Um mecanismo para identificação, representação e consulta de versões de objetos XML oriundos de bibliotecas digitais

Eduardo Nunes Borges, Renata de Matos Galante

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{enborges, galante}@inf.ufrgs.br

Programa: Programa de Pós-Graduação em Computação – UFRGS

Nível: Mestrado

Ano de ingresso: 2006 / 1

Conclusão prevista: 2007 / 2

Etapas concluídas: Defesa da Proposta, Seminário de Andamento

Resumo

Em sistemas de informação distribuídos é possível existir múltiplas representações de um mesmo recurso. Um artigo científico indexado por diversas bibliotecas digitais é um exemplo desse recurso. Visando eliminar resultados redundantes das consultas, este trabalho apresenta um mecanismo para detectar, representar, e consultar versões de objetos XML provenientes de diferentes bibliotecas digitais. A abordagem proposta visa representar as versões sem perder informações sobre a proveniência desses objetos, de forma que seja possível realizar consultas identificando a fonte da informação dos metadados XML através de um modelo de proveniência de dados. Foram definidas funções de similaridade aplicadas ao domínio das bibliotecas digitais e realizados experimentos responsáveis por avaliar a eficiência do mecanismo de detecção de versões. Pretende-se comparar a abordagem proposta com outros trabalhos na área que não consideram versões ou proveniência. Além de eliminar a redundância dos resultados de consultas a objetos XML obtidos através da Web, este trabalho também visa manter informações a respeito da origem dos dados recuperados.

Palavras-chave: XML, similaridade, bibliotecas digitais, versões, proveniência.

1. Introdução

Em sistemas de informação distribuídos é possível existir múltiplas representações de um mesmo recurso. A replicação de objetos na rede pode ser uma grande vantagem no desempenho de consultas sobre ambientes distribuídos. Neste contexto, é necessário um processo de identificação ou detecção destas múltiplas representações do mesmo recurso compartilhado. Exemplos desses objetos são artigos científicos indexados por diversas bibliotecas digitais como a Biblioteca Digital Brasileira de Computação (BDBComp) [Laender, Gonçalves e Roberto 2004] e a *Digital Bibliography & Library Project* (DBLP) [University of Trier 2007].

Considere um ambiente de integração de bibliotecas digitais formado pela BDBComp e DBLP, por exemplo. Um usuário submete uma consulta por nomes de autores: "Edleno Silva de Moura, Altigran Soares da Silva". A representação da referência a um determinado artigo é diferente para cada biblioteca digital. Os metadados associados aos artigos, freqüentemente representados no formato XML, são heterogêneos (Figura 1). Os elementos `creator` (linhas 03-04) e `date` (linha 05) presentes nos metadados da BDBComp correspondem, respectivamente, aos elementos `author` (linhas 10-11) e `year` (linha 14) na DBLP. As representações, apesar de diferentes, fazem referência à mesma informação.

BDBComp

```
01 <oaidc:dc>
02    <title>Detec&#231;&#227;o de Sitios Replicados Utilizando Conte&#250;do e Estrutura</title>
03    <creator>Edleno Silva de Moura</creator>
04    <creator>Altigran Soares da Silva</creator>
05    <date>2005</date>
06    <identifier>http://www.sbbd-sbes2005.ufu.br/arquivos/artigo-02-novo_Carvalho.pdf</identifier>
07    <language>por</language>
08 </oaidc:dc>
```

DBLP

```
09 <inproceedings>
10    <author>Edleno Silva de Moura</author>
11    <author>Altigran Soares da Silva</author>
12    <title>Detec&amp;ccedil;&amp;atilde;o de R&amp;eacute;plicas Utilizando Conte&amp;uacute;do e Estrutura.</title>
13    <pages>25-39</pages>
14    <year>2005</year>
15    <booktitle>SBBD</booktitle>
16    <ee>http://www.sbbd-sbes2005.ufu.br/arquivos/artigo-02-novo_Carvalho.pdf</ee>
17 </inproceedings>
```

BDBComp + DBLP

```
18 <metadata>
19    <title>Detecção de Réplicas Utilizando Conteúdo e Estrutura</title>
20    <author>Edleno Silva de Moura</author>
21    <author>Altigran Soares da Silva</author>
22    <year>2005</year>
23    <fulltext>http://www.sbbd-sbes2005.ufu.br/arquivos/artigo-02-novo_Carvalho.pdf</fulltext>
24    <language>Portuguese</language>
25    <booktitle>SBBD</booktitle>
26    <pages>25-39</pages>
27 </metadata>
```

Figura 1. Metadados heterogêneos associados ao mesmo artigo científico.

