Integração e Interoperabilidade de Conteúdos em Portais Semânticos

Ana Maria de C. Moura[1] e Maria Cláudia Cavalcanti[2]

[1]Coordenação de Ciência da Computação
Laboratório Nacional de Computação Científica
Av. Getúlio Vargas 333 – Quitandinha - Rio de Janeiro

[2]Departamento de Engenharia de Computação
Instituto Militar de Engenharia
Praça General Tibúrcio 80 - Praia Vermelha - Rio de Janeiro
{anamoura, yoko}@ime.eb.br

Abstract

E-gov portals are a practical and effective way to disseminate information on the Web. However, most portals are built as independent sites, lacking support to deal efficiently with processes and services, and to provide an integrated view of similar and complementary information extracted from other institutional portals. From the citizen's side, searching information in these portals may become a hard task, since it is spread out along innumerous sites. This work proposes the development of enriched e-gov semantic portals, in order to facilitate citizen´s navigation, by integrating content extracted from other sites, considering similar or complementary domains. This work focuses on interdisciplinary work involving semantic web, and will certainly benefit from a research cooperation with INRIA, since this institute develops important projects on semantic web and maintains effective collaboration with the W3C working group.

1. INTRODUÇÃO

Nos últimos anos as organizações corporativas e setores públicos de vários países vêm sofrendo fortes pressões quanto à necessidade de revitalizarem suas administrações, tornando-as mais pró-ativas, integradas, eficientes, transparentes e, especialmente, mais orientadas a serviços. Apesar do grande desafio que essa mudança de paradigma representa, estes setores vêm introduzindo gradativamente inovações em sua estrutura organizacional de forma a mobilizar, implantar e utilizar o capital humano e de informação, recursos tecnológicos e financeiros para a entrega de serviços aos cidadãos em todo o mundo.

Nesse contexto, as iniciativas de governo eletrônico (e-Gov) referem-se ao uso da Web e outras tecnologias de informação pelo corpo governamental para interação com seus cidadãos. A prática de gestão do conhecimento aliada à adequada utilização das Tecnologias de Informação e Comunicação (TIC) desempenha um papel crucial em iniciativas privadas e públicas, tornando-as mais eficientes e competitivas. Um ótimo exemplo de iniciativa pública no Brasil é a arquitetura do e-Ping[1] (Padrões de Interoperabilidade de Governo Eletrônico), cujo objetivo é definir um conjunto mínimo de premissas, políticas e especificações técnicas que regulamentam a utilização

[1]http://www.governoeletronico.gov.br/acoes-e-projetos/e-ping-padroes-deinteroperabilidade.

de TIC no governo federal, estabelecendo as condições de interação com os demais poderes e esferas de governo e com a sociedade em geral. Nesta arquitetura, fica evidenciada a forte preocupação em oferecer portais intuitivos e de fácil acesso ao cidadão em todas as esferas do governo. Os portais oferecidos pelo governo, podem ser categorizados como: Governo para Cidadãos (G2C), Governo para Negócios (G2B), Governo para Empregados (G2E) e Governo para Governo (G2G) [13].

Apesar do grande avanço das iniciativas públicas e privadas no uso de várias modalidades das TICs na construção de seus artefatos, raras são as iniciativas que consideram a Web Semântica como uma forte tecnologia de integração e reuso de recursos. Grande parte dos portais governamentais disponibilizados até o momento não consideram o uso da Web semântica.

Com relação às iniciativas das diferentes instituições governamentais no Brasil observa-se que: nos diferentes níveis de governo, existe uma grande disparidade no grau de informatização dos serviços oferecidos, incluindo-se órgãos com apenas uma tênue ou nenhuma referência na Internet; existe, até o momento, uma grande independência e diversidade entre os órgãos que oferecem informações e serviços via Web; muito pouco existe em termos de suporte a processos ou serviços envolvendo mais de uma instituição. O mais usual é que estes se limitem às fronteiras de um único órgão governamental; em geral, cria-se um portal de referência às informações dos diversos órgãos do governo, mas que, longe de oferecer uma visão integrada destas informações, apenas se apresenta como um serviço de páginas amarelas de instituições[2], cabendo ao usuário navegar por estas instituições em busca das informações de seu interesse; na maior parte dos sites governamentais, não é fácil a localização de informações, sendo usual o usuário percorrer um longo caminho até encontrar o que deseja; com freqüência, as informações são disponibilizadas segundo níveis de detalhe, unidades e formatos bastante diversos, existindo, em geral, pouca ou nenhuma facilidade para que o usuário possa adequar o resultado de sua busca às suas necessidades.

