

Aperfeiçoando a navegação hiperbólica em um repositório de documentos por meio das tecnologias da Web Semântica

kadu Neves Batista Pereira , Ivan L. M. Ricarte (Orientador)

Departamento de Engenharia de Computação e Automação Industrial (DCA)

Faculdade de Engenharia Elétrica e de Computação (FEEC)

Universidade Estadual de Campinas (Unicamp)

Caixa Postal 6101, 13083-970 – Campinas, SP, Brasil

Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA)

Caixa Postal 6141, 13083-886 – Campinas, SP, Brasil

{eng_kadu,ricarte}@dca.fee.unicamp.br

Abstract – Access to a set of documents whose content was classified according to a hierarchical structure of topics can be facilitated with the use of hyperbolic trees. However, this mechanism restricts the navigation to the topics selected for the organization of the hierarchy, eventually leaving to take for shipping other similarities or relationships between documents that have been classified into distinct branches of the hierarchy. The objective of this study is to explore this possibility, using for that purpose metadata associated with documents. Considering the particular case of documents on the Web, the appropriate mechanism for the organization of metadata is the use of Resource Description Framework (RDF), one of the key technologies of the Semantic Web. The adoption of such technology could allow, in future, the integration of this application to other Semantic Web.

Keywords – Semantic Web, Metadata, Resource Description Framework, Hyperbolic Tree.

1. Introdução

A representação e descrição dos recursos eletrônicos, tais como áudio, vídeo, imagens e textos podem ser feitos por meio de metadados. Dentre os tipos de formatos existentes para descrição dos metadados, estão o *Dublin Core* e o *Resource Description Framework* (RDF). Sua utilização favorece a representação de recursos eletrônicos, tornando-os mais visíveis aos motores de busca e sistemas de recuperação [2]. A Agência de Informação da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), serviço de informação voltado ao tratamento, qualificação e gestão da informação, possui um grande volume de informação técnico-científica, resultante de atividades de pesquisa e desenvolvimento agropecuários. Na implementação atual do sistema tais recursos são disponibilizados através dos hipertextos. Além disso, todos os hipertextos da Agência de Informação já possuem embutidos em seus códigos HTML metadados no padrão *Dublin Core*. Isso permite a recuperação de seus conteúdos pelos robôs de busca da web.

Uma das formas de acesso aos recursos disponibilizados pela Agência na *Web* utiliza a navegação por árvore hiperbólica [8], cujo conteúdo foi classificado de acordo com uma estrutura hierárquica de tópicos. No entanto, esse mecanismo restringe a navegação aos tópicos selecionados para a organização da hierarquia, deixando eventual-

mente de aproveitar para a navegação outras similaridades ou relações entre documentos que tenham sido classificados em ramos distintos da hierarquia. Nesse contexto, o trabalho visa explorar tal possibilidade, usando para tal fim metadados associados aos documentos. Considerando o caso particular de documentos na *Web*, o mecanismo apropriado para a organização desses metadados é o uso de RDF, uma das tecnologias básicas da Web Semântica.

2. Web Semântica

A Web Semântica é uma extensão da Web atual que permitirá aos computadores e humanos trabalharem em cooperação. Interliga significados de palavras e, neste âmbito, tem como finalidade conseguir atribuir um significado (sentido) aos conteúdos publicados na Internet de modo que seja perceptível tanto pelo ser humano como pelo computador [1]. RDF provê uma infraestrutura para representar informações sobre recursos na *Web* por meio de arquivos com metadados. É uma tecnologia endossada e recomendada pelo *World Wide Web Consortium* (W3C), tendo como principais objetivos criar um modelo simples de dados com uma semântica formal, usar o vocabulário baseado em *Uniform Resource Identifier* (URI) e uma sintaxe baseada em XML. Os arquivos RDF têm três componentes básicos formando uma tripla: Sujeito, propriedade e objeto, o que torna a linguagem altamente escalável [7]. No domínio de aplicação do sistema Agência

de Informação, o sujeito representa uma instância de entidade, representada por uma URI. A propriedade representa um atributo da entidade, enquanto o objeto representa o valor da propriedade.

2.1. Trabalhos Relacionados

No que se refere a mecanismos de busca, uma das estratégias mais utilizadas é a expansão de consultas por meio de informações contidas em estruturas conceituais. Em [5] os autores propõem um algoritmo de expansão de consultas semânticas para recuperação de informações médicas. Sua abordagem consiste em identificar conceitos em um conjunto de termos MeSH (de Medical Subject Headings) em consultas de usuários e aplicar um algoritmo de expansão para incluir outros termos relacionados. Também é possível aumentar a eficiência de busca por conteúdos multimídia por meio da extensão de consultas feitas por usuários com a adição de informação semântica [9], bem como agregar termos a uma consulta de usuário por meio de uma expansão baseada em relações semânticas entre estruturas conceituais de domínios de conhecimento distintos [3].

A busca de um recurso por meio de navegação também pode ser melhorada com a associação de informação conceitual aos documentos. Tecnologias da Web Semântica como ontologias e metadados baseados em RDF podem ser utilizadas para prover melhor acesso e navegação em conteúdos disponibilizados na rede. Em [6], os autores abordam o desenvolvimento de um sistema de navegação por documentos baseado em similaridade semântica, a qual é representada por anotações semânticas obtidas automaticamente pela análise de conteúdos. Com a base de conhecimento gerada com a extração dessas informações, o sistema oferece ao usuário, na forma de um serviço Web, possibilidade de navegação pelos documentos organizados pela similaridade dos conceitos. No domínio de informações médicas, é abordado o desenvolvimento de um browser semântico que habilite, além de busca, navegação contextualizada com a integração de conhecimento obtido de portais de informações médicas. O sistema proposto utiliza ferramentas linguísticas para analisar artigos científicos e gerar anotações semânticas no formato RDF [4].