Outros problemas como diferentes representações de conteúdo e variações na codificação de caracteres são identificados. O metadado `title` assume o valor "Detecção de `Sítios Replicados` Utilizando Conteúdo e Estrutura" na BDBComp (linha 02), enquanto "Detecção de `Réplicas` Utilizando Conteúdo e Estrutura" na DBLP (linha 12).

Para o suporte a integração e consulta em diversas bibliotecas digitais é necessário retornar ao usuário uma resposta única (linhas 18-27), sem perda de informação relevante das diversas fontes de dados – bibliotecas digitais – envolvidas na consulta. Além disso, esta resposta deve ser livre de redundância. Dentro deste contexto, o objetivo deste trabalho

de mestrado é especificar um mecanismo para detectar, representar e consultar versões de objetos XML[1] provenientes de diferentes bibliotecas digitais. A abordagem proposta visa identificar as versões utilizando funções de similaridade e representá-las sem perder informações sobre a proveniência desses objetos, de forma que seja possível realizar consultas identificando a fonte da informação dos metadados XML através de um modelo de proveniência de dados.

A principal contribuição deste trabalho é fornecer ao usuário uma resposta livre de redundância para consultas a objetos XML obtidos através da *Web*, sem perder informações da origem dos dados. A principal aplicação da abordagem proposta é na área de bibliotecas digitais, na consulta a metadados de artigos científicos.

O restante do texto está organizado da seguinte forma. Na seção 2 é apresentada uma curta revisão bibliográfica sobre técnicas de similaridade aplicadas a documentos XML e proveniência de dados. A seção 3 define o mecanismo proposto para detecção, representação e consulta a objetos XML oriundos de bibliotecas digitais. Por fim, na seção 4, são expostas as considerações parciais e próximas atividades.

2. Trabalhos Relacionados

Esta seção apresenta o contexto no qual o trabalho está inserido: funções de similaridade aplicadas a XML e proveniência de dados. As áreas de pesquisa relacionadas são apresentadas a seguir visando delimitar o escopo do problema a ser tratado.

As técnicas de similaridade entre documentos XML geralmente utilizam algoritmos de detecção de diferenças (*diff*). Grande parte destes algoritmos retorna um documento ou *script* (*delta*) contendo as operações básicas necessárias para transformar um documento em outro. Este *script* não é suficiente para a análise semântica do grau de similaridade entre os documentos analisados. É necessário um coeficiente de similaridade que possa classificar os documentos XML como versões ou instâncias diferentes. Algumas propostas tradicionais consideram somente o conteúdo textual dos documentos [Baeza-Yates & Ribeiro-Neto 1999]. Toda informação estrutural é descartada. Outras abordagens mais recentes analisam a estrutura dos documentos [Nierman & Jagadish 2002] [Joshi et al 2003] [Dorneles et al 2004] [Flesca et al 2005].

Apesar de várias abordagens terem sido propostas, a análise semântica do grau de similaridade entre dois documentos XML – de modo a classificá-los como versões de um mesmo documento ou instâncias diferentes – ainda é um problema em aberto pela comunidade científica.

Com o grande número de conjuntos de dados aparecendo em domínio público, torna-se cada vez mais necessário determinar a veracidade e qualidade destes dados. Um histórico detalhado dos dados permite aos usuários avaliar se estes dados são aceitáveis e confiáveis. Uma nova linha de pesquisa denominada *data provenance* tem proposto soluções para este problema. Conhecida também como *data pedigree* ou *data lineage*, a proveniência de dados é a descrição das origens de uma porção de dados e o processo pelo qual ela é obtida [Buneman, Khanna e Tan 2001]. Greenwood [et al 2003] diz que a proveniência de dados é caracterizada por metadados que descrevem os processos de *workflows* e anotações sobre experimentos.

Duas principais características compõem a proveniência de dados: *where-provenance* – de onde os dados são obtidos, ou seja, a origem de uma porção de dados – e

[1] O termo *objeto XML* utilizado neste artigo refere-se a quaisquer informações representadas no formato XML, as quais podem estar contidas tanto em parte quanto em mais de um arquivo XML.

why-provenance – por que esta porção de dados está em um determinado banco de dados. Uma taxonomia das técnicas de proveniência é definida em [Simmhan, Plale e Gannon 2005].

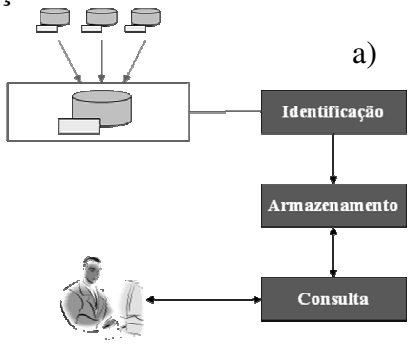
O uso de proveniência de dados em sistemas de integração de bibliotecas digitais, de modo a rastrear a origem das informações recuperadas nas consultas à metadados, é um objetivo específico do estudo desenvolvido neste trabalho de mestrado.

