Incorporando Dados Espaciais Vagos em Data Warehouses Geográficos: A Proposta do Tipo Abstrato de Dados VagueGeometry

Anderson Chaves Carniel[1],
Prof. Dr. Ricardo Rodrigues Ciferri[1]

[1]Pós-Graduação em Ciência da Computação
Departamento de Computação
Universidade Federal de São Carlos (UFSCar)
Rodovia Washington Luís, km 235 – SP-310 – 13565-905 – São Carlos – SP – Brasil

{anderson.carniel, ricardo}@dc.ufscar.br

Nível: Mestrado
Ingresso no programa: Março de 2012
Exame de qualificação: Abril de 2013
Época esperada de conclusão: Março de 2014

Abstract. *A data warehouse is a solution for organizing and storing multidimensional data related to decision-making processes in companies, generating a historical, highly voluminous, subject-oriented and nonvolatile database. A geographic data warehouse (GDW) stores spatial data (represented by crisp geometries) as attributes in dimension tables or as measures in fact tables. Thus, spatial data have exact location in the space and well-defined boundaries. However, modern geographic applications require the storage of vague spatial data, which have inaccurate location or uncertain boundaries. This master's project aims at incorporating vague spatial data to GDWs. More specifically, we address the implementation of a new abstract data type (ADT) called VagueGeometry to represent vague spatial data in the Spatial Database Management System PostgreSQL with the PostGIS extension. The proposal of the ADT VagueGeometry encompasses the issue of physical storage and the management of vague spatial data over GDW. Although there are few studies in the literature that implement vague spatial data, these studies have several limitations, for instance, the related work produces low performance in processing analytical queries with vague spatial predicates. This master's project, therefore, aims to investigate this gap in the literature of GDW related to the management of vague spatial data.*

Palavras-Chave. *Data Warehouse, Data Warehouse Geográfico, Dados Espaciais Vagos, Tipo Abstrato de Dados.*

Os autores agradecem o apoio financeiro das seguintes agências de fomento à pesquisa do Brasil: FAPESP, CNPq e CAPES.

1. Introdução

Data warehouse é um importante componente da inteligência de negócio (*business intelligence*) por ser uma base de dados voltada para a tomada de decisão visando a compreensão dos dados para elaboração de estratégias e melhorar a lucratividade nos negócios [Kimbal e Ross 2002]. Adicionalmente, dados geográficos podem ser armazenados, sendo formado um *data warehouse* geográfico (DWG) [Malinowski e Zimányi 2008]. Enquanto sobre o DW incidem consultas OLAP (*Online Analytical Processing*) para a análise das respostas [Kimbal e Ross 2002], sobre o DWG incidem consultas SOLAP (*Spatial OLAP*), que viabilizam a análise multidimensional e a análise espacial [Malinowski e Zimányi 2008].

Comumente no DWG, os dados espaciais têm natureza vetorial e usam tipos de dados geométricos para representar objetos espaciais que sejam simples, tais como ponto, linha e polígono, ou complexos, tais como multiponto, multilinha e multipolígono. Esses objetos frequentemente têm sua localização exata no espaço, ou seja, assume-se que suas coordenadas geográficas definem com clareza a posição geográfica do objeto. Além disso, uma região é representada no espaço com fronteiras bem definidas expressando com exatidão os limites da região. Tais objetos são denominados *crisp*. Contudo, fenômenos podem ter a localização inexata ou fronteiras incertas, sendo, portanto denominados objetos espaciais vagos. Na literatura, ainda se discute a respeito de padrões para a modelagem de dados espaciais vagos e para a definição de predicados topológicos envolvendo dados espaciais vagos. Consequentemente, não há suporte aos dados espaciais vagos em DWGs. De fato, também inexiste suporte nativo a dados espaciais vagos em Sistemas Gerenciadores de Banco de Dados (SGBD) Espaciais, tais como no PostgreSQL com a extensão PostGIS. Nesse sentido, a pesquisa sobre dados espaciais vagos é cada vez mais importante, desde que, com o avanço tecnológico, o uso de dados espaciais vagos em DWGs modernos é cada vez mais requerido para representar situações comumente encontradas no mundo real. Assim, este projeto de mestrado objetiva propor um novo tipo abstrato de dados (TAD), denominado *VagueGeometry*, que engloba uma forma de armazenamento interno para os dados espaciais vagos, os quais são complexos e podem possuir diversas partes disjuntas. Além de modificações na camada de processamento de consultas espaciais do SGBD PostgreSQL/PostGIS para permitir o tratamento de relacionamentos topológicos envolvendo dados espaciais vagos.

