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

MARCELO PROVENZI FROSI

Uma Ferramenta para Análise Automatizada da Produção Científica de Pesquisadores

Trabalho de Graduação.

Prof. Dr. Carlos Alberto Heuser
Orientador

Porto Alegre, novembro de 2009.

AGRADECIMENTOS

Agradeço, de todo o coração, a todo apoio e carinho demonstrados pela minha família. Sem sua constante presença, tanto nos bons momentos quanto nas horas mais difíceis, teria sido complicado chegar até aqui.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CSV	Comma Separated Values
HTTP	Hypertext Transfer Protocol
HTML	Hyper Text Markup Language
URL	Uniform Resource Locator
API	Application Programming Interface
WEB	World Wide Web
ASP	Active Server Pages

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Este trabalho descreve o desenvolvimento de uma ferramenta cujo objetivo é realizar uma avaliação automatizada da relevância das publicações de um pesquisador. Essa avaliação é realizada através do cruzamento de informações obtidas no Currículo Lattes do pesquisador com informações referentes às suas publicações, obtidas através do Google Scholar. Essa integração de dados possibilita o cálculo de índices, como o *g*-index e o *h*-index, que servem como métricas para avaliar a relevância de seus trabalhos publicados.

A ferramenta, apresentada em detalhes, consiste em um aplicativo capaz de realizar o cálculo do *h*-index e *g*-index de um pesquisador tomando como base o seu Currículo Lattes. Os resultados obtidos com a utilização dessa ferramenta sobre dados reais se mostraram bastantes satisfatórios.

Palavras-Chave: *h*-index, avaliação de pesquisadores.

A Tool for Automated Analysis of the Scientific Production of Researchers

ABSTRACT

This work describes the development of a tool whose objective is to perform an automated evaluation of the relevance of a researcher's publications. That evaluation is done by merging information obtained in the Currículo Lattes of the researcher with information regarding their publications obtained through Google Scholar. This integration of data makes possible the calculation of evaluation indexes, as the *h*-index and the *g*-index, which serve as metric for the relevance of your published works.

The tool, presented in details, consists of an application capable of performing the calculation of the h-index and g-index of a researcher, taking as base his Currículo Lattes. The results obtained with the use of that tool on real data they were shown plenty satisfactory.

Keywords: *h*-index, researcher's evaluation

1 INTRODUÇÃO

A excelência de uma instituição de ensino depende de uma série de fatores, dentre eles a qualidade de seus integrantes. Uma das maneiras de se realizar uma avaliação desses integrantes é através da análise da relevância de suas pesquisas. O *h*-index (HIRSCH, 2005) é uma das maneiras de medir essa relevância. Ele utiliza a quantidade de citações dos trabalhos publicados pelo pesquisador para criar um índice que serve de métrica para a avaliação da relevância de sua pesquisa.

Uma das possíveis maneiras de se calcular o *h*-index é realizar uma busca pelo nome do autor em uma base de dados de publicações científicas, como o *Google Scholar* (SCHOLAR, 2009) ou o *Journal Citation Reports* (JCR, 2009), e, a partir dos resultados obtidos, calcular a avaliação do pesquisador. Essa abordagem é utilizada, por exemplo, pelo software "*Publish or Perish*" (POP, 2009). O problema dessa metodologia é que publicações de pesquisadores que possuem um nome semelhante podem aparecer no resultado da busca, sendo necessário um trabalho de revisão por parte do usuário, que deve excluí-las manualmente. Isso pode requerer uma grande quantidade de iterações, o que torna muito trabalhosa, ou praticamente inviabiliza, a avaliação de um grupo grande de pesquisadores. Além disso, esse processo de revisão, por ser uma atividade realizada pelo usuário, sempre está sujeita a erros e enganos, que podem levar a resultados inexatos.

Neste contexto, uma importante fonte de informações sobre pesquisadores é a Plataforma Lattes (LATTES, 2009). Ela contém o currículo da grande maioria dos pesquisadores em atividade no país, com informações detalhadas acerca da produção bibliográfica desses pesquisadores.

A idéia desenvolvida neste trabalho de conclusão é realizar um cruzamento das informações obtidas em uma fonte de dados que conheça os trabalhos publicados pelo pesquisador, neste caso a Plataforma Lattes, e as citações obtidas para cada publicação, obtidas em uma fonte de dados de publicações, neste caso o Google Scholar. Esse cruzamento de dados pode nos dar resultados mais precisos e que necessitem menos intervenção humana em sua análise. O grande problema desta metodologia é que os títulos das publicações, cadastrados na Plataforma Lattes, podem ser ligeiramente diferentes dos seus reais títulos, encontrados no Google Scholar. Essa discrepância entre as fontes de dados tende a dificultar o processo de integração de informações.

Tendo em vista essa idéia e seus problemas recorrentes, o objetivo final desse trabalho é desenvolver uma ferramenta que realize uma avaliação automatizada de pesquisadores, com base nas publicações obtidas de seu Currículo Lattes e no número de citações de cada publicação, adquirido através do Google Scholar.

Os capítulos seguintes são organizados conforme descrito abaixo.

O segundo capítulo discorre sobre os conceitos e fontes de dados utilizados neste trabalho. São apresentadas noções fundamentais sobre os conceitos e as fontes de dados utilizados neste trabalho, como o Google Scholar e a Plataforma Lattes.

O terceiro capítulo apresenta a ferramenta desenvolvida neste trabalho do ponto de vista do usuário. São detalhados os casos de uso que permitem realizar a avaliação de um pesquisador ou um grupo de pesquisadores.

O quarto capítulo introduz e descreve em detalhes o sistema desenvolvido neste trabalho de conclusão. São discutidos nesse capítulo as questões técnicas relevantes e as soluções adotadas para contornar os problemas encontrados.

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 AVALIAÇÃO DE PRODUÇÃO CIENTÍFICIA: BASES DE DADOS E FERRAMENTAS

O problema modelado neste trabalho de conclusão envolve uma variedade de conceitos e definições básicas que devem ser explicitados antes que as soluções adotadas possam ser expostas. Este capítulo pretende introduzir e esclarecer esses aspectos fundamentais, com ênfase nos pontos mais relevantes para o desenvolvimento deste trabalho.

2.1 Plataforam Lattes

A Plataforam Lattes é uma base de dados de currículos acadêmicos e instituições de pesquisa na área de Ciência e Tecnologia (LATTES, 2009). Mantida pelo CNPq, possui o currículo da grande maioria dos pesquisadores em atividade no país. Os currículos que fazem parte dessa base, conhecidos como Currículos Lattes, são o foco de interesse deste trabalho.

