Metodologia para projeto de fragmentação de dados XML sobre bases distribuídas

Tatiane Lima da Silva[1],Vanessa Braganholo[1,3], Marta L. Queirós Mattoso[2]

[1]Programa de Pós-Graduação em Informática (PPGI/UFRJ)
[2]Programa de Engenharia de Sistemas e Computação (COPPE/UFRJ)
[3]Instituto de Computação (UFF)

tatynany@gmail.com, vanessa@ic.uff.br, marta@cos.ufrj.br

Abstract. *The large amount of XML data available on the Web and within organizations brings new challenges in query processing over distributed environments. In this paper we propose a methodology for XML fragmentation. The main goal is to allow XML documents to be fragmented so that queries submitted by users can be processed in parallel over the fragments, thus increasing the performance of queries.*

1. Introdução

Devido ao grande volume de dados predominantemente armazenado em bancos de dados relacionais, há uma grande preocupação quanto ao processamento de consultas nestes ambientes e, consequentemente, inúmeros estudos nesta área. Dentre as abordagens utilizadas para resolver o problema de processamento sobre um volume grande de dados está a fragmentação e distribuição dos fragmentos em diversos nós de uma rede [10]. Isso permite que consultas possam ser segmentadas e direcionadas a um determinado nó da rede e, por conseguinte, o paralelismo nas execuções das consultas proporciona aumento de desempenho.

Em contrapartida ao cenário do ambiente relacional, documentos XML deixaram de ser utilizados apenas para troca de dados, e se tornaram um importante formato de representação de dados, permitindo o desenvolvimento de aplicações web flexíveis, integração de dados oriundos de varias fontes, manipulação de dados de múltiplas aplicações, entre outros [9]. Este fato faz surgir a necessidade de desenvolvimento de metodologias para processamento eficiente de consultas sobre dados XML. De fato, no panorama dos modelos de distribuição em XML, um dos pontos mais explorados na literatura é justamente o processamento de uma consulta em um ambiente distribuído [1,5,8], enquanto que o projeto de fragmentação de dados XML ainda é um ponto pouco explorado. Conforme discutido em [6], de nada adianta uma metodologia para processamento de consultas distribuídas se a base de dados não estiver fragmentada adequadamente, para que as consultas mais frequentes se beneficiem da fragmentação.

Por definição, um projeto de distribuição descreve quais e como os dados na base devem ser fragmentados e alocados sobre os nós de uma rede [5]. Os principais objetivos no projeto de distribuição são: remoção de acesso de dados irrelevantes durante a execução de uma consulta e redução de troca de dados entre os diversos nós da rede. É importante ressaltar que não existem muitas metodologias para projetos de distribuição e nem para fragmentação para XML. As propostas disponíveis na literatura assumem que o projetista já sabe de que forma a base deverá ser fragmentada [2,4,5,10].

Em um projeto de distribuição de dados *top-down*, a fragmentação é a primeira etapa, sendo seguida pela etapa de alocação. A fragmentação é fundamental para que o projeto como um todo atenda os requisitos principais: diminuição do tempo de resposta das consultas e aumento da vazão do sistema. Contudo, nos modelos existentes na literatura (relacional e orientado a objetos) [3,10], o grande desafio é o desenvolvimento de uma metodologia que contemple a fase de análise, onde uma avaliação é realizada sobre um conjunto de variáveis que permite determinar a forma mais adequada para se fragmentar uma base. Por isso, a proposta desse trabalho é desenvolver uma metodologia de fragmentação para XML baseado no modelo relacional, visando preencher a lacuna referente à etapa de análise.

Este artigo foi estruturado da seguinte forma. Começamos com a fundamentação teórica na seção 2. Na seção 3 apresentamos a metodologia proposta para a fragmentação de dados XML. A seção 4 aborda uma visão comparativa dos trabalhos relacionados. Já a seção 5 descreve as considerações finais e as próximas etapas do trabalho.

