Ontologias e Heterogeneidade de Bases de Dados no Domínio da Geoinformação

Neide dos Santos, Oscar Luiz Monteiro de Farias
Programa de Pós-graduação em Computação - ênfase Geomática
Universidade do Estado do Rio de Janeiro

Resumo: A principal motivação da pesquisa em ontologia no Programa de Pós-Graduação em Engenharia de Computação - UERJ é a busca de soluções para a integração de bancos de dados heterogêneos, baseada no uso de ontologias de domínio, que representam o conhecimento no domínio das informações dos BD. Neste texto, descrevemos os antecedentes da pesquisa atual, os trabalhos já realizados e os em andamento e descrevemos as expectativas futuras. Em ontologias, temos duas dissertações defendidas, duas em andamento e uma tese de doutorado, na COPPE/Civil, em fase de conclusão, além de contarmos com artigos em congressos.
Palavras-chave: Ontologia de domínio. Interoperabilidade. Bases de dados heterogêneas.

Ontology and Heterogeneity of Data bases in Geoinformation Domain

Abstract: At the Program of Computer Engineering – UERJ, the main motivation of our research in ontology is finding solutions to the integration of heterogeneous database systems, based on the use of domain ontology, which represent the knowledge in the field of the databases information. In this paper, we describe the background of the current research, the work already done and in progress. We also describe our expectations for the future. In ontology, we supervised two master' dissertations, have two other dissertations and a doctoral thesis (at COPPE / Civil) in progress, besides some papers in national and international conferences.
Keywords: Domain ontology. Interoperability. Heterogenous dada bases.

1. Introdução

O Programa de Pós-Graduação em Engenharia de Computação (PGEC), concentração em Geomática da Universidade do Estado do Rio de Janeiro, é um programa interdisciplinar, resultante da fusão das áreas do conhecimento de Engenharia de Sistemas e Computação e de Engenharia Cartográfica, estendendo-se às demais áreas do conhecimento das Geociências.

Desde 2004, trabalhamos com ontologias, no PGEC, oferecendo a disciplina FEN06754: Ontologias e Web Semântica (www.ime.uerj.br/~neide) e desenvolvendo a linha de pesquisa Ontologias e a Integração de Bancos de Dados Heterogêneos, visando propor soluções para problemas de interoperabilidade. A pesquisa na área é motivada pela crescente heterogeneidade em sistemas computacionais, dificultando a interoperabilidade. Há diferentes níveis de heterogeneidade entre bases de dados, gerando conflitos computacionais, sintáticos, estruturais e semânticos (Sheth, 1999). Os conflitos semânticos relacionam-se ao significado dos termos usados nas bases de dados e são os de mais difícil solução. Neste sentido, a modelagem dos termos empregados nas bases de dados em um modelo de ontologias pode ser útil (Cui, Jones & O'Brien, 2002; Fonseca, Engenhofer, Agouris & Câmara, 2002). Nossa pesquisa caminha basicamente nesta direção.

2. Trabalhos Concluídos e Trabalhos em Andamento

Descrevemos a seguir, sucintamente, a) a origem de nossas pesquisas; b) dois trabalhos concluídos na área de ontologias e interoperabilidade; e três trabalhos andamento.

a) Antecedentes de nossa pesquisa atual: em 2002, iniciamos pesquisa conjunta entre a UERJ e o Departamento de Ciência da Computação - Universidade Federal de Juiz de Fora. Em trabalhos anteriores (Santos, Campos e Braga, 2004, 2005, 2009; Santos et al, 2005; Silva et al, 2005), descrevemos o produto desta pesquisa: o desenvolvimento de uma ontologia para a área de sistemas de educação a distância (EaD), para construir um repositório de componentes de software para a área e uma biblioteca virtual também para esta área. Usamos o conceito de modelo de *features*, que especifica os termos do domínio, formando uma rede semântica de termos. O modelo engloba termos e funcionalidades usados pela comunidade e permite a reutilização e o compartilhamento de um vocabulário comum entre desenvolvedores e usuários de software, permitindo uma classificação mais precisa dos tipos de software relacionados ao domínio e de componentes de software a serem reutilizados em novas aplicações.

