Um sistema de recomendação para fóruns de discussão na *web* baseado na estimativa da *expertise* e na classificação colaborativa de conteúdo

Aluno: Fernando M. Figueira Filho[1]**,**
Orientador: Paulo Lício de Geus[1]**,**
Co-orientador: João Porto de Albuquerque[2]

Nível: Doutorado. Ingresso: Out/2006. Término: Jun/2010.
Etapas concluídas: Exame de qualificação, com aprovação.

[1]Instituto de Computação – Universidade Estadual de Campinas (UNICAMP)

[2]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (USP)

{fmarques,paulo}@ic.unicamp.br, j.porto@usp.br

Abstract. *In the latter years, we noticed a paradigm change in the World Wide Web. New web application functionalities incite a growing change in the user's role from a mere information consumer to an active knowledge producer. Web forums are a good example of knowledge repositories which have been collaboratively constructed. However, finding information in these environments could be a time-consuming task, given that sometimes the user cannot suitably express his/her search task in a set of keywords, issue that is specially relevant to novice users. Recommender systems have obtained good results in these cases, by guiding users through the informational space and suggesting useful content without posing the requirement that all relevant terms must be known in advance. This paper proposes a recommendation model based on expertise and metadata co-occurrence, which will be evaluated in an experiment conducted with user participation.*

Palavras-chave Sistemas de recomendação, Web 2.0, comunidades de prática, análise de redes sociais, aspectos sociais e humanos em sistemas de informação, gestão de conhecimento.

1. Introdução

A *World Wide Web* tem sido considerada uma plataforma efetiva e de baixo custo para produção de conhecimento. A crescente participação do usuário da *web* em sistemas colaborativos com o propósito de compartilhar informações é revelada como um importante processo social, impulsionando, assim, o interesse científico para o que tem sido chamado de "*web* social" ou "Web 2.0" [O'Reilly 2005]. Web logs (blogs), fóruns de discussão e ferramentas *wiki* são alguns exemplos de aplicações nesse contexto.

Neste trabalho, consideramos o caso dos fóruns de discussão da *web* (ou fóruns, daqui em diante). Fóruns estão disponíveis em várias línguas, atraindo pessoas com problemas e interesses sobre uma grande variedade de tópicos. Um tipo popular de fórum é aquele que trata sobre tópicos em computação relacionados ao *software* livre, a exemplo das distribuições do sistema operacional Linux. Porém, os atuais fóruns não são os mesmos em comparação com os grupos da *Usenet* de tempos atrás. Discussões técnicas no passado realizavam-se entre especialistas, e.g. analistas de sistemas, programadores e acadêmicos. Hoje em dia observa-se duas situações que oferecem um cenário completamente diferente para comparação. Em primeiro lugar, o *software* livre não apresenta mais a complexidade de instalação, manuseio e gerenciamento dos primeiros anos. Em segundo lugar, o *software* livre tornou-se competitivo, como observamos na implantação de distribuições Linux em estações de trabalho governamentais como forma de reduzir custos com licenças de *software* e suporte. Como consequência, o usuário de *software* livre contemporâneo não é necessariamente um especialista em computação.

Tomando a perspectiva do usuário novato, a maneira utilizada para organização da informação em fóruns oferece uma barreira de difícil transposição. Diante de centenas de mensagens postadas a cada minuto, usuários em busca de informação podem optar por uma nova postagem ou por uma busca baseada em palavras-chave. No primeiro caso, é necessário aguardar por resposta, que por sorte pode vir em cerca de minutos. No segundo caso, usuários novatos enfrentam um outro problema: como expressar as intenções de busca em palavras-chave sem conhecimento suficiente no assunto? Discussões empregam muitos termos técnicos que são necessários na recuperação de conteúdo relevante pelos motores de busca atuais. O emprego de termos genéricos como palavras-chave pode funcionar como um agravante, já que na ausência de termos específicos, o usuário é normalmente sobrecarregado com resultados de busca irrelevantes para o contexto de pesquisa. Fóruns precisam, portanto, de meios alternativos para navegar e pesquisar informações, permitindo que os usuários inexperientes possam encontrar o que precisam sem a exigência de conhecerem de antemão todos os termos que correspondem a conteúdo relevante.

