UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

RAFAEL BARNI MUNARETTI

Qualificador Lattes: uma ferramenta para a padronização de dados utilizando junção por similaridade textual

Trabalho de Conclusão apresentado como requisito parcial para a obtenção do grau de Bacharel em Informática

Prof. Dr. Carlos Alberto Heuser
Orientador

Porto Alegre, junho de 2008

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Munaretti, Rafael Barni

Qualificador Lattes: uma ferramenta para a padronização de dados utilizando junção por similaridade textual / Rafael Barni Munaretti. – Porto Alegre, 2008.

50 f.: il.

Trabalho de Conclusão (graduação) – Universidade Federal do Rio Grande do Sul. Curso de Bacharelado em Ciência da Computação, Porto Alegre, BR–RS, 2008. Orientador: Carlos Alberto Heuser.

1. Integração de informações. 2. Junção por similaridade. 3. Métricas de similaridade. I. Heuser, Carlos Alberto. II. Título.

“Computer Science is no more about computers
than astronomy is about telescopes.”
— EDSGER W. DIJKSTRA

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CSS	Cascading Style Sheets
CSV	Comma–Separated Values
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ISBN	International Standard Book Number
ISSN	International Standard Serial Number
JSP	JavaServer Pages
MVC	Model–View–Controller
PRG	Post–Redirect–Get
SGBDR	Sistema de Gerenciamento de Bancos de Dados Relacionais
SQL	Structured Query Language
XML	Extensible Markup Language

LISTA DE SÍMBOLOS

Σ	Um alfabeto
Σ^*	O conjunto de todas as palavras possíveis sobre o alfabeto Σ
Φ	Um limiar de similaridade
$\approx_\Phi$	Relação de similaridade para o limiar Φ
$\tilde{\bowtie}_\Phi$	Junção por similaridade para o limiar Φ

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Este trabalho apresenta um estudo e uma aplicação de técnicas de integração de dados textuais utilizando junção por similaridade no escopo de currículos acadêmicos da Plataforma Lattes. Duas técnicas são analisadas: junção por similaridade aproximada baseada em amostragem, executada em um SGBDR; e junção por similaridade exata, executada em memória. A superioridade da segunda técnica para o domínio considerado, demonstrada através de experimentos, leva à sua utilização no sistema Qualificador Lattes. Essa aplicação, apresentada em detalhes, consiste em uma ferramenta com interface web para realizar a padronização da documentação de produção científica de pesquisadores de forma tão automática quanto possível. Os resultados obtidos com a associação de participações em eventos e periódicos científicos contidas em currículos Lattes com as bases de dados oficiais da CAPES são discutidos e avaliados como satisfatórios.

Palavras-chave: Integração de informações, junção por similaridade, métricas de similaridade.

Lattes Qualifier: a tool for data standardization using textual similarity join

ABSTRACT

This work presents an analysis and an application of textual data integration techniques using similarity join applied to the scope of academic résumés on the Lattes Platform. Two techniques are analyzed: the approximate similarity join based on sampling, which is executed inside a relational database management system; and the exact similarity join, which is executed in the main memory. The superiority of the latter, established through experimentation, leads to its utilization in the Lattes Qualifier system. This application, presented in detail, consists of a web tool to perform the standardization of scientific production reports of researchers as automatically as possible. The results obtained with the association of participations in scientific events and periodicals described in Lattes résumés with the official databases from CAPES are discussed and considered satisfactory.

Keywords: information integration, similarity join, similarity metrics.

1 INTRODUÇÃO

A integração de dados textuais heterogêneos é um problema de importância crescente no desenvolvimento de sistemas de informação e na área da Ciência da Computação em geral. Identificar entidades comuns em informações captadas de fontes diversas, abstraindo divergências em suas representações, pode ser interessante por uma série de motivos. A eliminação de ocorrências incompletas e/ou imprecisas na presença de outras mais relevantes e a simplicidade de trabalhar com bases de dados homogêneas são alguns deles.

Nesse contexto, entende-se a junção por similaridade textual como uma forma de integração de dados através da associação de pares de registros que, por alguma métrica previamente definida, podem ser considerados similares. Os registros associados em um par devem, idealmente, representar uma mesma entidade ou conceito, apesar de suas representações potencialmente não serem idênticas.

Este trabalho é focado no problema de integração de dados textuais de duas bases específicas e bastante conhecidas no meio acadêmico: a Plataforma Lattes, do CNPq, e a base Qualis, da CAPES. A Plataforma Lattes é um conjunto de currículos de pesquisadores que descrevem detalhadamente o histórico da produção técnica e científica dos mesmos. A base Qualis é a listagem oficial dos eventos e periódicos científicos reconhecidos pela CAPES, e contém informações importantes como nome padronizado, nível de circulação e avaliação dos diferentes veículos de divulgação de produções.

O objetivo final do trabalho é o desenvolvimento de uma ferramenta que seja capaz de realizar, com intervenção mínima por parte do usuário, a integração das informações sobre participações em eventos e publicações em periódicos contidas em currículos Lattes com as informações oficiais da listagem Qualis. Tal integração é desejável para que seja possível eliminar erros de transcrição e formatação dos currículos acadêmicos e para torná-los mais completos, agregando todas as informações adicionais disponíveis na base Qualis. Uma vez padronizados com os dados oficiais, os currículos podem ser mais facilmente processados por outras ferramentas a fim de armazenar, organizar e gerar informações sobre as produções e os próprios pesquisadores.

Os capítulos seguintes são organizados conforme descrito abaixo.

O segundo capítulo discorre sobre os conceitos e definições da área de integração de dados por similaridade textual. São apresentadas noções fundamentais sobre similaridade de strings, funções de similaridade e associação, junção por similaridade e tokenização.

O terceiro capítulo apresenta em maiores detalhes o contexto deste trabalho e trata das diferentes técnicas de junção por similaridade disponíveis. Os resultados obtidos em execuções de duas técnicas de junção são avaliados e comparados para auxiliar na escolha da técnica a ser utilizada na ferramenta desenvolvida.

O quarto capítulo introduz e descreve em detalhes o sistema Qualificador Lattes e o escopo no qual está inserido. Sua arquitetura, sua utilização e seu desempenho no cumprimento do objetivo para o qual foi proposto são expostos e discutidos.

Finalmente, o quinto capítulo apresenta as conclusões do trabalho e sugere o que pode ser feito no futuro para expandi-lo e aperfeiçoá-lo.

2 CONCEITOS E DEFINIÇÕES

O problema da integração de dados textuais por similaridade envolve uma variedade de conceitos e definições básicas que precisam ser explanados antes que soluções possam ser procuradas ou expostas. Este capítulo pretende introduzir e esclarecer os aspectos fundamentais da área, com ênfase nos pontos mais relevantes para o desenvolvimento deste trabalho.

2.1 Similaridade de strings

Strings, ou cadeias de caracteres, são as unidades primitivas processadas por qualquer aplicação que envolve noções de similaridade textual. Em última análise, a solução para o problema de integração passa por instâncias de uma pergunta aparentemente simples: dadas duas strings quaisquer, elas representam uma mesma entidade ou não?

Computacionalmente, essa é uma decisão extremamente difícil de ser tomada. Duas strings podem ser diferentes e ainda assim representar a mesma entidade. De fato, esse é um caso muito comum em bases de dados reais e a origem de todo o problema. Como exemplo, as strings “CLEI - Conferencia Latinoamericana de Informatica” e “Conferência Latino Americana de Informática (CLEI)”devem claramente representar a mesma entidade, mas diferenças de grafia e formatação fazem com que elas divirjam. Por outro lado, há casos em que diferenças nas representações realmente indicam referências a entidades distintas.

Imediatamente, outro problema se apresenta: o que são entidades distintas? Essa é claramente uma questão semântica dependente do domínio analisado. Em um determinado escopo, as strings “Conferência Latino Americana de Informática (CLEI)” e “ CLEI’2001 - Conferencia Latinoamericana de Informatica” podem representar a mesma entidade: um evento científico realizado periodicamente com as mesmas características. Em outro escopo possível, as entidades referidas pelas strings podem não coincidir por serem ocorrências diferentes de um mesmo evento.

Peculiaridades semânticas à parte, nota-se que a resposta para a questão inicial passa por uma avaliação do nível de semelhança, ou similaridade, existente entre duas strings. Intuitivamente, strings altamente similares têm mais chance de representarem a mesma entidade. É desejado, portanto, um método para estabelecer o grau de similaridade de duas strings — preferencialmente através de um valor numérico. Diversos algoritmos foram projetados para esse propósito e são conhecidos como funções de similaridade.

2.2 Função de similaridade

Uma função de similaridade sim pode ser definida formalmente como

$$sim : \Sigma^* \times \Sigma^* \rightarrow [0, 1]$$

onde Σ^* é o conjunto de todas as palavras, ou cadeias, possíveis sobre um alfabeto Σ.

Assim, quaisquer duas strings têm um grau de similaridade no intervalo fechado contínuo de 0 a 1, sendo o valor 0 equivalente a nenhuma semelhança e o valor 1 equivalente a semelhança absoluta. A tabela 2.1 demonstra valores possíveis de graus de similaridade para algumas strings utilizando uma função de similaridade arbitrária.

Tabela 2.1: Exemplos de graus de similaridade para uma função de similaridade arbitrária

String a	**String** b	$sim(a, b)$
Rafael Barni Munaretti	Rafael Barni Munaretti	1.00
Rafael Barni Munaretti	Rafael Barni	0.80
Rafael Barni Munaretti	Rafael Barni Silva	0.66
Carlos Alberto Heuser	Carlos Alberto Silva	0.66
Carlos Alberto Heuser	Carlos da Silva	0.33
Carlos Alberto Heuser	João da Silva	0.00

O grau de similaridade entre duas strings pode variar conforme a função de similaridade específica utilizada. A escolha de uma função que avalie apropriadamente as strings do domínio sendo tratado, considerando a forma como suas representações interferem em suas propriedades semânticas, é uma parte importante do problema a ser resolvido. A tabela 2.2 apresenta valores para pares de strings avaliados com algumas funções de similaridade bastante conhecidas.

2.3 Associação de strings

O grau de semelhança atribuído por uma função de similaridade a um par de strings não é, por si só, suficiente para definir se as strings correspondem ou não à mesma entidade do domínio. O grau computado deve ser comparado a um limiar de similaridade (*similarity threshold*) previamente escolhido para que o resultado final seja obtido.

