XSimilarity : Uma Ferramenta para Consultas por Similaridade embutidas na Linguagem XQuery

Maria Estela Vieira da Silva, Eduardo Nunes Borges e 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 – Brasil

{mevsilva, enborges, galante}@inf.ufrgs.br

Abstract. *This article presents XSimilarity, a tool for executing similarity queries embedded in XQuery language. Relevant similarity functions for database and information retrieval fields are implemented in the tool. The main contribution of this article is presenting a tool for similarity queries in XML databases. This allows for information retrieval optimization in these databases. Besides that, as many different similarity functions were implemented, the tool allows comparing the effectiveness of these funcions in different contexts. The examples are constructed over Computer Science scientific articles data provided by the DBLP digital library.*

Resumo. *Este artigo apresenta o XSimilarity, uma ferramenta para a realização de consultas por similaridade embutidas na linguagem XQuery. São implementadas funções de similaridade relevantes na área de banco de dados e recuperação de informações. A principal contribuição desse artigo é apresentar uma ferramenta que permite consultas por similaridade em bases de dados no formato XML, possibilitando a melhora da recuperação de informações nessas bases de dados. Além disso, como foram implementadas várias funções de similaridade diferentes, a ferramenta também permite comparar a eficácia dessas funções para diferentes contextos de utilização. A ferramenta é validada através de experimentos com dados de artigos científicos da área da Ciência da Computação oriundos da biblioteca digital DBLP.*

1. Introdução

Os mecanismos tradicionais de pesquisa em banco de dados [Groff and Weinberg 2002] utilizam uma busca por coincidência exata. Em uma base de dados de cidades, por exemplo, "Rio Grande" e "R. Grande" são considerados objetos diferentes. Mas ambos os objetos fazem referência a uma mesma cidade. A recuperação da informação por meio de consultas exatas pode ser ineficaz neste caso. Portanto, torna-se necessária a utilização de um mecanismo de pesquisa mais elaborado, com suporte a consultas imprecisas ou aproximadas, também conhecidas como consultas por similaridade [Jagadish, Mendelzon and Milo 1995].

A linguagem XML (*Extensible Markup Language*) [W3C 2006] foi criada em 1996 pelo W3C (*World Wide Web Consortium*). A principal característica da linguagem XML é a flexibilidade da linguagem, possibilitando ao usuário definir suas próprias tags e suas próprias linguagens de marcação, porém de uma forma bem mais simples, gerando uma aceitação parecida com o HTML (*Hypertext Markup Language*), e leve, possibilitando o uso da linguagem na Web. Durante os últimos anos, devido à sua simplicidade, flexibilidade, independência de plataforma, e outras vantagens, a linguagem XML tem sido o principal mecanismo para representação de dados semi-estruturados e troca de informações por aplicações Web.

A XQuery [W3C s.d.] é uma linguagem criada a partir da necessidade de se realizar consultas sobre documentos XML. XQuery está para o XML assim como SQL está para os bancos de dados

relacionais, pois embora suas sintaxes sejam diferentes, a semântica é a mesma. Entretanto, a sintaxe XQuery permite a realização de pesquisas mais complexas e de leitura mais fácil que a sintaxe SQL.

Algumas abordagens foram propostas para permitir que consultas aproximadas fossem executadas em SGBDs relacionais [Calado 2000, Borges e Cony 2005, Borges e Dorneles 2006]. Outros trabalhos estendem as linguagens de consulta a documentos XML [Padilha 2005]. Entretanto, existe uma lacuna nos SGBDs XML, como, por exemplo, eXist, Tamino, XIndice, entre outros, quanto ao suporte a consultas aproximadas. Torna-se evidente, portanto, o interesse em desenvolver mecanismos de consulta por similaridade em bases XML. A linguagem XQuery, desenvolvida especificamente para trabalhar sobre XML, é uma excelente alternativa para a inclusão de tal mecanismo.

Este artigo descreve o projeto e a implementação da ferramenta XSimilarity, que permite executar consultas por similaridade em documentos XML, considerando documentos armazenados em um sistema de arquivos. As funções de similaridade são embutidas na linguagem de consulta XQuery. A ferramenta é validada através de experimentos com dados de artigos científicos da área da Ciência da Computação oriundos da biblioteca digital DBLP [Trier s.d.].

A principal contribuição deste artigo é apresentar uma ferramenta que permite consultas por similaridade em bases de dados no formato XML, possibilitando a melhora da recuperação de informações nessas bases de dados. Além disso, como foram implementadas várias funções de similaridade diferentes, a ferramenta também permite comparar a eficácia dessas funções para diferentes contextos de utilização.

O presente trabalho pode ser aplicado na consulta a informações em ambientes integrados na *Web* como CiteSeer e bibliotecas digitais. Tais ambientes integram metadados de várias fontes, os quais geralmente estão estruturados em documentos XML, e consultas exatas não seriam suficientes para mostrar resultados satisfatórios. Além disso, a ferramenta desenvolvida soluciona uma das deficiências dos SGBD XML no que tange a execução de consultas por similaridade utilizando XQuery.