Um Currículo Lattes contém dados sobre a vida acadêmica de um pesquisador, informações detalhadas sobre sua produção bibliográfica e técnica, orientações, participações em bancas e eventos.

Atualmente, não existem meios públicos de se obter os dados de pesquisadores cadastrados na Plataforma Lattes, além do formato HTML. Esse fato faz com que seja necessário extrair esses dados através de um *parser* específico para a tarefa. Outra ferramenta que também faz a extração de dados de um Currículo Lattes, mas com objetivos distintos, é o scriptLattes (SCRIPTLATTES, 2009).

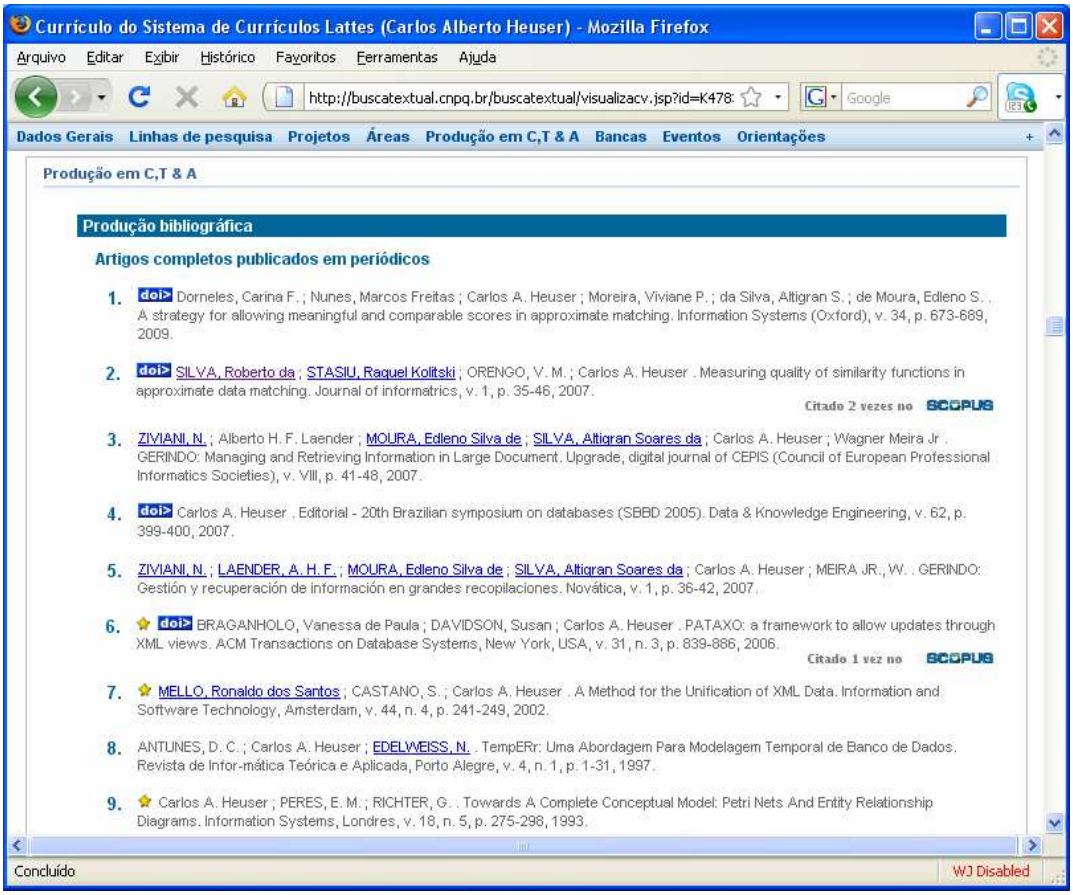

Figura 2.1: Exemplo de Currículo Lattes.

2.2 Google Scholar

O Google Scholar (SCHOLAR, 2009) é um serviço de pesquisa lançado pela Google em 2004, tendo como objetivo facilitar a pesquisa em livros, trabalhos e artigos acadêmicos, jornais científicos, teses, etc. Hoje em dia é capaz de indexar a maior parte dos trabalhos acadêmicos publicados.

Além de fornecer informações bastante úteis sobre produção científica, como versões disponíveis, trabalhos relacionados e artigos que referenciam um trabalho especifico, também possui opções de busca bastante versáteis, como nome do autor, veículo de publicação, data de publicação, etc.

Como atualmente o Google não disponibiliza, nem tem planos de disponibilizar, uma API para utilização do Scholar por terceiros, é necessário extrair os dados relevantes à ferramenta desenvolvida nesse trabalho diretamente da página HTML retornada por buscas realizadas no Scholar, através de um *parser* específico.