Tabela 2.2: Graus de similaridade obtidos com as funções (a) Cosine Similarity, (b) Levenshtein Distance e (c) Overlap Coefficient

String A	String B	(a)	(b)	(c)
Rafael Barni Munaretti	Rafael Barni Munaretti	1.00	1.00	1.00
Rafael Barni Munaretti	Rafael Barni	0.70	0.54	1.00
Rafael Barni Munaretti	Rafael Barni Silva	0.61	0.59	0.68
Carlos Alberto Heuser	Carlos Alberto Silva	0.70	0.71	0.72
Carlos Alberto Heuser	Carlos da Silva	0.31	0.38	0.38
Carlos Alberto Heuser	João da Silva	0.00	0.14	0.00

A função limiar lim_Φ, definida para $0 \leq \Phi \leq 1$ como

$$lim_\Phi : [0,1] \rightarrow \{V, F\} \ tal\ que\ lim_\Phi(x) = \begin{cases} V, & se\ x \geq \Phi \\ F, & se\ x < \Phi \end{cases}$$

em composição com uma função de similaridade sim, resulta na função de associação $assoc_\Phi$

$$assoc_\Phi = lim_\Phi \circ sim : \Sigma^* \times \Sigma^* \rightarrow \{V, F\}$$

cujo resultado indica se duas strings representam (T) ou não (F) uma mesma entidade, para o limiar Φ definido. A partir dessa função, pode-se definir a endorrelação $\langle \Sigma^*, \approx_\Phi \rangle$ como

$$\approx_\Phi : \Sigma^* \rightarrow \Sigma^* \ tal\ que\ a \approx_\Phi b \leftrightarrow assoc_\Phi(a, b) = T$$

que relaciona strings consideradas similares no domínio. Em geral, tal relação é reflexiva, simétrica e não–transitiva. A tabela 2.3 e a figura 2.1 ilustram essas propriedades com alguns exemplos.

É importante perceber a não–transitividade da endorrelação, exemplificada pelos nodos a, b e c. O nodo a se relaciona com b, e b se relaciona com c. Isso significa que a string "Rafael Barni" é considerada similar à string "Rafael Barni Munaretti" e que esta é considerada similar à string "Rodrigo Barni Munaretti". Contudo, a primeira string não é considerada similar à terceira, já que o nodo a não se relaciona com o nodo c.

A escolha do valor a ser utilizado como limiar de similaridade, determinante para que resultados apropriados sejam obtidos, deve ser feita com base no escopo considerado. Valores muito baixos têm grandes chances de gerar funções que associam strings de pouca ou nenhuma concordância semântica. Em contrapartida, valores excessivamente altos podem fazer com que representações díspares da mesma entidade não sejam associadas.

2.4 Junção por similaridade

A não–transitividade da endorrelação $\approx_\Phi$, já demonstrada, dificulta o processo de integração de dados. Por não ser transitiva, $\approx_\Phi$ não pode ser considerada uma relação de

Tabela 2.3: Resultados de uma função de similaridade arbitrária e da função de associação $assoc_{0.5}$ correspondente

Nodo	String	a	b	c	d
a	Rafael Barni	1.00 / T	0.80 / T	0.40 / F	0.00 / F
b	Rafael Barni Munaretti	0.80 / T	1.00 / T	0.80 / T	0.00 / F
c	Rodrigo Barni Munaretti	0.40 / F	0.66 / T	1.00 / T	0.00 / F
d	Carlos Alberto Heuser	0.00 / F	0.00 / F	0.00 / F	1.00 / T

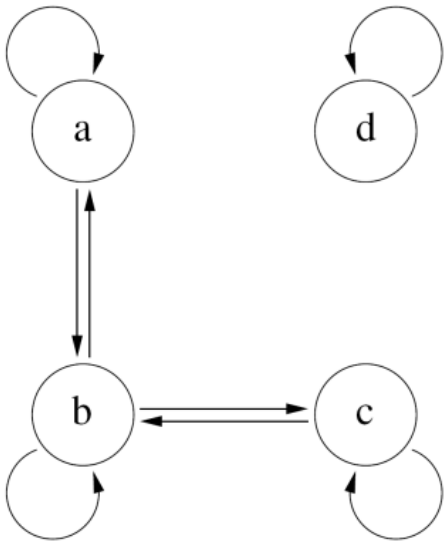

Figura 2.1: Grafo com a endorrelação $\approx_{0.5}$ formada pelos resultados da tabela 2.3

equivalência e, portanto, não induz uma partição de Σ^* em classes de equivalência. Caso isso ocorresse, cada classe de equivalência poderia corresponder a uma entidade específica do domínio e cada elemento de uma classe de equivalência seria considerado uma representação possível para a entidade definida por sua classe. O problema da integração de dados estaria mais próximo de uma solução.

Para contornar essa deficiência, resultado da escolha da função de similaridade, a integração dos dados é realizada através da junção por similaridade. Para dois subconjuntos de Σ^*, A e B, define-se a junção por similaridade $A\tilde{\bowtie}_\Phi B$ como

$$A\tilde{\bowtie}_\Phi B = \{\langle a, b\rangle \mid a \in A \wedge b \in B \wedge a \approx_\Phi b\}$$

Dessa forma, é obtido um conjunto $S \subseteq A \times B$ cujos elementos são todos os pares de elementos de A e B que podem ser considerados similares. Diretamente, tem-se que

$$a \in A \wedge b \in B \wedge assoc_\Phi(a, b) = T \leftrightarrow \langle a, b\rangle \in A\tilde{\bowtie}_\Phi B$$

As tabelas 2.4 e 2.5 demonstram a junção por similaridade.

Por envolver a comparação de cada par de strings pertencente a $A \times B$, a complexidade

Tabela 2.4: Conjuntos A e B da junção por similaridade $A\tilde{\bowtie}_{0.75}B$ apresentada na tabela 2.5

Conjunto	**Elemento**
A	Rafael Barni
A	Rafael Barni Munaretti
A	Rodrigo Barni Munaretti
A	Carlos Alberto Heuser
B	Rafael Munaretti
B	Rafael Barni Munaretti
B	Carlos Alberto
B	Carlos Heuser

Tabela 2.5: Resultados da junção por similaridade $A\tilde{\bowtie}_{0.75}B$ com uma função de similaridade arbitrária para os conjuntos da tabela 2.4

$a \in A$	$b \in B$	$sim(a, b)$
Rafael Barni	Rafael Barni Munaretti	0.80
Rafael Barni Munaretti	Rafael Munaretti	0.80
Rafael Barni Munaretti	Rafael Barni Munaretti	1.00
Carlos Alberto Heuser	Carlos Alberto	0.80
Carlos Alberto Heuser	Carlos Heuser	0.80

do algoritmo que computa $A\tilde{\bowtie}_{\Phi}B$ de forma exata é $O\left(n^2\right)$. Esse é um fator complicador quando as bases de dados envolvidas são especialmente grandes. Uma alternativa ao cálculo exato da junção é discutida no próximo capítulo.

2.5 Tokens

Em algumas funções de similaridade (como, por exemplo, *Cosine Similarity*), a comparação das strings é feita a partir da comparação dos *tokens* das mesmas. *Tokens*, nesse contexto, podem ser entendidos como os segmentos atômicos que formam uma string. A escolha da forma de definição dos tokens, ou *tokenização*, influencia diretamente o resultado da função de similaridade e, portanto, a qualidade da junção por similaridade correspondente.

2.5.1 Palavras

Uma forma de tokenização possível é a simples divisão por palavras. Palavras, aqui, são definidas como seqüências de caracteres separadas por espaços em branco.

A escolha de palavras como tokens em uma função de similaridade pode levar a resultados indesejados. Erros de transcrição e formatação fazem com que instâncias sejam facilmente ignoradas na junção por similaridade, apesar de aparentemente serem bastante semelhantes. A tabela 2.6 demonstra essas possíveis discrepâncias.

Tabela 2.6: Valores de uma função de similaridade arbitrária com tokenização por palavras

String a	**String** b	$sim(a, b)$
Rafael Barni Munaretti	Rafael Barni Munaretti	1.00
Rafael Barni Munaretti	Rafael Barni Munareti	0.66
Rafael Barni Munaretti	Raphael Barni Munareti	0.33
Rafael Barni	Raphael Barni	0.50
Rafael Barni	Raphael Barney	0.00

2.5.2 Q–gramas

Q–gramas são seqüências sobrepostas de q caracteres. A tokenização de uma string por q–gramas gera um conjunto cujos elementos são todas as seqüências de q caracteres existentes na string. Caracteres especiais podem ser utilizados para marcar o início e o final da string.

Ao contrário da tokenização por palavras, o uso de q–gramas em uma função de similaridade pode torná-la mais permissiva em relação a erros de transcrição e formatação. Isso acontece porque tais erros tendem a se concentrar em alguns poucos tokens, não causando uma interferência negativa acentuada em trechos maiores (como palavras). Resistência a esses tipos de erros é especialmente desejável em aplicações cujos dados são obtidos através de interação direta com o usuário. A tabela 2.7 exemplifica a tokenização por q–gramas e a tabela 2.8 demonstra sua utilização em uma função por similaridade arbitrária.

Tabela 2.7: Q–gramas ($q = 3$) geradas para a string “Rafael Barni Munaretti”

Posição	**Q–grama**
1	“##R”
2	“#Ra”
3	“Raf”
4	“afa”
5	“fae”
6	“ael”
7	“el ”
8	“l B”
9	“ Ba”
10	“Bar”
11	“arn”
12	“rni”
13	“ni ”
14	“i M”
15	“ Mu”
16	“Mun”
17	“una”
18	“nar”
19	“are”
20	“ret”
21	“ett”
22	“tti”
23	“ti%”
24	“i%%”

Tabela 2.8: Valores de uma função de similaridade arbitrária com tokenização por q–gramas ($q = 3$)

String a	**String** b	$sim(a, b)$
Rafael Barni Munaretti	Rafael Barni Munaretti	1.00
Rafael Barni Munaretti	Rafael Barni Munareti	0.93
Rafael Barni Munaretti	Raphael Barni Munareti	0.79
Rafael Barni	Raphael Barni	0.75
Rafael Barni	Raphael Barney	0.53

3 CONTEXTO E TÉCNICAS

Conforme a discussão do capítulo anterior, o sucesso de uma tentativa de integração de dados textuais heterogêneos depende de diversas escolhas, como:

- função de similaridade;
- limiar de similaridade;
- método de tokenização;
- técnica de junção.

Este capítulo discute duas técnicas de junção por similaridade, aplicadas ao contexto de currículos acadêmicos e veículos de divulgação. A primeira delas, bastante direta, computa o resultado da junção através da simples aplicação da função de associação de strings; a segunda utiliza algoritmos sofisticados, executados totalmente dentro de SGBDRs, para obter uma aproximação do resultado final.

O objetivo desse estudo é a aquisição de dados comparativos para que seja feita uma escolha consciente da técnica a ser utilizada no sistema Qualificador Lattes. As técnicas devem ser analisadas tanto em relação à qualidade dos resultados como em relação à eficiência da execução. Preocupações com a eficiência, principalmente em termos de tempo de execução, são justificadas pela natureza da ferramenta proposta, discutida no próximo capítulo.

Antes de aprofundar a discussão das diferentes técnicas de junção por similaridade, o contexto no qual elas se inserem no presente trabalho deve ser apresentado.

