INTEGRAÇÃO DE ESQUEMAS RELACIONAIS E XML COM REALIZAÇÃO DE CONSULTAS NA BASE DE DADOS INTEGRADA

Tatiana S. Gois[1], Aglaê P. Zaupa[2]

[1]Faculdade de Informática – Universidade do Oeste Paulista (UNOESTE) 19.050-920 – Presidente Prudente – SP – Brasil. [2] Faculdade de Informática – Universidade do Oeste Paulista (UNOESTE). 19.050-920 – Presidente Prudente – SP – Brasil.
tatisgs@unoeste.edu.br, aglae@unoeste.br

RESUMO

A modelagem de estrutura de dados em um Sistema Gerenciador de Banco de Dados (SGBD) normalmente é feita de acordo com as regras de negócio da realidade que está sendo representada. Organizações que possuem o mesmo negócio nem sempre terão as mesmas estruturas de dados, por terem sido modeladas por pessoas diferentes e também para atender a necessidades diferentes. Atualmente, a identificação da correspondência existente entre esquemas relacionais é feita manualmente na maioria dos casos e tem sido tema de muitas pesquisas na área de Banco de Dados. Este processo é fundamental para possibilitar a integração de esquemas criados de forma independente e de bases de dados diferentes, uma vez que o problema da integração é o transporte de dados armazenados de um esquema para o outro. A proposta para este trabalho é desenvolver uma interface que, após a identificação dos atributos correspondentes entre os esquemas selecionados, realize a integração dos esquemas permitindo a busca de dados na base integrada. Mesmo com o processo de integração e casamento criado, a correspondência entre os esquemas é semi-automática porque os atributos podem ser relacionados sem possuir similaridade real.
Palavras-chave: Esquema XML, Integração de Esquemas, Esquema Relacional, Casamento de Esquemas.

INTEGRATING XML AND RELATIONAL SCHEMAS WITH CONSULTATIONS ON THE INTEGRATED DATABASE

ABSTRACT

Database Management System (DBMS) is usually done according to the business rules of the represented reality. Organizations that have the same business do not always have the same data structures, because they have been designed by different people and to meet different requirements. Currently, the identification of the correspondence between relational schemas is done manually in most cases and has been a research subject in the Database area. This process is essential to enable the integration between schemes created independently and from different databases, since the problem of integration is the transport of data stored in one scheme to another. After identifying the correspondents attributes between different schemas using many techniques, it is necessary to integrate these information to facilitate the data search in both schemes. Even with the integration and relationship approach created, the correlation between schemes is semi-automatic because attributes can be connected without having real similarity.
Keywords: XML schema; relational schema; schema integration; Schema matching.

1 INTRODUÇÃO

O casamento de esquemas é um processo fundamental em áreas que trabalham com dados que se encontram em bases diferentes, Mergen [1]. Neste trabalho, foi focado a importância e necessidade do casamento de esquemas Relacionais e XML (eXtensible Markup Language), bem com a integração. Normalmente, cada um dos esquemas é construído de forma independente e possui terminologias diferentes dos demais, uma vez que eles são elaborados por diferentes pessoas para suprir determinadas necessidades.

O casamento entre os esquemas é realizado a partir de técnicas de casadores, as quais podem ser estruturais, (inclui casador de nomenclatura (Carla), relação, vizinhança e cardinalidade) ou lingüísticos (inclui técnicas como QGram, Dice, Ledit, Overlap, Jaccard) que são aplicadas, inferindo assim uma similaridade para cada um dos atributos comparados de ambos os esquemas.

A partir da similaridade obtida para os atributos casados, torna-se possível a integração das bases de dados, que será realizada levando em consideração o Sistema Gerenciador de Banco de Dados (SGBD) Firebird.

A integração envolve dois passos, Miller [11]: o primeiro, é identificar e classificar as correspondências entre os dois esquemas e, depois, especificar o mapeamento entre estes esquemas. O mapeamento é um conjunto de expressões que especificam como os dados de um esquema podem ser traduzidos para dados do outro esquema. De acordo com Marques [3], Ferradin [13], Marqueto [14], é importante salientar que casamento de esquemas e mapeamento são termos distintos, Hernandes [12], uma vez que no casamento, são encontrados elementos nos dois esquemas que correspondam à mesma informação. Já no mapeamento são utilizados os elementos casados para definir como instâncias de um esquema podem ser transformadas em instâncias do outro esquema.