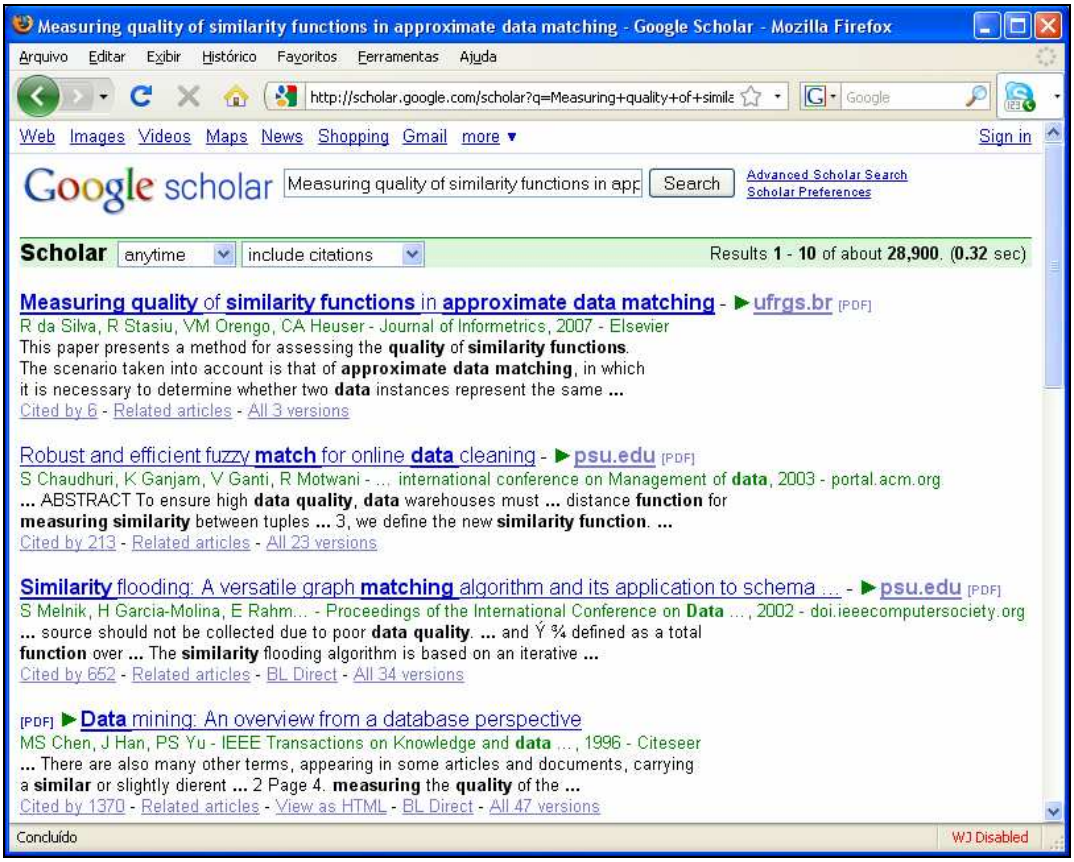

Figura 2.2: Exemplo de busca no Google Scholar.

2.3 *h*-index

O *h*-index é um índice proposto por Jorge E. Hirsch (2005) que tenta medir a produtividade e o impacto do trabalho científico de um pesquisador. É calculado com base em um subconjunto dos trabalhos científicos mais citados em outras publicações, tentando equilibrar tanto o número de citações de cada publicação quanto a quantidade de publicações.

A definição de *h*-index utilizada neste trabalho é a seguinte: "*Um pesquisador possui um índice h se h de seus N trabalhos possuem ao menos h citações cada, e os outros (N - h) trabalhos tenham no máximo h citações cada*".

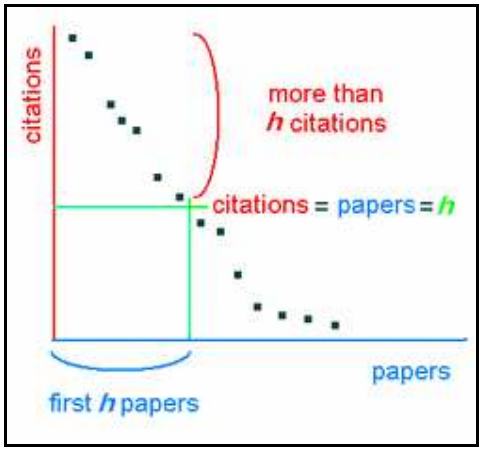

Figura 2.3: Decréscimo de citações por publicação (WIKIPEDIA, 2009).

Por exemplo, para um pesquisador que possua três publicações, e cujos números de citações sejam, em ordem decrescente: 3, 2, 1. Seu *h*-index, como mostra a tabela à seguir, é igual à 2.

Tabela 2.1: Primeiro exemplo de cálculo do *h*-index

Publicação	**Citações**	***h*-index**	**Citações > *h*-index**	**Novo *h*-index**
Publicação 1	3	0	Sim	1
Publicação 2	2	1	Sim	2
Publicação 3	1	2	Não	2

O próximo exemplo mostra o cálculo do *h*-index para um pesquisador com uma quantidade maior de publicações, mas com poucas citações em cada trabalho. Conforme se pode notar, em uma situação como esta, o valor do *h*-index acaba limitado pelo numero de citações o que o torna baixo.

Tabela 2.2: Segundo exemplo de cálculo do *h*-index

Publicação	**Citações**	***h*-index**	**Citações > *h*-index**	**Novo *h*-index**
Publicação 1	40	0	Sim	1
Publicação 2	2	1	Sim	2
Publicação 3	2	2	Não	2
Publicação 4	1	2	Não	2
Publicação 5	1	2	Não	2
Publicação 6	1	2	Não	2
Publicação 7	0	2	Não	2
Publicação 8	0	2	Não	2

O exemplo seguinte mostra o cálculo do *h*-index para um pesquisador com uma quantidade pequena de publicações, mas com muitas citações em cada trabalho. Conforme se pode notar, em uma situação como esta, o valor do *h*-index acaba limitado pela quantidade de trabalhos o que o torna baixo.

Tabela 2.3: Terceiro exemplo de cálculo do *h*-index

Publicação	**Citações**	***h*-index**	**Citações > *h*-index**	**Novo *h*-index**
Publicação 1	40	0	Sim	1
Publicação 2	25	1	Sim	2
Publicação 3	12	2	Sim	3

Conforme demonstrado nos exemplos, para se obter um *h*-index alto é necessário uma grande quantidade de trabalhos e uma boa quantidade de citações em cada um deles.

2.4 *g*-index

O *g*-index (EGGHE, 2006) é um índice alternativo ao *h*-index, tendo, portanto, as mesmas finalidades. A grande diferença entre os dois é o fato de que o *h*-index não considera uma quantidade de citações mínima para se obter um valor especifico. Esse fato acaba gerando certas injustiças. Pense, por exemplo, dois pesquisadores com *h*-index 5. Um deles pode ter um total de 25 citações, e o outro mais de 1000 citações. Esse tipo de situação não é comum, mas acaba tendo no *g*-index um índice complementar que ajuda na analise dos resultados.

A definição de *g*-index utilizada neste trabalho é a seguinte: "*Dado um conjunto de artigos ordenados de maneira descendente com relação ao seu número de citações, o g-index é o maior valor de g em que os primeiros g artigos receberam juntos ao menos* g^2 *citações*".

Por exemplo, para um pesquisador que possua quatro publicações, e cujos números de citações sejam, em ordem decrescente: 5, 3, 1, 1. Seu *g*-index, como mostra a tabela à seguir, é igual à 3.

Tabela 2.4: Exemplo de cálculo do *g*-index

Publicação	**Citações**	**Total de Citações (T)**	***g*-index**	***g*-index²**	**T ≥ *g*-index²**
Publicação 1	5	5	1	1	Sim
Publicação 2	3	8	2	4	Sim
Publicação 3	1	9	3	9	Sim
Publicação 4	1	10	4	16	Não

3 SOLUÇÃO – CASOS DE USO

Como descrito no primeiro capítulo deste trabalho, a ferramenta desenvolvida tem como objetivo de realizar um processo de avaliação automatizada de pesquisadores. Tendo em vista esse objetivo, dois casos de uso foram implementados: "Avaliação de um pesquisador" e "Avaliação de um grupo de pesquisadores".