3.1 Contexto

Este trabalho trata da integração de informações das bases de dados da Plataforma Lattes e do Qualis. Essas bases são, atualmente, bastante importantes e conhecidas na comunidade acadêmica nacional.

3.1.1 Plataforma Lattes

A Plataforma Lattes é a base de dados do CNPq que armazena informações sobre currículos acadêmicos e instituições das áreas de ciência e tecnologia (LATTES, 2008). Os currículos que fazem parte dessa base, conhecidos como *Currículos Lattes*, são o foco de interesse do presente trabalho.

Um Currículo Lattes contém dados sobre a vida acadêmica de um pesquisador. Informações detalhadas sobre produção bibliográfica e técnica, orientações e participações em bancas e eventos fazem parte do modelo padronizado Lattes.

3.1.2 Qualis

O Qualis é o resultado do processo de classificação de veículos de divulgação de produção intelectual concebido pela CAPES (QUALIS, 2008). Ele é disponibilizado no formato de listagens com nomes e informações de classificação de eventos e periódicos.

Cada veículo que consta na base Qualis é classificado em um nível de qualidade e em um nível de circulação. A qualidade pode ser definida por um de três níveis, A, B e C, sendo o nível A atribuído aos veículos de maior reconhecimento. O nível de circulação de um veículo pode ser local, nacional ou internacional. Dessa forma, nove combinações de classificação são possíveis.

Por ser produzido por um órgão do governo brasileiro, o Qualis pode ser considerado uma fonte de informações oficiais acerca dos veículos de divulgação de produção intelectual. Isso leva à sua escolha como parâmetro para a ferramenta de padronização automática de Currículos Lattes apresentada neste trabalho.

3.1.3 Integração das bases

A integração de dados das bases Lattes e Qualis é desejada para que Currículos Lattes possam ser padronizados de forma automática. A padronização pretendida é específica para as informações relativas a trabalhos publicados em periódicos ou apresentados em eventos científicos. O objetivo final é a eliminação de nomes não oficiais de veículos de divulgação do conteúdo de um Currículo Lattes e a incorporação de dados novos e relevantes, como níveis de classificação e identificadores ISBN e ISSN.

Para realizar a integração das duas bases, deve ser feita a junção por similaridade entre os conjunto dos nomes de veículos de divulgação contidos em um Currículo Lattes (L_e e L_p, para *e*ventos e *p*eriódicos) e o conjunto dos nomes de veículos de divulgação contidos na base Qualis (Q_e e Q_p). Assim, serão obtidos os conjuntos $J_e = L_e \tilde{\bowtie}_\Phi Q_e$ e $J_p = L_p \tilde{\bowtie}_\Phi Q_p$, cujos elementos $\langle l, q \rangle$ representam substituições possíveis de nomes de eventos e periódicos não oficiais l por nomes oficiais q.

3.1.4 Trabalhos relacionados

As peculiaridades de diferentes funções de similaridade aplicadas ao escopo Lattes/Qualis são discutidas em (BERRETA, 2006), juntamente com avaliações e comparações detalhadas dos resultados obtidos.

(MARTINS, 2007) aborda a integração das duas bases de dados. Contudo, a ferramenta apresentada não realiza a junção por similaridade de forma automática, requerendo intervenção por parte do usuário para cada entidade a ser associada. Além disso, a arquitetura utilizada inviabiliza a manutenção e expansão do sistema, o que torna sua aplicação em ambientes reais bastante difícil. Para que fosse possível incluir uma nova técnica de junção por similaridade (ou substituir a existente), praticamente todo o sistema precisaria ser refeito. Isso corrobora a importância de uma nova ferramenta.

3.2 Junção por similaridade exata

3.2.1 Visão geral

A técnica de junção por similaridade exata nada mais é do que a aplicação da função de associação descrita no capítulo anterior (seção 2.3) para todos os pares de strings que podem ser formados com elementos dos dois conjuntos envolvidos.

Um algoritmo projetado para esse fim depende de alguns parâmetros essenciais, como a função de similaridade a ser utilizada, o método de tokenização de strings (que neste contexto pode ser abstraído e considerado interno à função de similaridade) e o limiar de similaridade desejado. A listagem 3.1 apresenta tal algoritmo.

Listagem 3.1: Junção por similaridade exata

Entradas: A e B, subconjuntos de Σ^*; função de similaridade sim; limiar Φ
Saída: $R = A \tilde{\bowtie}_{\Phi} B$
1: $R \Leftarrow \varnothing$
2: **para cada** $a \in A$
3: **para cada** $b \in B$
4: **se** $sim(a, b) \geq \Phi$
5: **então** $R \Leftarrow R \cup \{\langle a, b \rangle\}$
6: **retornar** R

3.2.2 Resultados obtidos

Para os experimentos descritos a seguir, a técnica de junção por similaridade exata foi implementada através da linguagem de programação Java (JAVA, 2008). As funções de

similaridade utilizadas foram providas pela biblioteca *SimMetrics* (CHAPMAN, 2007), que contém implementações em Java de diversas das mais conhecidas funções de similaridade existentes.

Todas as execuções foram realizadas com o Currículo Lattes real de um pesquisador da área de Ciência da Computação e a versão mais recente da base Qualis disponível. O currículo contém 85 referências a eventos científicos e 7 referências a periódicos. A base Qualis contém cerca de 500 eventos (foram utilizados somente os eventos da área de Ciência da Computação) e mais de 33.000 periódicos (todas as áreas de pesquisa foram incluídas).

A tabela 3.1 apresenta os tempos de execução obtidos com a técnica de junção por similaridade exata para diferentes funções de similaridade. Já a tabela 3.2 contém alguns dados sobre a qualidade das junções obtidas. Como a técnica gera, por definição, o resultado exato da junção por similaridade, esses dados são relevantes apenas como fator de avaliação das funções de similaridade utilizadas. Um valor $x/y/z$ na tabela 3.2 significa que, de um total de z nomes não oficiais de veículos de divulgação, y receberam uma ou mais sugestões de substituição através da junção por similaridade e x nomes podem ser corretamente substituídos.

Tabela 3.1: Tempos de execução da junção por similaridade exata

Função de similaridade	**Eventos**	**Periódicos**
Cosine Similarity	1560 ms	3940 ms
Levenshtein Distance	5520 ms	10200 ms
Overlap Coefficent	1640 ms	3630 ms

Tabela 3.2: Quantidades de acertos da junção por similaridade exata

Função de similaridade	**Eventos**	**Periódicos**
Cosine Similarity	30 / 46 / 85	6 / 7 / 7
Levenshtein Distance	22 / 33 / 85	6 / 7 / 7
Overlap Coefficent	31 / 61 / 85	5 / 7 / 7

3.3 Junção por similaridade aproximada com amostragem

Para grandes bases de dados, a obtenção de um resultado exato da operação de junção por similaridade pode ser inviável. A complexidade quadrática do algoritmo básico faz com que o tempo de execução cresça rapidamente com o aumento da quantidade de dados envolvidos. A técnica de junção por similaridade aproximada com amostragem (GRAVANO et al., 2003) pretende ser uma alternativa mais adequada para essas situações. Para tanto, são aproveitadas a robustez e a escalabilidade dos SGBDRs, ambientes altamente adaptados para lidar com grandes volumes de dados.

A visão geral da técnica, apresentada a seguir, tem caráter meramente introdutório. Não é seu objetivo substituir o texto da publicação original. Em particular, as expressões SQL utilizadas na implementação do algoritmo não são reproduzidas.

3.3.1 Visão geral

A técnica de junção por similaridade aproximada com amostragem é executada totalmente dentro de um SGBDR, através de expressões SQL padrão. Ela é baseada na função de similaridade *Cosine Similarity* e no esquema de ponderação *tf.idf* (*term frequency, inverse document frequency*).

O esquema *tf.idf*, neste contexto, é utilizado para atribuir um peso a cada token de cada string existente na base de dados. O peso de um token t para uma string s de um conjunto C é elevado se t aparece diversas vezes em s e poucas vezes em C. Dessa forma, tokens raros em um conjunto de strings recebem pesos maiores do que tokens muito comuns, refletindo o grau de diferenciação elevado gerado por eles. No escopo de nomes de veículos de divulgação de produção científica, o token relativamente raro "XML", por exemplo, deve receber um peso maior, em cada string em que aparece, do que o token bastante comum "Conferência".

A função *Cosine Similarity* calcula a semelhança entre duas strings através do ângulo formado entre os vetores dos pesos de seus tokens. Uma string é mapeada em um vetor n-dimensional cujos componentes são os pesos, a ela associados, de todos os tokens contidos na base de dados. Assim, a proximidade de dois vetores representa o grau de similaridade existente entre suas strings.

Pela utilização conjunta da ponderação *tf.idf* e da função de similaridade *Cosine Similarity*, pares de strings que compartilham tokens raros recebem graus de similaridade maiores do que pares de strings que compartilham apenas tokens freqüentes nas bases de dados. Além disso, tokens com pesos reduzidos têm pouca influência no resultado da função de similaridade. Como o vetor dos pesos relacionados a uma string é, em geral, esparso (porque uma base de dados tende a ter muito mais tokens do que cada uma de suas strings e tokens não existentes em uma string recebem peso zero), conclui-se que a maior parte dos tokens de uma string tem influência mínima na definição do seu grau de semelhança a outras strings.

É possível, portanto, obter um resultado aproximado da função de similaridade considerando apenas os tokens com pesos relativamente altos. O restante dos tokens pode ser ignorado, o que reduz drasticamente a quantidade de informações que devem ser processadas e, logo, o tempo de execução da tarefa. Como a influência dos tokens desconsiderados para o grau de similaridade é baixa, a margem de erro da aproximação obtida também o é. Em junções por similaridade com valores de limiar suficientemente grandes,

uma margem de erro pequena deve interferir pouco no resultado final.

A definição do conjunto dos tokens que são considerados no cálculo do grau de similaridade é feita por amostragem. A probabilidade de um token ser escolhido para fazer parte da amostra é proporcional ao seu peso. Com isso, strings que contêm apenas tokens comuns na base de dados (e que, assim, dificilmente seriam avaliadas como altamente similares a outras strings) são descartadas da junção. Em geral, a amostragem é feita unilateralmente, ou seja, em apenas um dos dois conjuntos envolvidos.

3.3.2 Implementação em SQL

Em um banco de dados relacional, os conjuntos de strings envolvidos em uma junção por similaridade são, normalmente, atributos de tabelas distintas. Para que a técnica de junção aproximada com amostragem possa ser realizada, diversas tabelas auxiliares devem ser criadas e populadas com resultados intermediários em diferentes etapas do processamento. Essas etapas, cada uma implementada por uma ou várias expressões SQL executadas em seqüência, são descritas brevemente a seguir.

Na primeira etapa, os tokens de cada conjunto são criados e inseridos em uma tabela auxiliar, juntamente com atributos identificadores de suas strings. A implementação é necessariamente específica para o método de tokenização desejado.