O modelo de features é subdividido em dois sub-modelos: conceitos e funcionalidades. Na área de EaD, o modelo de conceitos foi denominado de modelo de categorias e definido dois sub-modelos: Categorias, onde se encontram os principais conceitos relacionados ao domínio e Funcionalidades, onde estão as funcionalidades gerais do domínio. Os conceitos mais específicos encontram-se em uma ordem decrescente, em Setores e Sub-Setores, e as Funcionalidades, em Sub-Funcionalidade e em Funcionalidade Específica**.** Cada um dos termos do domínio foi descrito e inter-relacionado. Uma das funções das ontologias é ajudar a imprimir semântica nas buscas na Web. Para usarmos esta ontologia na construção da máquina de busca da nossa biblioteca virtual, construímos um pequeno conjunto de regras para apoiar o processo de descoberta de novas relações dentro da mesma ontologia ou dentro de ontologias relacionadas: Sinônimo (A, B) ➞ Sinônimo (B, A): Se o termo A tem termo B como sinônimo então termo B tem termo A como sinônimo; Hipônimo (A, B) ^ Hipônimo (B, C) ➞ Hipônimo (A, C): Se termo A tem termo B como hipônimo (termo mais especializado) B como termo C como hipônimo então termo A tem termo C como hipônimo. Este linha de trabalho está encerrada, mas ela foi um grande aprendizado para os pesquisadores envolvidos e serviu de embrião para os trabalhos seguintes.

b) Trabalhos concluídos em Ontologias e Interoperabilidade: como mencionado, a principal motivação para nossa pesquisa é a busca de solução para a integração de bancos de dados heterogêneos, baseada no uso de ontologias, que descrevem e representam o conhecimento no domínio das informações dos bancos de dados heterogêneos. Nesta linha, orientamos duas dissertações de mestrado. No primeiro trabalho (Aparício, 2005; Aparício, Santos e Farias, 2005; 2006), desenvolvemos uma solução conjugando a modelagem de ontologias do domínio e sua utilização apoiada por um suporte de software. A solução propõe a especificação de uma ontologia no domínio específico, a ser compartilhada por vários sistemas de banco de dados construídos *a posteriori* da especificação da ontologia. A interoperabilidade entre os diversos sistemas é alcançada através de um esquema global, desenvolvido como uma camada de software entre os diversos sistemas em questão. A solução foi testada construindo-se uma ontologia para a classificação dos solos brasileiros (ClassSolos), organizados por suas características morfológicas. A ontologia foi modelada com Protege, a partir dos seis principais conceitos de solos: Morfologia, Perfil, Atributos Diagnósticos, Horizontes Diagnósticos Superficiais, Horizontes Diagnósticos Sub-superficiais e Classificação.

Para testar nossa solução para o problema de heterogeneidade semântica, foi elaborado um estudo de caso, a partir de consultas hipotéticas a uma base de dados distribuída e heterogênea – situada em três diferentes sistemas gerenciadores de banco de dados (*Oracle*, *SQL Server* e *Access*) que poderiam também estar localizados em