Este artigo analisa o problema baseado em estudos anteriores, desenvolvidos utilizando o conceito de comunidades de prática [Wenger 1999] e das teorias em cognição distribuída [Lave 1988, Hutchins 1995]. Baseado nesse arcabouço teórico, [Bowker and Star 1999] exploram os impactos da classificação da informação em vários casos do mundo real. No entanto, pouca investigação tem sido dedicada na direção dos sistemas de informação colaborativos na *web*, tais como os fóruns, e nas contingências derivadas das diferenças de nível de experiência entre os usuários nesses ambientes. Em vista disso, o presente trabalho propõe um novo modelo de recomendação baseado na análise das redes sociais e na inferência de relações semânticas a partir dos metadados

produzidos colaborativamente nos fóruns. O modelo será implementado em um protótipo e avaliado com a participação de usuários. Este artigo está organizado da seguinte forma: a Seção 2 explora o referencial teórico. A Seção 3 propõe o modelo de recomendação e o artigo conclui com a Seção 4, que apresenta as expectativas futuras de trabalho e um plano de avaliação.

2. Comunidades de prática e classificação

O conceito de comunidades de prática (CoP) tem suas origens nas teorias da aprendizagem desenvolvidas em meados dos anos 80 por antropólogos como Lave [Lave 1988, Lave and Wenger 1991]. De acordo com Wenger [Wenger 1999], "comunidades de prática são grupos de pessoas que partilham um interesse ou uma paixão por algo que fazem, e aprendem como fazê-lo melhor na medida que interagem regularmente". Essas comunidades não são necessariamente filiadas a uma organização e, conseqüentemente, não são definidas com base em princípios como a rígida divisão do trabalho e normas de conduta formais. Além disso, CoPs não são necessariamente co-localizadas geograficamente e seus membros podem interagir usando apenas um meio virtual (e.g., sistemas *web*). O conceito foi concebido primeiramente baseado na observação de processos de aprendizagem informal [Lave 1988, Lave and Wenger 1991]. Esse trabalho vê uma profícua relação entre o conceito e o tipo de comunidade encontrada nos fóruns.

Como ilustração, suponha alguém que trabalha como um programador de sistemas. A interação diária com diversas categorias e termos relacionados à prática de programação molda a experiência e o vocabulário do indivíduo. Por esse motivo, supõe-se que programadores tenham um maior conhecimento sobre termos especificamente relacionados à atividade de programação. Na verdade, quanto mais à vontade alguém está em uma comunidade de prática, mais esse alguém esquece a natureza contingente e estranha de suas categorias como vistas de fora [Bowker and Star 1999]. Neste sentido, a naturalização é um resultado do processo de adesão (*membership*) a uma CoP. No entanto, aqui a adesão não é uma relação binária (i.e. membro ou não membro), como em um clube convencional. Em vez disso, a adesão é um processo informal, que geralmente inicia como uma 'legítima participação periférica' e culmina com a 'adesão completa' (*full-membership*), tal como definem [Lave and Wenger 1991]. Termos técnicos, jargões e expressões relacionadas às práticas regulares de uma CoP colocam-se cada vez mais em uma condição naturalizada na medida que eles funcionam como artefatos que mediam as atividades dos membros de uma mesma CoP.

No caso dos fóruns, alguns usuários são membros com adesão completa (e.g. moderadores), participando ativamente da gestão da comunidade e contribuindo com discussões técnicas. Novatos, por outro lado, têm uma participação periférica legítima na medida que adquirem conhecimento com a comunidade. Eles também participam respondendo a novas questões e auxiliando usuários menos experientes. A discrepância observada no nível de experiência tem efeitos na maneira como o conteúdo do fórum é efetivamente classificado e descrito. Usuários mais experientes possuem um repertório maior de categorias e um vocabulário mais rico, o que lhes permite navegar com maior facilidade pelo conteúdo do fórum. Novatos, por outro lado, possuem um vocabulário mais limitado, o que dificulta a expressão das suas intenções de busca através de palavras-chave e o reconhecimento de conteúdo relevante.

Em vista disso, a classificação colaborativa de conteúdo mostra-se como uma alternativa interessante às abordagens que empregam algum tipo de padronização (e.g. abordagens baseadas em ontologias) e, ao mesmo tempo, são capazes de incorporar o vocabulário da própria comunidade, evidenciando os termos e as categorias que melhor classificam uma determinada discussão no fórum. Esses termos muitas vezes não estão contidos no conteúdo textual das mensagens trocadas nos fóruns, sendo assim considerados como meta-informação ou metadados. Tais termos são úteis na medida que geralmente não são indexados pelos motores de busca tradicionais. Esses metadados podem também auxiliar no preenchimento do *gap* cognitivo existente entre os membros de uma mesma comunidade, uma vez que qualquer usuário, e.g. experiente ou novato, pode contribuir com a classificação de conteúdo. A próxima seção introduz a abordagem para produção colaborativa de metadados.