Na segunda etapa, define-se o peso de cada token para cada string em que ele aparece. Para a aplicação do esquema *tf.idf* devem ser utilizadas três outras tabelas, que armazenam a freqüência do token na string, a freqüência inversa do token no conjunto e, por fim, o resultado final.

Na terceira etapa, é realizada a amostragem de um dos conjuntos de tokens. Um maior ganho de eficiência será obtido se o conjunto amostrado for aquele que contiver mais tokens.

Finalmente, na quarta etapa, a junção por similaridade é efetivamente executada. O resultado é um conjunto de tuplas nos mesmos moldes do conjunto de elementos $\langle a, b \rangle$ gerado pela técnica de junção por similaridade exata.

3.3.3 Resultados obtidos

Os testes para a avaliação da técnica de junção por similaridade aproximada com amostragem foram executados com uma implementação das expressões SQL definidas em (GRAVANO et al., 2003) para o SGBDR MySQL (MYSQL, 2008). A etapa de tokenização utiliza a técnica de geração de q–gramas através de SQL introduzida em (GRAVANO et al., 2001), que precisou ser adaptada para os recursos disponíveis no SGBDR escolhido.

Os testes foram executados com as mesmas bases de dados utilizadas para a avaliação da técnica de junção por similaridade exata (seção 3.2.2). As tabelas 3.3 e 3.4 apresentam os resultados. A tabela 3.5 apresenta as quantidades de tuplas geradas nas tabelas auxiliares em cada etapa da execução.

Tabela 3.3: Tempos de execução da junção por similaridade aproximada

Etapa de execução	**Eventos**	**Periódicos**
Etapa 1	94 ms	1312 ms
Etapa 2	17437 ms	4338969 ms
Etapa 3	8297 ms	826187 ms
Etapa 4	11313 ms	2719 ms
Total	37141 ms	5169187 ms

Tabela 3.4: Quantidades de acertos da junção por similaridade aproximada

Eventos	**Periódicos**
24 / 39 / 85	0 / 1 / 7

Tabela 3.5: Quantidades de tuplas armazenadas em tabelas auxiliares

Etapa de execução	**Eventos**	**Periódicos**
Etapa 1	32972	971737
Etapa 2	68276	1925832
Etapa 3	14218	60663
Total	115466	2958232

3.4 Comparação das técnicas

Com os dados coletados durante as execuções das duas técnicas de junção por similaridade e apresentados nas seções anteriores, a decisão de qual técnica utilizar em uma ferramenta de integração de dados não é difícil de ser tomada. Os tempos de execução e a qualidade dos resultados podem ser avaliados de forma direta e objetiva para guiar a escolha.

A avaliação da eficiência da execução, mesmo considerando apenas o tempo utilizado e desprezando a quantidade de memória consumida, sugere fortemente que a técnica de junção aproximada com amostragem não é adequada para a integração das bases Lattes e Qualis. Contudo, deve-se notar que a maior parte do processamento é feita sobre os dados da base Qualis e independe do Currículo Lattes específico que deve ser padronizado. Assim, pode-se executar as etapas de criação de tokens, atribuição de pesos e amostragem das informações Qualis uma única vez, utilizando os mesmos resultados para todas as padronizações posteriores. Infelizmente, a etapa final, que deve obrigatoriamente fazer parte de cada execução, ainda é mais demorada do que todo o processamento necessário para a técnica de junção exata.

Quando é avaliada a qualidade dos resultados obtidos, a junção aproximada fica novamente em desvantagem. Porém, pela própria motivação da técnica, a perda de qualidade é esperada. Por definição, nenhuma técnica que envolve aproximações poderia gerar resultados melhores do que aqueles gerados pela junção exata.

Conclui-se, com esse breve estudo comparativo, que a técnica de junção por similaridade exata é mais apropriada para a integração das bases abordadas neste trabalho. A superioridade dessa técnica em relação à técnica de junção aproximada com amostragem é evidente tanto em termos de eficiência como em termos de qualidade. Dessa forma, é justificada sua utilização no sistema Qualificador Lattes, introduzido no próximo capítulo.

4 QUALIFICADOR LATTES

Conforme descrito no capítulo anterior, o foco deste trabalho é a padronização de Currículos Lattes segundo o Qualis, a base de dados oficial de veículos de divulgação de produção intelectual. Este capítulo introduz o sistema Qualificador Lattes, uma ferramenta que faz uso do conceito de similaridade de strings e, mas especificamente, de técnicas de junção por similaridade textual para realizar automaticamente a padronização desejada através da integração das bases envolvidas.

4.1 Arquitetura e tecnologias empregadas

Dado o cenário atual da computação, bastante focado em modelos distribuídos através da Internet, optou-se por desenvolver o Qualificador Lattes como uma aplicação Web. As vantagens desse modelo sobre o modelo tradicional de software executado na máquina do usuário são diversas e já bem conhecidas: distribuição e instalação são facilitadas (normalmente, não necessárias), atualizações atingem todos os usuários automaticamente, processamento independe da capacidade das máquinas que acessam o sistema, etc.

Uma aplicação Web é caracterizada pela arquitetura cliente–servidor. Cada um dos dois componentes, no entanto, tem sua própria arquitetura interna e utiliza tecnologias distintas. Detalhes sobre o lado cliente e o lado servidor do Qualificador Lattes são apresentados a seguir.

4.1.1 O lado cliente

Como acontece com todas as aplicações Web, o usuário do Qualificador Lattes deve acessar o sistema através de um *browser*. A interface gráfica é composta por páginas HTML que permitem que o usuário interaja com links e controles como botões e campos de entrada de dados e visualize os resultados do processamento.

Para tornar a experiência do usuário mais amigável, reduzindo tempos de resposta da interface e adicionando dinamismo a páginas HTML, que são normalmente estáticas,

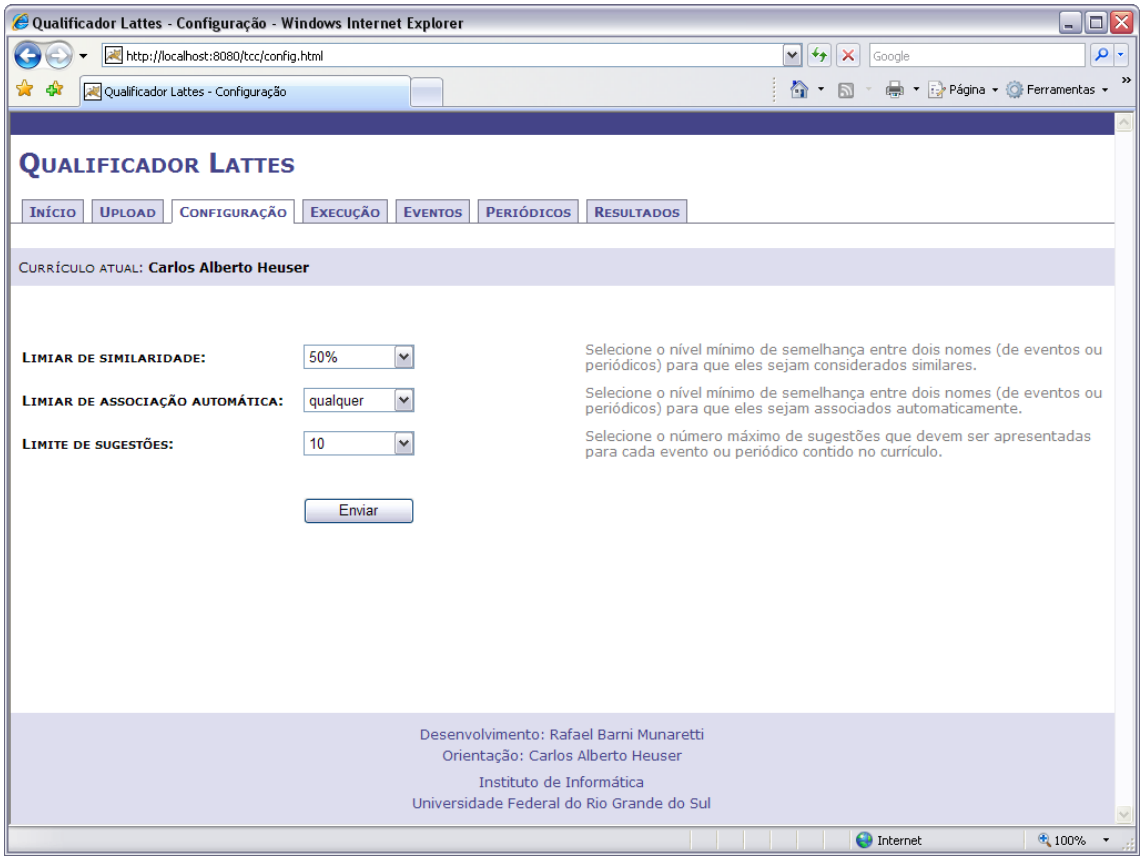

Figura 4.1: Interface do sistema Qualificador Lattes

são utilizadas a linguagem de scripting *JavaScript* e técnicas *Ajax*. Além disso, uma padronização visual, importante para a familiarização do usuário, é atingida com o uso de CSS. A figura 4.1 demonstra o aspecto gráfico da interface com uma das telas do sistema.

4.1.2 O lado servidor

Apesar de preocupações com interface e experiência do usuário não poderem ser ignoradas, é no lado servidor que o Qualificador Lattes realmente se concentra. Ao contrário de páginas Web tradicionais, o papel do servidor em uma aplicação Web envolve mais do que apenas receber requisições HTTP e enviar como resposta documentos HTML. Fica sob responsabilidade do servidor as tarefas de manutenção de sessões de usuário, invocação da camada de aplicação para processamento dos dados e geração do conteúdo dinâmico HTML que deve ser transferido para os clientes.

