# Consulta por similaridade a dados semi-estruturados

*Carina Friedrich Dorneles*
*Orientador: Carlos A. Heuser*
Universidade Federal do Rio Grande do Sul - UFRGS
Instituto de Informática
e-mail: {dorneles,heuser}@inf.ufrgs.br

**Resumo**

Instâncias de dados semi-estruturados, presentes no ambiente *Web*, não possuem identificadores. Consultas, neste caso, não podem ser baseadas em predicados binários, pois não há como determinar com precisão se duas instâncias representam o mesmo objeto do mundo real. A proposta aqui apresentada é a aplicação de um mecanismo de consulta por similaridade a dados semi-estruturados na *Web*.

**Palavras-chave:** similaridade, consulta por similaridade, XML, modelo de similaridade, modelo de interação

## 1 Introdução

Em bancos de dados tradicionais, como relacionais e orientado a objetos, a questão de redundância de dados é controlada através de identificadores. Além disso, as consultas são executadas através de operação binária, ou seja, cada objeto é selecionado ou não de acordo com um predicado de consulta. Quando estes bancos de dados são integrados, identificadores globais podem ser gerados a partir do mapeamento dos identificadores locais [10]. Com estes identificadores globais, os dados resultantes do processo de integração podem ser armazenados uma única vez em um banco de dados central e as consultas são executadas sobre os dados já integrados.

Por outro lado, quando dados resultantes da integração de fontes *Web* são considerados, dados semi-estruturados, a mesma informação pode estar representada mais de uma vez, devido à grande heterogeneidade dos dados nas diversas fontes. Por exemplo, informações em *sites* como DB&LP ou páginas de congressos podem conter informações sobre o mesmo conjunto de autores, mas com diferentes representações. Neste contexto, instâncias de dados não possuem identificadores únicos. Para este tipo dado, as consultas baseadas em operadores binários são inadequadas, pois não havendo identificadores, não há como determinar que duas instâncias representam o mesmo objeto do mundo real.

Consultas sobre estes tipos de dados têm sido baseadas em busca por palavra-chave ou *full text search*. Técnicas de recuperação de informação têm sido propostas neste sentido [1]. O principal objetivo é encontrar *strings* em documento, ou parte dele, a partir de padrões fornecidos como entrada. Um dos problemas desta abordagem, é que as técnicas de indexação utilizadas fornecem informação abundante e inútil. Outro problema é a imprecisão de resultados decorrente da subjetividade das consultas, esta devido à falta de um esquema no qual as consultas possam ser construídas. Esta deficiência dificulta a consulta semântica por parte dos usuários. Assim, uma possível alternativa é o uso de um modelo conceitual, ou até mesmo uma ontologia, representativo do domínio que se deseja consultar, uma vez que este serve para descrever o domínio independente de implementação e de quaisquer operações efetuadas sobre os dados.

Como consultas a dados desta natureza não podem ser baseadas em predicados binários e as técnicas usadas em recuperação de informação são, até o momento, sintáticas, outra aboragem deve ser discutida. Propostas de consulta baseada na similaridade das características de objetos [3] têm sido desenvolvidas em diferentes áreas, como recuperação de informação visual e podem ser adaptadas aos dados semi-estruturados.

Dentro deste contexto, torna-se necessário o uso de uma abordagem que utilize um modelo de similaridade para a identificação dos objetos entre as diversas fontes de dados semi-estruturados, já que o uso de identificadores não é possível. A proposta apresentada aqui é a aplicação de um mecanismo de consulta semântica por similaridade a dados semi-estruturados, uma vez que usa um modelo conceitual para construção das consultas e o processo de identificação dos objetos para o resultado é baseado em um modelo de similaridade. No que diz respeito à representação dos dados semi-estruturados, XML foi escolhida. O processo da consulta propriamente dito assemelha-se ao proposto em [7]

O texto está organizado como segue. Na seção 2 são apresentados alguns trabalhos relacionados à consulta por similaridade englobando diversas áreas. A definição do trabalho em andamento na tese é detalhada na seção 3. A metodologia utilizada para o desenvolvimento da tese de doutorado é descrita na seção 4. Por fim, na seção 5 as principais contribuições são brevemente apresentadas.

## 2 Trabalhos relacionados

O uso de modelos de similaridade pode variar dependendo da aplicação e do domínio. Em recuperação de informação textual, modelos de similaridade são propostos com o objetivo de medir a semelhança de um documento com termos fornecidos como entrada nas consultas. Os modelos clássicos de recuperação utilizados são boleano, vetorial e probabilístico [1], que utilizam funções distâncias como *Hamming* e *Levenshtein* (ou *edit distance*) [1] para medir a (dis)similaridade. A linguagem de consulta WHRIL [5] foi desenvolvida para acessar relações STIR (*Storing Text In Relations*). Em banco de dados que armazenam seqüências de tempo [6], ou seqüências de DNA[2], a pesquisa por seqüências similares é feita com base em um modelo de consulta por similaridade. Em banco de dados de imagens [8, 13], o uso de descrições textuais para cada imagem do banco não é uma tarefa trivial. Desta forma, a recuperação é baseada no conteúdo das características visuais, por exemplo, o histograma de cores pode ser usado para representar a característica cor.