O restante do artigo está estruturado da seguinte forma: A seção 2 apresenta os trabalhos relacionados. A seção 3 descreve detalhadamente a ferramenta XSimilarity enquanto a seção 4 apresenta um estudo de caso com experimentos com dados oriundos da biblioteca digital DBLP. A seção 5 encerra o trabalho com considerações finais, relacionando trabalhos futuros.

2. Trabalhos relacionados

Algumas abordagens foram propostas para permitir que consultas aproximadas fossem executadas em SGBDs relacionais. Calado et al. [Calado 2000] desenvolveu uma interface que requisita dados de um SGBD relacional através de consultas exatas e calcula, de maneira independente do SGBD, a aproximação destas tuplas em relação à resposta ideal. [Borges e Cony 2005] anexam funções de similaridade ao SGBD PostgreSQL e criam novos operadores que implementam estas funções. A similaridade é processada internamente ao SGBD, facilitando o uso e agilizando o cálculo das funções. O pacote de funções e operadores de similaridade são disponibilizados através da ferramenta PgSimilar [Borges e Dorneles 2006].

Além das propostas para bases relacionais, outros trabalhos estendem as linguagens de consulta a documentos XML. Padilha [Padilha 2005] faz uma extensão do processador da linguagem XPath para suportar funções de similaridade como operadores *booleanos*, possibilitando a execução de consultas aproximadas.

No presente trabalho, buscou-se preencher uma lacuna dos SGBDs XML, como Tamino [Software Ag s.d.], por exemplo, implementando uma ferramenta para a execução de consultas

aproximadas, inexistente nos atuais SGBDs. O trabalho proposto neste artigo difere dos demais relacionados nos seguintes aspectos: (i) foi desenvolvida uma ferramenta para consultas por similaridade em documentos XML; (ii) as funções de similaridade são embutidas na linguagem XQuery que é mais dinâmica e poderosa do que XPath; (iii) o XSimilarity implementa funções de similaridade que atuam de formas variadas, sendo uma ferramenta muito útil para realizar buscas; (iv) o conjunto de funções de similaridade implementado envolve métricas baseadas em caracteres, palavras (*tokens*) ou híbridas. Isto torna a ferramenta bastante abrangente, permitindo maior flexibilidade para experimentos dos mais diversos tipos. Além disso, o XSimilarity implementa algumas das principais funções de similaridade utilizadas nas áreas de bancos de dados e recuperação de informações.

3. XSimilarity - Uma ferramenta para consultas por similaridade embutidas na linguagem XQuery

XSimilarity é uma ferramenta desenvolvida para a realização de consultas por similaridade embutidas na linguagem XQuery. Esta seção apresenta uma visão geral da ferramenta e as tecnologias usadas para a sua implementação. Também são apresentadas mais detalhadamente as principais funcionalidades da ferramenta XSimilarity.

3.1 Arquitetura da ferramenta

A ferramenta foi implementada em Java, sendo utilizada a biblioteca NUX [Berkeley Lab s.d.] como mecanismo de execução XQuery e a biblioteca SimMetrics [Chapman s.d.] como fonte de funções de similaridade. Também é utilizado o pacote Java swing para a definição da interface.

Internamente, é construída uma classe para permitir a disponibilização das funções de similaridade em um *namespace* da XQuery. O nome desta classe aparece na declaração do *namespace*, que fica como "java:xsimilarity.Similaridade".

Na Figura 1 apresenta, de forma gráfica, a arquitetura da ferramenta. Considerando que a ferramenta utiliza a API Java Swing para a sua interface e a Nux como mecanismo de execução XQuery e ambas são externas a ferramenta, elas são representadas abaixo do módulo XSimilarity. Como não há comunicação direta entre as bibliotecas *Swing* e *Nux*, elas não são ligadas entre si na arquitetura. A ferramenta implementa um *namespace*, denominado Similaridade, que utiliza a biblioteca de funções *SimMetrics* que é externa à ferramenta.

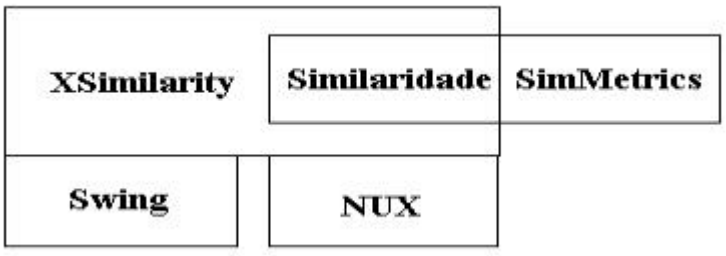

Figura 1: Arquitetura da Ferramenta.

3.2 Descrição da ferramenta

A ferramenta XSimilarity foi desenvolvida para suprir a necessidade de um mecanismo de consulta por similaridade em bases de dados XML. Para tanto, a ferramenta permite a execução de consultas XQuery em que funções de similaridade podem ser embutidas através de um mecanismo de *namespace*. O uso das funções de similaridade possibilita a melhora da recuperação de informações nessas bases de dados.