O objetivo deste trabalho é aplicar técnicas específicas para o casamento de esquemas Relacionais e XML e, a partir da similaridade obtida através destas, realizar a integração de ambas as bases de dados através de um modelo de dados comum capaz de representar o conteúdo dessas fontes. A partir da criação deste modelo, o usuário poderá realizar consultas SQL (Structured Query Language) através de uma interface desenvolvida, que retornará como resultado, informações de ambas as bases que anteriormente foram integradas.

O presente artigo está estruturado em sete seções. A seção 2 aborda uma visão geral de casamento de esquemas, bem como os passos necessários para sua realização. A seção 3 apresenta as técnicas de casadores, bem como os casadores implementados. A seção 4 aborda um estudo de caso do trabalho desenvolvido. Na seção 5, é apresentado o casamento com base no estudo de caso. Na seção 6, é abordado a integração das bases. Na seção 7, é abordada a conclusão e mencionados os trabalhos futuros.

2 CASAMENTO DE ESQUEMAS

De acordo com Mergen [1], Wirt [2] o casamento de esquemas relacionais e XML, tem como finalidade unir duas bases de dados heterogêneas através de elementos que são descritos de forma distintas, mas que contenham a mesma informação. Atualmente o casamento de esquemas, é realizado de maneira semi-automática, onde atributos de ambos os esquemas são analisados através de técnicas de casadores para definir quais elementos podem ser casados, ou seja, quais elementos são correspondentes entre si e a partir daí é necessária a intervenção do usuário, onde o mesmo definirá quais atributos farão parte do

modelo de dados escolhido para mapear os atributos correspondentes entre ambos os esquemas.

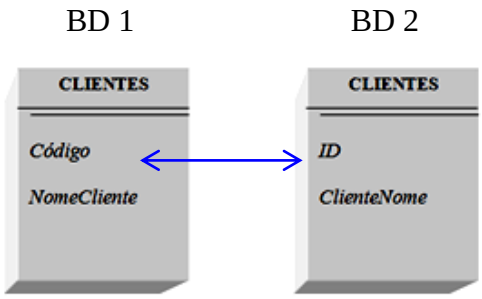

Figura 1: Exemplo de Casamento de Esquemas

De acordo com a Figura 1, o casamento de esquemas pode ser identificado em duas situações: Clientes.Codigo = Clientes.ID ou seja, o atributo Código da base de dados BD 1 é equivalente ao ID da base de dados BD 2 e Clientes.NomeCliente = Clientes.ClienteNome, da mesma forma, NomeCliente da base BD 1 é equivalente a ClienteNome da base BD 2. Assim, para estas situações serão calculadas similaridades através das técnicas de casamento, onde esta similaridade resultará ou não no casamento entre os atributos das bases de dados.

O processo de casamento de esquemas é composto por três etapas:

- Conversão de Esquemas: o usuário interage com o sistema, informando quais os esquemas que ele deseja que sejam casados;
- Execução dos Casadores: alimenta os algoritmos de casamento com os dois esquemas provindos da etapa anterior, Marques [3]. O objetivo de cada algoritmo é inferir uma similaridade entre cada par de atributos, sendo que cada par é composto por um atributo do esquema relacional e um nó (como é chamado um atributo em esquemas XML) XML;
- Filtro de resultados: tem por objetivo eliminar casamentos com base em alguma evidência que caracteriza como casamentos incorretos. Considera-se neste trabalho, que os casamentos entre os esquemas possuem cardinalidade local e global 1:1, ou seja, um nó de um esquema pode ser casado com apenas um nó do outro esquema.

2.1 Aplicações com Casamento de Esquemas

Casamentos de esquemas podem ser utilizados em várias aplicações, tais como: intercâmbio de dados e integração de dados. Daí a necessidade de se criar técnicas para o casamento, Wirt [2].

2.1.1 Intercâmbio de Dados

Com a especificação do XML (eXtensible Makup Language), o mesmo passou a ser utilizado para intercâmbio de dados, uma vez que torna capaz agregar ao seu conteúdo informações que o descrevem, tornando possível a representação de dados que não poderiam ser representados através do modelo relacional utilizado pela maioria dos SGBDs, Mergen [1].