Neste contexto, uma das poucas iniciativas no sentido de oferecer portais semânticos do tipo G2C, encontra-se em andamento no IME-RJ, um projeto de pesquisa que tem como objetivo demonstrar que o uso de tecnologias da Web Semântica contribui sobremaneira para uma melhoria na organização, integração e gestão das informações em portais. Em [7] [8] foi desenvolvida uma arquitetura que serviu como infra-estrutura básica para a construção de portais semânticos, capaz de integrar e instanciar informações a partir do uso intensivo de ontologias. Essa arquitetura permite que informações distribuídas na Web sejam recuperadas com base em uma ontologia de domínio e, através de mapeamentos definidos entre essas informações e a ontologia de domínio, novos conteúdos sejam dinamicamente categorizados e adicionados ao portal. De forma a evidenciar a aplicabilidade dessa arquitetura, foi desenvolvido o POrtal SEmântico EDUcacional (POSEDU[3]), [9], voltado para o domínio educacional.

No entanto, ao longo desse trabalho, observou-se um número muito restrito de ontologias na Web, mesmo em domínios de grande relevância, a exemplo de educação. A grande maioria dos portais educacionais das universidades e centros de pesquisa tanto no Brasil quanto no exterior não são portais semânticos. Verificou-se também que algumas instituições de ensino, a exemplo das universidades de Lehigh[4] (Pensilvânia, Estados Unidos) e Munique[5] (Alemanha), apresentam suas estruturas organizacionais através de taxonomias representadas em OWL, porém com poucas ou quase nenhuma instância. Portanto, observa-se que o uso de ontologias e ferramentas de anotação ainda é bastante insipiente.

Além disso, a informação oriunda da Web profunda não é contemplada pelos serviços de busca oferecidos pelos portais, uma vez que estes conteúdos não detectados por robôs, a exemplo de formulários HTML ou banco de dados. Diversos trabalhos têm sido realizados com o objetivo de definir mecanismos e algoritmos visando extrair informações (estruturadas ou não) relevantes da Web profunda e de sites sociais [1] [3].

2. PROPOSTA E RELEVÂNCIA DO PROJETO

O presente projeto se propõe a explorar e desenvolver um conjunto de ferramentas que permitam integrar conteúdos de portais, de modo a facilitar a navegação entre portais, e com isso contribuir para melhorar a qualidade de buscas e serviços em atendimento às demandas do cidadão, levando-se em conta a sua diversidade e diferenças culturais, regionais e sócio-econômicas.

A partir da experiência obtida em [8], é possível destacar algumas questões importantes, ainda em aberto, no cenário atual de pesquisa em portais semânticos, que o projeto em questão pretende responder:

[2] www.redegoverno.gov.br

[3] www.comp.ime.eb.br/~posedu

[4] http://www.lehigh.edu/~zhp2/2004/0401/univ-bench.owl

[5] http://wwwradig.in.tum.de/ontology/organisation

- Como integrar conteúdos em portais G2C sobre vários domínios do conhecimento, a partir de sites e portais tradicionais desprovidos de metadados ou anotações semânticas?
- Como anexar um mecanismo de busca ao portal de forma a integrar conteúdos extraídos da Web profunda?
- Como construir ontologias de domínio a partir da organização e/ou estruturação de sites convencionais?
- Os algoritmos existentes na literatura para realização de casamento (*matching*) entre ontologias são eficientes para selecionar e classificar conteúdos em portais? Em [4] é feita uma comparação entre essas ferramentas, porém em aplicações e uso distintos dos pretendidos no escopo do trabalho em questão;
- Seria viável a construção de um repositório público de metadados com a descrição e localização de ontologias em diversos domínios, para uso em portais semânticos?

