

Interfaces Web baseadas em Conhecimento para Anotação de Recursos de Informação e Gerenciamento de Repositórios

Wanderson Rigo¹, Renato Fileto¹, Divino Inácio Ribeiro Júnior², Vinícius de Araújo Oliveira³, Vilmar César Pereira Júnior¹, Ricardo Azambuja Silveira¹

¹Departamento de Informática e Estatística (INE) – Programa de Pós-Graduação em Ciência da Computação (PPGCC) – Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88.040-900 – Florianópolis – SC – Brasil

²Centro de Ciências Humanas e da Educação (FAED) – Universidade do Estado de Santa Catarina (UDESC) – Florianópolis – SC – Brasil

³Ministério da Saúde – Brasília – DF – Brasil

{wander, fileto, vilmar.pereira, silveira}@inf.ufsc.br, divino@udesc.br, vinicius@unasus.net

Abstract. *The current repositories of information resources do not provide proper semantic support for tagging, retrieving and reusing these resources. This paper proposes the use of knowledge visualization techniques to support the selection of metadata for describing information resources and to analyze the contents of repositories. A case study in the health domain shows the potential of the proposed approach.*

Resumo. *Os repositórios de recursos de informação atuais não oferecem suporte semântico apropriado para a anotação, recuperação e reuso desses recursos. Este artigo propõe o uso de técnicas de visualização de conhecimento para apoiar a seleção de metadados na descrição de recursos de informação e para analisar o conteúdo de repositórios. Um estudo de caso no domínio da saúde mostra o potencial da abordagem proposta.*

1. Introdução

À medida que cresce o volume de informações disponíveis, mais cresce a dificuldade dos usuários encontrarem os recursos de informação que procuram (artigos, objetos multimídia, etc.) em sistemas de recuperação de informação, acervos digitais e repositórios de conteúdo [Dias 2007]. Em vista disso, a descrição e a organização da informação tornaram-se fundamentais. Diversas pesquisas têm sido desenvolvidas visando o desenvolvimento de mecanismos mais eficazes para catalogação e recuperação de informação [Faloutsos and Oard 1995] [Shah 2002] [Silva and Lima 2007]. Um suporte adequado à anotação (*tagging*) de recursos de informação [Peters 2009] faz-se necessário, pois recursos mal descritos em decorrência das limitações das linguagens naturais (e.g., uso de sinônimos, homônimos, termos genéricos) podem dificultar ou até inviabilizar a sua recuperação [Souza 2008]. Uma anotação adequada, baseada em vocabulários controlados e levando em consideração conceitos que podem ser referenciados por diferentes termos facilita a recuperação da informação e o reuso de recursos de informação em contextos diferentes daqueles para os quais eles foram

criados. Como muitos desses recursos demandam esforço e são caros para produzir, seu reuso evita gastos financeiros e atrasos decorrentes de produção replicada. Sendo assim, este trabalho analisa técnicas e ferramentas de visualização de conhecimento para amparar usuários na escolha dos termos de vocabulários controlados mais adequados para descrever seus recursos de informação. Este é o primeiro passo para se definir um processo de catalogação, busca e reuso de recursos, amparados por técnicas da Web semântica [Berners-Lee 2001]. Ainda propõe uma arquitetura e uma abordagem para adequação de conhecimento de domínio e utilização deste conhecimento com técnicas de visualização para efetuar anotação e gerenciar o conteúdo de um repositório.

2. Catalogação de Informação

O processo de anotação de recursos de informação visa descrever as características relevantes dos recursos para suportar a sua recuperação (*IR – Information Retrieval* [Baeza-Yates, R. and Ribeiro-Neto 1999]). Isso é feito por meio da associação de metadados aos recursos, com o objetivo de municiar os sistemas de IR com informações que facilitem a tarefa de recuperar tais recursos [Souza 2008]. É fundamental definir campos adequados para descrever os recursos (e.g., título, autor, formato, descrição). Diversos padrões definem campos de metadados genéricos (e.g., Dublin Core [Hillmann 2005]) ou para domínios específicos (e.g., *Learning Objects Metadata – LOM* [IEEE 2002]). Todavia, tão importante quanto definir os campos de metadados a utilizar em uma aplicação de IR é definir possíveis valores para os campos usados na catalogação e garantir o seu preenchimento correto, levando em consideração conhecimento específico do domínio.

2.1 Vocabulários Controlados

Na visão de [Warner 2002], vocabulários controlados “são listas organizadas de palavras e frases, ou sistemas de notação, que são usados inicialmente para marcação de conteúdo, e em seguida para encontrar esse conteúdo através da navegação ou pesquisa.”