2.1.2 Integração de Dados

A integração de informações tem sido amplamente abordada pela literatura, Mergen [1], Marques [3], Ferradin [13], Marqueto [14], Wirt [2], Bortoleto [5]. A integração de bases de dados heterogêneas vem se tornando peça fundamental no meio governamental e empresarial, Bortoleto [5]. Dessa forma, pode-se fornecer uma representação uniforme e flexível dos dados de fontes distintas.

O principal objetivo da integração é fornecer ao usuário uma interface uniforme para as duas bases de dados heterogêneas, ou seja, de acordo com a Figura 2, em nenhum momento

o usuário tem a visão de estar realizando uma consulta em várias bases de dados ao mesmo tempo. Além de poupá-lo de localizar essas fontes de dados, interagir com cada uma delas isoladamente e combinar manualmente dados vindos dessas múltiplas fontes.

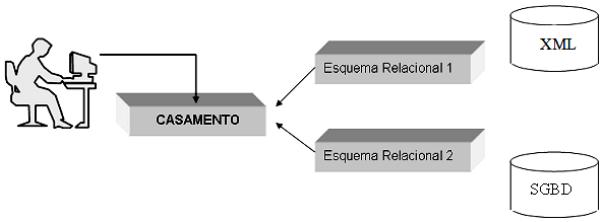

Figura 2: Visão Geral do modelo proposto

De acordo com a Figura 2, a partir do esquema relacional e XML selecionados, é possível, através da intervenção do programador, realizar o casamento entre estes esquemas.

3 TÉCNICAS DE CASADORES

De acordo com Meirelles [4], considera-se que o casamento pode ser implementado de diferentes maneiras, envolvendo o uso de um conjunto de casadores. A forma como os algoritmos são utilizados permite imaginar a solução do problema em dois níveis. No primeiro, são considerados os casadores individuais, ou seja, apenas um algoritmo é aplicado. No segundo, vários algoritmos individuais são combinados.

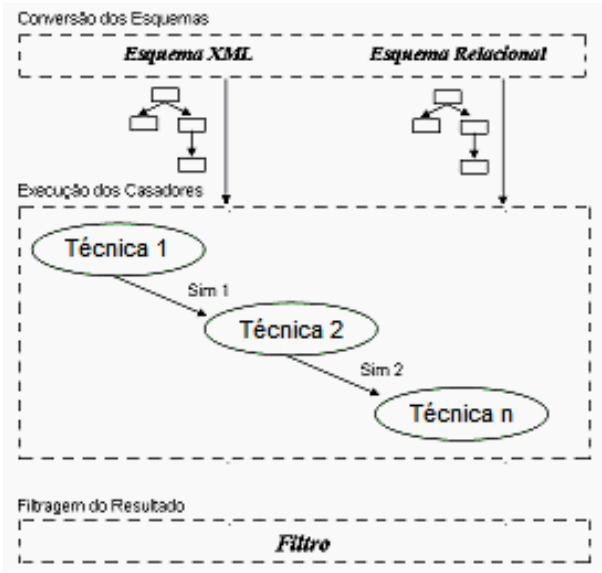

Figura 3: Visão geral do processo de casamento

De acordo com a Figura 3, após entrar com o Esquema Relacional e com o Esquema XML, diversas técnicas de casadores devem ser aplicadas para inferir uma similaridade entre os atributos comparados.

Para o trabalho proposto, foi implementado o casador lingüístico Carla, Dice e QGram pelo fato de que, no esquema XML não é possível fazer verificação e correspondência das restrições e tipo de dados com o Banco de Dados Relacional, uma vez que o mesmo não possui em sua estrutura, tal definição.

3.1 Técnicas de Casadores Lingüístico

Um casador lingüístico possui um algoritmo de reconhecimento de caracteres em comum entre duas cadeias, e uma métrica que permite computar a similaridade entre os caracteres em comum, Mergen [1], Marques [3], Ferradin [13], Marqueto [14]. Existem técnicas também que possibilitam inferir a similaridade entre duas cadeias verificando os termos que elas têm em comum.

Após a aplicação dos algoritmos, faz-se uma combinação das similaridades obtidas de cada uma das técnicas. A combinação é dada

pela média aritmética das similaridades obtidas por estas técnicas.