As funcionalidades incluídas na interface de usuário podem ser vistas na Figura 2, a qual apresenta o diagrama de casos de uso da ferramenta. A ferramenta apresenta as funcionalidades de

criar nova consulta, abrir consulta, salvar consulta, editar consulta, executar consulta, inserir função de similaridade e ajuda. As principais funcionalidades constantes no diagrama de casos de uso são detalhadas a seguir.

Figura 2: Funcionalidades da ferramenta.

A funcionalidade de "Editar consulta" permite que a consulta seja editada em um editor de texto simples. É possível realizar todas as operações básicas de edição, incluindo copiar, recortar, colar, selecionar e deletar as *strings* desejadas.

A funcionalidade de "Inserir função de similaridade" estende a funcionalidade de "Editar consulta", permitindo que sejam inseridas funções de similaridade na consulta que está sendo editada. Mesmo sem utilizar esta funcionalidade, o usuário pode editar a consulta diretamente e utilizar as funções de similaridade. Entretanto, ao utilizar a funcionalidade de "Inserir função de similaridade", o usuário tem auxílio da ferramenta na montagem da função a ser inserida, incluindo a descrição das funções disponíveis, suas assinaturas, parâmetros e *namespace* que deve ser utilizado.

Por fim, a funcionalidade de "Executar consulta" permite que a consulta Xquery que está sendo editada na ferramenta seja executada. Consultas com a sintaxe adequada são executadas e seu resultado é retornado na própria ferramenta.

Na parte superior da interface principal (Figura 3) há uma série de botões. Os três primeiros apresentam as funções clássicas "novo", "abrir" e "salvar", respectivamente. No caso desta ferramenta, estes botões atuam sobre a consulta. Ao clicar no botão "novo", a caixa que contém a consulta a ser executada é esvaziada e o usuário pode criar uma nova consulta. Os arquivos de consulta são salvos em formato texto. A extensão "xq" é utilizada para diferenciá-los de outros arquivos.

Dois botões oferecem as funcionalidades centrais da ferramenta. O botão com a imagem de um "S" é o que permite inserir funções de similaridade (detalhes na seção 3.2.3). O botão com a figura de um funil é o que permite executar uma consulta XQuery. Os arquivos a serem consultados são referenciados na própria consulta XQuery, assim como as funções de similaridade utilizadas.

O último botão da interface, que contém a imagem de uma interrogação, permite ao usuário acessar a ajuda da ferramenta.

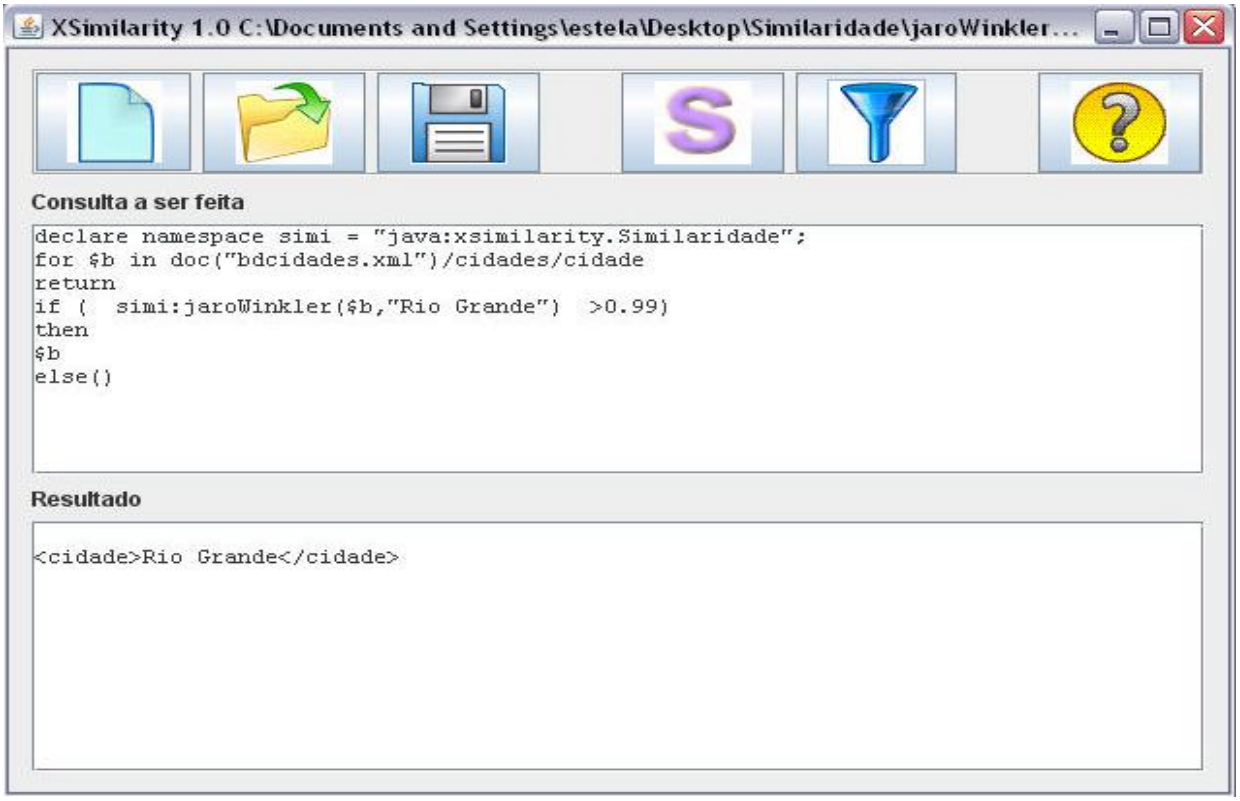

Figura 3: Interface principal da ferramenta.