Este artigo está organizado como segue. Na seção 2 são descritos fundamentos sobre DWG e dados espaciais vagos. Na seção 3 são descritos os principais trabalhos correlatos. Na seção 4 é apresentada a proposta deste trabalho de mestrado, bem como a sua validação. Atividades realizadas e em andamento são descritas na seção 5. Por fim, na seção 6 são feitas as considerações finais e descritas as atividades futuras.

2. Data Warehouse Geográfico e Dados Espaciais Vagos

Um DWG [Malinowski e Zimányi 2008] assim como um DW convencional [Kimbal e Ross 2002] é uma base de dados histórica, integrada, orientada a assunto e não volátil que objetiva auxiliar na tomada de decisão estratégica. Um DWG pode ser implementado utilizando o modelo relacional por meio de tabelas de fatos e de dimensão. Enquanto as tabelas de fato armazenam as medidas numéricas, as tabelas de dimensão contêm os atributos descritivos que contextualizam essas medidas. Ademais,

um DWG armazena dados espaciais como atributos específicos em tabelas de dimensão ou como medidas em tabelas de fatos [Malinowski e Zimányi 2008].

Contudo, em DWGs, os atributos espaciais armazenados podem ter características de dados espaciais vagos. Existem diversos modelos de representação de dados espaciais vagos. Este trabalho se concentra na investigação de dois modelos: os modelos exatos e os modelos baseados na teoria de conjuntos *fuzzy*. Os modelos exatos tem o objetivo de reutilizar a implementação de dados espaciais *crisp* já existente, e os seus principais modelos existentes na literatura são: Egg-Yolk [Cohn e Gotss 1995], QMM [Bejaoui et al. 2009] e VASA [Pauly e Schneider 2010]. O modelo Egg-Yolk de Cohn e Gotss (1995) somente define regiões vagas o qual utiliza duas sub-regiões em sua representação, uma sub-região denominada clara (parte que engloba a vagueza) e outra denominada gema (parte que representa a exatidão), a qual está contida na clara. Já o modelo QMM (*Qualitative Min-Max model*) de Bejaoui et al. (2009) define dados espaciais vagos em dois limites, o limite mínimo (que se refere a parte que certamente pertence ao objeto espacial) e o limite máximo (que engloba o limite mínimo e o estende com a parte que possivelmente pertence ao objeto espacial). Além disso, utiliza classificações qualitativas para diferenciar "níveis de vagueza", tais como: completamente *crisp*, parcialmente vago e completamente vago. Por fim, o modelo VASA (*Vague Spatial Algebra*) de Pauly e Schneider (2010) propõe uma álgebra para definir tipos de dados espaciais vagos com base nos modelos exatos. Um dado espacial vago é definido por um par de objetos complexos *crisp* do mesmo tipo de dado. Seja $\alpha \in \{ponto, linha, região\}$, um tipo de dado espacial vago é definido formalmente como $v(\alpha) = \alpha \times \alpha$, o qual para $w = (w_n, w_c) \in v(\alpha)$ deve-se respeitar $disjunto(w_n, w_c) \vee toca(w_n, w_c)$, onde w_n é o núcleo e w_c a conjectura. Enquanto o núcleo se refere à porção conhecida e determinada, a parte da conjectura se refere à porção vaga. A álgebra VASA se torna mais ampla que os outros modelos exatos, por também definir vários operadores topológicos, numéricos e específicos para dados espaciais vagos.

