Ontologias no Suporte a Evolução de Conteúdos em Portais Semânticos

Débora Alvernaz Corrêa[1], Maria Cláudia Cavalcanti[1], Ana Maria de C. Moura[2]

[1]Departamento de Sistemas e Computação
Instituto Militar de Engenharia (IME) - Rio de Janeiro, RJ – Brasil
[2]Extreme Data Lab (DEXL Lab)
Laboratório Nacional de Computação Científica (LNCC) - Petrópolis – RJ – Brasil
{deboradac,anamaria.moura}@gmail.com, yoko@ime.eb.br

Abstract. *In a semantic portal, contents are described and organized based on domain ontologies. However, with the increasing amount of information generated each day on the web, dynamic publishing in these portals, whose contents are obtained from a large and diversified number of sites, still represents a major challenge, since this task lacks mechanisms to update and integrate information automatically. This paper presents an architecture that facilitates the population of a domain ontology from web sites that have a certain semantic feature. These instances may be used later in the process of a semantic portal automatic updating.*

Resumo. *Em um portal semântico, conteúdos são descritos e organizados com base em ontologias de domínio. Entretanto, com a quantidade crescente de informações geradas a cada dia na web, a publicação dinâmica nesses portais, cujos conteúdos são oriundos de um grande e diversificado número de sites, ainda representa um grande desafio, uma vez que essa tarefa carece de mecanismos para atualizar e integrar informações automaticamente. Este artigo apresenta uma arquitetura que facilita a população de uma ontologia de domínio a partir de sites que apresentam alguma característica semântica. As instâncias recuperadas destes sites podem ser utilizadas posteriormente no processo de atualização automática de um portal semântico.*

1. Introdução

A Web Semântica surgiu com a finalidade de suprir as deficiências da web atual. De acordo com [Berners-Lee et al., 2001], significa disponibilizar informações com significados adicionais, de forma a contextualizá-los e torná-los interpretáveis por máquina, permitindo que agentes e pessoas possam trabalhar em cooperação.

Neste contexto surgiram os portais semânticos que, ao contrário dos portais tradicionais, agregam valores semânticos que ajudam na classificação e organização dos seus conteúdos, facilitando os mecanismos de busca a recuperarem informações mais úteis ao usuário final, isto é, com maior precisão. Os portais semânticos utilizam-se de ontologias [Gruber, 1995] como mecanismo básico para fornecer expressividade semântica a seu conteúdo. A tendência atual é adicionar às suas funcionalidades a capacidade de realizar consultas aos conteúdos da ontologia que embasam o portal via *endpoints*, utilizando a linguagem SPARQL[1]. No entanto, para alimentar portais semânticos é preciso buscar formas automatizadas de modo a mantê-los sempre atualizados. Muitos ainda contam com mecanismos manuais, como formulários.

[1] http://www.w3.org/TR/rdf-sparql-query/

Em [Latchim, 2008], é proposta uma arquitetura para recuperar informações da web baseadas em ontologias de domínio. Com essas informações recuperadas, é possível integrar o conteúdo do portal com as informações obtidas a partir de diferentes portais semânticos. No entanto, já que boa parte da informação que se deseja recuperar encontra-se em portais tradicionais ou simplesmente na web aberta, voltamos a enfrentar os problemas e limitações já conhecidos. Cada página web tem sua estrutura própria, textos mal formatados e carentes de metadados que possam descrevê-los, não havendo uma separação nítida entre o conteúdo e a apresentação da informação, o que interfere sobremaneira na qualidade dos serviços de busca.

Assim, o trabalho aqui proposto dá continuidade ao trabalho de [Latchim, 2008], tendo como objetivo a especificação de uma arquitetura que permita coletar conteúdos a partir de outros sites e/ou portais da web aberta, considerando um domínio específico, e instanciar à ontologia já existente que serve de base para um portal semântico, desse mesmo domínio, com tais conteúdos. Dessa forma, estaremos facilitando a alimentação deste portal semântico, e ajudando-o a manter-se atualizado.

O restante desse artigo está estruturado da seguinte forma. A seção 2 descreve alguns trabalhos relacionados. Na seção 3 é descrita a arquitetura proposta para alimentar portais semânticos através de sites e/ou portais web. A seção 4 apresenta um estudo de caso no qual, conteúdos de portais no contexto educacional, servem de subsídio para atualizar e popular uma ontologia nesse mesmo domínio. E por fim, a seção 5 conclui o artigo com alguns comentários e sugestões para trabalhos futuros.