diferentes computadores, com dados sobre solos, para identificar os tipos de solos mais apropriados ao cultivo de um determinado cítrico. Criamos cinco tabelas dispersas pelos três SGBD´s acima referidos: Solo, Textura, Profundidade, Drenagem, Relevo. Usamos *ClassSolos* e desenvolvemos um protótipo de sistema para dar suporte às consultas, de acordo com o conceito de esquema global, conceito este bem descrito na literatura como uma solução eficiente para apoiar a solução de conflitos sintáticos. Os termos e relacionamentos presentes em *ClassSolos* refletiram-se nas tabelas pertencentes aos diferentes bancos de dados, segundo os termos ontológicos definidos. Foram levantadas, então, as informações técnicas para o plantio de cítricos, usando dados disponíveis da EMBRAPA. Ou seja, as características dos solos desejáveis para o plantio de cítricos são: textura arenosa, média (mais adequada) ou argilosa, com relevo plano, suave ondulado ou ondulado, de pouco profundo a profundo e fortemente drenado ou bem drenado. Estas informações foram usadas para modelar os campos do banco de dados do software desenvolvido. Formulada a consulta, a resposta obtida foi que Latossolos é o solo mais apropriado para o cultivo de cítricos. O trabalho completo encontra-se em http://www.geomatica.eng.uerj.br/dissertacoes, item 1.

Em Manhães (2009); Manhães, Santos e Farias (2009, 2006), apresentamos uma proposta de modelagem de elementos de zoneamento urbano, baseada em uma ontologia de domínio. A ontologia é representada por classes e atributos do domínio e, na proposta, as classes são exportadas para o formato XMI, compondo um repositório de classes e permitindo reutilização. Como exemplo, foi construída uma ontologia do Zoneamento Urbano do município de Macaé-RJ, a partir do Plano Diretor Municipal, usando Protégé. Como a motivação do trabalho era a interoperabilidade entre sistemas de informação geográfica (SIG) de um mesmo domínio, a ontologia foi exportada para o formato XMI, sendo a seguir criado um diagrama de classes, obtido através da importação das classes (figura 1).

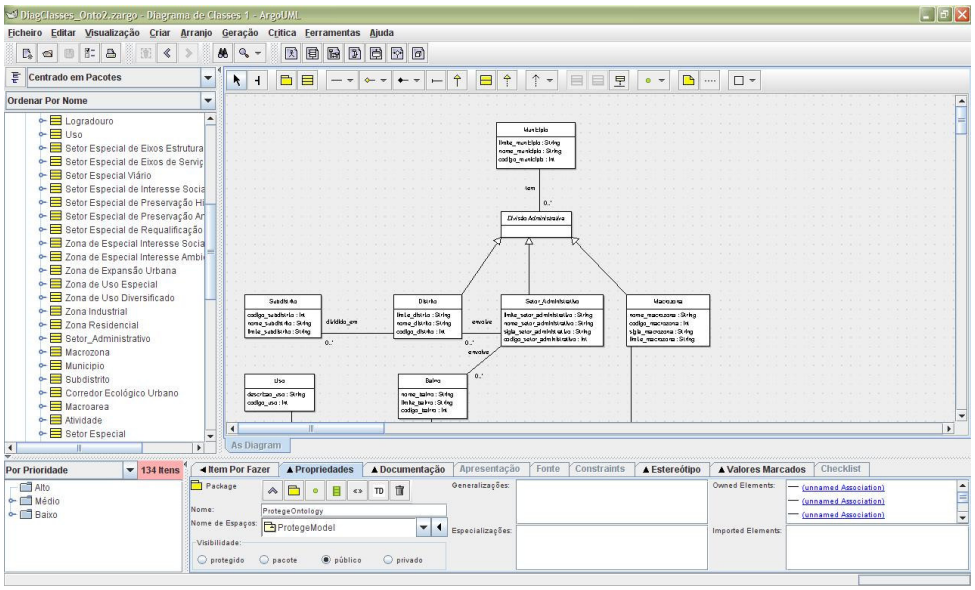

Figura 1 – Classes resultantes do arquivo XMI migradas para o ArgoUML.

