autor, pretende-se, inicialmente classificar os usuários em categorizadores e descritores, e utilizar as *tags* dos usuários categorizadores, pois elas devem facilitar o consenso durante os experimentos de modelagem.

Quanto ao uso da autoria das *tags*, segundo Wilson [12], entidades consideradas autoridades em determinado assunto tendem a organizar melhor suas informações, possuírem conteúdos de qualidade, e manterem contato com pessoas que entendam ou tenham interesse no mesmo assunto. O autor define o conceito de **autoridade cognitiva** – uma autoridade fundamentada na competência e nas capacidades intelectuais de quem a recebe e cuja concessão é compreendida como o reconhecimento e o mérito por estas capacidades – uma autoridade que define "quem sabe o quê sobre o quê". Pressupõe-se, portanto, que pessoas que conhecem melhor determinado assunto tendem a organizar melhor as suas *tags*. Por isso, outro aspecto a ser considerado pelo algoritmo proposto é saber se quem realizou a etiquetagem possui autoridade no domínio de interesse.

Pretende-se avaliar se as folksonomias produzidas com base nestes parâmetros realmente auxiliam a atenuar o gargalo da aquisição de conhecimento na construção de modelos conceituais. Serão realizados experimentos com diferentes parâmetros de geração das folksonomias a fim de determinar em quais condições elas podem ser consideradas como conceitualizações compartilhadas.

Para fins de comparação, um algoritmo de indução de folksonomias, que não utiliza as informações de origem e de motivação, servirá de referência na avaliação (em princípio, será implementado o algoritmo de Hamasaki [7] – que utiliza a autoria das *tags*, mas não o conceito de autoridade). Um segundo algoritmo de controle fundamentado na técnica *TF-IDF* também fornecerá dados para comparação. Os três algoritmos (o proposto e os dois de comparação) utilizarão o mesmo conjunto de anotações como entrada e produzirão um conjunto de termos como saída, que será utilizado nos experimentos. O algoritmo baseado em *TF-IDF* gerará um conjunto de termos a partir das *URLs* encontradas nas mesmas anotações dos algoritmos de indução de folksonomias. No caso dos algoritmos de indução de folksonomias este conjunto de termos representa uma folksonomia em uma estrutura plana.

Os experimentos serão realizados com três grupos: o grupo de teste, que utilizará a folksonomia obtida pelo algoritmo proposto; o grupo de controle I, que utilizará a folksonomia obtida pelo algoritmo de Hamasaki; e o grupo de controle II, que utilizará o conjunto de termos obtido pelo algoritmo de *TF-IDF*. Espera-se que o grupo de teste se depare com um número menor de divergências, durante o processo de construção colaborativa do modelo conceitual, do que os grupos de controle.

6. ESTÁGIO ATUAL DO TRABALHO

Após a revisão de literatura e a especificação do modelo proposto, o algoritmo proposto e os algoritmos de comparação estão sendo implementados para gerar os conjuntos de termos/*tags* que devem ser utilizados nos experimentos iniciais. Para isso, alguns problemas de eficiência ao lidar com grandes bases de dados estão sendo resolvidos. A base de dados utilizada para testar os algoritmos possui aproximadamente 3,7 milhões de triplas usuário-*tag*-recurso, o que implica que o acesso a estes dados deve ser feito de forma muito eficiente para não se tornar um gargalo.

7. COMPARAÇÃO COM TRABALHOS RELACIONADOS

Como já citado, várias abordagens [1, 2, 3, 6, 8, 11, 13, 14, 15, 16, 18] também implementam algoritmos que induzem algum tipo de estrutura dos dados de *tagging*. Inclusive algumas delas derivam taxonomias ou ontologias leves [1, 3, 15, 16], muito próximas de modelos conceituais. O fato é que várias delas consideram a estrutura derivada como consenso [1, 8, 11, 13], mas nenhuma delas avalia este aspecto. O que se pretende avaliar neste trabalho é se esta estrutura derivada é realmente consensual e útil na especificação de um modelo conceitual a ponto de diminuir o número de divergências entre os atores envolvidos no processo. Algumas abordagens citadas até avaliam a estrutura resultante, mas em sua maioria, de forma empírica [8, 11, 13, 14], no contexto de busca [16] ou recomendação [6] para sistemas de *tagging*.

