Enriquecendo Data Warehouses Espaciais com Descrições Semânticas*

Renato Deggau[1,2]

Renato Fileto[1] **(orientador)**

[1]Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Universidade Federal de Santa Catarina, Caixa Postal 476, 88.040-900, Florianópolis-SC

[2]Epagri – Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina
Rod. Admar Gonzaga 1347, Itacorubi, 88.034-901, Florianópolis-SC

{rdeggau,fileto}@inf.ufsc.br
Nível: Mestrado

Ingresso: Março de 2008

Previsão de conclusão: Julho de 2010

Etapas já concluídas: Pesquisa bibliográfica e seminário de andamento.

Resumo: Data warehouses e sistemas de informação geográfica estão entre as tecnologias mais usadas em sistemas de suporte à decisão. A conjugação dessas tecnologias em um data warehouse espacial (DWE) possibilita a análise de grandes volumes de dados integrados e orientados a assuntos, levando em consideração aspectos geográficos e permitindo a visualização de resultados em tabelas, gráficos e mapas. Todavia, o desenvolvimento de um DWE envolve estreita colaboração de especialistas em tecnologia da informação com especialistas do domínio da aplicação, para se expressar precisamente e atender adequadamente os requisitos de análise, usando as possibilidades oferecidas pela tecnologia. Este trabalho propõe o uso de ontologias para descrever semanticamente a estrutura, o conteúdo e as possibilidades de análise de informação de um DWE, visando a geração automatizada de data marts espaciais (DMEs) para atender necessidades específicas de análise espaço-dimensional. O foco desta proposta é a definição do modelo conceitual para a descrição semântica de DWEs e DMEs. Ele inclui uma ontologia de domínio, que pode ser trocada para permitir a adaptação a diferentes áreas de aplicação, além de uma ontologia de estruturas, tipos de dados e operações de análise de dados dimensionais e espaciais. A validação do modelo proposto será realizada sobre um DWE e uma ontologia do domínio agrícola, especialmente desenvolvidos para servir como estudo de caso.

Palavras-chave: data warehouses espaciais, OLAP, dados e operadores geográficos, ontologias, descrição e geração de data marts.

* Este trabalho foi parcialmente financiado pela Fapesc (contrato 12552-2007-0) e pelo CNPq (contrato 48139212007-6).

1. Introdução e Motivação

Suporte à decisão frequentemente envolve a integração e a análise de grandes volumes de dados. A análise de informação em domínios como planejamento urbano, manejo de recursos naturais e agricultura requer também o tratamento de aspectos geográficos. Um data warehouse espacial (DWE) visa atender esse tipo de demanda, integrando o modelo dimensional e processamento analítico on-line (OLAP) com recursos para representação e manipulação de objetos espaciais e sistemas de informação geográfica (SIG).

A identificação das medidas e das operações de manipulação de dados a utilizar em um DWE fica a cargo do seu projetista, mas depende das necessidades dos usuários. A comunicação entre usuários e projetistas no levantamento de requisitos de análise para compor, por exemplo, um data mart espacial (DME) é um desafio. De um lado, os potenciais usuários têm dificuldade para entender e usar os recursos oferecidos pela tecnologia da informação (TI). De outro, o desenvolvimento de um DME exige a coleta de conhecimento do domínio pela equipe de TI, que acaba codificando tal conhecimento nas soluções de análise desenvolvidas, as quais precisam ser refeitas sempre que aparecem novos requisitos.

Propostas recentes propõem a adoção de técnicas da Web semântica para descrever a estrutura e o conteúdo de um data warehouse convencional (DW), de modo a apoiar processos de extração, transformação e carga de dados (ETC) e a geração de data marts [Skoutas and Simitsis, 2006; Sell *et al*., 2008; Xie *et al.*, 2008]. Explicitar a semântica permite aos usuários especialistas do domínio organizar o seu conhecimento e expressar os seus requisitos de análise, e aos usuários de TI focarem no mapeamento desses requisitos em aspectos técnicos.