2.2 Folksonomias

Thomas Vander Wal [Vander Wal 2007] cunhou o termo folksonomia para definir a técnica de escolha de palavras-chave de forma livre e de acordo com a linguagem, conhecimento, interesse, opinião ou interpretação do usuário em relação ao recurso em marcação. Em [Echarte 2007] são citados inconvenientes de folksonomias para a catalogação de recursos. Tais problemas podem ser sanados pelo uso de vocabulários controlados bem definidos [Mathes 2004].

3. Visualização de Conhecimento

Vocabulários controlados e outras formas de denotação de conhecimento, tais como ontologias, apresentam um emaranhado de termos, conceitos e relações semânticas. Assim, é essencial permitir aos usuários explorar e navegar sobre a estrutura de conhecimento do domínio, tanto durante a catalogação quanto na recuperação dos recursos de informação.

3.1 Princípios Básicos

3.1.1 Cognição e Percepção Humana

Segundo [Dias and Carvalho 2007], o entendimento sobre como funciona a cognição do ser humano mostra-se valioso quando da elaboração de bens de consumo em geral, estruturas de visualização de informação, softwares e suas interfaces. Ainda, o oferecimento de recursos gráficos com a finalidade de apresentar informação “produz a compreensão da mensagem transmitida, pois esta se torna mais natural e exige menos esforço cognitivo”. [Judelman 2004] defende que “visualizações podem idealmente ser ferramentas de pensamento e aprendizado, estendendo o processo cognitivo por permitir a ativa exploração de um espaço do conhecimento”.

3.1.2 Visualização

Segundo [Freitas 2001], a visualização da informação tem por objetivo o estudo de representações gráficas para a apresentação de informações e visa ajudar a dedução de novos conhecimentos baseados no que está sendo apresentado.

3.2 Técnicas de Visualização em GUI

Interfaces homem-computador podem incluir facilidades gráficas e de navegação que influenciam na recuperação da informação. Então, quais dessas facilidades são as mais adequadas para se desenvolver um dado sistema de recuperação de informação?

3.2.1 Mapas Conceituais

Mapas conceituais podem ser definidos como um conjunto de conceitos e relacionamento entre conceitos. Eles têm seu referencial teórico baseado na teoria de aprendizagem [Silva and Lima], a qual afirma que o conhecimento apreendido por um indivíduo é armazenado na estrutura cognitiva deste indivíduo. Baseando-se nessa teoria, Joseph D. Novak [Novak 2006] criou a abordagem de Mapas Conceituais.

3.2.2 Mapas Hiperbólicos

Segundo [Peters 2009], embora o cérebro possa perceber vários alvos visuais simultaneamente, ele não pode processá-los em paralelo. A solução então é restringir os objetos apresentados aos olhos. Mapas hiperbólicos, assim como mapas conceituais, apresentam hierarquias de conceitos. Porém, a interface de mapas hiperbólicos enfatiza a apresentação do foco (ao centro) em detrimento do seu entorno, cujo nível de detalhamento é decrementado segundo uma função hiperbólica à medida que os objetos apresentados encontram-se afastados do centro em direção à periferia [Lamping 1995].

3.2.3 Diagramas Hierárquicos

Técnicas de navegação, como mapas hiperbólicos, embora úteis para explorar grandes massas de conhecimento, mostram-se ineficazes na localização de conceitos específicos. Os usuários podem se perder em visões caóticas, preferindo visualizações que oferecem uma navegação mais gradativa e ordenada. Neste contexto destacam-se as visualizações hierárquicas, as quais aproveitam a estrutura semântica para orientar o acesso à hierarquia de conceitos [Katifori 2007]. Além disso, os diagramas hierárquicos

permitem visualizações condensadas, possibilitando que o usuário veja apenas o conhecimento relevante [Silva and Lima 2007].

3.3 Ferramentas de Visualização: Comparativo

Diversas ferramentas atualmente disponíveis permitem a visualização de conhecimento declarativo (e.g, Prefuse, Treebolic, SpaceTree, VisNomad, Last.forward, yWorks, ManyEyes, Baobab, TreeDyn, Otter e TreePlus). Este trabalho avalia 5 dessas ferramentas, para as quais encontrou-se documentação e ao menos código executável para permitir testes. Os critérios utilizados na análise comparativa dessas ferramentas, adaptados de [Silva and Lima 2007], são divididos em aspectos funcionais e aspectos não-funcionais.