Além dos modelos exatos, têm-se os modelos baseados na teoria de conjuntos *fuzzy* [Zadeh 1965]. Nesse sentido, um dado espacial vago é composto por uma função de pertinência associado a um objeto espacial que irá determinar o grau de pertinência de cada ponto do objeto espacial. Em Dilo, de By e Stein (2007) e Schneider (2008) são definidos tipos de dados espaciais *fuzzy*. Um ponto *fuzzy* contém um grau de pertinência associado a um par de coordenadas, sendo definido como $\mu P(x, y) = [0, 1]$, onde (x, y) é um par de coordenadas no espaço Euclidiano bidimensional. Uma linha *fuzzy* é definida por uma função contínua de pertinência que tem transições suaves entre seus pontos vizinhos ao longo da linha e não pode ter auto intersecção. A função contínua pode ser um homeomorfismo *h* de [0,1] para uma linha em $\mathbb{R}^2$, como definido em Dilo, de By, e Stein (2007). As regiões *fuzzy* podem representar fenômenos que contêm fronteiras indefinidas. Para uma região *fuzzy*, é definida uma função de pertinência que determina, para cada ponto, o quanto ele pertence a uma região, sendo ela contínua [Dilo, de By, e Stein 2007; Schneider 2008].

Uma representação de regiões *fuzzy* que reutiliza dados espaciais *crisp* é a região *plateau* [Kanjilal, Liu e Schneider 2010]. Cada região *plateau* é representada por uma sequencia finita de pares, onde cada par é formado por um objeto espacial *crisp* do tipo região e um grau de pertinência associado. Cada região *crisp* de uma região *plateau* é

chamada de sub-região. As sub-regiões estão topologicamente relacionadas com os predicados de "disjunto" ou "toca".

3. Trabalhos Correlatos

Na literatura existem poucos trabalhos que implementam dados espaciais vagos em sistemas gerenciadores de banco de dados (SGBD). Apesar da existência do iBLOB (*Intelligence Binary Large Objects*) de Chen et al. (2010) para definir TADs genéricos, esta estrutura não foi utilizada para definir dados espaciais vagos. Além disso, o iBLOB não foi testado exaustivamente para avaliar seu desempenho comparando com outras técnicas de implementações de TADs. Aspectos relacionados ao armazenamento de dados espaciais vagos em DWGs são investigados nos trabalhos de Siqueira et al. (2011; 2012). Testes de desempenho foram efetuados para investigar o impacto de manter dados espaciais vagos em uma única dimensão ou de separá-los em outra tabela. Já em Siqueira et al. (2012) foram propostos esquemas específicos no nível lógico de DW para permitir a representação de dados espaciais vagos a partir da implementação de dados espaciais *crisp* em SGBDs baseados em modelos relacionais. Porém, nestes trabalhos, não foi implementado um TAD específico para tratar dados espaciais vagos, o qual é o objeto deste trabalho.

Uma implementação de dados espaciais vagos baseado em modelos *fuzzy* é proposta em Dilo et al. (2004), o qual implementam ponto vago, linha vaga e região. Um ponto vago é armazenado como um tripla (x, y, λ) e uma linha vaga como um conjunto de triplas $((x_1, y_1, \lambda_1), ..., (x_N, y_N, \lambda_N))$, onde $(x, y) \in \mathbb{R}^2$ fornece a localização e $\lambda \in (0, 1]$ o grau de pertinência. Uma região vaga é composta por várias linhas vagas e pela triangulação de Delaunay. Esta implementação foi realizada no *software* GRASS e não em um SGBD. Além disso, é importante enfatizar que somente a operação de união foi implementada. Assim, os outros operadores geométricos de conjuntos intersecção e diferença, predicados topológicos e operadores numéricos não foram implementados. Limitações que não existirão neste projeto de mestrado, pois estes aspectos serão investigados e implementados no SGBD PostgreSQL.