Consulta por similaridade a dados semi-estruturados, mais especificamente XML, tem sido alvo de trabalhos recentes. Na sua grande maioria, o foco é na estrutura das árvores de documentos XML. A proposta descrita em [4] apresenta uma solução para encontrar aninhamentos aproximados em estruturas XML diferentes. O trabalho apresentado em [12] considera similaridade entre *tags* de elementos. Assim, dois documentos são estruturalmente similares quando dois documentos compartilham muitas *tags* em comuns, bem como quando estas *tags* possuem os mesmos relacionamentos com seus descendentes e antecessores.

## 3 Definição do trabalho

O objetivo da tese é o estudo e a validação de um processo de consulta semântica baseada em similaridade a dados semi-estruturados armazenados no ambiente *Web*. A construção da consulta é baseada em um modelo conceitual. Este modelo é uma abstração de DTDs originadas de fontes diferentes. O modelo[1] é construído através de um método chamado BInXS[2]. O MCC é uma variante do modelo conceitual ORM/NIAM (*Object with Role Model/Natural Language Information Analysis Method*) pois adota convenções gráficas do modelo ER (Modelo Entidade-Relacionamento) e lida apenas com relacionamentos binários. No modelo MCC, existem dois tipos de conceitos: os conceitos léxicos e os conceitos não léxicos.

As consultas são construídas pelo usuário usando o esquema fornecido pelo modelo. Os resultados desta consulta fornecem uma visão XML das fontes de dados, a qual é gerada de acordo com a estrutura da consulta fornecida pelo usuário. Como resposta, uma lista ordenada de instâncias, obtida a partir da visão, é gerada. A ordem é dada de acordo com a similaridade de cada instância com

[1] O MCC, modelo conceitual canônico, foi definido em tese de doutorado na UFRGS por Mello [11]

[2] *Bottom-Up Integration of XML Schemata*

a consulta fornecida pelo usuário. O grau de similaridade de cada instância também é apresentado.

A figura 1 apresenta uma visão geral da proposta. O trabalho é proposto no nível de uma camada de mediação. Cada *wrapper* é responsável por disponibilizar à camada de mediação os dados no formato XML [14, 9]. A camada de mediação possui dois módulos: i) módulo de **integração de dados**, que recebe os dados resultantes das consultas e aplica o modelo de similaridade sobre as instâncias; e ii) o **módulo de interação**, que fornece ao usuário a possibilidade de realimentar a consulta efetuada inicialmente, usando as respostas desta. Os modelos são apresentados nas subseções abaixo. O processador de consultas é responsável por decompor a consulta do usuário em subconsultas para cada fonte de dado. O processador está fora do escopo desta proposta e está sendo desenvolvido em dissertação de mestrado na UFRGS.

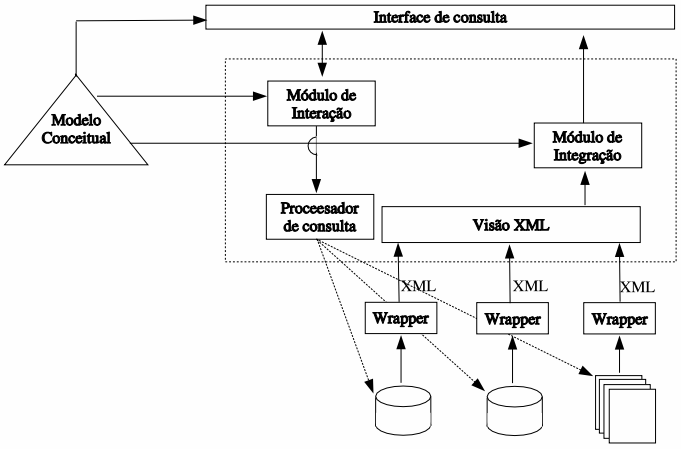


Figura 1: Visão geral da proposta

### 3.1 Modelo de similaridade

O processo de integração de dados é baseado em um modelo de similaridade, tendo como suporte o modelo conceitual. Este suporte é dado através da especificação das medidas de similaridade que devem ser usadas, pesos de cada conceito no cálculo da similaridade, e principalmente quais as propriedades que devem ser consideradas. As propriedades de um conceito são indicadas pelos seus relacionamentos. A fim de que o modelo possa ser definido, alguns problemas devem ser considerados ao construí-lo. Tais problemas são apresentados abaixo através de exemplificação, tendo como domínio "artigos científicos".