3.3.1 Aspectos Funcionais

Os aspectos funcionais englobam as capacidades de representação conceitual e as facilidades de navegação das ferramentas (1-Prefuse, 2-Treebolic, 3-SpaceTree, 4-yWorks, 5-TreePlus). A Tabela 1 apresenta as capacidades de representação e visualização e a Tabela 2 apresenta as facilidades de navegação suportadas por estas ferramentas.

Tabela 1. Capacidades de representação e visualização das ferramentas avaliadas.

Capacidade	1	2	3	4	5
Mapas conceituais	X				
Visualização hiperbólica	X	X			
Diagramas hierárquicos	X		X	X	X
Representar diferentes tipos de conceitos e diferenciá-los		X		X	
Representar diferentes tipos de relações entre conceitos e diferenciá-los		X		X	
Exibir informações adicionais sobre cada conceito	X	X			
Herança múltipla		X		X	X

Tabela 2. Facilidades de navegação das ferramentas avaliadas.

Facilidade	1	2	3	4	5
Expansão e contração da estrutura	X	X	X		X
Busca textual e destaque nos resultados	X	X	X	X	X
Foco ou ênfase no conceito observado (<i>fish-eye</i>)	X	X	X		X
Ordenação alfabética dos conceitos subordinados					
Visualização do caminho da cadeia observada até a raiz	X	X	X	X	X

3.3.2 Aspectos Não-Funcionais

A Tabela 3 apresenta os aspectos não-funcionais das ferramentas analisadas. Esses dados foram coletados de artigos, páginas Web e documentação técnica sobre as ferramentas. O símbolo de interrogação (“?”) em algumas células indica que não foi possível aferir (por falta de informação) a respectiva característica para a ferramenta. A coluna “Status” indica o estado do desenvolvimento da ferramenta. A coluna “Customização” indica as possibilidades de customização da ferramenta no que se refere à representação e apresentação dos nós, arestas e dados a eles associados. A

possibilidade de se aprimorar o comportamento ou incluir novas funcionalidades é contemplada pela coluna “Extensível” (“Ext.”), que apresenta certa correlação com a disponibilidade do código fonte (coluna “Código Aberto”). A coluna “Web” indica se a ferramenta executa no ambiente Web. A disponibilidade de documentação é retratada pela coluna “Documentação” (“Doc.”). A coluna “Linguagem” (“Ling.”) indica a linguagem em que a ferramenta foi desenvolvida. Também foram realizados experimentos com dois alunos de graduação da área de tecnologia da informação para medir o tempo gasto para construir a visualização apresentada na Figura 7. As colunas “t1” e “t2” apresentam o tempo gasto (em minutos) por cada um dos dois avaliadores.

Tabela 3. Características não-funcionais das ferramentas.

Nome	Status	Customização			Ext.	Código Aberto	W e b	Doc.	Ling.	t1	t2
		Nós	Arestas	Dados							
<i>Prefuse</i>	versão Beta	via código fonte	via código fonte	via parâmetro	via código fonte	sim	sim	sim	Java	30	40
<i>Treebolic</i>	concluído em 2008	via parâmetro	via parâmetro	via parâmetro	via código fonte	sim	sim	sim	Java	28	35
<i>SpaceTree</i>	concluído em 2003	?	?	via parâmetro	?	não	sim	sim	Java	12	21
<i>yWorks</i>	ativo	?	?	?	?	não	sim	?	Java.Net Flex	20	18
<i>TreePlus</i>	concluído	via código fonte	via código fonte	via parâmetro	via código fonte	sim	?	sim	C#	40	19

3.3.3 Ferramentas Seleccionadas

*Treebolic*¹ é considerada mais adequada para navegação hiperbólica porque preenche a maioria dos requisitos funcionais e não funcionais. *Prefuse*² apresenta os melhores recursos para visualização e navegação em hierarquias e mapas conceituais.

4. Abordagem Proposta

A Figura 1 ilustra a arquitetura da abordagem proposta. A base de conhecimento é inicialmente criada pela adaptação de um vocabulário controlado de um domínio. Ela é gradativamente enriquecida pela inserção de anotações semânticas relacionando recursos de informação do repositório com termos do vocabulário controlado. Ferramentas de visualização apresentam visões da base de conhecimento para apoiar a seleção dos termos a serem utilizados como valores de metadados na anotação e busca de recursos armazenados no repositório. As visualizações utilizadas (indicadas pelas letras A, B e C na Figura 1) são apresentadas em detalhes na seção 5. A abordagem proposta é composta de quatro passos enumerados na Figura 1:

1. Adaptação inicial do conhecimento: Recortes temáticos em vocabulários controlados podem ser feitos visando elencar somente porções de conhecimento convenientes para catalogação e recuperação de informação. Estas porções devem ser

¹ <http://treebolic.sourceforge.net/en/index.html>

² <http://prefuse.org>

posteriormente armazenados em um repositório como o DSpace³. Dentre os diversos campos de metadados do LOM, este trabalho trata especificadamente do campo *keyword*, cujo preenchimento depende de conhecimento específico do domínio da saúde. A idéia é utilizar o DeCS (Descritores em Ciências da Saúde) [BIREME 2010], um vocabulário controlado muito difundido na área de saúde, para prover o conhecimento necessário para apoiar a escolha dos valores de metadados.

A Figura 3 e a Figura 4 mostram trechos do DeCS exibidos com os recursos de visualização das ferramentas Treebolic e Prefuse, respectivamente. Além das técnicas de visualização hierárquica, quando o usuário já tem conhecimento sobre o domínio, ele pode usar um componente de interface capaz de auto-completar e sugerir termos com base em correspondências léxicas e semânticas com o que foi digitado até então. A Figura 5 ilustra tal componente, que permite inclusive verificar a descrição e os sinônimos de cada termo.

Finalmente, a Figura 6 apresenta o componente para a visualização do conteúdo de acordo com conhecimento de domínio, que colore e amplia o tamanho dos nodos de acordo com a quantidade de OAs anotada com cada termo. Todos esses componentes foram acoplados ao repositório DSpace³ customizado para a UnA-SUS (Figura 5).

6. Trabalhos Relacionados

[Silva and Lima 2007] analisam três interfaces baseadas em hipertexto para organização e representação de informação. [Judelman 2004] apresenta o estado da arte em estratégias de visualização e preconiza que “o grande desafio hoje não é, necessariamente, produzir novos conhecimentos, mas desenvolver modos de melhor trabalhar com ele e dar sentido àquele conhecimento que nós já possuímos”. O trabalho de [Katifori 2007] estuda métodos e técnicas de visualização de ontologias e categoriza suas características a fim de apoiar a seleção da técnica mais adequada, além de promover pesquisas. Em [Freitas 2001] os autores apresentam uma introdução à visualização de informações, abordando aspectos considerados fundamentais e técnicas que ilustram esses aspectos. Em [Souza 2007], a autora também faz um comparativo sucinto entre ferramentas e aponta a Prefuse para a implementação do mapa de um site.

7. Conclusões

Este trabalho propõe o uso de visualizações de conhecimento de domínio para apoiar a catalogação, busca, reuso e gerenciamento de conteúdo. Suas principais contribuições são: (i) uma revisão dos fundamentos da catalogação de recursos de informação baseada em conhecimento; (ii) uma análise comparativa de técnicas e ferramentas de visualização; (iii) uma arquitetura de sistemas baseados em conhecimento e visualização para a anotação, busca e reuso de recursos de informação; (iv) uma abordagem para apoiar desde a adequação do conhecimento de domínio disponível, até a gerência da base de conhecimento e do conteúdo dos repositórios. O próximo passo desta pesquisa é a realização de testes de aceitação e aprimoramentos da proposta. Também precisam ser realizados experimentos com diversos algoritmos para processamento das buscas semânticas. Além disso, pretende-se analisar as possibilidades de ordenação dos

³ <http://www.dspace.org>

resultados, navegação sobre os mesmos e enriquecimento colaborativo da base de conhecimento, visando facilitar e potencializar a recuperação e o reuso dos recursos.

Agradecimentos

A CAPES (programa de bolsas), Ministério da Saúde (programa UnA-SUS) e ao Ministerio de Ciencia e Innovación (Espanña), projeto TIN2010-21288-C02-01.