3.2.1 A tela de abrir consulta

Quando o usuário clica no botão com a imagem de uma pasta amarela sendo aberta na interface principal, a ferramenta exibe a tela de abrir consulta, que pode ser visualizada na Figura 4. Pode-se clicar sobre o nome do arquivo com extensão "xq" ou digitar o nome do arquivo desejado diretamente na caixa "*File Name*". Ao clicar no botão "*Open*", a consulta selecionada será aberta na ferramenta e estará pronta para ser editada na interface principal. Para abrir arquivos em outros diretórios, o usuário pode utilizar os recursos de navegação, existentes na parte superior da tela.

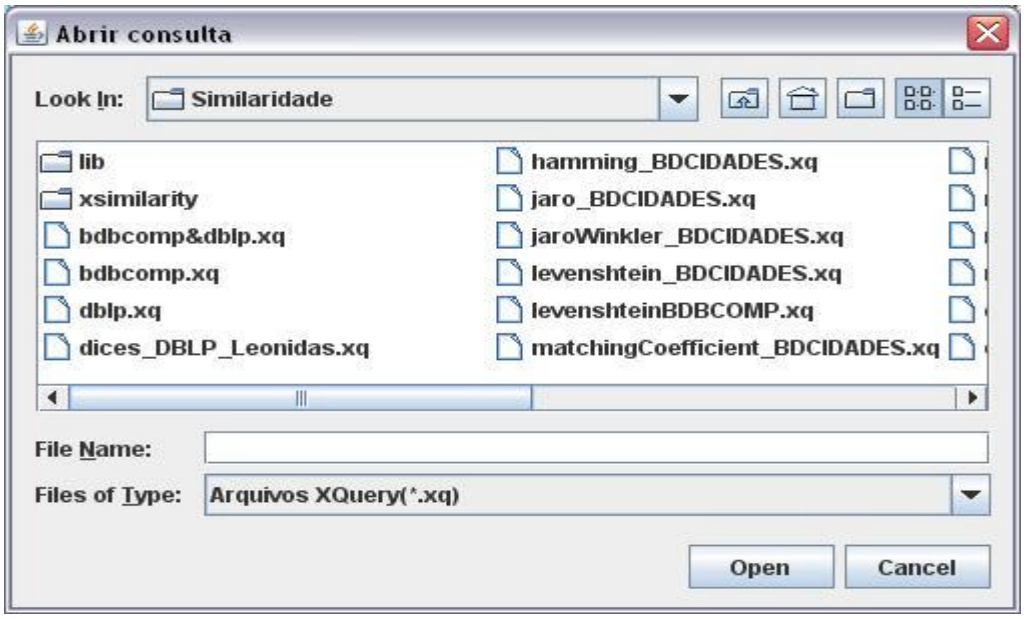

Figura 4: Janela de abrir consulta

3.2.2 A tela de salvar consulta

Quando o usuário clica no botão com a imagem de um disquete azul na interface principal, a ferramenta exibe a tela de salvar consulta, que pode ser visualizada na Figura 5. Pode-se clicar sobre o nome do arquivo com extensão "xq" ou digitar o nome do arquivo desejado diretamente na caixa "*File Name*". Ao clicar no botão "*Save*", a consulta sendo editada será salva no diretório escolhido com o nome selecionado. A escolha do diretório onde a consulta será salva é feita através dos recursos de navegação, existentes na parte superior da tela.