**1) Consulta-exemplo incompleta:** Em consulta por similaridade, o usuário fornece como entrada um objeto exemplo e o sistema recupera os objetos similares encontrados no banco. Na maioria das vezes, o objeto fornecido como consulta possui todas as propriedades consideradas na medida de similaridade. No caso abordado aqui, as consultas fornecidas como entrada são incompletas. Por exemplo, o usuário executa a seguinte consulta: "Recuperar os artigos do autor de nome "Navathe". A única propriedade com valor fornecida nesta consulta é `nome`, no entanto o usuário deseja que o objeto `artigo` seja recuperado. Desta forma, o modelo de similaridade deve levar em conta que certos valores de determinadas propriedades não serão informados. O modelo desenvolvido deve tratar de forma diferenciada as características que não apresentam valor, não ignorando-as simplesmente.

**2) Similaridade recursiva:** Para que a similaridade entre dois conceitos seja calculada, primeiramente a similaridade entre suas propriedades deve ser obtida. Estas propriedades, como já dito anteriormente, são fornecidos pelo modelo conceitual. Uma das propriedades de um conceito não léxico pode ser, por si só, outro conceito não léxico. Por exemplo, uma propriedades do conceito não léxico `artigo` é o conceito não léxico `evento`. Assim, para calcular a similaridade entre conceitos `artigo`, deve-se calcular a similaridade entre os conceitos não léxicos `evento` relaci-

onados. Por outro lado, um dos relacionamento do conceito `evento` é um conjunto de conceitos `artigo`. Neste contexto, o modelo de similaridade deve ser capaz de tratar a definição de recursividade de similaridade. No caso apresentado, quando a similaridade entre objetos do tipo `evento` é considerada, provavelmente a propriedade `artigo` não deve ser mais considerada.

**3) Propriedades podem ser conjuntos de objetos:** Conceitos que podem possuir "n" ocorrências de uma determinada propriedade devem possuir um tratamento diferenciado, pois são conjuntos. Por exemplo, um `artigo` pode possuir como propriedade `autor`, que pode haver n ocorrências. Neste caso, uma propriedade a ser considerada quando a similaridade entre conceitos `artigo` estiver sendo calculada é o conjunto `autor`. Assim, para calcular a similaridade entre dois conceitos, não apenas a similaridade entre suas propriedades é considerada, mas também a cardinalidade de suas propriedades deve ser levada em conta.

O modelo de similaridade deverá ser aperfeiçoado durante doutorado sanduíche. Durante este estágio, pretende-se executar atividades de validação, discussão e definição de propostas para alguns dos problemas descritos na seção de Contribuições (seção 5). Estas atividades serão desenvolvidas com o Grupo de Pesquisa em Banco de Dados da *University of Washington* sob a coordenação do Professor Dr. Alon Halevy. Basicamente, duas tarefas serão feitas: i) validação da proposta de integração de dados XML; e ii) discussões de novas alternativas.

### *3.2 Modelo de interação*

Resolvidos os problemas apresentados na seção 3.1, o modelo de similaridade é aplicado e a resposta é apresentada ao usuário de forma ordenada. A ordem é dada pela similaridade de cada instância em relação à consulta solicitada pelo usuário. Baseado neste resultado, o usuário pode interagir de forma a construir novas consultas a partir dos resultados fornecidos.

Considerando que os dados resultantes da consulta citada no item 1 da seção 3.1 seja um lista de autores, seguido pela lista de artigos publicados por ele e o respectivo evento de cada um. Neste caso, o usuário pode construir uma nova consulta solicitando, por exemplo, o conceito `bibliografia` de um dos artigos da lista.

Neste caso, a interação pode ser feita da seguinte forma:

1. o usuário deve selecionar nesta lista a instância que representa o `artigo` mais similar à resposta por ele desejada;
2. o sistema habilita o usuário a indicar no modelo conceitual quais as propriedades relacionadas à instância escolhida que podem participar desta nova consulta. Considerando o resultado do exemplo, o usuário estaria apto a escolher qualquer propriedade do conceito `artigo`;
3. o usuário envia a consulta novamente ao sistema que executa o processo de busca novamente.

O caso apresentado demonstra que os modelos de interação definidos em recuperação de informação visual são inadequados devido às características estruturais dos dados semi-estruturados. Um modelo de interação especial deve ser definido.

## 4 Metodologia

Durante os próximos meses de doutorado a metodologia a ser empregada é a seguinte: levantamento bibliográfico adicional que se julgar importante, discussões com o grupo de pesquisa do Brasil (e durante o doutorado sanduíche com o grupo de pesquisa da *University of Washington*) e validação das propostas discutidas.