2.1. Classificação colaborativa e a produção de metadados

"Tags" são termos utilizados para rotular o conteúdo da *web*. Essa abordagem de classificação vem se tornando cada vez mais popular nas ferramentas da "Web 2.0". *Tags* funcionam como termos adicionais que podem suplementar o conteúdo textual indexado pelos motores de busca (e.g. Google[1]). O conjunto de todos os conteúdos e suas respectivas *tags* normalmente recebe o nome de folksonomia [Wal 2008]. Ao contrário das abordagens baseadas em ontologias geradas por especialistas, folksonomias baseiam-se em vocabulários não-controlados que são criados pelas próprias pessoas que utilizam o conteúdo. No caso dos fóruns, *tags* são atribuídas às discussões e representam a perspectiva de classificação, ainda que heterogênea, da própria comunidade.

A partir dessas folksonomias, muitas aplicações da *web* tem oferecido uma maneira alternativa para navegar pelo conteúdo disponível, usando as chamadas nuvens de *tags*. A nuvem do fórum da distribuição livre Ubuntu[2] é mostrada na Fig. 1 abaixo.

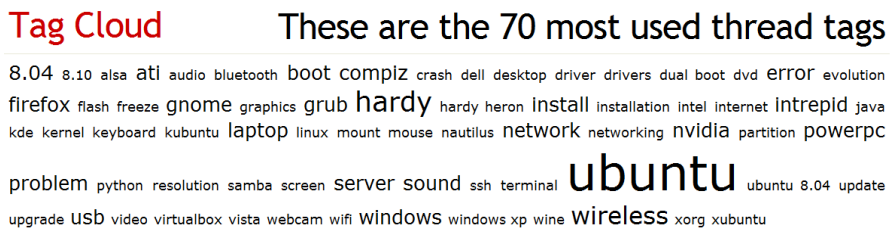

Figura 1. Nuvem das *tags* mais populares no fórum da distribuição Ubuntu

[Sinclair and Cardew-Hall 2008] mostram em estudo empírico que nuvens de *tags* podem ser úteis quando não há precisão quanto aos termos que devem ser utilizados na pesquisa. Como pode-se observar na Fig. 1, a nuvem funciona de modo similar a uma lista ponderada: *tags* que foram utilizadas com maior frequência na classificação de discussões no fórum são mostradas. No entanto, chamamos a atenção para o número de termos técnicos (e.g. *ati*, *nvidia*, *wireless*) que são apresentados com um maior tamanho de fonte. Embora esses termos possam ajudar os usuários mais experientes na busca por conteúdo, novatos provavelmente estariam perdidos. Outros termos, como *ubuntu*

[1]http://www.google.com
[2]http://ubuntuforums.org

ou *install*, se considerados isoladamente, são muito genéricos para descrever o contexto semântico de qualquer pesquisa. Além disso, as implementações atuais de nuvens permitem ao usuário filtrar informações selecionando apenas uma *tag* por vez. Entretanto, na maioria dos casos, o contexto semântico de pesquisa não pode ser descrito utilizando uma única palavra. Precisamos de uma abordagem mais inteligente que permita ao usuário informar progressivamente o seu contexto de pesquisa, através da escolha de múltiplas *tags*. Com esse objetivo, a seção seguinte apresenta o modelo de recomendação proposto nesse trabalho.

3. Um modelo de recomendação baseado no ranqueamento pela *expertise* e no refinamento progressivo do contexto semântico de pesquisa

O modelo proposto tenta resolver dois problemas no âmbito dos fóruns: (a) o vocabulário limitado dos novatos e (b) o ranqueamento de discussões no fórum de acordo com a estimativa do nível de *expertise* dos usuários no fórum. Cada questão é abordada nas próximas seções.

3.1. Sugerindo *tags* baseado em relações de subsunção

Como novatos têm pouco conhecimento sobre termos técnicos no fórum, este trabalho aborda o problema utilizando uma abordagem mais inteligente para navegação em nuvens de *tags*. Em vez de sugerir termos baseado exclusivamente na freqüência, nós propomos que os termos da nuvem sejam progressivamente atualizados para refletir um refinamento no contexto semântico de pesquisa do usuário. Se o contexto semântico da pesquisa ainda não está especificado, a nuvem mostra os termos mais gerais em primeiro lugar.