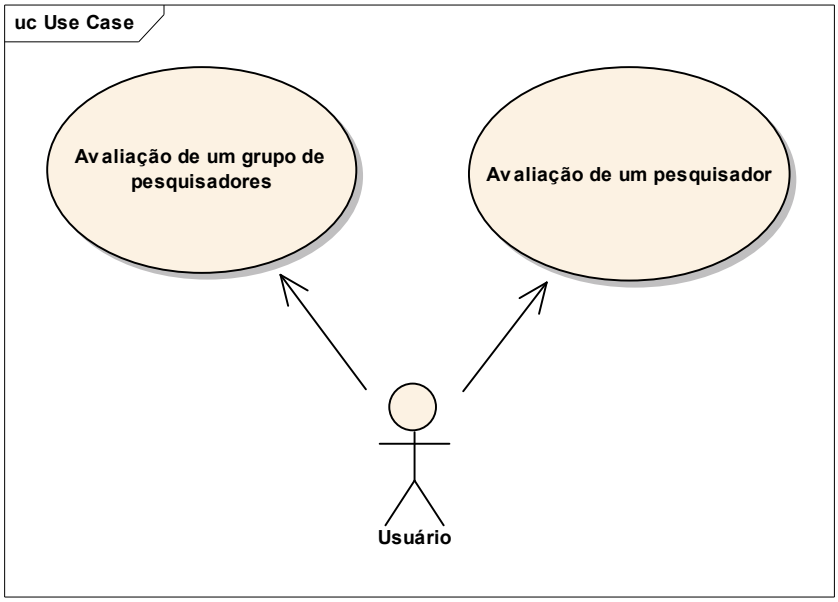

Figura 3.1: Casos de uso implementados

3.1 Avaliação um pesquisador

Neste caso de uso, o sistema apresenta uma tela que permite ao usuário realizar um busca pelo nome no pesquisador. Para realizar essa busca, é necessário informar parte do nome do pesquisador, se o pesquisador possui o título de doutor, e o código de segurança apresentado pelo sistema.

O código de segurança consiste em um conjunto de quatro caracteres apresentados em forma de uma imagem, e é exigido pela Plataforma Lattes como forma de evitar buscas realizadas de maneira automatizada, visto que esse tipo de pesquisa pode levar a uma sobrecarga do sistema. Caso essa imagem esteja ilegível demais e o usuário prefira trocá-la, pode fazer isso através de um botão específico.

Após o usuário preencher os campos solicitados e submeter sua busca, o sistema retorna uma lista de pesquisadores que preenchem os requisitos requeridos. Neste ponto o usuário deve selecionar o pesquisador desejado para que o sistema possa realizar o processamento de seu currículo.

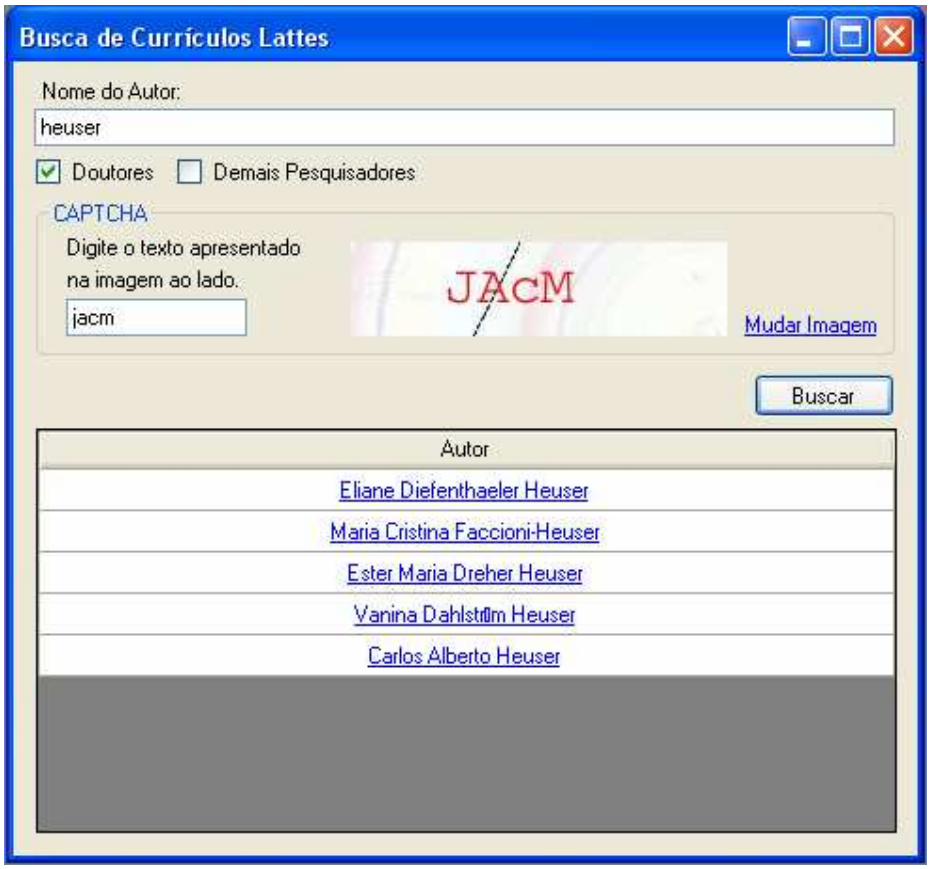

Figura 3.2: Tela de busca por Currículo Lattes

Em um cenário alternativo, caso o usuário já possua o endereço para o Currículo Lattes a ser pesquisado, a aplicação permite que se inicie o processamento à partir dessa URL, ao invés de realizar todo processo de busca.

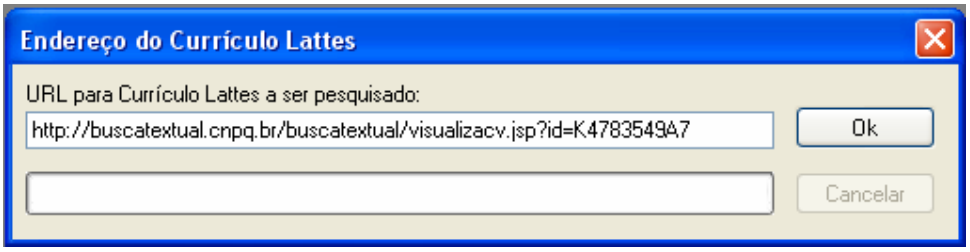

Figura 3.3: Tela de entrada de endereço de um Currículo Lattes

Após o usuário indicar qual vai ser o currículo a ser avaliado, é iniciado o seu processamento. Durante esse processamento, o sistema exibe uma barra de progresso informando qual o status da tarefa, e um botão que permite abortá-la. As ações executadas pelo sistema durante esse processamento serão detalhadas em um capítulo especifico.