Cada abordagem, dependendo do tipo de algoritmo utilizado para derivar a estrutura (clusterização, baseadas em ontologias existentes, ou outros) e dos parâmetros adotados, deriva uma estrutura que pode ser bem diferente de outras abordagens, o que implica que a estrutura pode não ser consensual. Praticamente todas elas utilizam como ponto de partida a única relação explícita entre duas *tags* – a relação de coocorrência (duas *tags* coocorrem se elas fazem parte de uma mesma etiquetagem), sem, no entanto, levar em consideração qual o conhecimento/especialidade do usuário que fez a etiquetagem ou quais foram os motivos que o levaram a etiquetar aqueles recursos (características que o algoritmo proposto neste trabalho pretende levar em consideração). Algumas abordagens também utilizam informações sobre a autoria das *tags* [7, 9, 12, 15, 19] (em termos de qual usuário utilizou qual *tag* para etiquetar qual recurso) para extrair a relação de coocorrência entre as *tags*, mas sem avaliar o conhecimento do usuário sobre o recurso que está sendo categorizado.

REFERÊNCIAS

[1] Angeletou, S. et al. 2007. Bridging the Gap Between Folksonomies and the Semantic Web: An Experience Report. *Workshop Bridging the Gap between Semantic Web and Web 2.0, European Semantic Web Conference* (2007), 93.

[2] Begelman, G. et al. 2006. Automated *Tag* Clustering: Improving search and exploration in the *tag* space. *Collaborative Web Tagging Workshop at WWW'06* (Edinburgh, Scotland, 2006).

[3] Damme, C.V. et al. 2008. Deriving a Lightweight Corporate Ontology from a Folksonomy: a Methodology and its Possible Applications. *Scalable Computing: Practice and Experience - Scientific International Journal for Parallel and Distributed Computing*. 9, 4 (2008), 293–301.

[4] Feigenbaum, E.A. 1984. Knowledge Engineering. *Annals of the New York Academy of Sciences* (1984), 91–107.

[5] García-Silva, A. et al. 2012. *Building ontologies from folksonomies and linked data: Data structures and Algorithms*.

[6] Guizzardi, G. 2005. *Ontological Foundations for Structural Conceptual Models*. University of Twente, Enschede.

[7] Hamasaki, M. et al. 2007. Ontology Extraction using Social Network. *Proceeding of International Workshop on Semantic Web for Collaborative Knowledge Acquisition* (2007).

[8] Hauagge, J. M., Tacla, C. A., Freddo, A. R., Molinari, A. H., Paraiso, E.C. 2011. The Use of Well-founded Argumentation on the Conceptual Modeling of Collaborative Ontology Development. *International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (2011).

[9] Jäschke, R. et al. 2008. Discovering shared conceptualizations in folksonomies. *Web Semantics: Science, Services and Agents on the World Wide Web*. 6, 1 (Feb. 2008), 38–53.

[10] Körner, C. et al. 2010. Of Categorizers and Describers: An Evaluation of Quantitative Measures for *Tag*ging Motivation. *Proceedings of the 21st ACM conference on Hypertext and hypermedia* (2010), 157–166.

[11] Maedche, A. and Staab, S. 2001. Ontology Learning for the Semantic Web. *IEEE Intelligent Systems*. 16, 2 (2001), 1–18.

[12] Mika, P. 2007. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*. 5, 1 (2007), 1–15.

[13] Patrick, W. 1983. *Second-hand knowledge: An Inquiry into Cognitive Authority*. Westport: Greenwood Press.

[14] Robu, V. et al. 2009. Emergence of consensus and shared vocabularies in collaborative *tag*ging systems. *ACM Transactions on the Web*. 3, 4 (Sep. 2009), 1–34.

[15] Schmitz, P. 2006. Inducing ontology from Flickr *tags*. *In Collaborative Web Tagging Workshop, 15th WWW Conference, Edinburgh.* (2006).

[16] Specia, L. and Motta, E. 2007. Integrating Folksonomies with the Semantic Web. *4th European Semantic Web Conference* (Berlin Heidelberg, Germany, 2007), 624–639.

[17] Strohmaier, M. et al. 2012. Evaluation of Folksonomy Induction Algorithms. *Transactions on Intelligent Systems and Technology*. (2012).

[18] Tempich, C. et al. 2005. An argumentation Ontology for DIstributed, Loosely-controlled and evolvInG Engineering processes of oNTologies (DILIGENT). *The Semantic Web: Research and Applications – Lecture Notes in Computer Science* (2005), 241–256.

[19] Wu, X. et al. 2006. Exploring social annotations for the semantic web. *Proceedings of the 15th international conference on World Wide Web - WWW '06*. (2006), 417.