Este trabalho propõe o uso de ontologias para descrever DWEs e necessidades específicas de análise de informação espaço-dimensional sobre as mesmas. Embora o objetivo final seja a geração automatizada de DMEs, esta proposta foca no modelo conceitual para descrição de DWEs e DMEs. Tal modelo inclui uma ontologia de domínio, configurável para propiciar adaptabilidade a diferentes áreas de aplicação, e uma ontologia de estruturas, tipos de dados e operações de análise de dados dimensionais e espaciais. Espera-se com este trabalho definir um modelo e uma arquitetura de sistema baseado em ontologias e descrições semânticas para facilitar o processo de criação de DMEs, propiciando ao usuário do domínio autonomia e agilidade para definir suas análises de informação.

O restante deste artigo tem a seguinte organização. A seção 2 apresenta fundamentos e discute trabalhos relacionados a DWEs e ao uso de semântica em DWs. A seção 3 apresenta a arquitetura do sistema proposto e a abordagem a ser adotada na concepção e validação do modelo para descrição semântica de DWEs e DMEs. A seção 4 apresenta alguns exemplos de DMEs a serem considerados no estudo de caso para o setor agrícola. Finalmente, a seção 5 fecha o artigo, com algumas conclusões e um breve relato sobre a situação atual do trabalho.

2. Fundamentos e Trabalhos Relacionados

2.1. Data Warehouses Espaciais

Data Warehouses Espaciais (DWEs) estendem Data Warehouses (DWs) tradicionais com suporte ao tratamento de objetos geográficos [Fidalgo, 2005]. Um DWE combina dados e operadores do modelo dimensional com representações geométricas e operações para manipulação de dados geográficos [Rao et al., 2003; Malinowsky and Zimányi, 2007]. Além da possibilidade de filtrar e agrupar informações com operadores OLAP tradicionais e

funções de agregação de dados escalares, um DWE permite a aplicação de operadores geográficos e funções de agregação de dados espaciais nas análises de dados.

Objetos geográficos podem aparecer nas dimensões do esquema de um DWE (e.g., polígonos representado estados e cidades na dimensão espaço) ou como medidas na tabela fato (e.g., pontos onde ocorrem intoxicações por agrotóxicos). A representação de objetos espaciais permite a utilização de operadores geográficos em análises e a apresentação de resultados em mapas (e.g., associando cores às cidades para indicar sua produtividade agrícola ou o número de intoxicações por agrotóxicos). Como diferentes dimensões e medidas podem referenciar os mesmos objetos geográficos, as representações de entidades geográficas complexas (e.g., perímetro de uma cidade) devem ser compartilhadas, evitando redundâncias.

Os problemas em aberto em DWEs incluem: (i) integração do modelo dimensional com o modelo espacial de dados, considerando aspectos de modelagem, operadores e a implementação de sistemas integrados; (ii) extração, transformação e carga de dados espaciais e (iii) apoio ao processo de geração de data marts espaciais (DMEs).

Data Marts Espaciais (DME) são componentes de um DWE voltados a atender necessidades específicas de análise de dados. A especificação de um DME deve incluir um ou mais cubos de análise de dados, cada qual definido por uma tabela fato e dimensões, juntamente com as possibilidades de análise dos dados, usando operadores e funções de agregação de dados escalares e espaciais. Vários DMEs podem ser criados sobre um DWE, compartilhando dimensões e medidas, inclusive aquelas que referenciam objetos geográficos.

A classificação e a descrição de membros de dimensões, medidas, operadores e funções de agregação de dados disponíveis em um DWE podem auxiliar os usuários a melhor entender os recursos disponíveis e utilizá-los em suas análises. Este trabalho pretende adaptar e usar a classificação proposta por [Silva, 2008], que por sua vez é baseada em classificações de operadores geográficos de [Egenhofer, 1992] e [Rigaux, 2002].