2. Fundamentação Teórica

Como mencionado anteriormente, o objetivo desse trabalho é propor uma metodologia para fragmentação para dados XML, utilizando como base o projeto de fragmentação proposto em [10] para o modelo relacional. A escolha da proposta de fragmentação definida para o modelo relacional se deve à consagração deste projeto na literatura, sendo utilizado como fundamentação para diversos projetos de distribuição sobre outros tipos de base dados [3]. Por isso, apresentamos na seção 2.1 uma breve revisão.

No entanto, antes que se possa definir a metodologia, é necessário escolher a definição de fragmentação para documentos XML que será utilizada. Ao contrário do modelo relacional, no modelo XML existem várias propostas [1,5,7,8]. No entanto, o PartiX [1] é a proposta que mais se aproxima ao modelo de fragmento definido para o relacional [10], e foi o escolhido por este trabalho. Sendo assim, apresentamos na seção 2.2 as definições de fragmentos XML e suas respectivas regras de correção propostas pelo PartiX [1].

2.1 Projeto de fragmentação para banco de dados relacional

Dentro do projeto de distribuição, a fragmentação é a primeira etapa que objetiva a geração de fragmentos que são alocados de forma distribuída e replicada numa etapa posterior. No modelo relacional os principais pontos que devem ser levados em consideração ao desenvolver um projeto de fragmentação são [10]: alternativas de fragmentação, que consiste na escolha do tipo de fragmentação e dos algoritmos de fragmentação correspondentes que geram as definições dos fragmentos; grau de fragmentação, que determina até que ponto se deve fragmentar o banco de dados; validação das regras de correção, que consiste em garantir que não haverá perda de dados durante a fragmentação.

No que diz respeito às alternativas de fragmentação, existem três modos alternativos de fragmentar uma tabela [10]: **horizontal**, no qual predicados de seleção são aplicados sobre as tabelas para definir os fragmentos; **vertical**, onde a fragmentação é realizada a partir de projeções definidas sobre as tabelas; e **híbrida**, que consiste na aplicação da fragmentação horizontal sobre a vertical ou vice-versa. De acordo com

[10], as principais etapas em um projeto de fragmentação são: **análise**, **fragmentação horizontal**, **vertical** e **híbrida**.

Análise: Consiste no levantamento e na avaliação das informações do banco de dados e das informações dos aplicativos, por exemplo, análise das consultas mais frequentes.
Fragmentação Horizontal: Consiste no algoritmo propriamente dito para a fragmentação horizontal, levando-se em consideração os pontos relevantes apresentados anteriormente. Resumidamente, a fragmentação horizontal é subdividida em dois tipos: **primária** (FHP), que usa os predicados de seleção definidos sobre uma relação; **derivada** (FHD), cuja fragmentação de uma relação resulta da definição de predicados pertencentes à outra relação [10].
Fragmentação Vertical: Neste tipo de fragmentação, o algoritmo analisa o uso dos atributos nas diversas consultas submetidas à base pelos usuários. Nesse tipo de fragmentação, todos os fragmentos possuem o atributo chave para que a reconstrução da tupla possa ser realizada posteriormente.

No projeto de fragmentação, em [10] e na literatura em geral, não há a definição de heurísticas que determinem o tipo de fragmentação deve ser aplicado a um dado banco de dados. Normalmente é assumido que o projetista já sabe qual tipo de fragmentação deve ser realizado. Quanto à fragmentação híbrida, a abordagem apresentada apenas faz referência a aplicações sucessivas dos outros tipos de fragmentação, não descrevendo em detalhe o seu funcionamento.

2.2. PartiX: Definição de fragmentação de dados XML baseado no modelo relacional

O processamento de consultas sobre grandes volumes de dados XML é uma questão que merece destaque, principalmente na Web. O PartiX [1] propõe a distribuição de consultas XQuery em um ambiente no qual os dados XML estão distribuídos. Os principais pontos desenvolvidos no projeto PartiX foram: a definição dos tipos de fragmentos que podem ser aplicados a XML; e a proposta de uma metodologia de processamento de consultas XML sobre ambientes distribuídos e fragmentados [6]. Dentre os pontos desenvolvidos neste trabalho, o que nos interessa é a definição de fragmentação e as regras de correção, pois ambos são baseados em álgebras e nas definições dadas no modelo relacional.