O diagrama apresentado na Fig. 2 ilustra esta idéia. Fizemos uma seleção das 70 *tags* mais populares representadas na Fig. 1. As *tags* estão organizadas em um espaço bidimensional, de acordo com a sua proximidade semântica que, na atual fase do trabalho foi estimada qualitativamente, com base na observação de três tópicos de discussão no fórum. *Tags* mais gerais, como *ubuntu* e *error*, estão posicionadas no topo, uma vez que não estão intrinsecamente relacionadas com qualquer tópico. As elipses representam os tópicos, que podem ter termos em comum na interseção. No núcleo de cada tópico pode-se ver termos muito especializados e fortemente relacionados entre si. À medida que nos afastamos do núcleo, termos tornam-se fracamente relacionados.

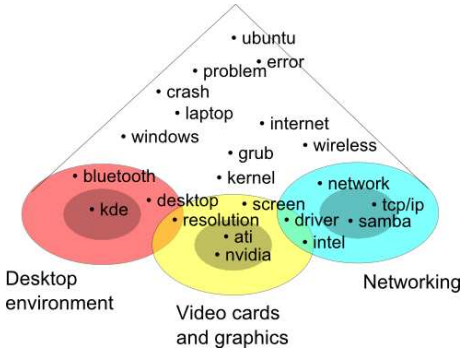

Figura 2. Hierarquia de termos

Como foi mencionado, os termos da nuvem são baseados nos metadados produzidos pela comunidade do fórum. No entanto, esses termos não são explicitamente

relacionados, considerando que nenhuma fonte exógena de conhecimento é utilizada (e.g. ontologias). Como conseqüência, se queremos mostrar os termos na nuvem dos mais gerais para os mais específicos (o que corresponde a descer do topo do triângulo da Fig. 2 para a base), precisamos de um método para explorar relações de subsunção nos termos de uma folksonomia.

[Schmitz 2006] induz essas relações da folksonomia do portal de compartilhamento de fotos Flickr[3], baseado em um modelo probabilístico de co-ocorrência de termos, i.e. *tags* que são atribuídas simultaneamente na classificação de um conteúdo. O modelo probabilístico foi adaptado a partir do modelo proposto em [Sanderson and Croft 1999]. São derivadas árvores que detém relações de subsunção entre termos quaisquer numa folksonomia. Os pais nessa árvore são, com elevada probabilidade, conceitualmente mais gerais que seus filhos. Em [Schmitz 2006], as árvores derivadas foram avaliadas manualmente e o trabalho constatou um erro médio de 23% na inferência de relações. Embora seja aceitável para os nossos objetivos iniciais, acreditamos que ainda seja necessária uma avaliação mais profunda deste modelo probabilístico, que ocorrerá nos próximos passos deste trabalho. A próxima seção aborda o segundo problema identificado nos fóruns.

3.2. Ranqueamento de conteúdo baseado na estimativa da *expertise*

Considerando-se um conjunto de discussões relevantes recuperadas através do *matching* de termos, precisamos de uma forma para diferenciar as 'boas' discussões das 'más'. Como uma 'boa' discussão, consideramos aquela composta por mensagens que podem efetivamente auxiliar no cumprimento de uma dada tarefa. Com esse objetivo, o presente trabalho encontrou uma direção na estimativa da *expertise* dos usuários como um indicativo de qualidade e perícia na produção de conteúdo no fórum.

Seguindo essa direção, [Zhang et al. 2007] mapeia as interações entre usuários numa rede de postagem-resposta, conforme representado na Fig 3. Interações são mapeadas em um grafo bipartido, através dos IDs únicos dos usuários participantes do fórum. O início de uma nova discussão é representado pelas setas tracejadas. As respostas são representadas por setas inteiras. Este grafo bipartido é então transformado em um grafo dirigido, no qual cada vértice representa um usuário no sistema, e as arestas são direcionadas do usuário que faz a postagem inicial para todos os que responderam ao mesmo. Vértices com um maior grau de entrada possuem um maior prestígio estrutural na rede, métrica que está correlacionada com o nível de *expertise* na medida que esses usuários apresentam uma maior participação na elaboração de respostas efetivas no fórum.