2.2. Semântica em Data Warehouses

Estudos sobre a aplicação das idéias da Web Semântica em DWs convencionais incluem [Skoutas and Simitsis, 2006], [Sell *et al*., 2008] e [Xie *et al*., 2008], entre outros. [Skoutas and Simitsis, 2006] usam ontologias para mapear atributos das fontes de dados a tabelas do DW no processo de ETC. [Sell *et al*., 2008] propõem uma arquitetura baseada em Web semântica para ferramentas de análise dimensional, usando ontologias para integrar a semântica do negócio, dos dados e dos serviços oferecidos, com o objetivo de apresentar as informações de acordo com o vocabulário do usuário.

[Xie *et al*., 2008] usam uma extensão de OWL para representar metadados do negócio, de descrições de DWs e data marts (DMs). Tal proposta permite aos analistas de negócio expressar suas necessidades de análise de informação através da especificação de um modelo de análise sobre o modelo conceitual do negócio. Especialistas em TI constroem mapeamentos dos termos utilizados no modelo de análise para o esquema do DW. Utilizando tais mapeamentos a ferramenta gera DMs automaticamente, segundo padrões da indústria.

A vantagem do processo de geração de DMs proposto por [Xie *et al*., 2008] é a separação dos conceitos do domínio de aplicação daqueles de TI. Especialistas do domínio de aplicação podem organizar seu conhecimento e expressar seus requisitos de análise usando termos que lhe são familiares. Usuários de TI, por outro lado, têm seu foco na resolução dos mapeamentos dos termos específicos do domínio de aplicação para a visão técnica.

3. Proposta de trabalho

Esta proposta de trabalho baseia-se principalmente em [Xie *et al.*, 2008], acrescentando o componente espacial. O objetivo geral é definir um modelo baseado em ontologias para descrever a estrutura, operações e conteúdo de DWEs, para permitir aos usuários melhor entendimento dos recursos disponíveis e expressar suas necessidades de análise específicas, de modo a facilitar a geração de DMEs para atender essas necessidades.

A Figura 1 apresenta a arquitetura proposta. O modelo conceitual inclui duas ontologias: de DWEs e do domínio de aplicação. A primeira é fixa (independente de domínio de aplicação) e descreve os conceitos a serem utilizados na descrição de qualquer DWE, incluindo medidas, dimensões e operações de manipulação de dados escalares e espaciais. A ontologia de domínio, por sua vez, pode ser trocada para permitir a adaptação do sistema a diferentes domínios. Ela define conceitos específicos do domínio a serem usados para descrever conteúdo e análises de informação sobre um DWE.