A camada de aplicação do servidor foi desenvolvida em Java e utiliza a tecnologia JSP para a geração dinâmica de conteúdo. Essa camada, chamada daqui em diante simplesmente de "aplicação", deve ser executada sobre um *Servlet container*, que também atua como servidor HTTP. Durante o desenvolvimento do Qualificador Lattes, o *container* utilizado foi o *Apache Tomcat*. Pequenas alterações na aplicação podem ser necessárias para

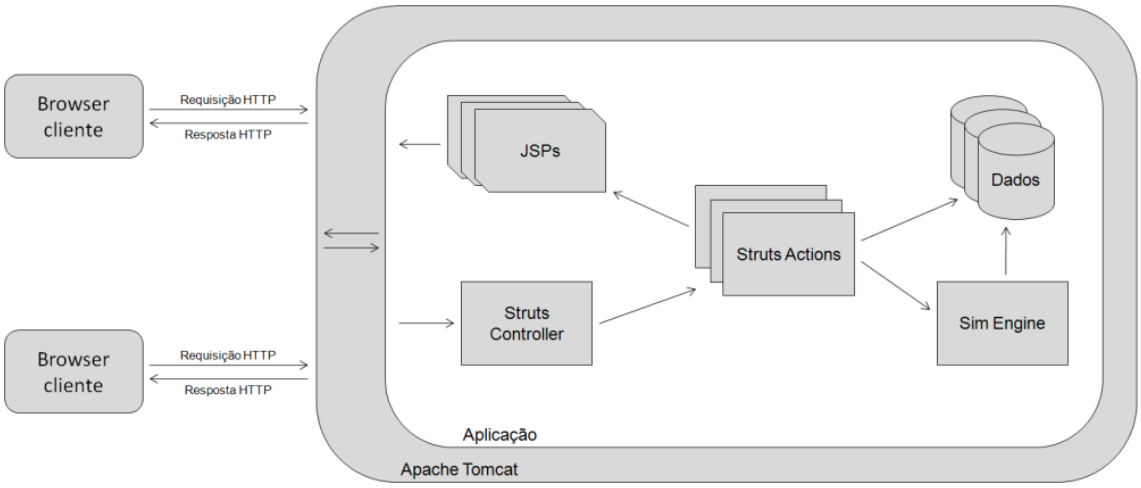

Figura 4.2: Diagrama arquitetural do sistema Qualificador Lattes

seu correto funcionamento em outros *containers*.

A arquitetura da aplicação segue o padrão MVC, com a utilização do *framework Apache Struts*. Também é empregado o padrão PRG, conhecido como *"redirect–after–post"*, para contornar alguns inconvenientes da utilização do protocolo HTTP pelos *browsers* atuais (JOURAVLEV, 2004). A figura 4.2 apresenta uma visão geral da arquitetura do sistema.

4.2 Integração de dados

O objetivo do Qualificador Lattes é eliminar de um Currículo Lattes referências imprecisas a veículos de divulgação de produção intelectual, substituindo-as pelos nomes oficiais que constam na base Qualis através da integração dos dados. Contudo, por motivos diversos, essa tarefa nem sempre pode ser realizada de forma totalmente automática. Entre os motivos mais comuns pode-se citar o estado bastante incompleto da base Qualis disponível atualmente e a utilização intensa, em Currículos Lattes, de siglas ambíguas no lugar de nomes de eventos e periódicos científicos.

Para contornar esse problema, o Qualificador Lattes utiliza a técnica de junção por similaridade exata (seção 3.2), a função de similaridade *Cosine Similarity* e o método de tokenização por q–gramas (com $q = 3$) para gerar *sugestões* de substituições possíveis de nomes não oficiais por seus correspondentes oficiais. O usuário tem a responsabilidade de avaliar as sugestões disponíveis e selecionar as associações válidas, ou seja, as substituições que realmente devem ser feitas no conteúdo do currículo. Antes disso, porém, algumas configurações da integração de dados devem ser definidas para direcionar a execução.

4.2.1 Configurações

Para tornar a integração de dados flexível e adaptável às necessidades de cada caso, o Qualificador Lattes permite que o usuário configure a execução de forma a realizar uma pré–seleção dos resultados. Podem ser definidos valores para o *limiar de similaridade*, o *limiar de associação automática* e o *limite de sugestões*.

O limiar de similaridade já foi abordado na seção 2.3. No Qualificador Lattes, os nomes oficiais de veículos de divulgação sugeridos como substituições têm, obrigatoriamente, graus de similaridade com os nomes oficiais correspondentes maiores ou iguais ao limiar de similaridade utilizado na execução da junção. O valor do limiar de similaridade é representado por uma percentagem e deve ser selecionado pelo usuário dentre algumas opções previamente fixadas. A utilização de valores elevados contribui para filtrar sugestões irrelevantes, mas também pode ser prejudicial por causar a omissão de sugestões potencialmente importantes para nomes que no currículo específico têm representações muito distintas daquelas contidas na base Qualis.

O limiar de associação automática serve como um mecanismo para reduzir a necessidade de intervenção do usuário. Ele é utilizado para definir quais situações caracterizam associações automáticas. O valor do limiar de associação automática pode ser selecionado entre algumas porcentagens válidas, mas, ao contrário do limiar de similaridade, sua utilização é opcional. Mais detalhes na seção 4.2.4.

O limite de sugestões é um simples filtro para a quantidade de sugestões apresentadas para cada nome de evento ou periódico do currículo. Se um determinado nome recebe um número de sugestões superior ao limite especificado, apenas as melhores fazem parte do resultado. A utilização de um limite de sugestões é obrigatória.

4.2.2 Sugestões

Conforme já foi mencionado superficialmente, a impossibilidade de padronização automática levou à necessidade de serem feitas sugestões de substituição para os nomes de eventos e periódicos contidos em Currículos Lattes. Uma sugestão, portanto, é um elemento de um dos conjuntos resultantes das junções por similaridade realizadas entre nomes que constam em um currículo e nomes oficiais.

Cada sugestão representa uma substituição possível para um nome, e múltiplas sugestões podem se referir ao mesmo nome. A existência de diversas sugestões para um só nome acontece quando mais de um nome oficial é suficientemente semelhante ao nome contido no currículo. É justamente essa relação de "um para muitos" existente entre nomes e sugestões que impossibilita a automatização total da integração dos dados.

Por outro lado, também é possível que determinados nomes não recebam quaisquer

sugestões de substituição. Isso acontece quando a representação do nome do evento ou periódico no currículo difere tanto de todos os nomes contidos na base Qualis que nenhum par recebe um grau de similaridade maior ou igual ao limiar especificado para a execução da junção. Claramente, nomes que não recebem sugestões não podem ser substituídos no currículo e devem ser mantidos mesmo não sendo considerados oficiais.

4.2.3 Associações

Em um conjunto de sugestões que se referem a um determinado nome, apenas uma pode ser efetivada como substituição real de um nome não–oficial, potencialmente impreciso, por seu correspondente oficial da base Qualis. A sugestão escolhida como correta é chamada "associação", por representar a ligação concreta entre os nomes.

É importante notar que a existência de uma ou mais sugestões para um nome não implica necessariamente na existência de uma associação para o mesmo. No caso de diversas sugestões de substituição para um nome existirem mas não corresponderem semanticamente ao mesmo evento ou periódico que o nome representa, não é obrigatória (ou aconselhada) a escolha de uma delas como associação verdadeira. O nome em questão simplesmente é mantido sem associação, o que não deixa de ser um cenário válido.

No momento da geração do resultado final do Qualificador Lattes, que é o Currículo Lattes padronizado, somente as associações realizadas são levadas em consideração para a substituição dos nomes de eventos e periódicos. Dessa forma, nomes que não receberam sugestões ou não tiveram nenhuma de suas sugestões associada são reproduzidos no resultado com a mesma representação em que se encontram no documento original.

4.2.4 Associação automática

Para evitar que o usuário seja obrigado a definir manualmente cada associação entre nomes a partir das sugestões geradas pela junção por similaridade, foi concebido o mecanismo de associação automática. Com sua utilização — opcional — o usuário pode definir um limiar acima do qual os graus de similaridade de sugestões são considerados suficientemente altos para que elas sejam tornadas, automaticamente, associações.

Como um nome pode receber muitas sugestões, também é possível que mais de uma delas tenha um grau de similaridade superior ao limiar de associação automática definido pelo usuário. Nesse caso, a sugestão que tiver o maior grau de similaridade é escolhida como associação automática.

Se nenhuma das sugestões de um nome tem um grau de similaridade igual ou superior ao limiar de associação automática, fica a cargo do usuário decidir qual (se alguma) das sugestões deve ser associada. Quando um nome não recebe quaisquer sugestões, tam-

pouco pode ser associado automaticamente.

A tabela 4.1 ilustra os conceitos de sugestão, associação e associação automática com um exemplo retirado de uma execução real da integração de dados realizada pelo sistema Qualificador Lattes. A execução foi configurada para utilizar um limiar de similaridade igual a 50%, um limiar de associação automática igual a 80% e um limite de 10 sugestões por nome de evento ou periódico.

Tabela 4.1: Sugestões geradas para o nome de evento "International Conference on Application and Theory of Petri Nets"

Nome oficial sugerido	**Similaridade**	**Associação automática**
International Conference on Applications and Theory of Petri Nets (ICATPN)	92,2%	✓
International Conference on Application and Theory of Petri Nets and Other Models Of Councurrency	83,9%	×
International Conference on Computer Vision Theory and Applications	73,0%	×
International Conference on Computer Graphics Theory and Applications	70,7%	×
International Conference on Internet and Web Applications and Services	64,4%	×
International Conference on the Practice and Theory of Automated Timetabling	61,7%	×
Application and Theory of Petri Nets and Other Models of Concurrency (ATPN)	61,5%	×
International Conference on Advanced Information Networking and Applications (AINA)	60,5%	×
International Conference on Field Programmable Logic and Applications	59,6%	×
International Conference on Virtual Reality Continuum and Its Applications	58,7%	×

4.3 Utilização

Os usuários–alvo do sistema Qualificador Lattes são os pesquisadores da área de Ciência da Computação que têm seus currículos cadastrados na Plataforma Lattes. Para poder utilizar a ferramenta, um pesquisador deve antes acessar seu Currículo Lattes e exportá-lo para o formato XML. Essa conversão está disponível na própria Plataforma Lattes e permite que todas as informações de um currículo sejam representadas em um formato independente de aplicação.

Em posse do arquivo XML contendo seu currículo, o pesquisador pode acessar o sistema Qualificador Lattes através de seu *browser*. Uma sessão de uso é iniciada quando o usuário faz o *upload* de um arquivo. Em uma sessão de uso apenas um currículo é analisado e, potencialmente, padronizado. Para trabalhar com outro currículo, o usuário deve voltar à etapa de *upload* e selecionar o novo arquivo.

Uma vez iniciada a sessão de uso, pode-se definir as configurações desejadas para a geração das sugestões e associações automáticas e iniciar a execução da junção por similaridade. Em qualquer etapa posterior, o usuário pode voltar à fase de configuração e iniciar uma nova execução, com valores de configuração diferentes, sem precisar refazer o *upload* do currículo. Contudo, sugestões e associações (automáticas ou não) da execução anterior serão perdidas.

Ao término da execução, o usuário tem a opção de revisar as associações automáticas realizadas para identificar possíveis incorreções. Quando o mecanismo de associação automática não é utilizado, o usuário deve analisar todas as sugestões e realizar as associações necessárias. As sugestões são apresentadas convenientemente divididas em grupos relativos a cada evento ou periódico referenciado no currículo e ordenadas por seus graus de similaridade com os nomes originais. Em qualquer momento antes do final da sessão de uso, associações podem ser feitas e desfeitas.

Quando o usuário estiver satisfeito com as associações realizadas, pode fazer o *download* da versão padronizada de seu currículo. Essa versão contém todas as informações do documento original, exceto os nomes de eventos e periódicos que tenham sido associados. Nos lugares desses nomes aparecem as representações oficiais selecionadas. Adicionalmente, são inseridas, nos locais apropriados, informações como ISBN (para eventos) e ISSN (para periódicos), nível de circulação e nível de qualidade dos veículos de divulgação de produção científica. Todas essas informações adicionais fazem parte da base Qualis.

