

Tagging: Uma Análise a Partir do Uso de Dados Pessoais para a Recomendação de Informação na Era da Web Semântica

Marcelo Rafael Borth¹, Carlos A. Meier Basso², Henrique Yoshikazu Shishido³

¹Instituto Federal de Mato Grosso do Sul (IFMS) - Campus Ponta Porã
Rua Itibiré Vieira, s/n – 79.907-414 – Ponta Porã – MS – Brasil

²Universidade Federal da Integração Latino-Americana (UNILA)
Av. Tancredo Neves, 6731 – PTI – Bloco 4 – 85.867-970 – Foz do Iguaçu – PR – Brasil

³Universidade Tecnológica Federal do Paraná (UTFPR) - Campus Cornélio Procópio
Av. Alberto Carazzai, 1640 – 86.300-000 – Cornélio Procópio – PR – Brasil

{marceloborth,meierbasso}@gmail.com, shishido@utfpr.edu.br

***Abstract.** This paper presents an analysis focused on the challenges of the use of tagging in an attempt to improve the information filtering in recommendation systems. Also discusses the use of tagging to generate data well organized in a controlled and efficient way, making the approach more meaningful to users and setting a balance between the freedom and the control in the process of categorization. Finally, we present an interface for the use on the recommendation process of the tagging-based systems that use semantic.*

***Resumo.** Este artigo apresenta uma análise focada nos desafios do uso do tagging na tentativa de melhorar a filtragem de informação nos sistemas de recomendação. Também discute-se a utilização do tagging para gerar dados mais bem organizados de forma controlada e eficiente, tornando a abordagem mais significativa aos usuários e estabelecendo um equilíbrio entre a liberdade e o controle no processo da categorização. Por fim, é apresentada uma interface como proposta para uso nos sistemas de recomendação baseados em tagging que utilizam semântica.*

1. Introdução

Devido à facilidade de publicação de novos conteúdos e a ausência de mecanismos para atuar no controle de novas informações ocorreu um crescimento descontrolado de conteúdos na Web, gerando uma **sobrecarga de informação** sobre os usuários. Nesse contexto, o *tagging* (MATHES, 2004) representa uma iniciativa para ajudar no processo de organização de conteúdos disponíveis na Web (SMITH, 2008). Esse processo é caracterizado por três elementos: o usuário – ator responsável por realizar a categorização; o objeto – o recurso a ser categorizado; e as *tags* – as palavras-chave atribuídas ao objeto com a finalidade de expressar uma descrição ao objeto. Assim, essa técnica visa facilitar o gerenciamento das informações pessoais de um usuário e auxiliar na recuperação desses objetos quando necessário (SMITH, 2008). Além disso, ela demanda pouco tempo e um baixo esforço cognitivo do usuário quando comparada a

esquemas de organização de informação mais rígidos, como as taxonomias (GOLDER *et al.*, 2006).

Nos sistemas baseados em *tagging*, os usuários escolhem livremente as *tags* (geralmente do seu próprio vocabulário) para descrever um respectivo recurso, criando-se assim o que é conhecido de **personomia** – um conjunto de recursos categorizados pelo usuário. Guy *et al.* (2006) referem-se que uma das características mais vantajosas do *tagging* é a personomia do usuário, cuja finalidade é armazenar as informações utilizadas nas categorizações do usuário, construindo o seu vocabulário e um modelo conceitual. Além disso, alguns sistemas adotam a estratégia de permitir que os usuários compartilhem suas personomias com outros usuários, produzindo o que é conhecido de **folksonomia** (SMITH, 2008) (MATHES, 2004), o qual reflete o ponto de vista de uma comunidade referente àquele objeto. O sistema mais popular que adota o uso da folksonomia é o *Delicious*¹, um *website* gerenciador de *bookmarks* social que permite a categorização de conteúdos que não são do próprio usuário. Esses tipos de sistemas oferecem vantagens interessantes no processo de organização da informação, atribuição de significado aos conteúdos (objetos) disponíveis na Web (*URLs*², fotos, vídeos, etc.), e no enriquecimento do perfil do usuário, pois a personomia reflete os interesses, terminologias, preferências, particularidades do usuário.