2. Trabalhos Relacionados

Algumas alternativas de solução para o problema de interoperabilidade de informações entre portais e instanciação automática de conteúdo têm sido alvo de pesquisa há alguns anos. A literatura apresenta alguns trabalhos, tais como [Lachtim, 2008], [Lachtim et al., 2009], [Suominem et al., 2009], [Yvon et al., 2009] e [Castaño, 2008].

Esses trabalhos apresentam como característica comum o uso de tecnologias da WS, porém utilizados em contextos diferentes. Lachtim [Lachtim, 2008] integra e instancia informações a partir de portais semânticos. Em [Castaño, 2008] a população da ontologia é feita através de páginas HTML de currículos, mas com o objetivo de gerar relatórios. Já em [Suo minem et al., 2009] metadados e documentos são recuperados de conteúdos publicados nos Sistemas de Gerenciamento de Conteúdos (*Content Management Systems*) ou por conteúdos anotados manualmente, através do editor de metadados SAHA [Kurki et al., 2010]. Posteriormente, estes metadados são conectados aos serviços da ontologia ONKI [Viljanen et al., 2010] para serem validados. Os metadados validados com sucesso são então publicados no portal. O portal apresentado por [Hyvönen et al., 2009] utiliza como processo de criação de conteúdos uma variedade de esquemas de metadados, ferramentas como ONKI, SAHA, POKA [Poka, 2011] e VERA [Vera, 2011], além de serviços da Web 2.0, como *Wikipedia* e *Panoramio*. Esse processo permite a produção e recuperação de conteúdos relacionados a museus, bibliotecas, arquivos e outras organizações, cidadãos individuais e de fontes nacionais e internacionais da Web 2.0.

O grande diferencial do trabalho aqui proposto em relação aos demais está na atualização de portais semânticos a partir de conteúdos de sites e/ou portais da web aberta, considerando apenas a estrutura de apresentação e a estrutura navegacional dos portais, como *links*, listas e tabelas. Dessa forma, as atualizações destes portais são feitas de forma simples, permitindo assim que estes deixem de ser simples páginas voltadas somente para usuários, e tornem-se capazes de integrar e instanciar informações a partir do uso de ontologias.

3. Arquitetura Proposta para Alimentar Portais Semânticos

No escopo desse trabalho, o termo portal com potencial semântico é designado a todo aquele que se beneficie de uma das seguintes características: (i) tenha algum tipo de organização e hierarquia em sua estrutura; e/ou (ii) parte de seu conteúdo seja apresentado na forma de uma taxonomia, isto é, bastando que algumas de suas páginas apresentem tais características.

A figura 1 apresenta a arquitetura proposta por esse trabalho. É constituída de vários componentes que, em termos gerais, contribuem para alimentar uma ontologia a partir de portais da web aberta. A estratégia adotada por esta arquitetura segue os seguintes passos: a partir de uma navegação realizada em portais web com potencial semântico (lista de portais definida previamente pelo usuário), tendo como base uma ontologia de domínio (OB), informações consideradas úteis são extraídas para enriquecer esta ontologia com novas instâncias. Lembrando que este artigo não contempla a construção de uma nova ontologia, mas sim uma extensão da mesma. A seguir é apresentada uma nova versão da ontologia inicial, aqui denominada OB', com as novas informações ali encontradas em formato de triplas RDF[2]. Essa nova versão da ontologia OB' servirá como entrada para o alinhamento com as informações do portal semântico em questão. Dessa forma, as categorias existentes no portal semântico considerado serão atualizadas, tendo como base os novos conteúdos adicionados à OB'.