Somente serão apresentados ao usuário, casamentos cuja similaridade tenha passado por uma etapa de filtro. A etapa utilizada neste trabalho é o *filtro de limiar*, ou seja, apenas serão mostrados aos usuários casamentos cujo valor de similaridade seja superior a um limiar pré-estabelecido. Dessa forma, o limiar é necessário para desconsiderar similaridades muito baixas. Este valor *de limiar* tende a maximizar os valores de precisão, pelo menos nos testes realizados. Não serão aqui estabelecidos os critérios e métodos para a melhor escolha do valor de limiar, uma vez que esta área por si só merece um estudo aprofundado a parte.

Neste trabalho, foram implementados algumas técnicas de casadores liguísticos, tais como: Dice, QGram e Carla. Cada um analisa a cadeia de caracteres de uma maneira diferente. A técnica QGram analisa as cadeias comparando seus qgrans, onde o mesmo representa blocos de caracteres agrupados com tamanho fixo 3:

$$sim = \frac{S - 1}{T - 1}$$

Onde:

S: quantidade de qgrans em comum entre as duas cadeias;

T: quantidade de qgrans da maior cadeia.

Ex:

Cadeia 1 : autor

Cadeia 2 : nomeautor

Cadeia 1: ##a #au aut uto tor or% r%%

Cadeia 2: ##n #no nom ome mea eau aut uto tor or% r%%

$$sim = \frac{5 - 1}{11 - 1} = 0{,}4$$

Dessa forma, será obtida uma similaridade entre a comparação dos qgrans das duas cadeias.

O casador Dice, por sua vez, utiliza:

$$sim = \frac{2 * |V_1 \cap V_2|}{|V_1| + |V_2|}$$

Onde:

$|V_1 \cap V_2|$ é a quantidade de termos em comum entre as duas cadeias;

V_1 é a quantidade de termos da primeira cadeia.

V_2 é a quantidade de termos da segunda cadeia.

No entanto, esta técnica de casador não pôde ser implementada no trabalho proposto, uma vez que seus termos são separados por um espaço em branco e, em esquemas relacionais e XML, não pode existir atributos e nem nós separados por espaço.

Em alguns casos, esse casador pode retornar como similaridade, um valor abaixo da similaridade real que realmente existe entre as cadeias, uma vez que dependendo do tamanho de cada termo, se torna cada vez menos provável que sejam encontrados termos iguais entre os dois atributos casados. Mas isso, na maioria das vezes, é equilibrado pela utilização de alguma outra métrica de casamento, que por sua vez, resultará em uma similaridade maior do que a similaridade real existente entre as cadeias. Daí a importância de se utilizar várias técnicas de casamentos, para obter uma similaridade o mais próxima da real possível.

Já o casador Carla utiliza a fórmula adaptada do Dice, passando a analisar cada caracter de forma individual ao invés de analisar seus termos. Assim, de acordo com Mergen [1], o casador Carla utiliza a seguinte fórmula:

$$sim = \frac{|C|}{|S_1|*\alpha + |S_2|*\beta}$$

Onde:

C: quantidade de caracteres em comum entre as duas cadeias;

S_1: cadeia 1

S_2: cadeia 2

α,β: 0,5

Ex:

S1 = autor

S2 = nomeautor

$$sim = \frac{|5| = 0,71}{|5|*0,5 + |9|*0,5}$$

4 ESTUDO DE CASO

Serão apresentados a seguir, exemplos de um Esquema de Banco de Dados Relacional e um Esquema XML, onde ambos servirão de apoio para o esclarecimento e visão da Integração desenvolvida ao longo deste trabalho.

4.1 Esquema Relacional:Livro

O exemplo utilizado para representar o Esquema Relacional, (Figura 4) é um estudo de caso sobre Livro, onde o mesmo possui as tabelas Autor, Editora, Livro e Capítulo e seus respectivos atributos.

```
CREATE TABLE Autor (
       id_autor              INTEGER NOT NULL,
       id_livro              INTEGER NOT NULL,
       id_editora            INTEGER NOT NULL,
       nome                  CHAR(40) NULL,
       PRIMARY KEY (id_autor)
);

CREATE TABLE Editora (
       id_editora            INTEGER NOT NULL,
       Nome                  CHAR(40) NULL,
       email                 CHAR(40) NULL,
       PRIMARY KEY (id_editora)
);

CREATE TABLE Livro (
       id_livro              INTEGER NOT NULL,
       id_editora            INTEGER NOT NULL,
       id_autor              INTEGER NOT NULL,
       titulo                CHAR(40) NULL,
       PRIMARY KEY (id_livro, id_editora, id_autor)
);

CREATE TABLE Capitulo (
       id_capitulo           INTEGER NOT NULL,
       id_livro              INTEGER NOT NULL,
       id_editora            INTEGER NOT NULL,
       Titulo                CHAR(40) NULL,|
       PRIMARY KEY (id_capitulo, id_livro, id_editora)
);
```