Definição 1: *Um fragmento F de uma coleção homogênea C é uma coleção representada por F:= <C, γ>, onde γ denota uma operação definida sobre C. F é horizontal se γ denota uma seleção (σ); vertical, se o operador γ é uma projeção (π); ou híbrido, quando ocorre uma composição de operadores de seleção e projeção* [1].

Resumidamente, a proposta feita por [1] consiste no agrupamento de documentos que atendem algum predicado de seleção na fragmentação horizontal e a quebra de um mesmo documento em diversos segmentos de acordo com algum predicado de projeção no caso da fragmentação vertical. A figura 1 apresenta os resultados das fragmentações horizontais e verticais aplicadas sobre uma base XML contendo dois documentos.

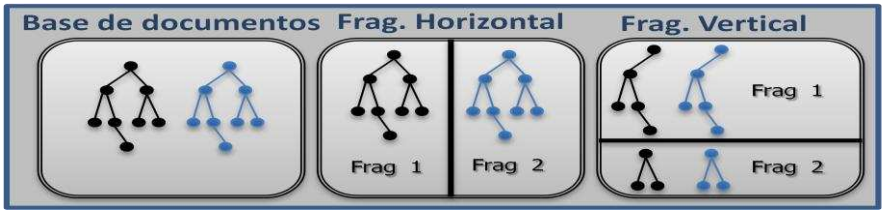

Figura 1 – Funcionamento da fragmentação Horizontal e Vertical

Todo projeto de fragmentação precisa garantir que não há perda de dados. Por isso, foram propostas em [1] três regras de correção para fragmentação. As regras definidas são: completude, que garante que cada item de dados de uma coleção decomposta está contido em pelo menos um dos fragmentos; reconstrução, que permite que uma coleção seja reconstruída a partir dos fragmentos; e, por último, a disjunção, que determina que um dado item de dados só deva aparecer em um dos fragmentos gerados.

3. Metodologia para Projeto de Fragmentação de Documentos XML

O objetivo deste trabalho é a apresentação de uma proposta de metodologia para fragmentação de dados XML baseada no modelo relacional, sendo complementada com a fase analítica que hoje é pouco explorada.

A metodologia proposta tem como informações de entrada os repositórios de dados XML e o conjunto de consultas mais frequentes. O projeto de fragmentação proposto é composto por três etapas: **análise**, **fragmentação horizontal e fragmentação vertical**. Este modelo considera as ideias propostas em [1,3,12] tanto nas questões relativas ao processo de fragmentação do modelo relacional quanto nas definições dos tipos de fragmentos para dados XML e suas respectivas regras de correção.

Análise: Tendo como entrada as informações dos repositórios de dados XML e as consultas mais frequentes, esta etapa consiste em avaliar o tipo de repositório ao qual a fragmentação é aplicada e verificar se o repositório deve ou não ser fragmentado. Esta questão é importante, pois dependendo do tipo de repositório existe a limitação quanto ao tipo de fragmentação que a base pode sofrer. A ideia inicial para essa fase consiste em um algoritmo que receba como parâmetros um conjunto de critérios que permitam definir o tipo de fragmentação que melhor se aplica a uma determinada base. Neste trabalho, iremos definir alguns fatores que nos permitirão escolher de forma mais coerente o tipo de fragmentação a ser aplicada a bases de dados XML. Dentre os possíveis critérios a serem levados em consideração estão:

> **Tipo da base de dados**: que consiste na avaliação da base que sofrerá a fragmentação. Dependendo da base existem limitações quanto ao tipo de fragmentação a ser aplicada.
>
> **Volume de dados**: consiste em avaliar o tamanho da base tanto em quantidade de documentos quanto em espaço em disco ocupado. Este tipo de analise permite verificar a viabilidade da fragmentação sobre a base.