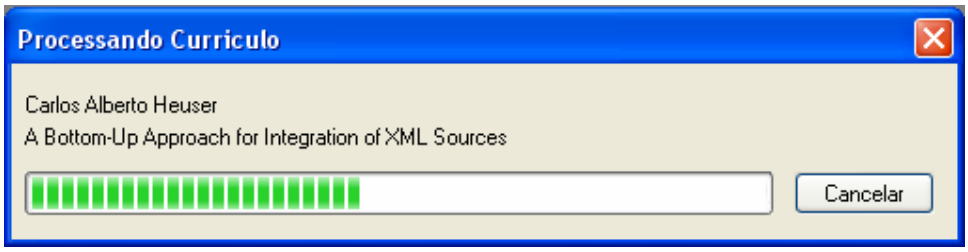

Figura 3.4: Tela de processamento de um Currículo Lattes

Após processar o currículo do pesquisador selecionado, o sistema exibe uma tela que apresenta as informações obtidas. Dentre essas informações estão:

- O nome do autor;
- O número de citações cada publicação possui;
- O *h*-index do pesquisador;
- O *g*-index do pesquisador;
- O total de citações das publicações encontradas;
- O total de citações que se encontram dentro do *h*-index;
- O total de citações que se encontram dentro do *g*-index;

Existem ainda nesta tela, um botão que permite exportar o resultado da pesquisa para um arquivo no formato CSV, e um campo de filtro, que permite ao usuário realizar buscas por algum trabalho específico dentre os apresentados.

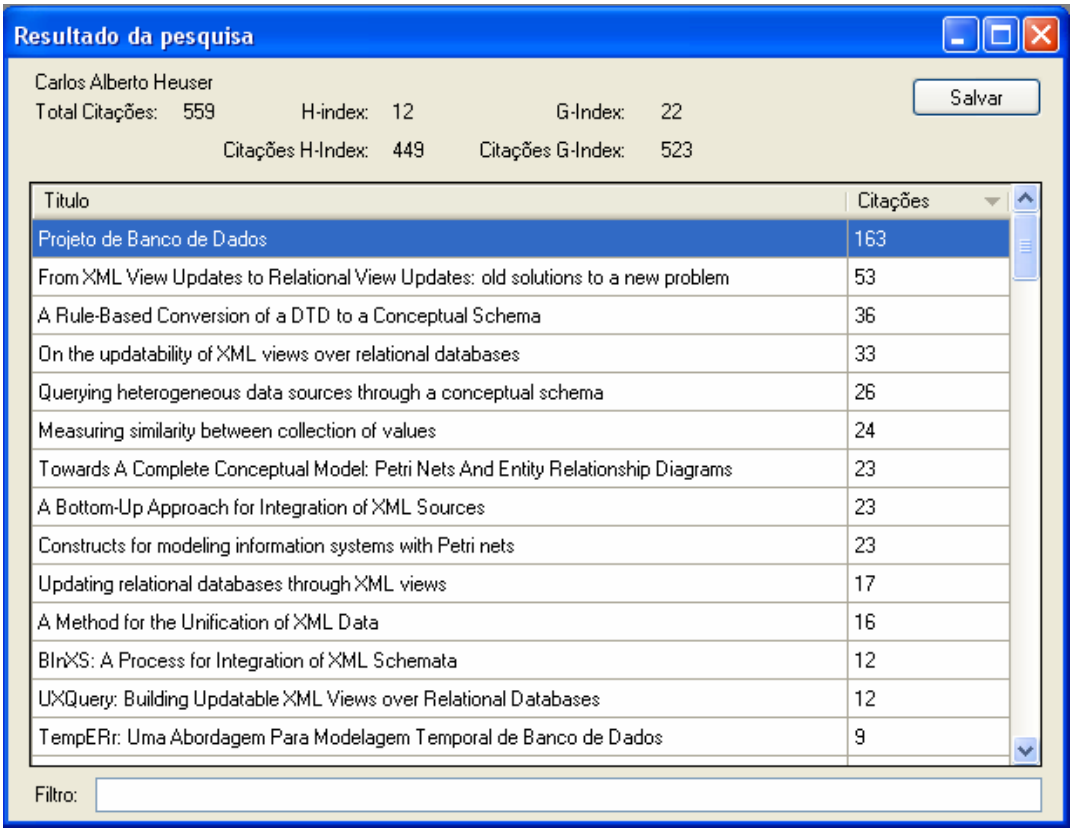

Figura 3.5: Tela de resultados

3.2 Avaliação de um grupo de pesquisadores

Neste caso de uso, o usuário requisita ao sistema o processamento de um arquivo que contêm os endereços de uma série de currículos. Esse arquivo utilizado como entrada é um arquivo texto simples, onde cada linha representa a URL de um Currículo Lattes.

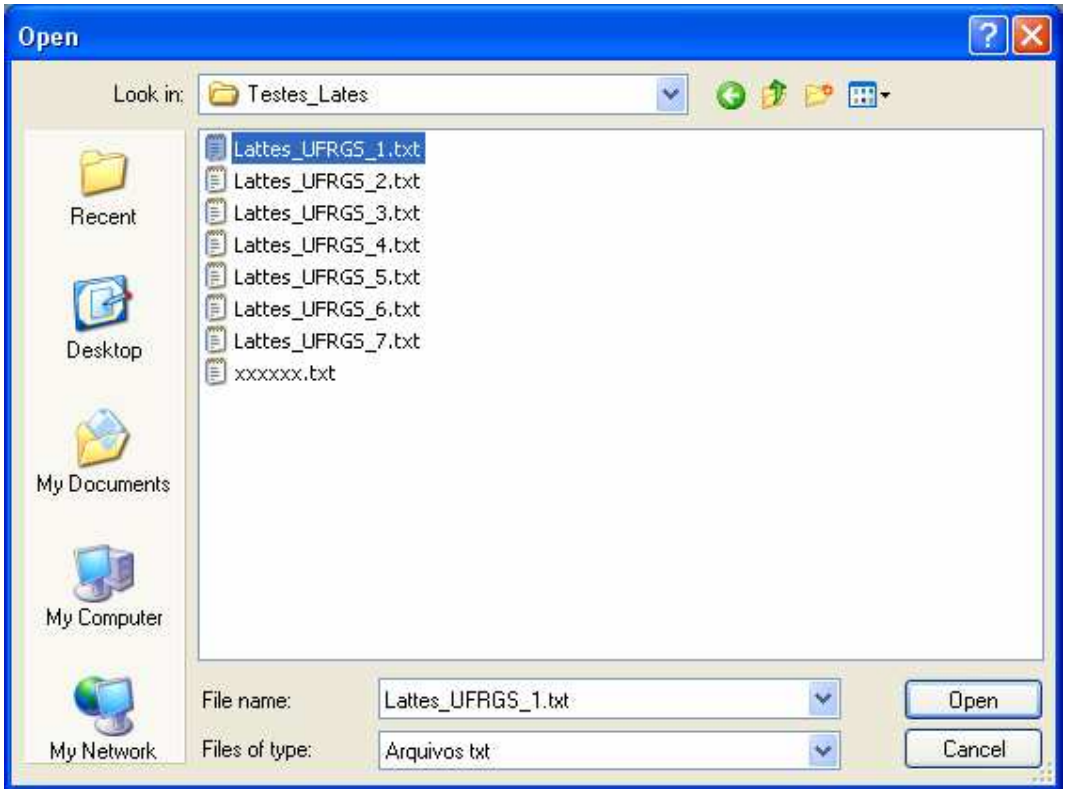

Figura 3.6: Tela de busca por arquivo de entrada.