Figura 4: Esquema Relacional de Livro

4.2 Esquema XML:Livro

O segundo exemplo apresentado (Figura 5), também contém nós que trazem informações de Livro, tais como: <autor>, <livro>, <info>, <nomeedit>, <email>, <capitulo>, <paginaini> e <titulo>.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<autores>
        <autor>01</autor>
        <livro>
                <titulo>Dom Casmurro</titulo>
                <info>
                        <nomeEdit>Epoca</nomeEdit>
                        <email>epoca@uol.com.br</email>
                </info>
                <capitulo>
                        <paginaini>10</paginaini>
                        <titulo>Capitulo I</titulo>
                </capitulo>
        </livro>
        <email>text</email>
</autores>
```

Figura 5: Esquema XML de Livro

5 CASAMENTO E INTEGRAÇÃO

5.1 O Casamento

Após calcular todas as similaridades entre os atributos das bases de dados, é necessário realizar o casamento. Para isso, é gerado um *ranking* com as cinco maiores similaridades de cada nó XML, conforme a Figura 6. A partir da interface, o usuário escolhe os pares que deseja integrar, sendo que nessa figura, a integração corresponde ao nó AUTOR do esquema XML e apresenta na lista o *ranking* de similaridades encontradas no esquema relacional, em ordem decrescente por similaridade. O atributo com maior índice de similaridade é o ID_AUTOR da tabela AUTOR e aparece na primeira posição da lista.

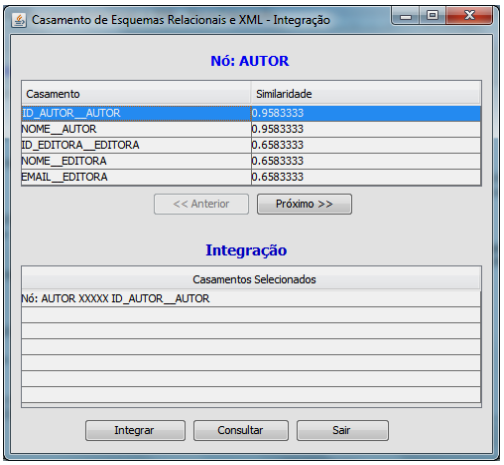

Figura 6: Interface de Casamento de Esquemas

Com a técnica Carla, para a análise entre o nó AUTOR do esquema XML e o atributo ID_AUTOR da tabela AUTOR, o cálculo realizado foi:

$$sim1 = \frac{5}{5*0,5 + 8*0,5}$$

$$sim1 = 0.7692$$

E para a técnica QGram, o cálculo realizado foi:

$$sim1 = \frac{4-1}{10-1}$$

$$sim2 = 0.4444$$

milaridade obtida (sim1 e sim2), é calculada a média ponderada entre essas similaridades, obtendo assim, a similaridade final *0,95833*, a mesma que aparece na Figura 6, na coluna “Similaridade”.

Para todos os demais casamentos exibidos no ranking, foram realizados os mesmos cálculos.

A partir daí, cabe ao usuário confirmar quais casamentos fazem realmente sentido quando analisados semanticamente e, após selecioná-los, clicar em “Integrar”.

Em alguns casos, as técnicas podem retornar similaridades pouco desejáveis. Por isso a importância do casamento ser *semi-automático*, possibilitando ao usuário intervir na integração e selecionar casamentos que realmente façam sentido.

5.2 Implementação da Integração

Para integrar as bases de dados, neste trabalho, foi escolhido um modelo de dados comum para representar os dados de ambas as fontes. Optou-se pelo modelo eXtensible Markup Language (XML), uma vez que o mesmo vem crescendo cada dia mais e conquistando o mercado, além oferecer uma grande flexibilidade de que qualquer tipo de dado pode ser convertido para XML, de acordo com Saccol [7].

O novo XML gerado, ilustrado na Figura 7, conterá a concatenação dos casamentos realizados entre os dois esquemas, ou seja, o nome da tabela juntamente com o atributo do esquema relacional concatenado com o caminho do nó que foi casado com esse atributo (Figura 7). Isso é feito para cada casamento selecionado pelo usuário para integração. Logo este XML conterá todos os caminhos, os mapeamentos entre os elementos casados entre esquema relacional e XML.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!--XML Integracao-->
<SCRIPT>
    <AUTOR_ID_AUTORxxxAUTORES_AUTOR> </>
    <LIVRO_TITULOxxxLIVRO_CAPITULO_TITULO> </>
</SCRIPT>
```