Fragmentação Horizontal: Dadas as definições de fragmentação horizontal, vertical e híbrida [1], um algoritmo de fragmentação efetivamente pode ser criado com o intuito de atender a um determinado cenário de dados. Nesta etapa, as fragmentações horizontais são aplicadas sobre um determinado grupo de dados. É importante ressaltar que diferentemente da fragmentação horizontal aplicada aos modelos relacional e

orientado a objetos [4,10], para XML apenas teríamos a fragmentação primária, uma vez que não há identificadores que relacionem um documento a outro (chaves estrangeiras).

Fragmentação Vertical: Consiste na aplicação da fragmentação vertical sobre um dado conjunto de documentos XML, reduzindo as informações irrelevantes acessadas pelas aplicações. É importante ressaltar que a fragmentação híbrida é uma etapa que consiste da alternância entre a fragmentação horizontal e vertical ou vice-versa, como representado na Figura 2.

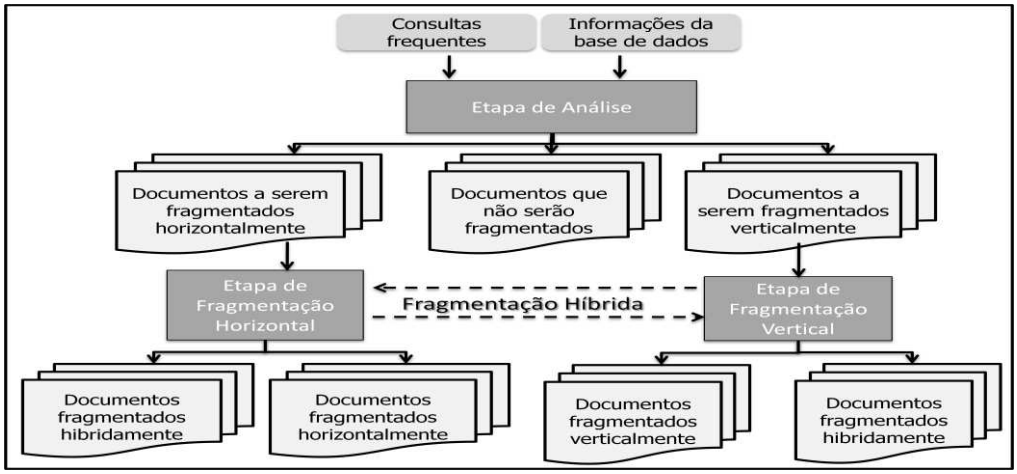

Figura 2 – Metodologia proposta para fragmentação de bases XML

As etapas da metodologia proposta para fragmentação de bases XML apresentada na Figura 2 estão definidas em alto nível, e os próximos passos deste trabalho consistem em definir detalhadamente os algoritmos de cada uma delas.

4. Trabalhos relacionados

Diversos trabalhos são encontrados na literatura sobre definições de fragmentos XML [1,4,6,7,11], mas sobre a perspectiva de projeto de distribuição desses tipos de dados, especificamente a projeto de fragmentação, ainda existem poucos trabalhos relacionados [4,5,7]. Especificamente para a etapa de análise do projeto de fragmentação, nenhum trabalho foi encontrado, nem para bases XML e para bases relacionais, sendo que este tipo de deficiência não permite definir um método decisório consistente quanto ao tipo de fragmentação mais aplicável, fazendo com que a fragmentação seja baseada na experiência dos projetistas.

Na arquitetura para banco de dados XML distribuídos proposta em [11], apenas as camadas gerenciamento de transações, processamento de consultas distribuídas e gerenciamento do esquema distribuído são apresentadas. Entretanto, dentro dessas camadas a etapa de fragmentação não é abordada. Já a abordagem para distribuição de documentos utiliza noções de fragmentação horizontal e vertical diferente da proposta convencional do modelo relacional. Além disso, não são apresentadas no trabalho as regras de correção referentes ao modelo de fragmentação aplicado e também não é descrito os critérios que definem quando cada tipo de fragmentação deve ser aplicado. Por último, o modelo de distribuição apresentado não estabelece uma ordem lógica de execução, sendo a forma de fragmentação diretamente ligada à forma de alocação dos fragmentos.