Após o usuário informar à aplicação o arquivo contendo os endereços dos currículos, o sistema pesquisa cada currículo individualmente, exibindo o nome do pesquisador que está sendo processado atualmente, uma barra de progresso informando qual o status do processamento e um botão que permite cancelar a tarefa atual.

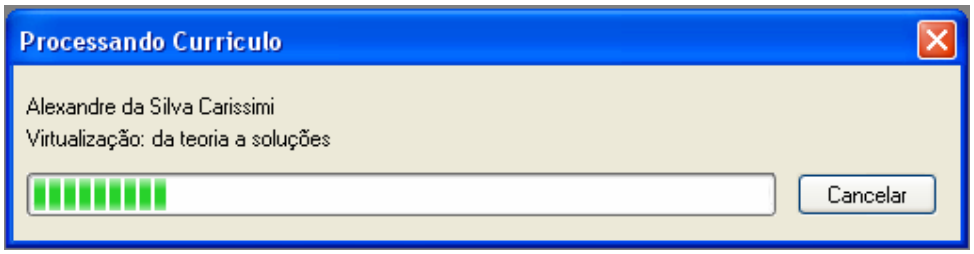

Figura 3.7: Tela de processamento de uma série de Currículos Lattes

Após o processamento de todos os currículos informados, é exibida a tela de apresentação dos resultados. Nessa tela é exibida:

- A lista de pesquisadores analisados;
- O *h*-index de cada pesquisador avaliado;
- O *g*-index de cada pesquisador avaliado;

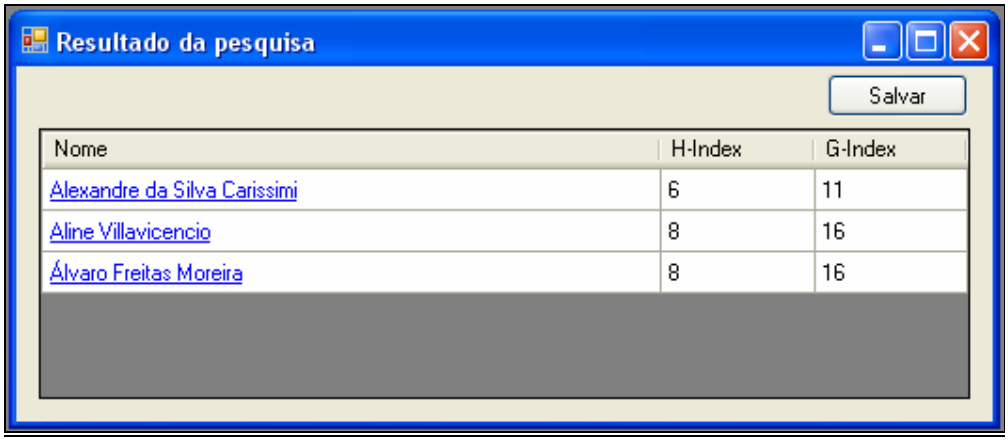

Figura 3.8: Tela de resultados do processamento de um grupo de pesquisadores.

É permitido ao usuário exibir a listagem de artigos de cada pesquisador e suas respectivas citações, através de um *hiperlink* inserido junto ao seu nome. Existe ainda nesta tela, um botão que permite exportar o resultado da pesquisa para um arquivo no formato CSV.

4 SOLUÇÃO - PONTO DE VISTA TÉCNICO

Essa seção discute alguns detalhes relevantes do projeto e implementação da ferramenta desenvolvida neste trabalho de conclusão. Não é seu objetivo, porém, descrever o processo de desenvolvimento integralmente.

4.1 Arquitetura

A arquitetura do sistema implementado foi baseada no agrupamento de tarefas similares em uma mesma camada lógica. Cada camada exerce funções bem definidas e provê uma interface para a camada superior.

Essa arquitetura em camadas reduz a coerção dos artefatos de software produzidos, o que permite que futuras manutenções e melhorias sejam feitas mais facilmente, já eu cada camada pode ser alterada, ou até mesmo substituída, sem que sejam necessárias maiores preocupações com o restante da aplicação.

Seguindo esse paradigma, quatro camadas foram projetadas: Model, Parser, Core e Apresentação.

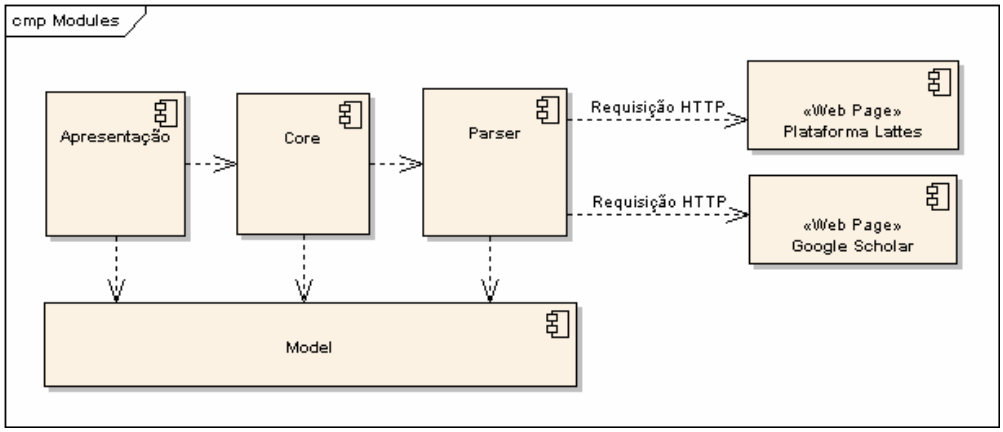

Figura 4.1: Camadas da aplicação

4.1.1 Model

As classes desse módulo representam as entidades do domínio modelado pela aplicação e suas características. Também estão implementadas nesse módulo as exceções disparadas pelo sistema.