A figura 4.3 contém o diagrama de casos de uso do sistema. O detalhamento dos casos de uso, ilustrados com as telas reais, é feito no Apêndice.

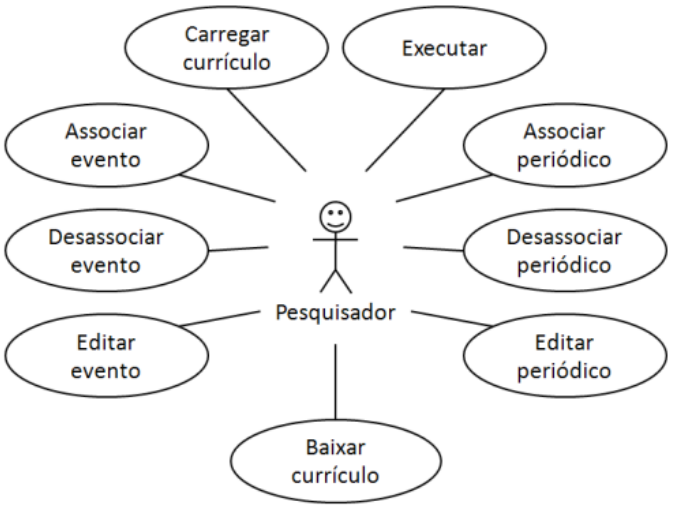

Figura 4.3: Diagrama de casos de uso do sistema Qualificador Lattes

4.4 Implementação

Essa seção discute alguns detalhes relevantes da implementação do Qualificador Lattes. Não é seu objetivo, porém, descrever o processo de desenvolvimento integralmente.

4.4.1 Otimizações

Como já mencionado, é empregada a função de similaridade *Cosine Similarity* e o método de tokenização por q–gramas na junção por similaridade do Qualificador Lattes. Inicialmente, a implementação da função provida pela biblioteca *SimMetrics* foi utilizada. Como essa biblioteca é de código aberto, a análise de seu código fonte levou à possibilidade de realizar otimizações no mesmo. Visto que uma só execução da junção faz dezenas ou centenas de milhares de chamadas à função de similaridade, mesmo alterações aparentemente insignificantes na função podem ser cruciais para a redução do tempo total de execução.

A primeira otimização realizada é puramente relativa à forma de programação. A função original tokeniza internamente as strings que estão sendo avaliadas e instancia alguns objetos intermediários para armazenar os resultados da tokenização. A construção desses objetos foi removida do código, fazendo com que a etapa de tokenização gerasse diretamente os objetos finais utilizados pela função de similaridade. Com isso, poupa-se o tempo de alocação de memória e transferência dos dados.

A segunda otimização vai além da programação da função de similaridade e se caracteriza por uma troca entre o tempo de execução e a memória utilizada pela junção por similaridade. Em uma execução da junção, a função de similaridade é chamada para cada par de strings dos conjuntos envolvidos. Se a geração dos tokens é realizada internamente à função de similaridade, o processo é desnecessariamente repetido muitas vezes para uma mesma string. Para evitar que esse tempo seja desperdiçado, todas as strings das duas bases passaram a ser tokenizadas apenas uma vez, em momentos apropriados. No caso da base Qualis, o momento da tokenização é a inicialização do servidor. Os tokens gerados são mantidos em memória e posteriormente compartilhados por todas as execuções de todos os usuários. No caso dos Currículos Lattes, a tokenização é feita durante a junção, de forma que os tokens de cada nome possam ser gerados uma única vez, utilizados em todas as comparações com a base Qualis e logo descartados.

A redução no tempo de execução da junção por similaridade causada pelas otimizações descritas chegou a 66%. A tabela 4.2 contém os dados de tempos de execução com a implementação original e com cada uma das otimizações. O consumo de memória pela aplicação depois da segunda otimização aumentou em 60 megabytes, cerca de 110%. Como o sistema deve rodar normalmente em um servidor dedicado, o maior consumo de

Tabela 4.2: Tempos de execução da junção por similaridade com a implementação original e com otimizações

Execução	Eventos	Periódicos
Original	1,56 s	5,53 s
Otimização 1	0,78 s	2,58 s
Otimização 2	0,52 s	1,88 s

memória não representa um problema grave.

4.4.2 Concorrência

Uma aplicação Web comum consegue atender múltiplas requisições simultâneas de forma natural. Isso acontece porque cada requisição pode ser tratada independentemente das demais, já que dificilmente o processamento de uma interfere nos resultados de outra. No Qualificador Lattes, contudo, a concorrência de requisições de um mesmo usuário precisou de tratamento específico.

A causa do problema é justamente o núcleo funcional do sistema: a execução da junção por similaridade. Por ser uma operação relativamente demorada, é possível que o usuário faça novas requisições enquanto a junção é executada. Para evitar que múltiplas requisições modifiquem constantemente o estado da execução, o componente que realiza a operação (*Sim Engine* na figura 4.2) foi modelado como uma máquina de estados finitos. As transições entre os estados podem corresponder a novas requisições ou ao término do processamento de requisições pendentes.

Toda vez que uma requisição precisa interagir com o componente de execução da junção, seja para tentar realizar uma transição de estados ou simplesmente para consultar o estado atual, ela deve disputar acesso a uma seção crítica. Dentro da seção crítica, transições de estados podem ser controladas de forma seqüencial, garantindo a integridade do estado do componente. Quando uma requisição consegue realizar uma transição de estados, ela tem a garantia de que nenhuma outra requisição interferiu simultaneamente no estado e nem irá conseguir interferir até o final de seu processamento.

5 CONCLUSÃO

Este trabalho apresentou um estudo de técnicas de junção por similaridade textual voltado ao desenvolvimento de uma ferramenta de padronização de dados em um domínio específico. Os resultados obtidos com a técnica de junção por similaridade exata e com a técnica de junção aproximada com amostragem levaram à conclusão de que a primeira é superior quando aplicada ao escopo de nomes de veículos de divulgação de produção científica das bases Lattes e Qualis.

A ferramenta apresentada tem o objetivo de integrar os dados das duas bases para realizar a padronização de Currículos Lattes, que podem conter referências imprecisas a eventos ou periódicos científicos. Com o conceito de similaridade de strings, imprecisões de representação podem ser abstraídas para identificar entidades semanticamente iguais.

Além de corrigir eventuais erros de grafia e formatação, o sistema Qualificador Lattes permite que uma variedade de informações oficiais disponíveis na base Qualis seja incorporada ao currículo do usuário. Entre essas informações, destacam-se os indicadores de nível de circulação e qualidade de eventos e periódicos.

Os resultados obtidos com testes do sistema utilizando currículos de pesquisadores reais mostram que grande parte das referências a veículos de divulgação pode ser padronizada com sucesso. Os tempos de resposta observados são bastante aceitáveis para uma aplicação Web, mantendo-se sempre na ordem de poucos segundos.

Um fator importante a ser considerado ao avaliar os resultados é o estado atual da base de dados Qualis. Infelizmente, as listagens de eventos e periódicos científicos estão longe de serem completas, o que prejudica a padronização dos currículos. Espera-se que, com o tempo, ambas as listagens sejam aperfeiçoadas com a adição de novos itens.

5.1 Trabalhos futuros

O sistema Qualificador Lattes pode ser aperfeiçoado com algumas funcionalidades que não foram implementadas simplesmente por restrições de tempo. A compressão de

arquivos, tanto para *upload* dos currículos originais como para *download* dos resultados, pode ser interessante para usuários que dependem de conexões com baixa largura de banda. A disponibilização de *web services* pode tornar viável a utilização do Qualificador Lattes por outras ferramentas.

Como tópico de pesquisa futura, pode-se mencionar um estudo detalhado das diferentes funções de similaridade aplicadas ao contexto Lattes/Qualis. Especialmente, pode-se imaginar composições de funções existentes para avaliar com maior precisão a semelhança entre strings desse domínio específico. Com os experimentos realizados, percebeu-se que detalhes como abreviaturas e omissão de palavras comuns são fatores importantes para o fracasso de algumas associações. O estudo ou desenvolvimento de técnicas específicas para esses casos pode melhorar significativamente os resultados obtidos com a junção por similaridade.

Por fim, deve-se ressaltar a necessidade de realizar testes do sistema desenvolvido com grupos de usuários reais. A conformidade da interface precisa ser validada a partir de sua apresentação a pessoal não treinado. Também é importante analisar a resistência da aplicação a cargas elevadas, geradas por múltiplos usuários ativos simultaneamente.

REFERÊNCIAS

BERRETA, E. C. **Uma aplicação de integração de dados por similaridade**. 2006. 48f. Trabalho de Conclusão (Bacharelado em Ciência da Computação). Instituto de Informática, UFRGS, Porto Alegre.

CHAPMAN, S. **SimMetrics - open source Similarity Measure Library**. Disponível em <http://www.dcs.shef.ac.uk/ sam/simmetrics.html>. Acesso em junho de 2008.

GRAVANO,L.; IPEIROTIS, P. G.; JAGADISH, H.; KOUDAS, N.; MUTHUKRISHNAN, S.; PIETARINEN, L.; SRIVASTAVA, D. **Using q-grams in a DBMS for approximate string processing**. In: IEEE Data Engineering Bulletin, 2001.

GRAVANO, L; IPEIRIOTIS, P. G.; KOUDAS, N.; SRIVASTAVA, D. **Text Joins in an RDBMS for Web Data Integration**. In: 12th International World Wide Web Conference, p. 90-101, 2003.

JAVA. **Developer Resources for Java Technology**. Disponível em <http://java.sun.com/>. Acesso em junho de 2008.

JOURAVLEV, M. **Enterprise Java Community: Redirect After Post**. Disponível em <http://www.theserverside.com/tt/articles/article.tss?l=RedirectAfterPost>. Acesso em junho de 2008.

LATTES. **Plataforma Lattes**. Disponível em <http://lattes.cnpq.br/>. Acesso em junho de 2008.

MARTINS, S. C. **Integração de registros de bancos de dados sem uma chave primária comum**. 2007. 44f. Trabalho de Conclusão (Bacharelado em Ciência da Computação). Instituto de Informática, UFRGS, Porto Alegre.

MYSQL. **MySQL :: The world's most popular open source database**. Disponível em <http://www.mysql.com/>. Acesso em junho de 2008.

QUALIS. **Classificação de Periódicos, Anais, Revistas e Jornais**. Disponível em <http://qualis.capes.gov.br/webqualis/>. Acesso em junho de 2008.