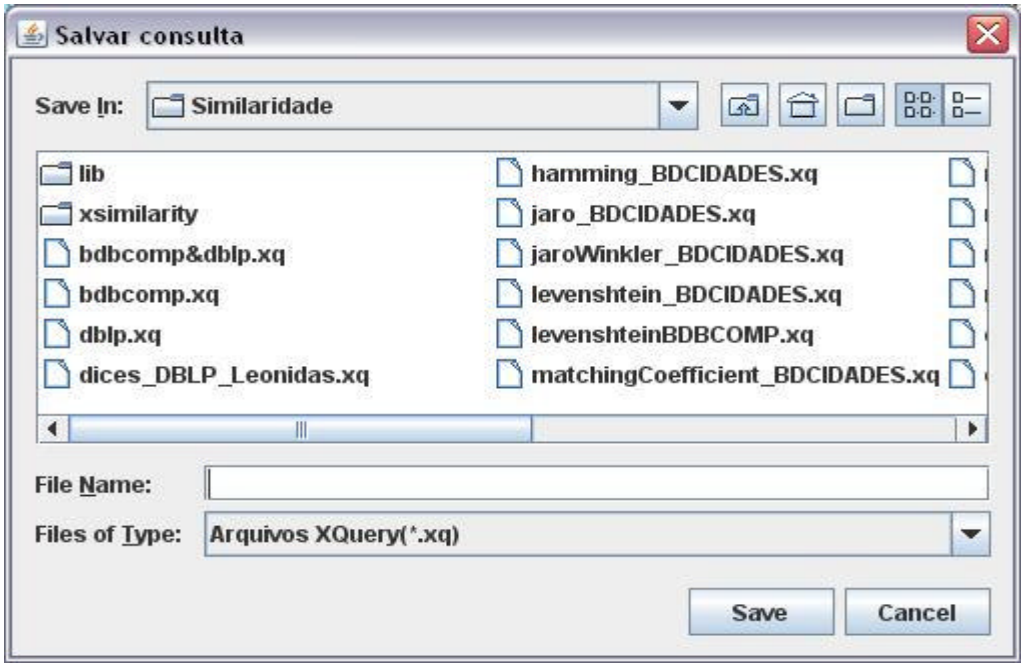

Figura 5: Janela de salvar consulta

3.2.3 A tela de inserção de função de similaridade

Quando o usuário clica no botão com a imagem de um "S" na interface principal, o sistema abre uma nova janela para permitir a escolha da função de similaridade a ser inserida. Esta janela pode ser vista na Figura 6.

Esta janela permite que o usuário visualize a assinatura da função de similaridade que deseja utilizar. O usuário seleciona a função desejada na lista à esquerda e o sistema preenche a assinatura, o *namespace* e a descrição da função nas caixas de texto à direita. Foram utilizadas as funções de similaridade, implementadas pelo SimMetrics: *Dice's Coefficient* [Dice 1945], *Jaccard Similarity* [Cohen, Ravikumar and Fienberg 2003], *Hamming Distance* [Hamming 1950], *Levenshtein Distance* [Levenshtein 1966], *Jaro Distance* [Jaro 1989], *Jaro Winkler Distance* [Winkler 1990] *e Monge Elkan Distance* [Monge and Elkan 1996].

O conjunto de funções de similaridade implementado nesta ferramenta envolve métricas baseadas tanto em caracteres, palavras (tokens) ou híbridas. Ainda possuem diferentes custos computacionais e aplicabilidade variada. Estas diferentes características fornecem maior flexibilidade para experimentos dos mais diversos tipos, o que justifica a escolha destas funções para a implementação da ferramenta . A ferramenta foi planejada de modo que possa ser facilmente estendida para a inclusão de novas funções de similaridade.

Antes de confirmar clicando em OK, o usuário pode editar a assinatura da função de similaridade para adaptá-la a sua consulta (por exemplo, inserindo valores para os parâmetros das funções). Ao editar a assinatura na janela de inserção de função, o usuário terá à sua disposição uma descrição detalhada da função escolhida e de seus parâmetros, o que facilita a adaptação da função.

Ao navegar por várias funções na lista à esquerda, o usuário poderá aprender mais sobre cada função de similaridade.

Outra funcionalidade importante dessa caixa de diálogo é permitir que o usuário marque a inserção do *namespace* padrão da função em sua consulta. Como o *namespace* das funções de similaridade é implementado em uma única classe Java, é necessário inseri-lo apenas na primeira vez que se insere uma função. Dessa forma, o *checkbox* responsável por habilitar essa inserção permanece desabilitado por *default*. Quando esse *checkbox* é habilitado, o *namespace* padrão é inserido na primeira linha do texto da consulta.