2. Discussão

A técnica do *tagging* começou a ser utilizada na tentativa de suprir algumas das limitações contidas nos sistemas de busca convencionais, principalmente pela sua característica de que os próprios usuários poderiam fornecer um suporte direto na qualidade da recomendação e, concedendo-os, assim, um papel mais ativo. Entretanto, a livre atribuição de *tags* é um aspecto responsável por uma série de fatores negativos (MATHES, 2004), uma vez que a associação de uma *tag* a um objeto pode ser realizada de várias maneiras devido a sinônimos, polissemia, formas léxicas de escrita distintas (singular e plural), erros ortográficos, abreviações vagas, palavras-chave cujos significados não estão diretamente relacionados ao recurso, acrônimos (siglas), rótulos com significado pessoal, diferentes níveis de precisão, alta quantidade de dados individuais, individualismo de rótulos e o fato dos sistemas não estarem adaptados a lidar com espaços nas *tags*, forçando-se, assim, o usuário inseri-la na sua forma unificada, usando *underscore* ou *CamelCase* (Ex.: “user_modeling”). Esses são os principais problemas que, se não tratados, aumentam a dificuldade de navegação, interação e gerenciamento das informações contidas na personomia do usuário. Além disso, o problema se maximiza na medida em que o conjunto de *tags* no sistema aumenta. Dentro desse contexto, trabalhos que lidam com o processamento de linguagem natural são de grande valia, uma vez que é necessário um pré-processamento das informações para elas sejam utilizadas nos sistemas de recomendação.

Em consequência dessa falta de coerência dos termos empregados na categorização, muitos recursos relevantes podem não ser recuperados em uma busca, pois em uma eventual tentativa de recuperar um objeto, a presença de termos polissêmicos, *tags* sinônimas, *tags* com variâncias no singular e plural, etc., poderão

¹ <http://del.icio.us>

² *Uniform Resource Locator*

retornar recursos que não condizem com o significado desejado. Um problema relevante nesse contexto é propor maneiras de evitar que os usuários informem *tags* com problemas, criando ferramentas que acusam tais erros ou, até mesmo, um processo que elimine ou sugira correções.

Por outro lado, os problemas na recuperação da informação acontecem em função das regras aplicadas pelos algoritmos, uma vez que muitos dos sistemas realizam a comparação da *string* informada pelo usuário com a escrita das *tags* lexicamente e não semanticamente. Entretanto, acreditamos que a maioria desses problemas estão relacionados e poderiam ser resolvidos caso o sistema utilizasse *tags* semânticas, com informações conceituais formais e sem ambiguidades. Conforme afirmam Silva e Borth *et al.* (2012), sistemas de recomendação de *tags* semânticos podem minimizar as dificuldades enfrentadas pelos usuários na recuperação da informação. Esses autores criaram uma abordagem de recomendação de *tags* semânticas, onde cada *tag* recomendada possui um conceito que está associado à base de dados da *WordNet*³. No entanto, uma das limitações da maioria desse tipo de abordagem é que elas não utilizam um banco de dados semântico com gírias, acrônimos, palavras específicas de um grupo de usuários - vocabulário regional, etc. Tratar essas limitações, envolvendo um banco de dados semântico juntamente com algum banco de dados de conhecimento geral tem grande referência na comunidade científica, pois pode retratar e/ou criar modelos de organização de informação para que sejam utilizados em futuras abordagens que utilizam semântica.