3. Complementação de informações

Nesta seção descreveremos brevemente a metodologia utilizada no desenvolvimento do trabalho, tendo como base um subconjunto de documentos cujo domínio de aplicação é o cultivo da cana-de-açúcar, disponibilizados pela Agência de Informação Embrapa com foco na complementação de informações para o sistema de navegação hiperbólica utilizando RDF. A figura 1 traz um exemplo do conteúdo de nó contendo os recursos de informação selecionados para a seção **Informações Complementares**.



Figura 1. Informações Complementares disponível em conteúdos de nó.

3.1. Metadados

Os recursos de informação catalogados estão associadas aos nós selecionados na árvore do conhecimento como um complemento à informação contida nesse nó. Todos os recursos têm associados metadados no formato RDF composto por um conjunto de descritores aderentes ao padrão Dublin Core. Segue um exemplo dos elementos descritores *Dublin Core* incorporado ao modelo RDF de metadados.

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://www.cnptia.embrapa.br/index.html">
    <dc:creator>Suzi</dc:creator>
    <dc:title>Embrapa Information Technology </dc:title>
    <dc:description>Research Unit Main Page.</dc:description>
    <dc:date>2001-01-20</dc:date>
```

```
</rdf:Description>
</rdf:RDF>
```

Para o gerenciamento eficiente dos dados RDF foi utilizado Banco de Dados Relacional para armazenar o conjunto de triplas RDF, permitindo inferir sobre os dados por meio de consultas SQL. Inicialmente o trabalho se propõe a relacionar recursos de informação que contenha similaridades como mesmo autor, categoria e palavra-chave, de forma que posteriormente possam ser utilizados outros descritores para relacionar os recursos de informação.

3.2. Visualização hiperbólica

Ao acessar o conteúdo de um nó e navegar até o item **Informações Complementares**, o usuário terá disponível o *hyperlink Documentos Relacionado*, que irá direcioná-lo a outra página Web contendo a listagem dos recursos relacionados, bem como a opção de visualização por árvore hiperbólica, conforme pode ser visto na figura 2.

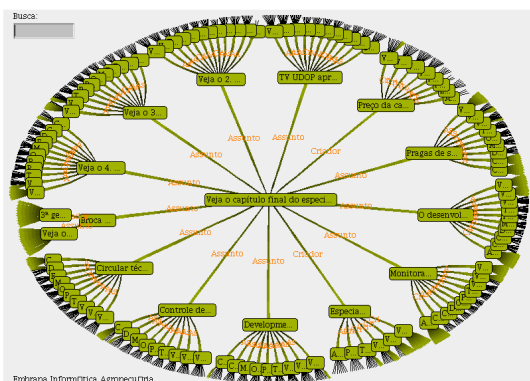


Figura 2. Visualização hiperbólica dos documentos relacionados.

4. Conclusão

Este trabalho apresentou uma abordagem para organização e visualização de recursos de informação na Web que tem por objetivo possibilitar ao usuário acesso a documentos relacionados através de similaridades e relações pertinentes ao contexto do ambiente informacional e que tenham sido classificados em ramos distintos da hierarquia.

O uso de metadados como mecanismo para estabelecer ligações semânticas entre recursos em diferentes ramos da árvore do conhecimento pode permitir, em trabalhos futuros, a integração dessa aplicação a outras da Web Semântica.

Referências

- [1] T. Berners-Lee, J. Hendler, and O. Lassla. The semantic web. 5:34–43, May 2001.
- [2] A. L. Boica and L. H. M. Oliveira. Conversão de metadados do padrão dublin core para o rdf. 1:11–17, 2008.
- [3] M.A. de Leite and I.L.M. Ricarte. Fuzzy information retrieval model based on multiple related ontologies. In *Proc. 20th IEEE International Conference on Tools with Artificial Intelligence ICTAI '08*, volume 1, pages 309–316, 2008.
- [4] G. Diallo, K. Khelif, O. Corby, P. Kostkova, and G. Madle. Semantic browsing of a domain specific resources: The corese-neli framework. In *Proc. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology WI-IAT '08*, volume 3, pages 50–54, 2008.
- [5] V. Jalali and M.R.M. Borujerdi. The effect of using domain specific ontologies in query expansion in medical field. In *Proc. International Conference on Innovations in Information Technology IIT 2008*, pages 277–281, 2008.
- [6] W. Kasper, J. Steffen, and Yajing Zhang. Using semantics for news navigation. In *Proc. IEEE International Conference on Semantic Computing*, pages 261–267, 2008.
- [7] G. Klyne and J. Carroll. Resource description framework: concepts and abstract syntax. <http://www.w3.org/TR/rdf-concepts/>.
- [8] John Lamping, Ramana Rao, and Peter Pirolli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 401–408, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.
- [9] M. Tvarozek and M. Bielikova. Personalized faceted navigation for multimedia collections. In *Proc. Second International Workshop on Semantic Media Adaptation and Personalization*, pages 104–109, 2007.