Figura 7: XML integrado

6 RECUPERAÇÃO DE DADOS

Depois de realizada a integração, o usuário consultará as bases heterogêneas através de Consultas em SQL.

Conforme exemplificado na Figura 8, foi utilizado no desenvolvimento deste trabalho, a criação de uma camada transparente ao usuário, onde a mesma é representada por um modelo de dados comum (Figura 7), capaz de representar as fontes de dados dos esquemas.

Assim, a recuperação das informações de forma integrada ocorre da seguinte forma:

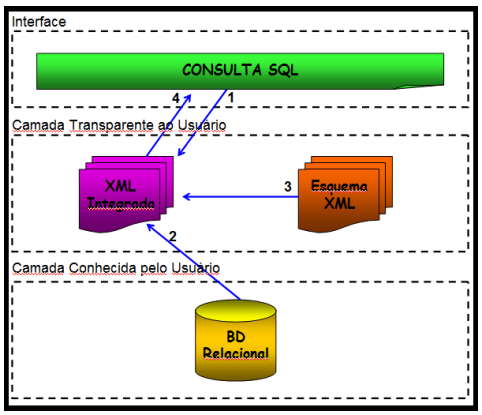

Figura 8: Recuperação dos Dados das Bases

- Passo 1: Os atributos e tabelas contidos na consulta serão buscados diretamente no XML integrado, contendo o mapeamento da

informação solicitada por esta consulta;

- Passo 2: As informações que atendem à consulta, serão buscadas no Banco de Dados Relacional;
- Passo 3: O esquema XML deve ser percorrido, de acordo com o caminho definido no XML integrado, e por fim retornar as informações correspondentes à consulta;
- Passo 4: Enfim, as informações de ambos os esquemas são apresentadas ao usuário de acordo com a consulta solicitada.