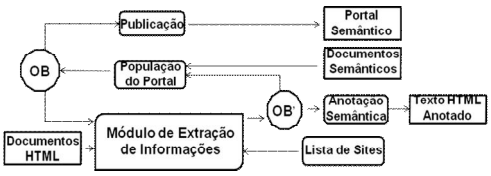

Figura 1. Arquitetura Proposta

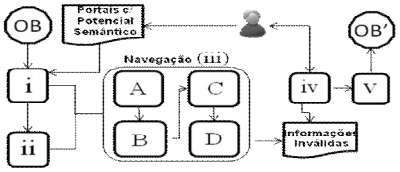

Figura 2. Módulo de Extração de Informações

O principal módulo dessa arquitetura diz respeito à Extração de Informações, que dará subsídios para a população de portais. Esse módulo, ilustrado na Figura 2, é composto por outros submódulos, cada um apresentando funcionalidades bem específicas, cujas características são descritas a seguir.

i. Recorte da OB. Esta etapa carrega uma lista de classes, instâncias e propriedades de relacionamentos existentes na Ontologia Base (OB) que servirão de base para a pesquisa de instâncias em cada página web visitada referente aos portais com potencial semântico. Os relacionamentos entre as classes são considerados para estabelecer a ordem de navegação pelas classes da ontologia. Além disso, é levado em consideração o nome real de cada classe (sempre começando pela classe mais abrangente definida pelo usuário), o *label* e as classes equivalentes para a navegação nas páginas (processo iii) dos portais. As instâncias de cada classe, bem como as instâncias equivalentes (definidas pela cláusula *same as*) às instâncias principais;

ii. Pré-categorização e identificação de página de início. Uma pré-categorização com base no título será efetuada para limitar a navegação estabelecida por (i). Caso a página inicial contenha no título um nome similar a uma instância de alguma classe da OB, a navegação se dará a partir da próxima classe a ser pesquisada. Caso contrário, a navegação se dará a partir da primeira classe. A identificação da página de início serve para definir a página a partir da qual será iniciada a navegação (processo iii). Se esta não for informada a navegação se inicia pela página principal do portal;

iii. Navegação pelas páginas. Este módulo realiza a navegação pelas páginas (de cada portal definido pelo usuário) a procura de *links*, que servirão para recuperar classes e instâncias. É

[2] RDF é uma linguagem ontológica. Triplas são declarações com a seguinte estrutura: "sujeito, predicado e objeto". Descreve a relação de um objeto a outro objeto ou literal através de um predicado. [RDF – W3C, 2011]

composto pelos subprocessos de *Recuperação de Classes, de Instâncias, de Pares de Instâncias* e *Análise de Hierarquia*, descritos a seguir.

A. *Recuperação de classes.* A navegação começa pela página de início pré-definida anteriormente pelo usuário no início do processo. O sistema deverá primeiramente ler a página a procura de *links* e *labels* que apresentem um grau de similaridade com a classe da OB procurada (definida no processo i). Estes *links* serão considerados prioritários para a navegação e deverão ser internos, i.e., do mesmo domínio de navegação. Quando um *link* que satisfizer as condições descritas anteriormente for encontrado, o sistema deverá verificar se este já foi visitado. Em caso afirmativo, irá para o próximo *link*, e em caso negativo, este deverá ser visitado e as instâncias recuperadas conforme o passo B;

B. *Recuperação de instâncias*. Para que instâncias sejam consideradas candidatas à alimentação de um portal semântico, estas deverão estar em *tags*[3] (*links*, listas e/ou tabelas). Além disso, devem ter o título com alguma similaridade (palavras semelhantes) com as instâncias existentes na OB (definida em i). Durante a navegação, as informações extraídas são recuperadas para posterior validação do usuário (processo iv), incluindo as triplas e os sites candidatos a portais com potencial semântico;

C. *Análise de hierarquia*. De modo a evitar a repetição de triplas na OB', as informações deverão estar compatíveis com a hierarquia definida pela OB. Por exemplo, uma instância de uma subclasse pode ser listada duas vezes, visto que é a mesma instância para superclasse;

D. *Recuperação de Relacionamentos*. Pares de instâncias em conformidade com os relacionamentos presentes na OB. Assim, por exemplo, na ontologia da Figura 3, as instâncias de "*Education_Program*" se relacionam com as de "*Academic_Research_Institution*" através da propriedade "*provided_By_Program*". Estes pares são então recuperados e armazenados como triplas RDF.