Esta importação permite que a ontologia fique disponível como um pacote de classes, que pode ser utilizado por aplicações de SIG análogos, baseadas em OO. Como forma de mostrar a utilização das classes, foi desenvolvido um protótipo que oferece a visualização destas classes, na Web, como mapas temáticos.

c) Trabalhos futuros: duas novas dissertações de mestrado estão em fase inicial de desenvolvimento, a partir de 2009/2. Uma retoma o trabalho de Aparício (2005), no que tange ao estudo e promoção da interoperabilidade entre bancos de dados heterogêneos e não relacionais com o apoio de ontologia. O estudo de caso irá estudar a integração das bases de dados do Instituto Estadual do Ambiente (INEA), órgão instalado em 2009, que unificou a ação dos três órgãos ambientais estaduais: a Fundação Estadual de Engenharia e Meio Ambiente, a Superintendência Estadual de Rios e Lagoas e o Instituto Estadual de Florestas. A outra dissertação reúne os conceitos de Web semântica, ontologias e agentes de software para tratar de aspecto de dados geo-referenciados a ser escolhido.

Outra frente de trabalho consolida a parceira UERJ-UFJF e busca utilizar a arquitetura orientada a serviços, que permite o armazenamento, a pesquisa, o reuso, a composição e a execução de modelos, descrita em Matos, Campos, Braga e Weber (2008). Um dos modelos a ser trabalhado e testado na arquitetura é a ontologia de zoneamento urbano descrita em Manhães (2009).

3. Contribuições esperadas

A pesquisa do PGEC visa aumentar o conhecimento na sua área de atuação e suprir a demanda de projetos por empresas e instituições de pesquisa, e prefeituras e órgãos governamentais que atuam no planejamento, manejo e gerenciamento de recursos naturais. Os trabalhos no tópico de pesquisa Ontologia consideram, sempre que possível, a ligação universidade-sociedade. Exemplos deste ponto são o trabalho desenvolvido para Macaé e o trabalho em desenvolvimento para o INEA.

A linha de pesquisa em ontologias está bem estabelecida no PGEC desde 2004. A oferta da disciplina Web Semântica e Ontologia tem ocorrido de forma regular desde 2004, tendo sempre em torno de 5 alunos matriculados/ano. Uma das alunas com mestrado terminado, Adriana Aparício, está concluindo seu doutorado da COPPE/Civil em geração semi-automática de ontologias utilizando mineração de texto, sob a orientação de Alexandre Evsukoff (COPPE/Civil) e Neide Santos (PGEC/UERJ). O trabalho visa o desenvolvimento de um módulo, que irá permitir a geração de ontologias a partir da mineração de textos. A mineração de texto irá apoiar a fase de classificação de documentos e de descoberta de padrões em textos, visando trabalhar com a similaridade entre termos, preocupando-se com a semântica dos documentos analisados. Os problemas semânticos serão resolvidos pela aplicação de ontologias. Para a validação do módulo, os documentos trabalhados estão restritos ao banco de teses da COPPE, relacionadas à mineração de texto.

A expectativa para 2009/2001 é captar mais dois alunos de mestrado, um para desenvolver trabalho em mapeamento de ontologias desenvolvidas para domínios análogos e outro para estender o trabalho de A. Aparício em ontologia e mineração de texto/conhecimento em domínio especifico da geomática.

4. Experiência e perfil do grupo

O grupo de pesquisa em Ontologia, no PGEC, é composto de dois docentes: Neide dos Santos e Oscar L. M. Farias e conta em 2009, com a colaboração do Prof. João Araújo, também do PGEC, no que se refere especificamente a Web semântica e agentes inteligentes. Conta ainda com dois alunos de mestrado: Marco Antonio Pessoa e Marcelo Rodrigues, que estão desenvolvendo dissertações na área, sob a supervisão destes professores. Os professores e alunos envolvidos têm formação acadêmica e experiência profissional em ciência da computação ou em áreas afins.

Referências Bibliográficas

APARÍCIO, A S, FARIAS, O L M, SANTOS, N. Applying Ontologies in the Integration of Heterogeneous Relational Databases In: Australasian Ontology Workshop - AOW 2005, 2005, Sydney. *Advances in Ontologies 2005 - Proceedings of Australasian Ontology Workshop* (AOW 2005). Australian Computer Society, 2005. v.58. p.11 – 16.