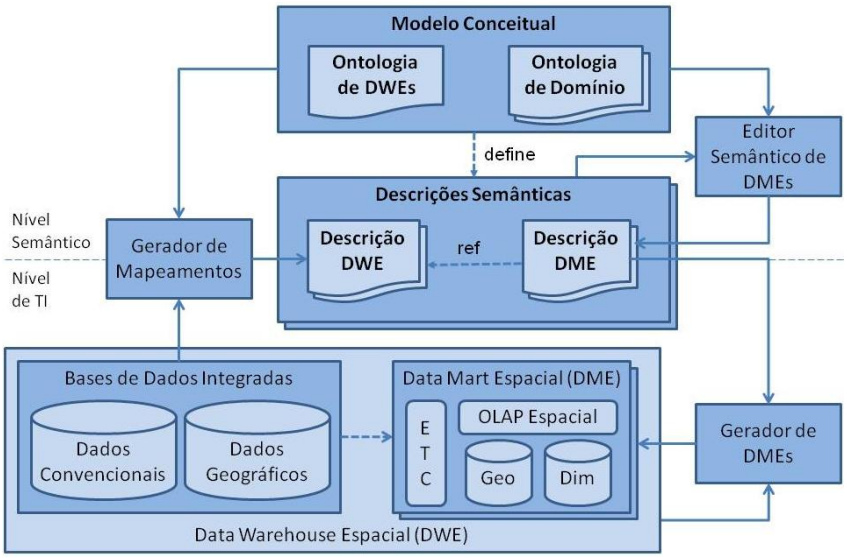

Figura 1. Arquitetura do Sistema

O Gerador de Mapeamentos permite que especialistas em tecnologia da informação e do domínio descrevam semanticamente porções de um DWE usando as ontologias do modelo conceitual. As descrições semânticas incluem mapeamentos entre estruturas e operações do DWE e conceitos específicos do domínio. O usuário especialista de domínio pode, então, utilizar o Editor Semântico de DMEs para descrever necessidades específicas de análise de informação usando conhecimento do domínio e os mapeamentos previamente estabelecidos para a DWE. As descrições de DMEs obtidas realimentam a base de conhecimento.

À medida que a base de conhecimento cresce fica mais fácil especificar novos DMEs, a partir de descrições semânticas de porções da DWE e descrições de DMEs acumuladas. A partir da descrição semântica formal de um DME podem ser geradas especificações de DMEs incluindo: (i) o esquema das bases de dados dimensionais e geográficos; (ii) análises que podem ser efetuadas com operadores e funções de agregação dimensional e espacial e (iii) definições dos processos de extração, transformação e carga dos dados do DWE para o DME.

Esta proposta de trabalho foca na definição e validação do modelo conceitual, visando atingir os seguintes objetivos específicos: (i) desenvolver uma ontologia para a descrição de DWEs; (ii) definir uma ontologia do domínio agrícola para um estudo de caso; (iii) validar as

ontologias considerando que elas serão usadas em mecanismos para a especificação semântica e a geração automatizada de DMEs em trabalhos futuros. A metodologia de trabalho inclui revisão bibliográfica, edição de ontologias sobre o Protégé[1], desenvolvimento de um protótipo de DWE para o setor agrícola sobre o PentahoBI[2] e a descrição de DMEs usando essas ontologias para validação dos resultados.

4. Estudo de caso: DMEs para o setor agrícola

A proposta será validada em um DWE para a área agrícola, criado para este fim. A figura 2 apresenta um extrato do esquema deste DWE, que contempla produção agrícola, dados populacionais e características das propriedades agrícolas de Santa Catarina.

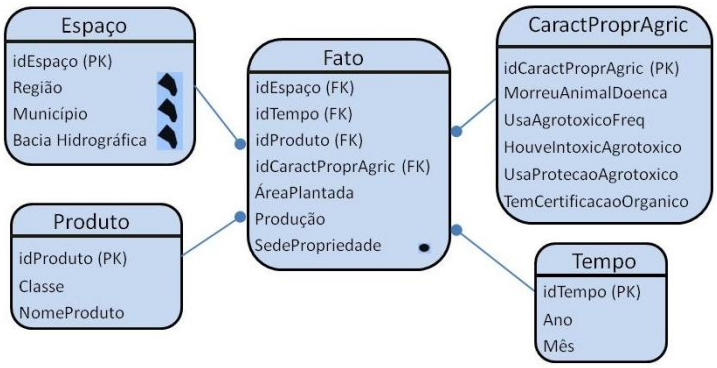

Figura 2. Visão do DWE agrícola

Os exemplos a seguir ilustram algumas necessidades específicas de análise de informação que podem ser contempladas por DMEs semanticamente especificados.

> ***Exemplo 1**: "Dado um produto agrícola, indicar a sua produtividade média em certos municípios e certo período de tempo, apresentando os resultados em um mapa para a fácil visualização dos locais onde a produtividade é maior/menor."*

Um DME para atender esta necessidade deve ter em sua tabela fato a medida produtividade, cuja definição (relação entre a quantidade produzida e a área plantada) pode ser expressa usando na ontologia de domínio. As representações dos perímetros dos estados e municípios precisam ser armazenadas e referenciadas na dimensão espaço para permitir a visualização da medida em mapas. As análises neste exemplo requerem apenas operadores OLAP convencionais e funções de agregação de dados escalares, gerando totais numéricos que podem ser mapeados a cores dos municípios nos mapas gerados.