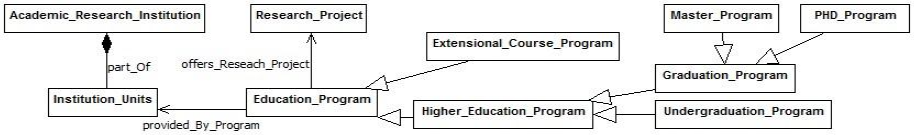

Figura 3. Recorte da OBEDU

iv. Validação. O usuário realiza a validação das informações extraídas. As que forem consideradas inválidas deverão ser armazenadas, para que posteriormente sirvam como uma pré-validação para as próximas informações recuperadas;

v. Transformação em triplas RDF. Este processo realiza a transformação das informações válidas em triplas RDF, que deverão ser adicionadas à nova ontologia, i.e., a OB'. Esta corresponde a um recorte vazio da OB (sem instâncias), que vai sendo atualizada com as novas instâncias recuperadas. Posteriormente a OB' deverá passar por um processo de alinhamento de ontologias com a OB, e suas instâncias poderão ser usadas para a população de um portal semântico cuja base seja a OB.

4. Estudo de Caso

A OB tem um papel muito importante na estratégia proposta para a alimentação de portais semânticos. Embora a arquitetura proposta tenha como foco uma aplicação genérica, essa seção apresenta um estudo de caso voltado para o domínio educacional, cuja ontologia base utilizada é

[3] *Tão* significa etiqueta e são utilizadas como breves instruções em linguagens de marcação.

a OBEDU [Lachtim et al., 2009], que serviu de ontologia base para a criação do portal semântico POSEDU[4].

A estratégia proposta na seção anterior foi adaptada para o contexto educacional. Nesta seção, foi utilizada apenas uma visão parcial da ontologia OBEDU, como mostrada na Figura 3. O recorte da OBEDU fornece classes e instâncias (recuperadas em i) para auxiliar a pesquisa nos portais com potencial semântico definidos pelo usuário. Inicialmente, o usuário realiza uma pré-categorização da página principal do portal (conforme mencionado anteriormente em ii), onde definirá a classe inicial de onde começará a navegação em busca de instâncias. As instâncias procuradas têm sempre como base as classes da ontologia. Assim, a página inicial é percorrida e seus *links* são extraídos começando pelos de maior prioridade em relação à OB (processo i). Os não prioritários deverão ser visitados posteriormente, até que as opções de extração de instâncias tenham sido esgotadas para uma determinada página. Um exemplo de navegação por um portal com potencial semântico é mostrado a seguir.

Neste exemplo, o portal do Instituto Militar de Engenharia - IME[5] foi escolhido como modelo para o estudo de caso (Figura 4). Na pré-categorização (processo ii), é verificado se o título da página principal do portal é similar a alguma instância da classe "*Academic_Research_Institution*". Com isso é possível verificar que este portal é restrito e fornece apenas informações específicas de uma única instituição (IME), o que torna desnecessária a navegação e a procura de instâncias desta classe. Assim, a navegação deverá ser iniciada pela próxima classe, ou seja, "*Institution_Units*". Para esta classe o portal IME não apresenta *links* com alguma similaridade (subprocesso A), i.e, não apresenta *links* prioritários, e por isso, os demais deverão ser visitados. Durante a navegação pelas páginas destes *links* é verificado se estas contêm tabelas, listas e outros *links*, com alguma similaridade às instâncias da classe "*Institution_Units*". A página referente ao primeiro *link* (não prioritário) visitado é ilustrada na Figura 5 (a). Nesta figura, o primeiro item com maior prioridade encontrado (Ensino de Pós-Graduação – SD/1) é comparado às instâncias da classe. Ao constatar tal similaridade, este deverá ser extraído e armazenado (subprocesso B). Esse processo é realizado para todos os itens (*links*, listas e tabelas) da página visitada.

Figura 4. Página Inicial para a Navegação

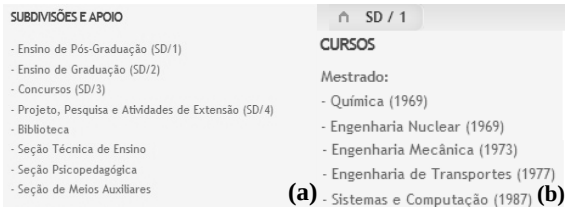

Figura 5. Páginas com Instâncias

Como a navegação pelas páginas do portal é feita de acordo com a OBEDU, após a extração das instâncias da classe "*Institution_Units*", a próxima classe a ter suas instâncias pesquisadas é a "*Education_Program"* e todas as suas subclasses. E por fim, a próxima classe é "*Research_Program"* (Figura 3), repetindo o processo para todas as classes da ontologia até que se esgotem todas as possibilidades de instâncias e o portal tenha sido totalmente visitado. Durante a navegação, a análise da hierarquia (subprocesso C) e a recuperação de pares de instâncias (subprocesso D) também deverão ser efetuadas. São analisadas primeiramente as instâncias das subclasses em relação às instâncias das suas superclasses para eliminar as instâncias repetidas. Como exemplo, podemos citar o programa de Sistemas e Computação, que por ser uma instância da Classe "*Graduation_Program"*, também é uma instância da classe "*Education_Program"*. Assim, a instância da classe mais específica é armazenada e a instância