APÊNDICE

CASOS DE USO ILUSTRADOS DO SISTEMA QUALIFICADOR LATTES

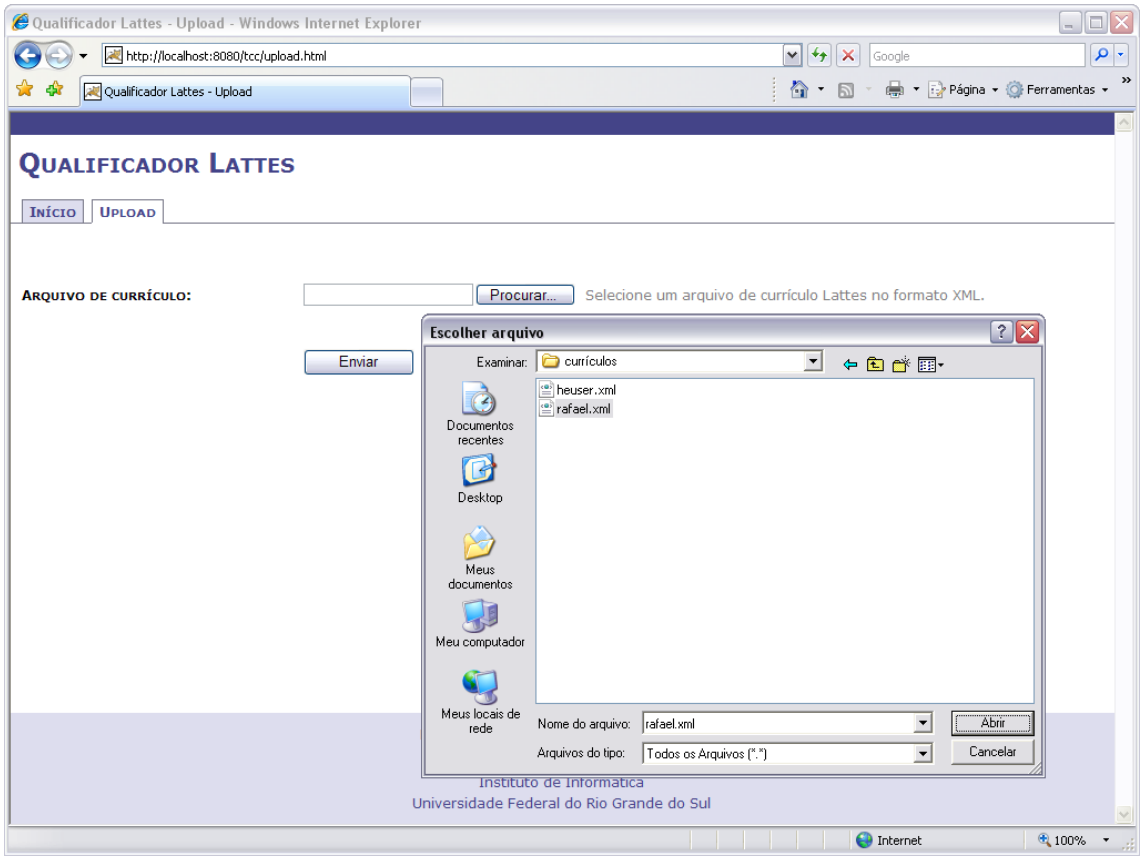

Figura A.1: Tela de upload de currículo, com janela de seleção de arquivo

A.1 Carregar currículo

Pré-condições: nenhuma.

Pós-condições: currículo carregado, tela de configurações de execução (figura A.2).

1. Acessar o sistema Qualificador Lattes;
2. Pelo menu, acessar tela de upload de currículo (figura A.1);
3. Na janela de seleção de arquivo, selecionar um currículo Lattes no formato XML;
4. Clicar no botão “Enviar”.

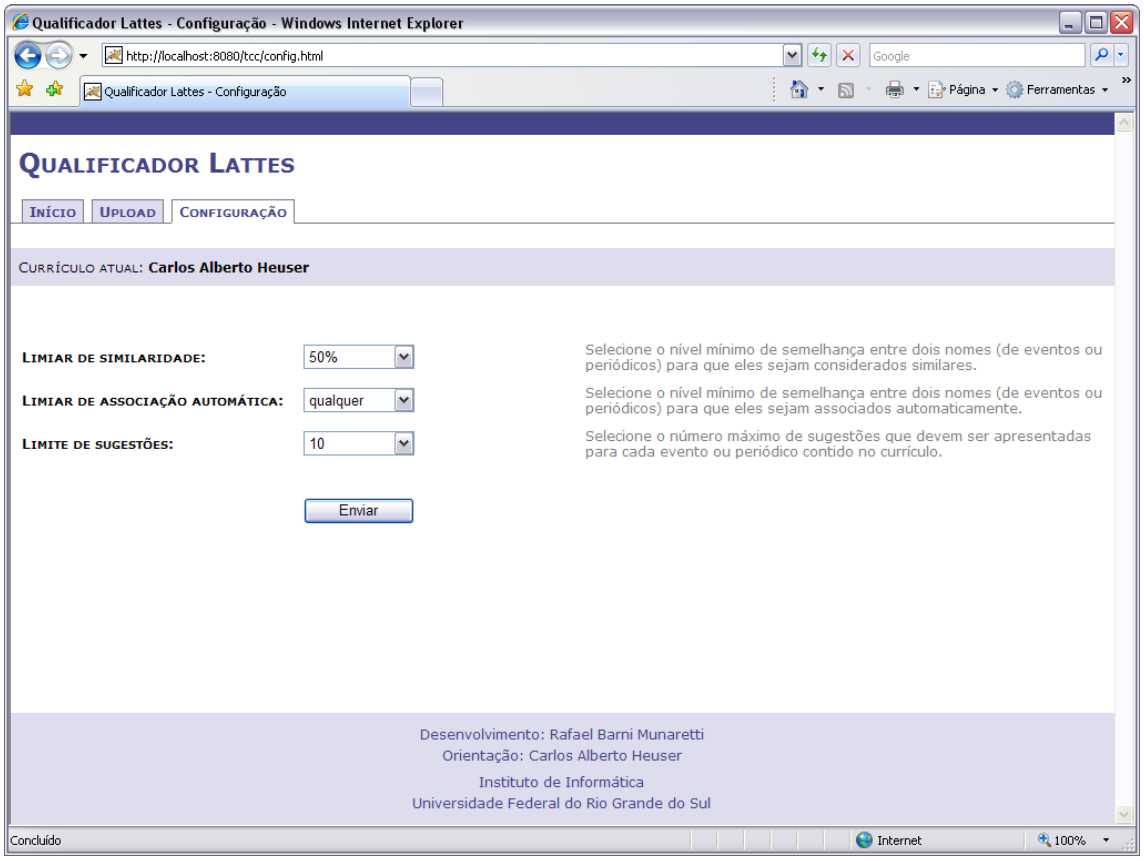

Figura A.2: Tela de configurações de execução

A.2 Executar

Pré-condições: currículo carregado, tela de configurações de execução (figura A.2).

Pós-condições: execução concluída, tela de progresso de execução (figura A.3).

1. Selecionar uma opção da lista para o limiar de similaridade;
2. Selecionar uma opção da lista para o limiar de associação automática;
3. Selecionar uma opção da lista para o limite de sugestões;
4. Clicar no botão “Enviar”;
5. Aguardar preenchimento completo das barras de progresso.