Referências

- Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. Scientific American, 2001.
- Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval. New York: ACM Press, 1999. 511p.
- Centro Latino-Americano e do Caribe de Informação em Ciência da Saúde (BIREME). Descritores em Ciências da Saúde (DeCS) <http://decs.bvs.br> (acessado em 3 de Maio de 2010).
- Chen, M. and Qin, J. Deriving Ontology from Folksonomy and Controlled Vocabulary. iConference 2008, University of California, Los Angeles, 2008.
- Crestani, F. Application of Spreading Activation Techniques in Information Retrieval, Artificial Intelligence Review, 11(6), pp.453-482, 1997a.
- Crestani, F. Retrieving documents by constrained spreading activation on automatically constructed hypertexts, Proc. of EU FIT 97 - Fifth European Congress on Intelligent Techniques and Soft Computing, Aachen, Germany, pp. 1210-1214, 1997b.
- Dias, M. P. and Carvalho, J. O. F. A Visualização da Informação e a sua contribuição para a Ciência da Informação. DataGramZero 8 (5), 2007.
- Echarte, F., Astrain, J. J., Córdoba, A., Villadangos, J. Ontology of Folksonomy: A New Modeling Method. Proceedings of the Semantic Authoring, Annotation and Knowledge Markup (SAAKM), 2007.
- Faloutsos, C. and Oard, D.W. A survey of information retrieval and filtering methods, Technical. Report CS-TR-3514, University of Maryland, College Park, MD, 1995.
- Freitas, C. M.D.S., Chubachi, O. M., Luzzardi, P. R. G., Cava, R. A. Introdução à Visualização de Informações. Revista de Informática Teórica e Aplicada, 8(2), pp.143-158, 2001.
- Gonçalves, V. A Web Semântica no Contexto Educativo: Um sistema para a recuperação de objectos de aprendizagem baseado nas tecnologias para a Web Semântica, para o e-Learning e para os agentes, Dissertação de Doutorado, FEUP, 2007.
- Hillmann, D. Using Dublin Core. Dublin Core Metadata Initiative, 2005. <http://dublincore.org/documents/usageguide> (acessado em 3 de Maio de 2010).
- IEEE Learning Technology Standards Comittee (LTSC). Learning Object Metadata (LOM), 2002. <http://ltsc.ieee.org/wg12> (acessado em 3 de Maio de 2010).

- Judelman, G. B. Knowledge Visualization - Problems and Principles for Mapping the Knowledge Space. M.Sc. thesis, University of Lübeck, Germany, 2004
- Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C. and Giannopoulou, E. Ontology Visualization Methods - A Survey. *ACM Comput. Surv.*, 39 (4), 2007.
- Lamping, J. , Rao, R. and Pirolli, P. A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. *Proc. ACM SIGCHI Conf. on Human Factor in Computing System*, 1995.
- Mathes, A. Folksonomies – Cooperative Classification and Communication Through Shared Metadata. *Computer Mediated Communication – LIS590CMC*, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign, December 2004.
- Universidade Aberta do SUS (UnA-SUS). <http://www.universidadeabertadosus.org.br> (acessado em 3 de Maio de 2010)
- Novak, J. D., Cañas. R. The theory underling concept maps and how to construct them. Technical Report IHMC CmapTools 2006-01, Florida Institute for Human and Machine Cognition, Pensacola, Florida, 2006.
- Peters, I. Folksonomies: Indexing and Retrieval in Web 2.0. Walter de Gruyter GmbH & Co., Berlin, 2009.
- Rodrigues Jr., J. F., Balan, A. G. R., Traina, A. J. M. and Traina Jr, C. The Visual Expression Process: Bridging Vision and Data Visualization, 9th International Symposium on Smart Graphics (SG 2008), vol. 5166/2008, Rennes, France, Springer Berlin / Heidelberg, pp. 207-215, 2008.
- Shah, U., Finin, T., Joshi, A., Cost, R. S. and Mayfield, J. Information Retrieval on the Semantic Web. 10th International Conference on Information and Knowledge Management, 2002.
- Silva, M. F and Lima, G. A. B. O. Estudo comparativo entre interfaces hipertextuais de softwares para a representação do conhecimento. 89f. Dissertação (Mestrado em Ciência da Informação) - Escola de Ciência da Informação, UFMG. Belo Horizonte, 2007.
- Souza, A. B. et al. Recuperação Semântica de Objetos de Aprendizagem: Uma Abordagem Baseada em Tesouros de Propósito Genérico. XIX Simpósio Brasileiro de Informática na Educação (SBIE). 2008.
- Souza, A. S. de. Avaliação de Técnicas de Visualização de Informações na Web: estudo de caso - mapa do site do UniRitter. Trabalho de Conclusão de Curso II apresentado à Faculdade de Informática - Centro Universitário Ritter dos Reis. Porto Alegre, 2007.
- Vander Wal, T. Folksonomy Coinage and Definition, 2007 <http://vanderwal.net/folksonomy.html> (acessado em 3 de Maio de 2010).
- Warner, A. A Taxonomy Primer. *Lexonomy*. 2002. <http://www.ischool.utexas.edu/~i385e/readings/Warner-aTaxonomyPrimer.html> (acessado em 3 de Maio de 2010);