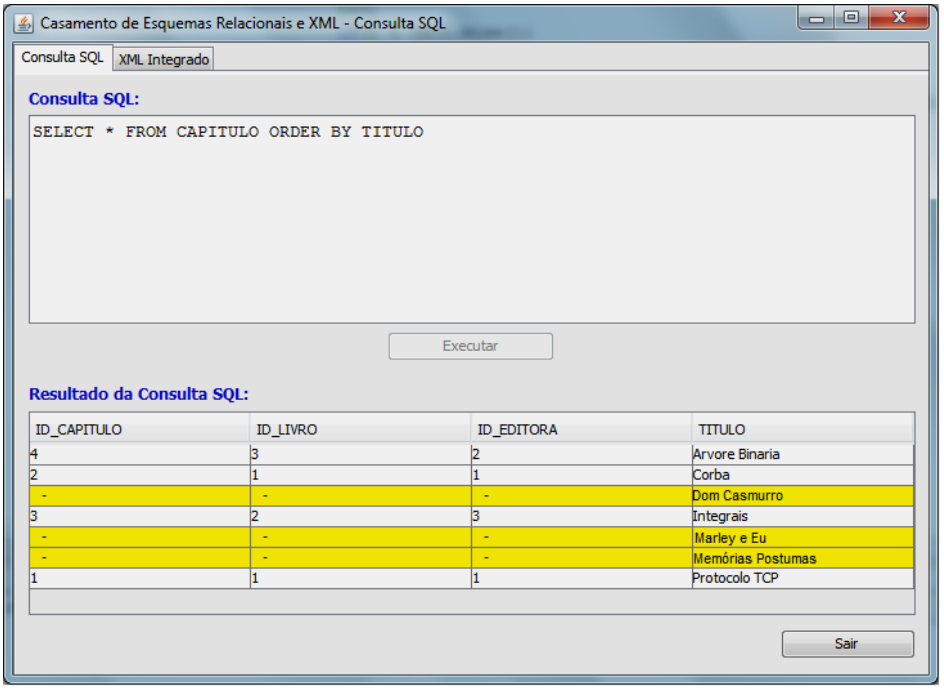

Figura 9: Interface de Consulta

A interface de consulta proposta neste trabalho está apresentada na Figura 9. Nela, o usuário deve digitar o SQL desejado na parte superior da tela e clicar no botão “Executar”. Nesse momento, os dados serão recuperados das bases relacional e XML, a partir do XML gerado na integração, conforme a Figura 8. O resultado da consulta aparece na tabela localizada na parte inferior da tela, onde as linhas em branco são as informações do esquema relacional e as linhas em amarelo são as informações do esquema XML.

7 CONCLUSÃO

No entanto, integrar bases de dados é algo muito importante e principalmente útil quando se trata de acesso a diversas bases de dados.

Atualmente, muitas empresas necessitam de uma integração de bases de dados heterogêneas. Esse é um assunto muito abordado e ainda em discussão sob muitos aspectos na área de Banco de Dados, de acordo com Marques [3], Ferradin [13]. Marqueto [14].

O desenvolvimento deste trabalho, trata a integração entre esquema relacional e XML, de

forma a facilitar o acesso a informações, através da Integração de ambas as bases.

A integração é realizada de forma semi-automática, ou seja, o usuário precisa escolher quais atributos ele deseja integrar, utilizando as diversas técnicas de casadores. Por isso ainda tantos questionamentos e discussões sobre o assunto, onde uma integração não pode ser automática, e a decisão final de quais serão os casamentos integrados pertence ao usuário.

Para trabalhos futuros, poderão ser estudadas formas de integração envolvendo Casamento entre dois esquemas relacionais, bem como a implementação de novas técnicas de casadores, além das enfocadas neste trabalho. Outra opção para estudos futuros pode ser algoritmos que tratam as restrições de chaves estrangeiras.

REFERÊNCIAS

C. A. Wirt, D. L. Musa. “Estudo de técnicas para casamento de Esquemas Relacionais”. Centro Universitário Unilasalle. Canoas. RS, 2005.

C. R. Marqueto. Uma proposta de mapeamento do Modelo XML Schema para o Modelo Relacional. Dissertação de Mestrado. Florianópolis. Agosto/2005. p. 94-120.

D. B. Saccol. Integração de Esquemas em Fontes Heterogêneas XML. Porto Alegre. Feveiro de 2005.

E. F. Marques. O uso do XML na Integração de Banco de Dados Relacionais. UNOESTE. 2007.

E. L. Daronco. XML Integrator: Interoperabilidade de Fontes de dados Heterogêneas, baseada no Mapeamento de Esquemas Conceituais. Porto Alegre. Fevereiro de 2003.

J. C. N. Raffaine, C. D. A. Ciferri. “Integração de Dados provenientes de fontes heterogêneas em uma base de dados relacional.” Instituto de Ciências e Matemáticas e da Computação. USP, São Carlos.

M. A. Hernandez. Clio: a semi-automatic tool for schema mapping. ACM Sigmod Record, New York, v.30, n.2, June 2001.

M. Ferrandin. Integrando Banco de Dados Heterogêneos através do padrão XML. Dissertação de Mestrado. Florianópolis. Outubro/2002.

M. S. Meirelles. Integração de Fontes de dados heterogêneas baseadas no modelo de dados XML. Fortaleza. Maio de 2004.

R. J. Miller. The Clio Project: managing heterogeneity. Sigmod Record, New York, v.30, n.1, 2001.

R. S. Melo. Uma abordagem Bottom-up para integração semântica de Esquemas XML. Porto Alegre. Julho de 2002.

S. Bortoleto. Integração de Multidatabases heterogêneos com aplicação de XML Schemas. Curitiba. 2004.

S. L. S. Mergen. Casamento de Esquemas XML e Esquemas Relacionais. PPGC da UFRGS. 2007.

S. L. S. Mergen, C. A. Heuser. Ferb: um framework para casamento de esquemas. Instituto de Informática – UFRGS.