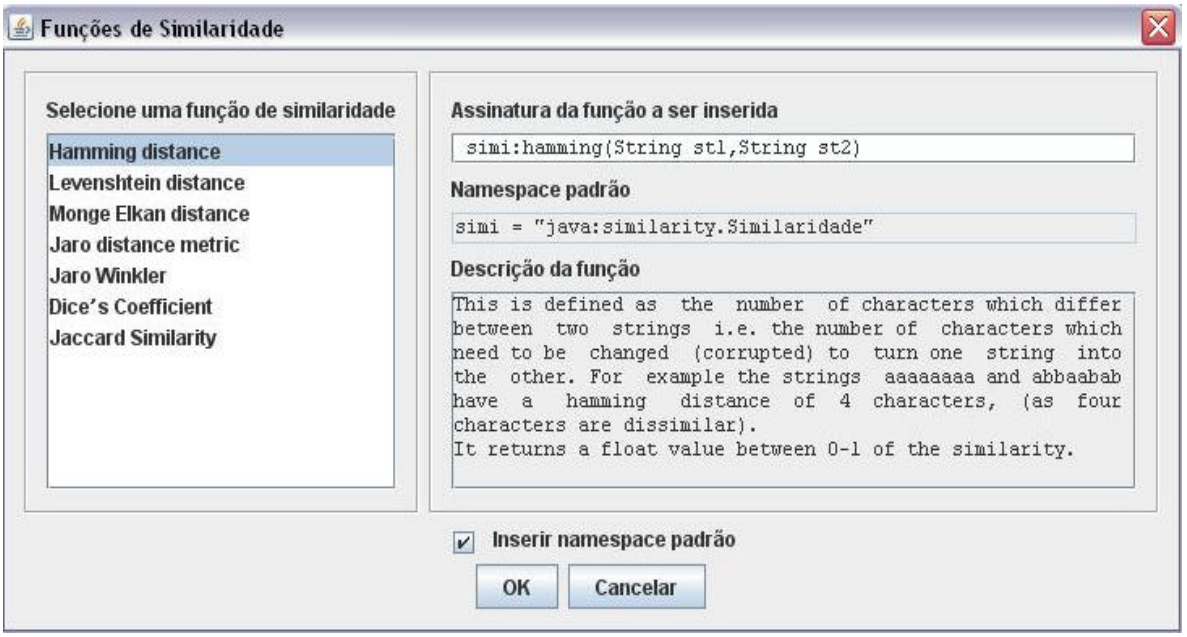

Figura 6: Janela de inserção de função de similaridade

Ao clicar em OK, a assinatura da função, como consta na caixa de assinatura, é inserida na posição do cursor na caixa de consulta da interface principal. Se desejar, o usuário também pode cancelar a inserção da função de similaridade. Nesse caso, a caixa de diálogo é encerrada sem nenhum efeito colateral.

4. Estudo de caso

Nesta seção são apresentados alguns exemplos de consultas que validam o funcionamento da ferramenta desenvolvida. As consultas são executadas sobre parte da base de dados DBLP [Trier s.d.], que é fornecida em formato XML. Foram utilizados 163 elementos *incollection* (que representam artigos científicos) da DBLP, e o tamanho em *bytes* total da base XML consultada era de 187 KB. O tamanho médio de cada elemento foi de aproximadamente 1,15 KB. Uma parte do arquivo de dados pode ser vista no Quadro 1.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE dblp SYSTEM "dblp.dtd">
<dblp>
<incollection mdate="2002-01-03" key="books/acm/kim95/AnnevelinkACFHK95">
<author>Jurgen Annevelink</author>
<author>Rafiul Ahad</author>
<author>Amelia Carlson</author>
<title>Object SQL - A Language for the Design and Implementation of
Object Databases.</title>
<pages>42-68</pages>
<year>1995</year>
<booktitle>Modern Database Systems</booktitle>
<url>db/books/collections/kim95.html#AnnevelinkACFHK95</url>
```