A relevância da pesquisa vislumbrada por essa proposta é justificada pelos seguintes fatores:

- Considerando o volume crescente de informações sendo disponibilizadas via Web pelas diversas instituições governamentais, é imperioso que se invista em iniciativas que tenham por objetivo a gerência e integração de informações provenientes de diferentes fontes, de modo a agilizar a utilização deste vasto acervo;
- Este trabalho está em consonância com dois dos Grandes Desafios da Pesquisa em Computação no Brasil[6]: gestão de informação em grandes volumes de dados multimídia distribuídos e acesso participativo e universal do cidadão brasileiro ao conhecimento;
- As questões abordadas por esse projeto são também alvo de pesquisa de equipes do INRIA, notadamente dos projetos WAM[7], GEMO[8] e EDELWEISS[9]. O primeiro foi responsável pelo desenvolvimento da ferramenta de anotação Amaya, em colaboração com a equipe W3C. O segundo estuda problemas ligados à mediação e integração de dados heterogêneos representados em XML, linguagem padrão da Web. Já o terceiro está ligado diretamente ao estudo de problemas ligados à interoperabilidade e contextualização semântica de informações, anotações semânticas de recursos de informação e interfaces Web baseadas em ontologias. Uma parceria internacional com um desses grupos seria de grande interesse científico para esse projeto, e permitiria ampliar e testar as técnicas desenvolvidas na equipe brasileira num outro contexto social.

[6] http://www.sbc.org.br/)
[7] http://www.inria.fr/recherche/equipes/wam.fr.html
[8] http://www.inria.fr/recherche/equipes/gemo.fr.html
[9] http://www.inria.fr/recherche/equipes/edelweiss.fr.html

3. OBJETIVOS E PLANO DE TRABALHO

Este projeto visa contribuir com a gestão dinâmica de informações em portais semânticos, por meio de técnicas e recursos da Web Semântica, notadamente no uso de ontologias para fins de descoberta, extração, classificação e integração de conteúdos de outros sites. Também é objetivo estratégico deste projeto integrar as proponentes a projetos de cooperação de natureza similar ou complementar com outras instituições.

Esse projeto será desenvolvido no período de dois anos, a partir de tarefas específicas, descritas a seguir.

i) Realizar o levantamento detalhado de requisitos do projeto, e definir o seu escopo e domínio de aplicação;
ii) Estudar e definir algoritmos e ferramentas específicos que poderão ser úteis para responder às questões identificadas na seção 2. Essa etapa deverá ser acompanhada de testes que comprovem a eficácia dos mecanismos e como poderão ser aplicados ao trabalho proposto;
iii) Desenvolver mecanismos para compor um portal semântico no domínio de aplicação escolhido, de modo a prover todas as facilidades especificadas na seção 2;
iv) Testar e validar as ferramentas e técnicas desenvolvidas, através da disponiblização de um portal semântico do tipo G2C.

5. EXPERIÊNCIA ANTERIOR

As professoras proponentes têm se dedicado ativamente no desenvolvimento de atividades de pesquisa e à preparação de material didático na área de Web Semântica.

A professora Ana Maria, nos últimos 10 anos de sua atuação no IME, foi responsável pela orientação de mais de 15 dissertações de Mestrado, com resultados publicados em periódicos, conferências nacionais e internacionais, além de ter participado e coordenado projetos de pesquisa na área. Dentre esses, destacam-se as teses orientadas e trabalhos realizados na linha de portais semânticos, a saber [12], [6], [5]. Trabalhou também durante três anos junto ao Proderj (Centro de Tecnologia da Informação e Comunicação do Estado do Rio de Janeiro), onde realizou atividades de consultoria e planejamento em governo eletrônico, e atualmente é pesquisadora junto ao LNCC.

A Profª Maria Cláudia vem trabalhando em conjunto com a Profª Ana Maria há mais de 4 anos, tendo participado das atividades e orientações relacionadas ao tema proposto. Além dos trabalhos já citados, vale destacar ainda o trabalho na linha de integração de ontologias [2]. Junto a grupos de pesquisa da Fiocruz, tem participado de outras iniciativas, envolvendo o uso, mapeamento e evolução de ontologias na área de Bioinformática.