As finalidades das classes desse módulo são o armazenamento de dados e a troca de informações entre as camadas do sistema.

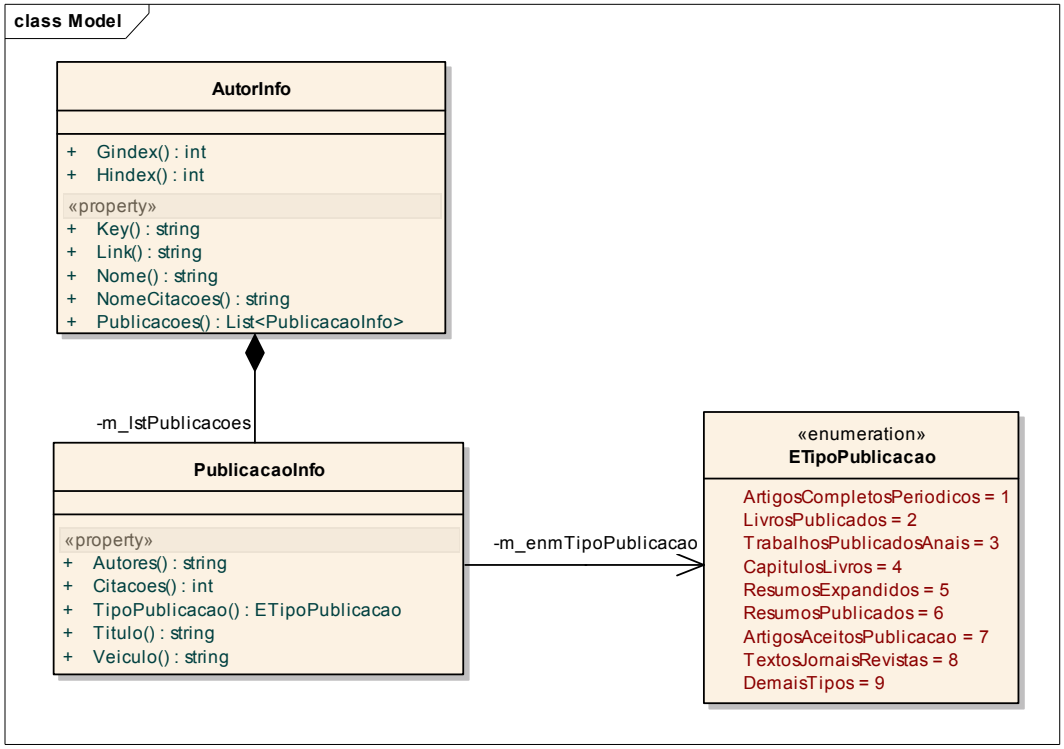

Figura 4.2: Principais classes implementadas na camada Model

4.1.2 Parser

Nesta camada estão implementadas as classes responsáveis por realizar a pesquisa das páginas da Plataforma Lattes e do Google Scholar e por transformar as informações recebidas como HTML em entidades tratáveis no sistema, implementadas na camada Model.

Essa camada está acessível apenas ao Core, não podendo ser referenciada em nenhuma outra parte da aplicação.

4.1.3 Core

Essa camada a responsável pela implementação das regras de negócio que compõe o sistema. Geralmente os métodos dessa camada são compostos de uma série de validações de dados e chamadas a métodos mais básicos, providos pela camada Parser.

Também é responsável por disparar os eventos utilizados para informar o progresso das tarefas realizadas pelo sistema.

4.1.4 Apresentação

Responsável por intermediar a iteração do usuário com o sistema e pelo controle de fluxo de atividades.

Foi implementada utilizando Windows Forms, a API do framework .NET (DOTNET, 2009) para construção de interfaces gráficas. No futuro, caso haja a necessidade de portar a ferramenta desenvolvida nesse trabalho para um ambiente WEB, basta substituir essa camada por outra compatível, desenvolvida, por exemplo, em ASP.NET.

4.2 Implementação

A ferramenta desenvolvida neste trabalho de conclusão foi implementada em C#, utilizando o framework .NET 2.0 (DOTNET, 2009). O ambiente de desenvolvimento escolhido foi o Microsoft Visual Studio 2008. A escolha desse ambiente e linguagem de programação específicos, devem-se puramente a uma escolha pessoal, já que estou familiarizado com essas ferramentas devido ao seu uso constante.

A única biblioteca adicional utilizada foi a *log4net* (LOG4NET, 2009), sendo esta utilizada para facilitar a criação de arquivos de logs e traces.

No início do desenvolvimento, uma das decisões tomadas, foi a de desenvolver a ferramenta como um aplicativo Windows, ao invés de aplicativo WEB. Essa decisão se deve principalmente à dificuldade de encontrar hospedagem livre para páginas .NET.

4.3 Processamento de um Currículo Lattes

Este tópico visa detalhar as tarefas realizadas pelo sistema durante o processamento de um currículo.

Após o sistema ser informado do endereço do Currículo Lattes a ser avaliado, é feita uma requisição HTTP direcionada a este endereço, a fim de obter as informações contidas nele. Como retorno dessa requisição é obtida uma página HTML de onde a aplicação extrai os dados necessários, fazendo uso de um *parser* especificamente implementado.

Os dados extraídos do currículo são:

- O nome do pesquisador;
- Os nomes utilizados em citações bibliográficas;
- A lista de seus trabalhos bibliográficos;

Dentre os trabalhos bibliográficos nem todos são utilizados, apenas os pertencentes a alguma das categorias abaixo:

- Artigos completos publicados em periódicos;
- Artigos aceitos para publicação;
- Trabalhos completos publicados em anais de congressos;
- Livros;
- Capítulos de livros publicados;

Todos os demais trabalhos bibliográficos, por não serem considerados relevantes para fins do calculo do *h*-index e *g*-index, são descartados pelo sistema.

Tendo em mão os dados do pesquisador, obtidos no Currículo Lattes, a aplicação deve agora obter as informações referentes ao número de citações de cada trabalho bibliográfico. Para isso é submetida uma requisição HTTP ao Google Scholar para cada um dos trabalhos. Nessas requisições, estão embutidas consultas que tentam indexar o trabalho desejado na *engine* de busca do Google Scholar.

Como os artigos retornados em uma busca no Google Scholar estão ordenados de acordo com sua relevância, apenas as citações do primeiro resultado são computadas ao artigo pesquisado.