```
</incollection>
<incollection mdate="2002-01-03" key="books/acm/kim95/Blakeley95">
<author>Jos&eacute; A. Blakeley</author>
<title>OQL[C++]: Extending C++ with an Object Query Capability.</title>
<pages>69-88</pages>
<booktitle>Modern Database Systems</booktitle>
<url>db/books/collections/kim95.html#Blakeley95</url>
<year>1995</year>
</incollection>
...
</dblp>
```

Quadro 1 : Base de dados DBLP.

Na Figura 7 tem-se um exemplo de consulta XQuery onde é utilizada a função de similaridade Monge Elkan para recuperar um registro na base da DBLP. Existia uma diferença entre o nome do autor consultado e o nome real desse autor armazenado na DBLP. Entretanto, devido ao uso da função de similaridade Monge Elkan, foi possível recuperar o registro mesmo sem o nome adequado de um dos autores.

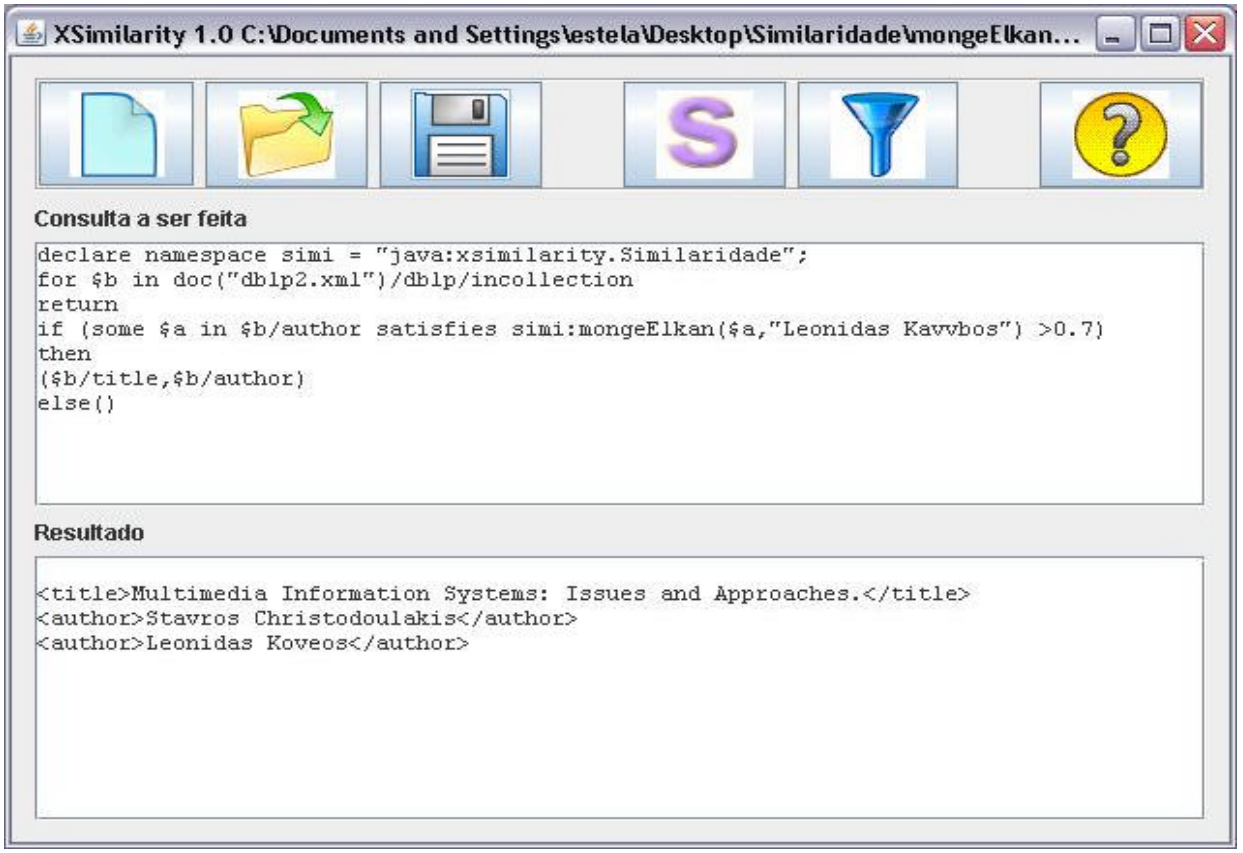

Figura 7 : Exemplo Monge Elkan

O resultado de uma consulta por similaridade depende da função utilizada, do limiar adotado e da base de dados a ser consultada. É possível recuperar resultados indesejados se a função utilizada não for adequada à consulta, ou se o limiar de similaridade não estiver corretamente ajustado. Na Figura 8, utilizou-se o limiar de 0,6 e a função Monge Elkan. A consulta retornou nomes que são considerados semelhantes ao nome “maria”. Entretanto, alguns nomes não eram exatamente esperados.

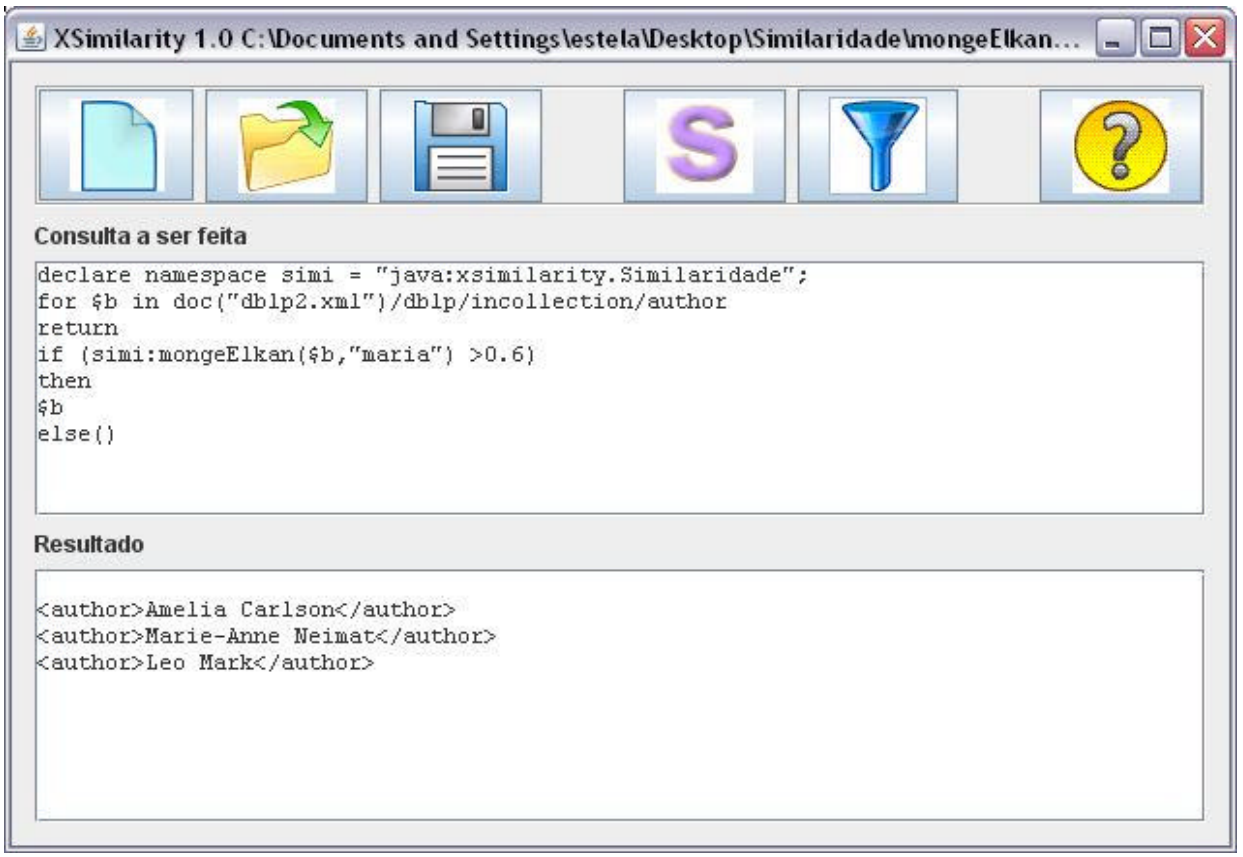

Figura 8: Resultados não esperados

Com o estudo de caso realizado e os resultados alcançados e apresentados, verificou-se o correto funcionamento da ferramenta XSimilarity e sua utilidade para realizar consultas aproximadas e testar funções de similaridade.

5. Conclusões e trabalhos futuros

Este trabalho apresentou o XSimilarity, uma ferramenta para consulta por similaridade embutidas na linguagem XQuery. O XSimilarity implementa funções de similaridade que atuam de formas variadas, sendo uma ferramenta muito útil para realizar buscas. As diversas funções de similaridade tornam a ferramenta bastante abrangente, permitindo maior flexibilidade para experimentos dos mais diversos tipos. O XSimilarity implementa algumas das principais funções de similaridade utilizadas nas áreas de bancos de dados e recuperação de informação.

Quanto à linguagem de consulta XQuery, observa-se que é uma linguagem bastante flexível, pois pôde-se facilmente estendê-la para suportar funções de similaridade. Além disso, funções de similaridade são muito importantes para encontrar registros quando não se conhece precisamente a *string* a ser encontrada e para detectar semelhanças ou diferenças em documentos maiores.

Como trabalhos futuros sugere-se a implementação de novas funções de similaridade na ferramenta e a utilização dessa ferramenta para comparar as funções em termos de eficácia na recuperação de informações. Também se pode aprimorar a funcionalidade de edição de consultas da ferramenta, incluindo recursos como *Syntax Highlight* e a funcionalidade de desfazer. A implementação de mensagens de erro e da funcionalidade de ajuda da ferramenta também são melhorias possíveis. Outra sugestão é a implementação de um mecanismo de paginação para dar suporte a consulta de documentos maiores.