[4] http://www.comp.ime.eb.br/~posedu
[5] http://www.ime.eb.br/

da classe mais abrangente é descartada. Na recuperação de pares de instâncias, são verificados os relacionamentos entre as instâncias de acordo com as propriedades do objeto. Como observado na Figura 5(b), pode-se dizer que Sistemas e Computação (instância da classe "*Education_Program")* é "*provided_By_Program*", SD/1, que é uma instância da classe "*Institution_Units*". Após esse processo, o usuário fará uma validação (processo iv) dessas informações obtidas e as informações válidas são transformadas em triplas RDF (processo v).

5. Conclusão

Este trabalho apresentou uma arquitetura genérica para a população de uma ontologia de domínio necessária para a atualização automática de um portal semântico. Dentre os módulos constituintes dessa arquitetura, foi dada ênfase ao módulo de Extração de Informações, componente fundamental no processo de atualização de portais. Um estudo de caso no domínio de educação permitiu ilustrar como conteúdos de um portal acadêmico na web aberta podem ser recuperados e integrados a uma ontologia básica de domínio, de modo a prover subsídios para posteriormente popular um portal semântico. Como etapa adicional desse trabalho, pretendemos especificar algumas métricas para validar os resultados obtidos, bem como realizar uma avaliação de usabilidade da ferramenta para identificar possíveis melhorias.

6. Referências

Berners-Lee, T. I.; Hendler, J.; Lassila, O. R. (2001). "The Semantic Web". Scientific American Magazine.

Castaño, A. C. (2008) "Populando ontologias através de informações em HTML - O caso do Currículo Lattes", Dissertação de Mestrado. Universidade de São Paulo. São Paulo, SP.

Kurki, J.; Hyvönen, E. (2010) "Collaborative Metadata Editor Integrated with Ontology Services and Faceted Portals.", Heraklion, Grécia.

Lachtim, F.A.; Cavalcanti, M.C.; Moura, A.M. (2009) "Ontology Matching for Dynamic Publication into Semantic Portals", Journal of the Brazilian Computer Society, ISSN: 0104-6500, vol 15. págs 27- 43, Março.

Lachtim, F.A; Ferreira, G.; Gama, R.; Moura, A.M.; (2009) Cavalcanti, M.C. "POSEDU: a Semantic Educational Portal", IEEE Multidisciplinary Engineering Education Magazine, Vol. 4, Nº 3.

Lachtim, F.A. (2008) "Organização e Instanciação Automática de Conteúdos em Portais Semânticos", Dissertação de Mestrado. Instituto Militar de Engenharia. Rio de Janeiro, RJ.

Hyvönen, E.; Mäkelä, E.; Kauppinen, T; Alm, O.; Kurki, J.; Ruotsalo, T.; Seppälä, K.; Takala, J.; Puputti, K.; Kuittinen, H.; Viljanen, K.; Tuominen, J.; Palonen, T.; Frosterus, M.; Sinkkilä, R.; Paakkarinen, P.; Laitio, J.; Nyberg, K. (2009) "CultureSampo - Finnish Culture on the Semantic Web 2.0. Thematic Perspectives for the End-user." Indianapolis, USA.

Poka (2011) "A framework for automatic annotation", http://www.seco.tkk.fi/tools/poka/, Abril.

RDF (2011) "Resource Description Framework (RDF): Concepts and Abstract Syntax", http://www.w3.org/TR/rdf-concepts/, Abril.

Suominem, O.; Hyvönen, E.; Viljanen, K.; Hukka, E. (2009) "HealthFinland - a National Semantic Publishing Network and Portal for Health Information". Finlândia.

Vera (2011) "Validation and quality assistant for Semantic Web data", http://www.seco.tkk.fi/services/vera/, Abril.

Viljanen, K.; Tuominen, J.; Hyvönen, E. (2010) "A Network of Ontology Repositories". Submitted for review.