3. O mecanismo para detecção e representação de versões

A Figura 2a mostra a arquitetura do mecanismo proposto, o qual é composto de três módulos: (i) Identificação – responsável por adaptar técnicas de similaridade entre documentos XML, considerando o conteúdo e a estrutura, a fim de detectar versões dos objetos XML referentes aos artigos científicos; (ii) Armazenamento – especifica um modelo de versões para o armazenamento das versões dos objetos XML. É prevista a especificação de um modelo de proveniência de dados, responsável por rastrear as informações de proveniência dos atributos dos objetos XML; (iii) Consulta – permite ao usuário final realizar consultas sem ter conhecimento sobre a existência de versões, obtendo uma resposta única (versões integradas, ou seja, a totalidade de informação relevante contida nas versões) para os artigos científicos pesquisados. Além disso, podem ser recuperadas informações de proveniência dos dados obtidos na resposta à consulta realizada.

O mecanismo proposto será validado através de dois tipos de experimentos. O primeiro conjunto de experimentos é responsável por avaliar o desempenho e qualidade da detecção de versões. O segundo conjunto compara a abordagem proposta com outros trabalhos na área que não consideram versões e/ou proveniência dos dados. As próximas subseções descrevem em detalhes os módulos do mecanismo e os experimentos realizados.

a)



b)

Nome	Iniciais
Eduardo N. Borges	ENB
E. Borjes	EB
Borges, Edward	BE
Borjes, E. Nuñes	BEN

Figura 2. a) Arquitetura do sistema proposto; b) Possíveis representações do autor "Eduardo Nunes Borges".

3.1. Detecção de versões

Para a implementação da detecção de versões é necessário mapear os diferentes metadados provenientes de cada biblioteca digital. O casamento entre esquemas XML [Rahm & Bernstein 2001] é considerado como um problema já resolvido, pois o módulo de identificação proposto é específico para o domínio das bibliotecas digitais e tem como escopo metadados relativos a artigos científicos. Após a fase de casamento, é necessário comparar cada par de referências disponível nos arquivos XML, utilizando técnicas de similaridade de conteúdo, no intuito de identificar as diferentes representações de um mesmo documento (artigo científico publicado).

Foram propostas duas métricas de similaridade entre os autores dos artigos, denominadas *nomesIni* e *simNomes*. A função de similaridade *nomesIni* recebe como parâmetro um par de palavras *a*, *b* (iniciais dos nomes de autores). A Figura 3a define a

função, onde a_i é a i-ésima letra da palavra *a*, b_i é i-ésima letra da palavra *b*, *m* é o tamanho da palavra *a*, e *n* é o tamanho da palavra *b*. As condições impostas pela função partem do princípio que o nome de um autor pode estar representado de diversas maneiras como mostra a Figura 2b.

a)

$$nomesIni(a, b) = \begin{cases} 1, & (a_1 = b_1 \wedge a_m = b_n) \vee \\ & (a_1 = b_2 \wedge a_m = b_1) \vee \\ & (a_1 = b_1 \wedge a_2 = b_2) \vee \\ & (a_1 = b_n \wedge a_2 = b_1) \\ 0, & \text{caso contrário} \end{cases}$$

b)

$$simNomes(K, L, t) = \begin{cases} 1, & \dfrac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(nomesIni\left(K_i, L_j\right)\right)}{\max(m,n)} \geq t \\ 0, & \text{caso contrário} \end{cases}$$

Figura 3. Definição das funções a) *nomesIni* e b) *simNomes*.

A função de similaridade *simNomes* recebe duas listas de palavras (iniciais dos nomes nos autores) *K* e *L* e um valor de limiar (*threshold*) de similaridade *t*. A Figura 3b define a função *simNomes*, onde *i* é a i-ésima palavra da lista *K*, *j* é a j-ésima palavra da lista *L*, *m* é o tamanho da lista *K*, e *n* é o tamanho da lista *L.* O casamento entre as iniciais é realizado através da função *nomesIni* definida anteriormente. Por exemplo, *simNomes((ENB, NI, OCM), (IFN, BE, CYQ), 0.75)* retorna *0*, pois *ENB* corresponde a *BE*, *NI* corresponde a *IFN*, *OCM* e *CYQ* não possuem correspondência. Portando somente 2/3 (66,6%) das ocorrências casaram, não atingindo o limiar mínimo de 75% passado como parâmetro.