> ***Exemplo 2**: "Indicar em um mapa as sedes das propriedades agrícolas de certo município que em certo período de tempo fizeram uso frequente de agrotóxicos, sem usar equipamentos de proteção, registraram casos de intoxicação por agrotóxicos e se encontrem a menos de 500 metros de corpos de água de certa bacia hidrográfica."*

Esta análise usa operadores OLAP e um operador espacial para selecionar pontos geográficos que se encontram a menos de 500 metros de corpos de água e devem ser exibidos no mapa. Um DME para atender esta necessidade deve referenciar em sua tabela fato pontos (ou polígonos) representando sedes de propriedades agrícolas. A descrição e classificação de operadores espaciais na ontologia de DWEs permite a sua seleção para compor o DME.

[1] http://protege.stanford.edu/

[2] http://www.pentaho.com/

5. Conclusões

O modelo para a descrição semântica de DWEs e DMEs proposto neste trabalho baseia-se em uma ontologia de DWEs e uma ontologia de domínio. Ele tem o objetivo de fornecer aos usuários especialistas de um domínio de aplicação um meio adequado para descrever suas necessidades, sem a exigência de constantes reuniões com a equipe de TI. Além disso, no caso de alteração de requisitos, espera-se que os usuários possam alterar a especificação do DME sem a necessidade de envolvimento da equipe de TI, caso os mapeamentos dessas especificações para o DWE já estejam criados na base de conhecimento. Esta separação de focos e a futura geração de DMEs devem facilitar a rápida adaptação a mudanças. Os resultados parciais obtidos até o presente momento são: pesquisa bibliográfica realizada, protótipo de um DWE do setor agrícola em operação e ontologias de DWE e da área agrícola parcialmente definidas. Os experimentos para a validação da proposta contarão com o apoio de especialistas em agricultura da Epagri – Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina.

Referências

Egenhofer, M., Herring J. (1992). Categorizing Binary Topological Relations Between Regions, Lines, and Points in Geographical Databases. Technical Report, Department of Surveying Engineering, University of Maine, Orono, ME.

Fidalgo, R. (2005). Uma Infra-estrutura para Integração de Modelos, Esquemas e Serviços Multidimensionais e Geográficos, Tese de Doutorado. Centro de Informática – UFPE.

Malinowski, E. and Zimányi, E. (2007). Logical Representation of a Conceptual Model for Spatial Data Warehouses. Geoinformatica 11, 4 (Dec. 2007), 431-457.

Rao, F., Zhang, L., Yu, X. L., Li, Y., and Chen, Y. (2003). Spatial hierarchy and OLAP-favored search in spatial data warehouse. In *Proceedings of the 6th ACM international Workshop on Data Warehousing and OLAP (DOLAP)*. ACM, New York, NY, 48-55.

Rigaux, P., Scholl, M., Voisard, A. (2002). Spatial databases with application to GIS. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA

Sell, D., da Silva, D. C., Beppler, F. D., Napoli, M., Ghisi, F. B., Pacheco, R. C., and Todesco, J. L. (2008). SBI: a semantic framework to support business intelligence. In *Proceedings of the First international Workshop on ontology-Supported Business intelligence* (OBI). ACM, New York, NY, 1-11.

Silva, J. (2008). GEOMDQL: Uma linguagem de consulta geográfica e multidimensional, Tese de Doutorado. Centro de Informática – Universidade Federal de Pernambuco.

Skoutas, D. and Simitsis, A. (2006). Designing ETL processes using semantic web technologies. In *Proceedings of the 9th ACM international Workshop on Data Warehousing and OLAP* (DOLAP). ACM, New York, 67-74.

Xie, G., Yang, Y., Liu, S., Qiu, Z., Pan, Y., Zhou, X. (2008). EIAW: Towards a Business-Friendly Data Warehouse Using Semantic Web Technologies. LNCS 4825, The Semantic Web, 857-870.