A primeira tarefa, descrita na subseção 3.1 está relacionada ao processo de integração. Nesta fase será desenvolvida a proposta de um modelo de similaridade que utiliza as propriedades dos objetos solicitados para "identificar" instâncias. Após a definição deste modelo, a validação deste deverá ser desenvolvida. Tal validação deverá considerar os seguintes aspectos: estudos de caso envolvendo domínios de aplicação e comparação com outros modelos existentes.

A segunda atividade, descrita na subseção 3.2 está relacionada com o modelo de interação com o usuário para refinamento dos resultados. Objetiva-se um estudo da literatura com enfoque nas abordagens usadas em recuperação de informação visual e textual. Concretamente, alternativas para o desenvolvimento deste processo serão comparados, sendo baseado em uma análise de vantagens e desvantagens das mesmas.

Atualmente, a aplicação de modelos de similaridade já existentes estão sendo testadas sobre documentos XML para validação da eficácia ou não de tais modelos. Ao término desta tarefa, um artigo deverá ser escrito, detalhando as idéias e apresentando os resultados dos protótipos.

## 5 Contribuições

A contribuição desta tese é um mecanismo de consulta semântica por similaridade baseado em um modelo conceitual para integração de dados XML. Este mecanismo, proposto na forma de uma camada de mediação, apresenta soluções a alguns dos problemas associados ao acesso a dados semi-estruturados. Estas soluções, são descritas abaixo.

**Problema:** Integração de instâncias semi-estruturadas resultantes de diversas fontes *Web*:
**Solução proposta:** Um modelo de similaridade para dados semi-estruturados é proposto a fim de possibilitar a integração de dados desta natureza vindos de diversas fontes. A definição de similaridade proposta aqui leva em conta o modelo conceitual do domínio.

**Problema:** Modelo de interação para realimentação das consultas:
**Solução proposta:** Modelo de interação que apresente flexibilidade no sentido de fornecer ao usuário facilidades na realimentação de novas consultas.

## Referências


[1] BAEZA, R., AND RIBEIRO-NETO, B. *Modern Information Retrieval*. Addison Wesley Longman and ACM Press Books, USA, NY, 1999.

[2] BAXEVANIS, ANDREAS D., O. B. F. F. *Bioinformatics a practical guide to the analysis of genes and proteins*. New York : Wiley-Interscience, 2001.

[3] BIMBO, A. D. *Visual Information Retrieval*. Morgan Kaufmann Publishers, 1999.

[4] CACCIA, P., AND PENZO, W. Relevance ranking tuning for similarity queries on XML data. In *First VLDB Workshop on Efficiency and Effectiveness of XML Tools, and Techniques (EEXTT2002)* (Hong Kong, China, 19 August 2002).

[5] COHEN, W. Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Informations Systems 18*, 3 (July 2000), 288–321.

[6] FALOUTSOS, C., RANGANATHAN, M., AND MANOLOPOULOS, Y. Fast subsequence matching in time-series databases. In *Proceedings 1994 ACM SIGMOD Conference, Mineapolis, MN*.

[7] IVES, Z. G., HALEVY, A. Y., AND WELD, D. S. Integrating network-bound XML data. *IEEE Data Engineering Bulletin 24*, 2 (2001), 20–26.

[8] JAGADISH, H. V., MENDELZON, A. O., AND MILO, T. Similarity-based queries. In *Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* (San Jose, California, May 22-25 1995), ACM Press, pp. 36–45.

[9] LAENDER, A., RIBEIRO-NETO, B., AND DA SILVA, A. S. Debye - data extraction by example. In *DKE* (2002), vol. 2, pp. 121–154.

[10] MANOLESCU, I., ET AL. Agora: Living with XML and relational. In *International Conference on Very Large Data Bases* (Cairo, Egypt, February 2000).

[11] MELLO, RONALDO, H. C. A. A bottom-up approach for integration of XML sources. In *Proceedings of International Workshop on Data Integration on the Web* (IME - Rio de Janeiro, 2001).

[12] NIERMAN, A., AND JAGADISH, H. V. Evaluating structural similarity in XML documents. In *Web Databases Workshop* (Madison, Wisconsin, June 6-7 2002).

[13] ORTEGA, M. Supporting similarity queries in mars. *ACM Multimedia 97, Electronic Proceedings* (November 8-14 1997).

[14] VITTORI, C. M., DORNELES, C. F., AND HEUSER, C. A. Creating xml documents from relational data sources. In *Electronic Commerce and Web Technologies, Second International Conference* (Munich, Germany, September 4-6 2001), Springer, pp. 60–70.