Bremer e Gertz [4,5] apresentam o projeto de distribuição em duas etapas: *Alternativas de fragmentação* e *Alocação de fragmentos***.** A primeira trata das regras de correção, dos tipos de fragmentação utilizados, e da linguagem de consulta utilizada, que neste caso especifico, é uma linguagem derivada do XPath. Contudo, não está nítida a forma como é feita a análise sobre as bases e os algoritmos de fragmentação.

Kling et al [7] discutem a importância do projeto de fragmentação de dados XML. Entretanto, essas questões não são o foco do trabalho. Os autores apenas sugerem os pontos considerados por eles como relevantes ao se desenvolver um projeto de fragmentação. Dentre os pontos apresentados vale ressaltar a importância sobre o conhecimento da técnica utilizada para distribuição das consultas da base, que segundo Kling et al [7] influenciam no modelo de fragmentação a ser adotado. Note que, na metodologia proposta nesta dissertação, estes dados serão levados em conta.

5. Considerações Finais

Este artigo apresenta uma proposta de metodologia de fragmentação de bases XML utilizando como base soluções bem sucedidas definidas para o modelo relacional. Esta abordagem irá utilizar a definição de fragmentos XML e suas respectivas regras de correção descritas no PartiX [1].

A principal contribuição deste trabalho é o desenvolvimento de uma abordagem ainda pouco explorada na literatura e que é de extrema importância num projeto de distribuição. Os próximos passos incluem: a definição dos critérios a serem utilizados na fase de análise e, desenvolvimento dos algoritmos para a etapa de análise e para os tipos de fragmentação definidos no projeto PartiX e suas respectivas regras de correção.

Referências

[1] ANDRADE, A.; “Efficiently Processing XML Queries over Fragmented Repositories with PartiX”. In: DATAX/EDBT, 2006, Munique, Alemanha, pp.150-163.

[2] ABITEBOUL, S.; GOTTLOB, G.; MANNA, M.; “Distributed XML Design”. In: Symposium on Principles of Database Systems, 2009, Estados Unidos, pp. 247-258.

[3] BAIÃO, F.;MATTOSO, M.;ZAVERUCHA, G.; “A Distribution Design Methodology for Object DBMS”. In: Distributed and Parallel Databases Journal, 2004, EUA, pp. 45-90.

[4] BREMER J.;GERTZ, M.;"On Distributing XML Repositories". In: WebDB, 2003, Estados Unidos, pp. 73-78.

[5] BREMER J.;GERTZ, M.;"Distributed XML Repositories: Top-down Design and Transparent Query Processing", Relatório Técnico,Universidade da Califórnia, EUA, 2003.

[6] FIGUEIREDO. G.; BRAGANHOLO, V.; MATTOSO, M.;“A Methodology for Query Processing over Distributed XML Databases”. Relatório Técnico, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, pp.1-24.

[7] KLING, P.; OZSU, M.; DAUDJEE, K.; “Optimizing distributed XML queries through localization and pruning”, Relatório Técnico. Universidade de Waterloo, Ontario, Canada.

[8] MA, H.; SCHEWE, D.; "Fragmentation of XML Documents". In: SBBD 2003, pp. 200-214.

[9] MOURA, D.; “XML”. Relatório Técnico, Universidade Federal do Rio de Janeiro (COPPE), Rio de Janeiro, Brasil.

[10] OZSU, M.; VALDURIEZ, P.; “Principles of Distributed Database Systems”. Prentice Hall, 1999.

[11] PAGNAMENTA, F. “Design and initial implementation of a distributed xml database”. Dissertação de Mestrado, Universidade de Dublin, Irlanda, 2005.

[12] SILVEIRA, F.; “Fragmentação e Decomposição de Consultas em XML”. Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil, 2006.