3.2. Armazenamento e proveniência

Pretende-se especificar um modelo de proveniência de dados para o módulo de armazenamento de versões de objetos XML. Este modelo deve levar em conta dados no formato XML oriundos de bibliotecas digitais. Informações de proveniência destes dados devem ser armazenadas a fim de identificar as origens dos dados e o processo pelo qual estes foram submetidos e assim derivados.

A proveniência do sistema é orientada a dados, com granularidade fina, ou seja, o produto de dados em questão é um metadado qualquer que descreve um artigo, como por exemplo, *title*. O modelo de proveniência previsto deve ser "escalável", no sentido de ser expansível a diversas bibliotecas digitais. Para representar a proveniência dos dados, pretende-se utilizar anotações, pois são semanticamente ricas. Ainda é necessário um estudo mais aprofundado quanto ao armazenamento da proveniência. Já a disseminação será realizada através de consultas. Ainda deve-se considerar se a proveniência de dados é imutável – onde a atualização das fontes de origem dos dados não implica em atualização da proveniência – ou se deve ser atualizada para refletir o estado atual de seus predecessores.

3.3 Experimentos.

Visando avaliar a identificação das versões, foi realizado um experimento preliminar sobre as bibliotecas digitais BDBComp (4 mil referências para artigos científicos) e DBLP (800 mil referências). Foi realizado um pré-processamento dos metadados XML com o objetivo de carregá-los em um banco de dados relacional. A função de similaridade *levenshtein* (*edit distance*), bem como as funções definidas na seção 3.1, foram implementadas internamente ao SGBD. Uma consulta utilizando a função *simNomes* sobre as iniciais dos autores e *levenshtein* sobre os títulos dos artigos identificou as versões do metadados com precisão de 99.54% e revocação de 91,58%.

4. Conclusões e trabalhos futuros

Este trabalho propõe um mecanismo para detectar, representar, e consultar versões de objetos XML oriundos de bibliotecas digitais. A abordagem proposta visa armazenar as versões dos objetos XML considerando a proveniência dos dados, sem perda de informação relevante obtida das diversas bibliotecas digitais, retornando ao usuário informações livres de redundância.

Foram especificadas funções de similaridade específicas para o domínio das bibliotecas digitais, onde experimentos preliminares já realizados indicam resultados satisfatórios em relação à qualidade da identificação de versões. Como trabalhos futuros destacam-se: (i) realizar mais testes com as funções de similaridade propostas, calculando precisão e revocação para análise da qualidade do resultado, além de medir o desempenho das funções. (ii) Especificar formalmente o modelo de proveniência de dados utilizado no módulo de armazenamento, aplicado ao formato XML. (iii) Comparar esta abordagem com outros trabalhos na área que não consideram versões e/ou proveniência de dados, validando assim a proposta.

Referências

Baeza-Yates R.; Ribeiro-Neto, B. Modern Information Retrieval. ACM Press Series/Addison Wesley, New York, 1999.

Buneman, P.; Khanna, S.; Tan, W. Why and Where: A Characterization of Data Provenance. In International Conference on Database Theory, (2001).

Dorneles, C.; Heuser, C; Lima, A; Silva A.; Moura, E.: Measuring similarity between collection of values. WIDM 2004: 56-63.

Flesca, S.; Manco, G.; Masciari, E.; Pontieri, L.; Pugliese, A. “Fast detection of XML structural similarity”, Knowledge and Data Engineering, IEEE Transactions on, vol.17, no.2pp. 160- 175, Feb. 2005.

Greenwood, M.; Goble, C.; Stevens, R.; Zhao, J.; Addis, M.; Marvin, D.; Moreau, L.; Oinn, T. “Provenance of e-Science Experiments - experience from Bioinformatics,” in Proceedings of the UK OST e-Science 2nd AHM, 2003.

Joshi, S.; Agrawal, N.; Krishnapuram, R.; Negi, S. A bag of paths model for measuring structural similarity in Web documents, Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, August 24-27, 2003, Washington, D.C.

Laender, A.; Gonçalves, M.; Roberto, P. BDBComp: Building a Digital Library for the Brazilian Computer Science Community. In: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tuscon, AZ, USA, pp. 23-24, 2004.

Nierman, A.; Jagadish, H. “Evaluating Structural Similarity in XML Documents”. In Int'l Workshop on the Web and Databases (WebDB), Madison, WI, Jun. 2002.

Rahm, E.; Bernstein, P. A survey of approaches to automatic schema matching. Very Large Database J., 10(4):334–350, 2001.

Simmhan, Y.; Plale, B.; Gannon, D. A survey of data provenance in e-science, ACM SIGMOD Record, v.34 n.3, September 2005.

University of Trier. Digital Bibliography & Library Project (DBLP). Disponível em <http://dblp.uni-trier.de/>. Acesso: abril, 2007.