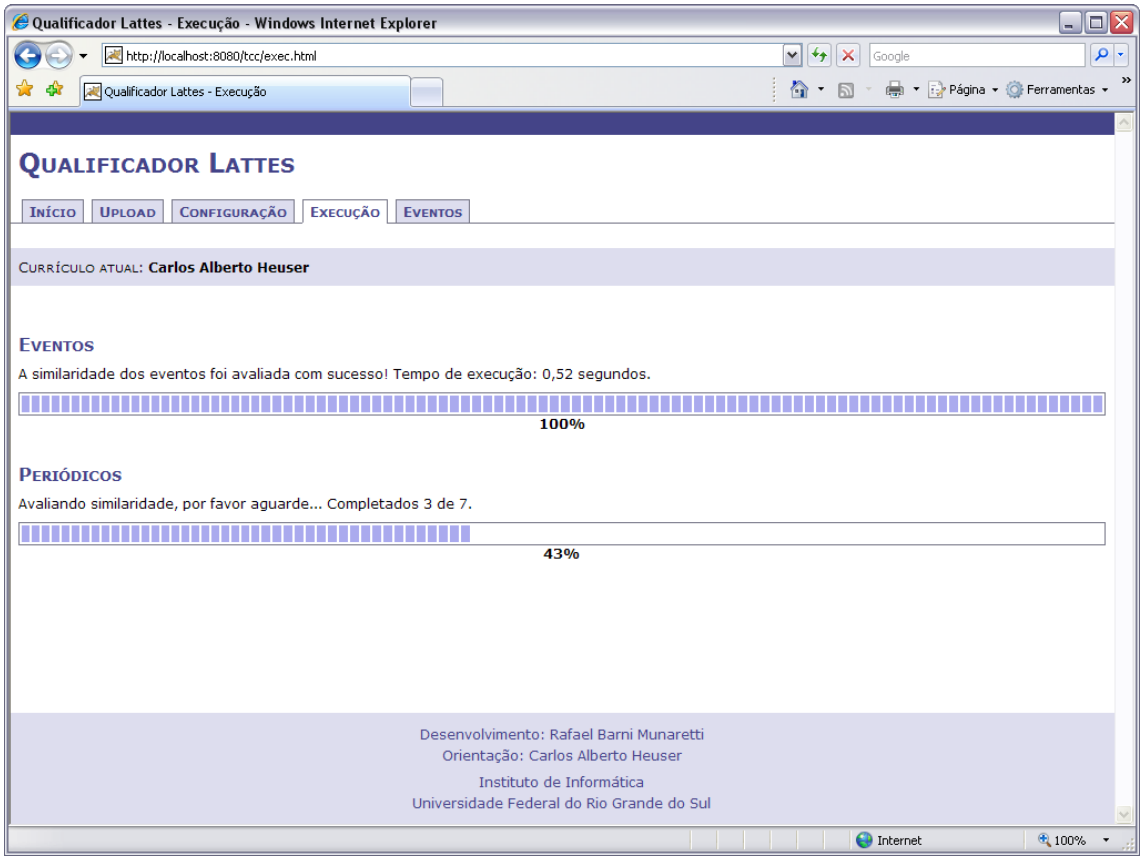

Figura A.3: Tela de progresso de execução

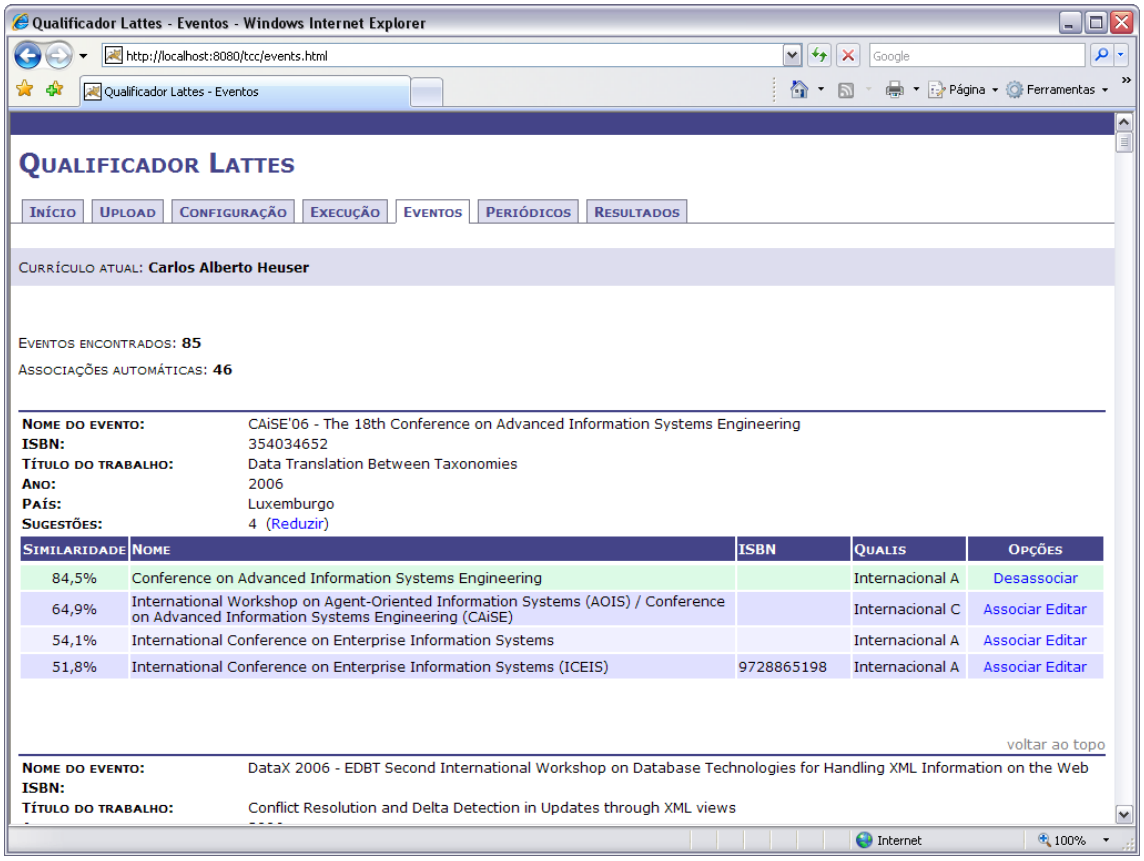

Figura A.4: Tela de sugestões de eventos

A.3 Associar evento

Pré-condições: execução concluída.

Pós-condições: associação de evento realizada.

1. Pelo menu, acessar tela de sugestões de eventos (figura A.4);
2. Clicar no link “Associar” da sugestão desejada.

A.4 Desassociar evento

Pré-condições: execução concluída.

Pós-condições: associação de evento desfeita.

1. Pelo menu, acessar tela de sugestões de eventos (figura A.4);
2. Clicar no link “Desassociar” da associação desejada.

A.5 Editar evento

Pré-condições: execução concluída.

Pós-condições: associação de evento corrigida realizada.

1. Pelo menu, acessar tela de sugestões de eventos (figura A.4);
2. Clicar no link “Editar” da sugestão desejada;
3. Preencher o campo de texto com o conteúdo desejado;
4. Clicar no link “Confirmar” da sugestão editada.

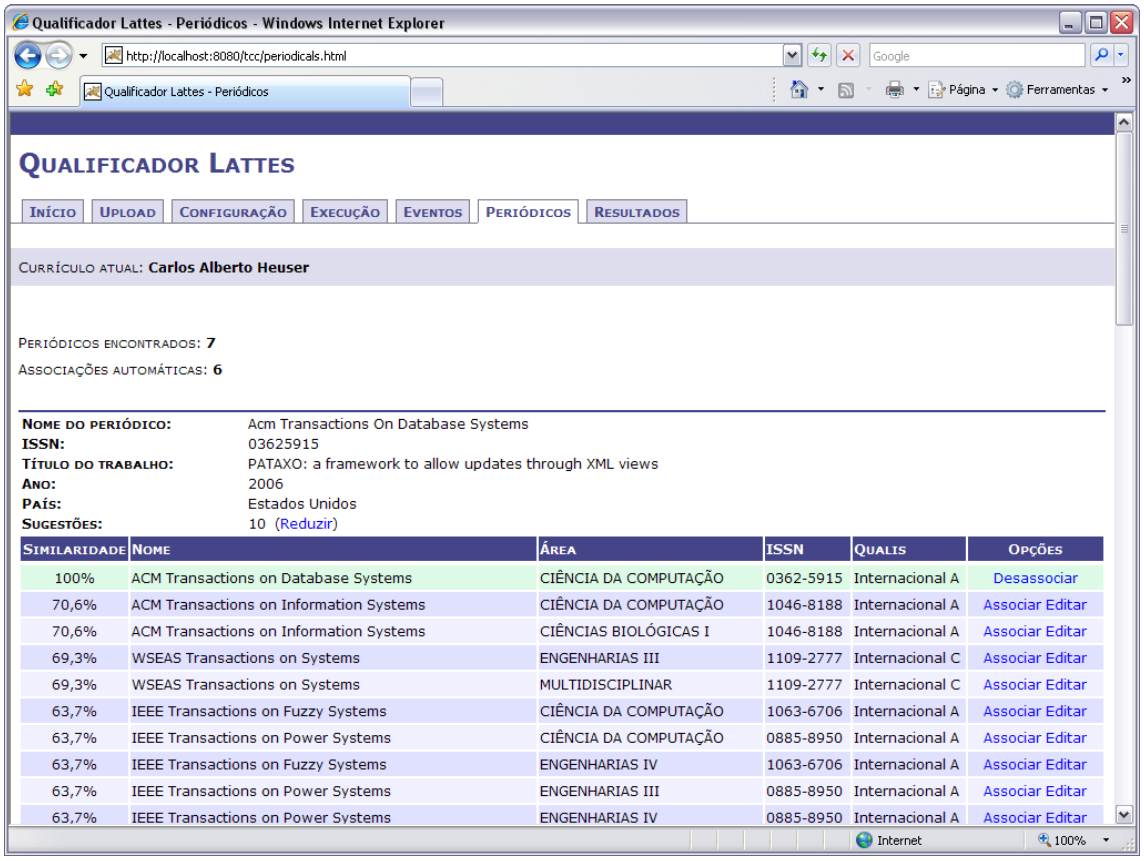

Figura A.5: Tela de sugestões de periódicos

A.6 Associar periódico

Pré-condições: execução concluída.

Pós-condições: associação de periódico realizada.

1. Pelo menu, acessar tela de sugestões de periódicos (figura A.5);
2. Clicar no link “Associar” da sugestão desejada.

A.7 Desassociar periódico

Pré-condições: execução concluída.

Pós-condições: associação de periódico desfeita.

1. Pelo menu, acessar tela de sugestões de periódicos (figura A.5);
2. Clicar no link “Desassociar” da associação desejada.

A.8 Editar periódico

Pré-condições: execução concluída.

Pós-condições: associação de periódico corrigida realizada.

1. Pelo menu, acessar tela de sugestões de periódicos (figura A.5);
2. Clicar no link “Editar” da sugestão desejada;
3. Preencher o campo de texto com o conteúdo desejado;
4. Clicar no link “Confirmar” da sugestão editada.

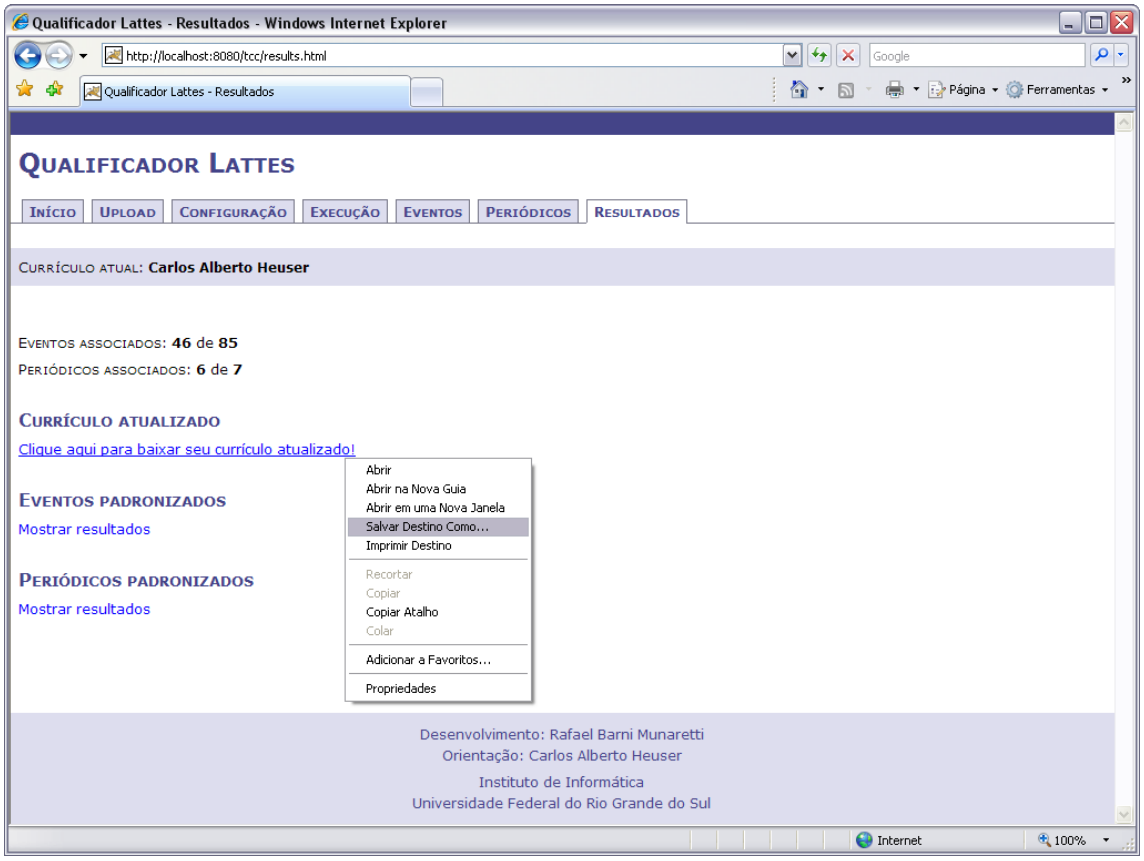

Figura A.6: Tela de resultados

A.9 Baixar currículo

Pré-condições: execução concluída.

Pós-condições: nenhuma.

1. Pelo menu, acessar tela de resultados (figura A.6);
2. Clicar no link “Clique aqui para baixar seu currículo atualizado!”.