A proposta discutida neste trabalho é que os sistemas baseados em *tagging* utilizem uma abordagem de recomendação de *tags* semânticas e, também, utilizem interfaces semânticas para favorecer a recuperação de informação. Com uma recomendação de *tags* semântica seria evitado que a única relação entre as *tags* de um recurso categorizado seja a co-ocorrência – problema presente nos sistemas baseados em *tagging* convencionais. Assim, é possível consolidar um vocabulário entre os usuários, acelerar o processo de categorização e, também, aumentar a eficácia da recuperação da informação em virtude da consistência do vocabulário utilizado (JÄSCHKE *et al.*, 2007). A utilização da semântica na recuperação da informação é importante, pois o usuário poderá informar um conceito definido por um dicionário de termos para recuperar um objeto categorizado, favorecendo também o serendipismo⁴. Com a utilização de um banco de dados que possui termos relacionados e com significados formais organizados em forma de uma ontologia, torna-se possível sugerir *tags* semânticas com termos alternativos, termos abstratos ou específicos, antônimos, etc., os quais também podem ser utilizados para refinar uma busca. Os conceitos informados na busca poderão ser *tags* previamente utilizadas na categorização ou conceitos relacionados ao conteúdo do objeto que não estão presentes no texto. Além disso, é possível que uma interface auxilie a pesquisa baseada em conceitos ou pela própria *tag* na sua forma léxica. Uma sugestão de interface que utiliza semântica para a recuperação de informação é apresentada na Figura 1. Nesse exemplo, as *tags* recomendadas poderiam ser geradas pela combinação de alguns atributos presentes na ontologia do

³Um grande banco de dados léxico eletrônico que possui relações formais entre seus termos. Disponível em: <http://wordnet.princeton.edu/>

⁴ Do inglês, *serendipity*, designa a arte de fazer descobertas felizes e não intencionadas.

sistema como, o nome do conceito, a descrição e suas formas diferentes de escrita. Com essa interface, o usuário pode continuar trabalhando da mesma forma como na maioria dos sistemas e, além disso, poderá informar mais de uma *tag* na pesquisa para refinar ainda mais a filtragem dos resultados.

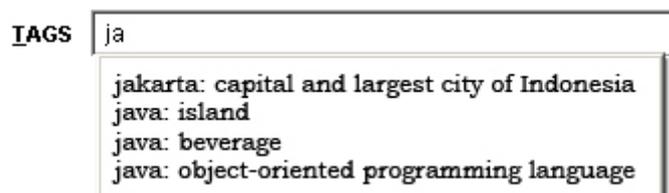


Figura 1. Sugestão de interface para a recuperação da informação em sistemas que possuem *tags* semânticas

3. Considerações Finais

Em uma época onde o efeito da sobrecarga de informação exige esforços extras para encontrar a informação desejada é interessante que seja observado com atenção o uso da técnica do *tagging*. No entanto, quando utilizada, o usuário pode ser prejudicado por vários fatores. Por essa razão, acreditamos que uma abordagem de categorização controlada possa fornecer mais benefícios ao usuário no momento de recuperar a informação desejada, de forma que o esforço cognitivo empregado na recuperação da informação seja minimizado, evitando, assim, que buscas adicionais sejam realizadas. Assim, se a técnica do *tagging* estiver aliada a uma ontologia, a filtragem da informação pode se tornar mais eficaz, permitindo uma recomendação organizada e evitando que as *tags* escolhidas se tornem polissêmicas e/ou ambíguas.

Referências

- GOLDER, S.; HUBERMAN, B. The Structure of Collaborative Tagging Systems. HP Labs Technical Report, 2006.
- GUY, M.; TONKIN, E. Folksonomies: Tidying up Tags?. D-Lib Magazine, v. 12, 2006.
- JÄSCHKE, R.; MARINHO, L.; HOTH, A.; SCHMIDT-THIEME, L.; STUMME, G. Tag Recommendations in Folksonomies. In Proceedings 11th Europ. Conf. on Principles and Practice of Knowledge Discovery in Databases, p. 506-514, 2007.
- MATHES, A. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Graduate School of Library and Information Science at University of Illinois Urbana-Champaign, Technical Report, 2004. Disponível em: <<http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>>. Acesso em 07 abr. 2013.
- SILVA, S. R. P.; BORTH, M. R.; FERREIRA, J. M. P.; FELTRIM, V. D. An Approach to Enrich Users' Personomy Using the Recommendation of Semantic Tags. Journal of the Brazilian Computer Society (Impresso), v. 18, p. 10-25, 2012.
- SMITH, G. Tagging: People-Powered Metadata for the Social Web. New Riders, Berkley, 2008.