Já em Pauly e Schneider (2007) foi implementada a álgebra VASA e implementados os predicados espaciais da seguinte forma. Um operador topológico *P* possui três funções definidas, recebendo dois objetos vagos A e B: (i) *true_P*(A, B), a qual retorna *true* se e somente se o predicado é verdadeiro e *false* caso contrário; (ii) *maybe_P*(A, B), a qual retorna *true* se e somente se o predicado talvez aconteça e *false* caso contrário; e (iii) *false_P*(A, B), a qual retorna *true* se e somente se o predicado é falso e *false* caso contrário. Esta adaptação foi necessária, pois os predicados espaciais da álgebra VASA podem retornar 3 valores lógicos: *true*, *false* ou *maybe*. Assim, o operador ~ foi proposto e que junto a um relacionamento topológico *P*, retorna *true* se o predicado com certeza ou talvez ocorra e *false* caso contrário. Porém, o operador ~ não foi implementado. Diferentemente deste projeto de mestrado, o trabalho de Pauly e Schneider (2007) não proporciona uma forma de representar os dados espaciais vagos internamente no SGBD, apenas oferecendo uma camada que adapta os operadores da álgebra para permitir o tratamento de relacionamentos topológicos e o uso de operações lógicas com três valores.

4. Proposta

4.1. Descrição

Este projeto de mestrado visa a implementação e definição de um TAD para dados espaciais vagos denominado *VagueGeometry*. Assim, pretende-se estender o SGBD PostgreSQL/PostGIS. O PostgreSQL/PostGIS foi escolhido por ser amplamente utilizado pela academia e indústria, ter um bom desempenho e ser de código fonte aberto. Mais especificamente, pretende-se investigar características de dados espaciais vagos, propor algoritmos de manipulação, e a proposta de operadores para a linguagem SQL no intuito de manipular dados espaciais vagos. Prioritariamente almeja-se implementar o TAD *VagueGeometry* usando o modelo exato da álgebra VASA e posteriormente, para o modelo baseado na teoria de conjuntos *fuzzy*.

4.2. Validação

A validação dos resultados obtidos será realizada por meio de testes de desempenho visando comparar o TAD proposto *VagueGeometry* com trabalhos correlatos existentes na literatura. Nos testes de desempenho serão considerados como fatores os tipos de dados espaciais vagos, a origem dos dados (sintético ou real), o volume dos dados, além do tipo e da seletividade das consultas. Os tipos de dados a serem usados serão dados espaciais vagos que podem ser ponto vago, linha vaga ou região vaga. Os demais fatores serão determinados ao longo do projeto.

Os testes enfocarão no uso do TAD *VagueGeometry*, o qual será implementado diretamente no SGBD PostgreSQL/PostGIS. As análises serão realizadas em termos do tempo gasto em segundos no processamento de consultas SOLAP sobre o DWG, utilizando dados espaciais vagos. Serão considerados esquemas diferentes de DWG (por exemplo, esquemas híbrido e convencional), a fim de investigar um esquema adequado para o processamento de consultas SOLAP utilizando o TAD proposto. Já com relação aos trabalhos correlatos a serem usados nos testes de desempenho, serão considerados os trabalhos descritos na seção 3, além de qualquer outro trabalho que por ventura seja proposto na literatura. Especialmente com relação ao SGBD PostgreSQL/PostGIS, pretende-se analisar formas de armazenamento e de processamento de consultas SOLAP com dados espaciais vagos reutilizando os tipos de dados espaciais definidos nesse SGBD, comparando-as com o TAD proposto.