Agradecimentos. Este trabalho foi parcialmente suportado pelo projeto Pronex FAPERGS número 0408933.

Referências

Berkeley Lab. (s.d.). "Nux", http://dsd.lbl.gov/nux/, Novembro de 2007.

Borges, E. N. e Cony, C. A. (2005) "Consultas por Similaridade em SGBDS Comerciais: Estendendo o PostgreSQL." 90f. Projeto de Diplomação (Engenharia de Computação), FURG, Rio Grande.

Borges, E. N. e Dorneles, C. F. (2006) "PgSimilar: Uma ferramenta open source para suporte a consultas por similaridade no PostgreSQL." In: Anais da III Sessão de Demos - Simpósio Brasileiro de Bancos de Dados, Florianópolis. p. 1-6.

Cohen, W., Ravikumar, P., and Fienberg, S. (2003) "A Comparison of String Distance Metrics for Name-Matching Tasks in IIWeb" , 73-78.

Calado, P.P. (2000) "Consultas Aproximadas em Bancos de Dados Relacionais." Dissertação de Mestrado, UFMG, Belo Horizonte.

Chapman, S. (s.d.). "Sam's String Metrics", http://www.dcs.shef.ac.uk/~sam/stringmetrics.html, Novembro de 2005.

Dice, L. R. (1945). "Measures of the Amount of Ecologic Association between Species", In *Journal of Ecology*, 26: 297-302.

Groff, J.R. and Weinberg, P.N. (2002) "SQL: the complete reference", Berkeley, CA: Osborne/ McGrawHill.

Hamming, R. W. (1950) "Error detecting and error correcting codes", In *Bell Syst. Tech. J.*, 29:147–160, 1950.

Jagadish, H. V.; Mendelzon, A. O. and Milo, T. (1995) "Similarity-Based Queries", Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems , May 22-25, San Jose, California. ACM Press. pages: 36-45.

Jaro, M. A. (1989). "Advances in record-linkage methodology as applied to matching the 1985 census of Tampa", Florida. In *Journal of the American Statistical Association*, 84:414–420.

Levenshtein, I. V. (1966). "Binary codes capable of correcting deletions, insertions and reversals", In *Cybernetics and Control Theory*, 10(8):707–710.

Monge, A. E. and Elkan, C. (1996). "The Field Matching Problem: Algorithms and Applications", In *KDD*, 1996: 267-270.

Padilha, A. B. A. (2005) "Suporte a argumentos de consulta vagos através da linguagem XPath", Dissertação de Mestrado, UFRGS, Porto Alegre.

Software Ag. (s.d.). "Tamino – The XML Database", http://www.softwareag.com/tamino, Setembro de 2007.

Trier, University of. (s.d.). "Digital Bibliography & Library Project (DBLP)", http://dblp.uni-trier.de, Novembro de 2007.

Winkler, W. E. (1990). "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage", *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 354-359.

W3C. (s.d.) "XML Query (XQuery) Draft", http://www.w3.org/XML/Query, Novembro de 2007.

W3C. (2006) "Extensible Markup Language (XML) 1.0 (Fourth Edition)", http://www.w3.org/TR/2006/REC-xml-20060816, Setembro de 2007.