Além disso, as proponentes participaram de experiências anteriores bem sucedidas com a França, através dos projetos ECOBASE[10] e KIWI[11], envolvendo o CNPq e INRIA. O projeto Ecobase teve como foco central o desenvolvimento de tecnologias para a integração de informações em sistemas ambientais. Já o projeto KIWI aproveitou a experiência adquirida anteriormente para definir e desenvolver ferramentas para a integração de informações na Web, cuja tecnologia envolveu o uso intensivo de metadados, gerência de workflows, gerência de dados semi-estruturados, técnicas de replicação e de processamento de consultas. Estas parcerias resultaram numa troca frutífera de conhecimento, com o intercâmbio de alunos e publicações [10] [11].

Portanto, os resultados obtidos ao longo da vida acadêmica e de pesquisa das proponentes demonstram a experiência e o amadurecimento requeridos para a realização de um projeto dessa natureza.

6. RESULTADOS ESPERADOS

Estes resultados de pesquisa serão produzidos no contexto de teses de mestrado e trabalhos de final de curso, que poderão ser orientadas conjuntamente por pesquisadores brasileiros e franceses. Adicionalmente, planeja-se produzir artigos de referência ao longo da pesquisa, publicados em periódicos e conferências internacionais. Os protótipos e técnicas desenvolvidas poderão ser utilizados pela comunidade acadêmica e instituições governamentais em geral, no sentido de facilitar o desenvolvimento de aplicações de governo eletrônico. Por fim, uma contribuição adicional deste projeto é a criação de um repositório de metadados sobre os sites e portais do domínio de aplicação estudados.

REFERENCES

[1] S. Amer-Yahia, L. Lakshmanan, C. Yu. SocialScope: Enabling Information Discovery on Social Content Sites (2009). 4th Bienal Conference on innovative data systems (CIDR), Asilomar, CA, EU, jan.

[2] H. Camargo, AM Moura, M C Cavalcanti. Ontologias Emergentes: Uma Nova Abordagem para Integração de Ontologias. XXIII SBBD, Campinas, S.P, out 2008

[3] M. Cafarella (2009). Extracting and Querying a Comprehensive web Database. 4th Bienal Conference on innovative data systems (CIDR), CA, EU, jan.

[4] J. Euzenat and P. Shvaiko. Ontology Matching. Springer-Verlag, Berlin Heidelberg, 2007.

[5] Ferreira e Gama (2008). Organização Automática de Páginas Web para Exibição em Portais Semânticos. Projeto de final de curso em Eng. da Computação, IME, agosto.

[6] F. A. Lachtim. Organization and Instantiation of Content in Semantic Portals (in Portuguese). *Master thesis*, IME, Rio de Janeiro, Brazil, May 2008.

[7] F. A. Lachtim, A. M. C. Moura, M. C. Cavalcanti. An Architecture for Dynamic Organization and Publication in Semantic Portals. *10th International Conference on Information Integration and Web-Based Applications and Services (IIWAS)*, Linz, Austria, Nov. 2008.

[8] F. Lachtim, A. M. Moura, M. C. Cavalcanti (2009). "Ontology Matching for Dynamic Publication into Semantic Portals," to appear.

[9] F. A. Lachtim, G. Ferreira, R. Gama, A. M. C. Moura, M. C. Cavalcanti (2008). POSEDU: um Portal Semântico Educacional *2nd Brazilian Workshop Semantic Web and Education*, Fortaleza, Brazil, Nov.

[10] L. Bouganim, M. C. Cavalcanti A. M. Moura, et al. (2000). The Ecobase environmental information system: applications, architecture and open issues. Network and Information Systems Journal (NISJ), Hermes, ISSN 1290-2926, Vol 3, No 5.

[11] L. Bouganim, M. C. Cavalcanti A. M. Moura, et al. (2001).The Ecobase Project: Database and Web Technologies for Environmental Information Systems. ACM Sigmod Record, 30(3), ISSN: 0163-5808, set.

[12] W. A. Pinheiro, Moura A.M.C. An Ontology Based-Approach for Semantic Search in Portals. *Proc of the WEBS* 2004, Saragoza – Spain, Sept.

[13] A. Saldhana (2007). Secure E-Government Portals - Building a web of trust and convenience for global citizens. *W3C Workshop on e-Gov and the Web, National Academy of Sciences, Washington DC.*

[10] Ecobase – "Database and Web Technologies for Environmental Information Systems", CNPq/INRIA Research Project, 1998.

[11] KIWI: Key Technologies for the Integration of Web Information, CNPq/INRIA Research Project, 2000.