APARÍCIO, A S. Ontologias, Uma Alternativa para Integração de Bases de Dados Heterogêneas. Programa de Pós-Graduação em Engenharia de Computação. Faculdade de Engenharia - UERJ. Fevereiro 2005.

CUI, Z.; JONES, D. & O'BRIEN, P. Semantic B2B integration: Issues in ontology-based applications. *SIGMOD Record* 31(1): 43-48, 2002.

FONSECA, F., ENGENHOFER, M., AGOURIS, P & CÂMARA, G. Using ontologies for integrated geographic information systems. *Transactions in GIS.* 6(3): 231-257, 2002.

MANHÃES, A.SANTOS, N, FARIAS, O L. Ontologias Aplicadas ao Desenvolvimento de SIGs:Estudo de Caso sobre Zoneamento Municipal In: *7o. Congresso Brasileiro de Cadastro Técnico Multifinalitário*, 2006, Florianópolis. v.1.

MANHÃES, A L P, SANTOS, N, FARIAS, O L M. Sistema de Informações Geográficas: Uma Abordagem Sobre Interoperabilidade Apoiada em Ontologias e XMI. In *II Congresso Tecnológico InfoBrasil.* Fortaleza, 2009.

MANHÃES, A L P. Uso de Ontologias e XMI como Instrumento de Modelagem para Zoneamento Urbano. Dissertação de mestrado. Programa de Pós-Graduação em Engenharia de Computação. Faculdade de Engenharia - UERJ. Abril 2009.

MATOS, E. ; CAMPOS, F. C. A. ; BRAGA, R. M. ; WEBER, R. Composição de modelos para a fisiologia: uma proposta de infra-estrutura de e-science baseada em ontologias. In: Semish Seminário Integrado de Software e Hardware, 2008, Belém. Anais do Congresso da SBC. Belém : SBC, 2008. v. 1. p. 46.

SHETH, A. P. Changing focus on interoperability in information systems. In: GOODCHILD, M.; EGENHOFER, M; FEGEAS, R. AND KOTTMAN, C. ed. *Interoperating geographic information systems*, Norwell, MA, Kluwer Academic Publishers, 1999. p. 165-180.

SANTOS, N., APARICIO, A S, FARIAS, O L. Integration of Heterogeneous Databases and Ontologies. *Cadernos do IME. Série Informática*, v.22, p.7 - 13, 2006.

SANTOS, N, CAMPOS, F C A, BRAGA, R M Digital libraries and Ontology. Encyclopedia on Digital Library Technologies and Applications. Yin-Leng Theng, Schubert Foo Dion Hoe-Lian Goh, Jin-Cheon Na (Eds). ed.Hershey, USA : Information Science Reference, 2009.

SANTOS, N, CAMPOS, F C A, BRAGA, R M . Desenvolvimento de uma Ferramenta para Avaliação de Alunos On-line baseada em Componentes de Software In: Taller Internacional de Software Educativo, 2004, Santiago.2004. v.1. p.127 - 138

SANTOS, N., CAMPOS, F C A, BRAGA, R M A Virtual Library for Lifelong Education on E-learning Domain In: World Conference on E-Learning in Corporate, Government, Healthcare, & High Education, 2005, Vancouver. v.4. p.3121 – 3128.

SANTOS, N., CAMPOS, F C, BRAGA, R M, OLIVEIRA, A, CIRILO, E.; SENADOR, T. An Ontology-Based Digital Library on the e-Learning Domain In: XVI Simpósio Brasileiro de Informática na Educação, 2005, Juiz de Fora., 2005. v.2. p.580 – 590.

SILVA, R T, CIRILLO, e J R, SIQUEIRA, T S, CAMPOS, F C, BRAGA, R M, SANTOS, N Web Semântica: Usando Ontologias para Buscas no Domínio da Educação Mediada pela Web. Principia (Rio de Janeiro). , v.10, p.46 - 56, 2005.