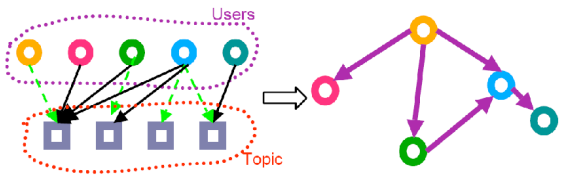

Figura 3. Rede de postagem-resposta em um fórum (retirado de [Zhang et al. 2007])

Os experimentos conduzidos por [Zhang et al. 2007] foram realizados em um conjunto de dados extraídos de um fórum popular da *web*. Eles aplicaram algoritmos

[3]http://www.flickr.com

do estado-da-arte na rede de postagem-resposta extraída do fórum, como o *Page Rank* [Brin and Page 1998]. Dois especialistas foram convidados para classificar o conjunto de dados manualmente. Ao comparar os resultados, eles encontraram uma boa correlação entre o ranqueamento proposto pelos algoritmos e o proposto pelos especialistas (em torno de 0,8). A avaliação indica que há uma forte relação entre o prestígio estrutural de usuários na rede de postagem-resposta e seus níveis reais de *expertise*. Como tal, este trabalho tem a intenção de utilizar a rede de postagem-resposta em conjunto com algoritmos de ranqueamento. A próxima seção explora os futuros passos deste trabalho.

4. Etapas futuras: Protótipo de recomendação e Avaliação

Nos novos passos deste trabalho, queremos aplicar as técnicas exploradas para desenvolver um protótipo recomendação com base no fórum de discussão da distribuição Ubuntu. O protótipo estará disponível através de uma aplicação *web* e sua interface é ilustrada na Fig. 4.

Figura 4. Ilustração da interface do protótipo de recomendação

No lado esquerdo, observa-se as discussões ranqueadas que possuem relação semântica com os termos previamente selecionados na nuvem de *tags* (à direita). No lado direito, o usuário pode observar o progresso da busca no topo. À medida que ele seleciona novos termos, a nuvem se modifica para retratar um conjunto mais refinado de termos semanticamente relacionados (Seção 3.1). Ao mesmo tempo, a cada seleção de um novo termo na nuvem, as discussões mostradas à esquerda também são atualizadas e ranqueadas de acordo com as técnicas apresentadas na Seção 3.2. Em qualquer momento o usuário terá a opção de alterar o conjunto de termos previamente selecionados se sentir que aquele conjunto não o está ajudando a encontrar a informação desejada.

A avaliação será baseada em um experimento de usabilidade realizado com um grupo de usuários. O grupo será selecionado a partir de alunos de graduação na nossa universidade. A experiência de cada usuário será avaliada com base em um questionário e em uma entrevista antes de cada sessão de testes. As tarefas serão divididas de acordo com a experiência de usuário e consistem basicamente em um cenário, i.e. um problema ou tarefa, cujas mensagens auxiliares estão disponíveis no fórum. As ações do usuário durante o experimento serão registradas pelo protótipo. Uma medida relevante é o tempo decorrido entre o início do experimento, i.e. quando o usuário é informado sobre a tarefa, e o término, i.e. quando o usuário estiver satisfeito, quando desistir ou, ainda, quando estourar o tempo limite do experimento. Outra medida relevante é o número de cliques entre o início e o final do experimento. Para uma base comparativa, o experimento será conduzido com o protótipo, e também sem ele, i.e. usando a interface atualmente disponível

no fórum do Ubuntu. O critério de sucesso (*successful completion criteria*) consistirá no usuário conseguir completar a tarefa que lhe foi atribuída no computador no qual o experimento será realizado.

Estamos particularmente interessados nas seguintes *perguntas de pesquisa*: (a) como a experiência de um usuário (medida no questionário/entrevista) afeta a interação do usuário usando o protótipo? A nuvem de termos auxilia apenas os novatos ou pode ser útil também para os mais experientes? Em que casos a busca convencional por palavra-chave foi melhor em comparação com a nuvem? (b) Em que medida o ranqueamento com base na *expertise* foi útil para encontrar respostas de alta qualidade? Esta questão, em particular, pode ser modelada em experimentos simulados sem a participação do usuário, uma vez que não dependem de dados recolhidos a partir de observação da interação do usuário. Os resultados dos experimentos serão publicados futuramente e o protótipo será colocado à disposição da comunidade de *software* livre.

Referências

Bowker, G. and Star, S. (1999). *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge, MA.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117.

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press, Cambridge, MA.

Lave, J. (1988). *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge University Press.

Lave, J. and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press.

O'Reilly, T. (2005). What is web 2.0 — o'reilly media. Online: `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html`. Acessado em: 10/8/2008.

Sanderson, M. and Croft, B. (1999). Deriving concept hierarchies from text. *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213.

Schmitz, P. (2006). Inducing ontology from flickr tags. *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May*.

Sinclair, J. and Cardew-Hall, M. (2008). The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1):15.

Wal, T. (2008). Folksonomy :: vanderwal.net. Online: `http://www.vanderwal.net/folksonomy.html`. Last access: 11/21/2008.

Wenger, E. (1999). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.

Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 221–230, New York, NY, USA. ACM.