5. Atividades em Andamento

As atividades de mestrado em andamento estão concentradas na etapa de definição e implementação do TAD *VagueGeometry* baseado no modelo *fuzzy* e na etapa de validação e refinamento do TAD *VagueGeometry* baseado no modelo exato da álgebra VASA. Assim, já existe uma versão preliminar do TAD *VagueGeometry* baseado no modelo exato da álgebra VASA que reutiliza estruturas de dados do PostGIS. Foram implementadas várias funções categorizadas como se segue: (i) **input/output**: recebe dados espaciais vagos em sua forma textual e os armazenam internamente, bem como o inverso; (ii) **métodos assessores**: edita, remove ou acessa partes dos dados espaciais vagos; (iii) **operações específicas de tipos**: manipula dados espaciais vagos de tipos pré-determinados (por exemplo, capturar a borda de regiões vagas); (iv) **operações geométricas de conjuntos**: operações de união, intersecção e diferença entre dados

espaciais vagos do mesmo tipo; (v) **operadores topológicos**: verifica os relacionamentos topológicos existentes entre dados espaciais vagos e retorna um objeto do tipo *VagueBool*, por exemplo, o predicado espacial "está contido" pode retornar *maybe*, *true* ou *false*; e, (vi) **operadores numéricos**: calculam medidas numéricas de dados espaciais vagos (por exemplo, a área de uma região vaga), bem como entre dados espaciais vagos (por exemplo, a distância entre duas regiões vagas), e retornam um objeto do tipo *VagueNumeric*, o qual contém um valor máximo e um valor mínimo. Operadores também foram propostos para manipular predicados espaciais entre dados espaciais vagos: (i) o operador unário **~**: retorna *true* se o predicado retorna *maybe* ou *true*, e *false* caso contrário; (ii) o operador unário **~~**: retorna *true* se o predicado retorna *maybe*, e *false* caso contrário; e, (iii) o operador unário **!**: retorna *true* se o predicado retorna *false*, e *false* caso contrário. Além desses operadores também foi implementado o operador binário **~** para comparar um *VagueNumeric* e um *Numeric*, o qual retorna *true* se o *Numeric* está entre o valor mínimo e o valor máximo do *VagueNumeric*. Por fim, também foram implementados os operadores da tabela verdade dos três valores lógicos da álgebra VASA: **&&** (**and**), **||** (**or**) e **!** (**not**). Assim, a validação destas operações está em fase de andamento, bem como a investigação de novos operadores e de refinamento das já existentes. A validação está na fase de configuração do ambiente para os testes, conforme descritos na seção 4.2.

Com relação a proposta do TAD *VagueGeometry* baseado no modelo *fuzzy*, a etapa em andamento é a implementação dos tipos de dados espaciais *fuzzy* e suas operações. Linhas *fuzzy* e pontos *fuzzy* foram implementados como segue: um ponto *fuzzy* é uma tripla *(x, y, u)* onde *(x, y)* é o par de coordenadas e *u* é o seu grau de pertinência; e uma linha *fuzzy* é um conjunto finito de pontos *fuzzy* e a interpolação linear é usada para calcular o grau de pertinência de um ponto na linha. Além disso, a implementação de regiões *fuzzy* está em andamento. A representação de região *fuzzy* considerada é a região *plateau*. As operações geométricas de conjuntos estão em desenvolvimento também. Os últimos aspectos a serem considerados são os operadores topológicos e numéricos, e a etapa da validação, conforme descrito na seção 4.2.

6. Considerações Finais e Próximas Atividades

Dados espaciais vagos são relevantes para a representação de vários fenômenos que contém fronteiras incertas ou localização inexata. Entretanto, no melhor do nosso conhecimento, inexiste um TAD responsável por manipular dados espaciais vagos bem como seus relacionamentos em um SGBD. Trabalhos correlatos que implementam este tipo de dado são limitados e incapacitam seu uso em contextos como a execução de consultas em DWGs com dados espaciais vagos. Dessa forma, este projeto de mestrado concentra-se na proposta de um TAD, denominado *VagueGeometry*, visando sua incorporação em DWGs.

As próximas atividades a serem realizadas envolvem a finalização da proposta do TAD *VagueGeometry* baseado no modelo *fuzzy*, bem como o refinamento da proposta do TAD *VagueGeometry* baseado no modelo exato da álgebra VASA. Em seguida, será focada a etapa de validação dessas propostas por meio de comparação com trabalhos correlatos ao se executar consultas SOLAP sobre DWGs com dados espaciais vagos. Neste sentido, está sendo submetido um pedido de bolsa BEPE para a FAPESP visando o estágio no exterior na University of Florida com o Prof. Markus Schneider. O objetivo deste estágio é usar o iBLOB para incorporar dados espaciais vagos e comparar

esta nova implementação com as implementações do TAD VagueGeometry propostas neste mestrado.

Referências

Bejaoui. L., Pinet, F., Bédard, Y. and Schneider, M. (2009) "Qualified topological relations between spatial objects with possible vague shape," International Journal of Geographical Information Science 23(7), p. 877-921.

Chen, T., Khan, A., Schneider M. and Viswanathan, G. (2010) "iBLOB: Complex Object Management in Databases Through Intelligent Binary Large Objects," In 3rd Int. Conf. on Objects and Databases, p. 85-99.

Cohn, A. G. and Gotts, N. M. (1995) "The Egg-yolk Representation of Regions with Indeterminate Boundaries," In P. A. Burrough, & A. U. Frank, Geographic Objects with Indeterminate Boundaries - GISDATA 2, p. 171-187.

Dilo, A., de By, R. A. and Stein, A. A. (2007) "A System of Types and Operators for Handling Vague Spatial Objects," International Journal of Geographical Information Science. v. 21, n. 4, p. 397-426.

Dilo, A., Kraipeerapun, P., Bakker, W. and de By, R. A. (2004) "Storing and handling vague spatial objects," In 15th Int. workshop on database and expert systems applications. p. 945-950.

Kanjilal, V., Liu, H. and Schneider, M. (2010) "Plateau Regions: An Implementation Concept for Fuzzy Regions in Spatial Databases and GIS," In 13th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems, p. 624-633.

Kimball, R. and Ross, M. (2002) "The Data Warehouse Toolkit". Wiley, 2nd ed.

Malinowski, E. and Zimányi, E. (2008). "Advanced Data Warehouse Design: From Conventional to Spatial and Temporal," Springer Publishing Company, Inc. 444 p.

Mateus, R., Siqueira, T., Times, V., Ciferri, R. and Ciferri, C. (2010) "How Does the Spatial Data Redundancy Affect Query Performance in Geographic Data Warehouses?," Journal of Information and Data Management, v.1, n.3, p. 519-534.

Pauly, A. and Schneider M. (2010). "VASA: An algebra for vague spatial data in databases," Inf. Syst. 35(1), p. 111-138.

Pauly, A. and Scheineder M. (2007) "Querying vague spatial objects in databases with VASA". In Int. Symposium on Spatial Data Quality.

Schneider, M. (2008) "Fuzzy Spatial Data Types for Spatial Uncertainty Management in Databases," Handbook of Research on Fuzzy Information Processing in Databases. p. 490-515.

Siqueira, T., Ciferri, C., Times, V. and Ciferri, R. (2012) "Towards Vague Geographic Data Warehouses," In 7th Int. conference GIScience. p. 173-186.

Siqueira, T., Mateus, R., Ciferri, R., Times, V. and Ciferri, C. (2011) "Querying Vague Spatial Information in Geographic Data Warehouses", In Advanced Geoinformation Science for a Changing World. p. 379-397.

Zadeh, L. A. (1965) "Fuzzy Sets," Information and Control, v.8, p. 338-353.