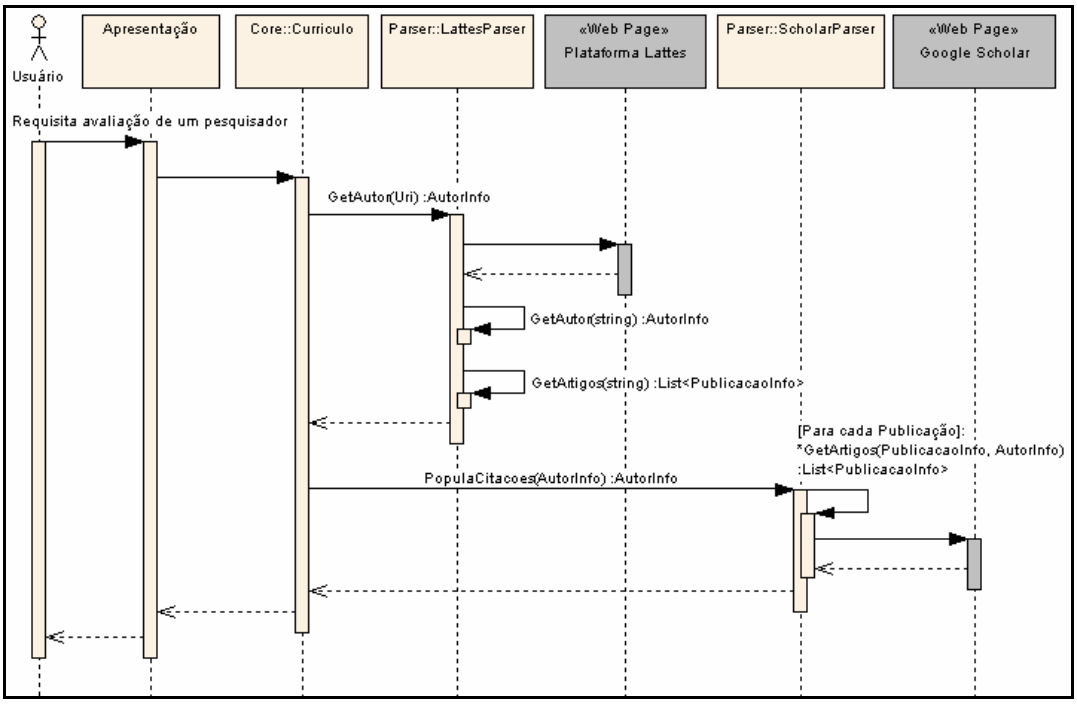

Figura 4.3: Diagrama de seqüência para processamento de um Currículo Lattes

4.4 *Parser* de uma página de um Currículo Lattes

A Plataforma Lattes não disponibiliza, de maneira pública, métodos que possibilitem exportar os dados de um currículo. Portando, fez-se necessário a implementação de um *parser* que extraia as informações relevantes diretamente da página HTML exibida durante a visualização de um Currículo Lattes.

As primeiras informações a serem obtidas pelo *parser* são o nome do pesquisador e os seus nomes utilizados em citações bibliográficas. Essas informações são exibidas em um Currículo Lattes conforme o exemplo a seguir:

Figura 4.4: Exemplo de exibição de nomes em um Currículo Lattes

Esse exemplo em particular, possui um código HTML conforme segue:

```
<em>Dados pessoais</em>
<table class="IndicProdTabela">
  <tbody>
    <tr class="IndicProdTabelaLinha">
        <td class="DadGerTabelaCelula26" width="148">
          Nome
        </td>
        <td class="DadGerTabelaCelula74" width="624">
          Carlos Alberto Heuser
        </td>
    </tr>
    <tr class="IndicProdTabelaLinha">
        <td class="DadGerTabelaCelula26" width="148">
          Nome em citações bibliográficas
        </td>
        <td class="DadGerTabelaCelula74" width="624">
          Carlos A. Heuser;Heuser, Carlos A.
        </td>
    </tr>
    <tr class="IndicProdTabelaLinha">
      <td class="DadGerTabelaCelula26" width="148">
        Sexo
      </td>
      <td class="DadGerTabelaCelula74" width="624">
        Masculino
      </td>
    </tr>
  </tbody>
</table>
```

Analisando esse trecho de código HTML, pode-se notar que sempre, antes da tabela que contêm os dados propriamente ditos, existe uma tag "*em*", cujo conteúdo é o nome da seção. A partir dessa constatação o *parser* procura pela *string* "*Dados pessoais*" como marcador de início para obtenção dos nomes do autor.

Continuando a análise deste trecho de código HTML, pode-se notar que as colunas da tabela possuem classes distintas. A primeira coluna, que contêm o nome do campo, possui a classe "*DadGerTabelaCelula26*", enquanto a segunda coluna, que possui o valor do campo, possui a classe "*DadGerTabelaCelula74*".

Baseado nessa analise, o parser percorre os elementos HTML fazendo uso de uma pequena máquina de estados: caso o conteúdo de um elemento "*td*", com classe igual à "*DadGerTabelaCelula26*", seja "*Nome*", o valor do próximo elemento "*td*" será considerado como o nome do pesquisador, caso seja "*Nome em citações bibliográficas*" será considerado como os seus nomes utilizados em citações bibliográficas.

A outra informação relevante a ser obtida em um Currículo Lattes são as publicações do pesquisador. Exibidas como o exemplo a seguir:

Produção bibliográfica

Artigos completos publicados em periódicos

1. doi> Dorneles, Carina F. ; Nunes, Marcos Freitas ; Carlos A. Heuser ; Moreira, Viviane P. ; da Silva, Altigran S. ; de Moura, Edleno S. . A strategy for allowing meaningful and comparable scores in approximate matching. Information Systems (Oxford), v. 34, p. 673-689, 2009.

Figura 4.5: Exemplo de exibição de publicação em um Currículo Lattes

Esse exemplo em particular, possui um código HTML conforme segue:

```
<a name="Producaobibliografica"></a>
<table style="margin-bottom: 0pt; padding-bottom: 0pt; border-
collapse: collapse;" class="IndiceTabela" cellpadding="0"
cellspacing="0">
  <tbody>
    <tr class="AtuaProfTabelaLinha" align="right">
      <td width="5"></td>
      <td class="AtuaProfTabelaCelula95" align="right">
        Produção bibliográfica
      </td>
    </tr>
  </tbody>
</table>


<table class="IndicProdTabela">
  <tbody>
    <tr></tr>
    <tr class="AtuaProfTabelaLinha" align="right">
      <td bgcolor="white" width="5"></td>
      <td class="agrupadorsub" align="left" width="702">
        Artigos completos publicados em periódicos
      </td>
    </tr>
  </tbody>
</table>

<table class="IndicProdTabela">
  <tbody>
    <tr class="AtuaProfTabelaLinha" align="right" valign="top">
      <td class="ProdBibliCelula10" width="78">1. </td>
      <td class="textoProducao" width="702">
        <a href="http://dx.doi.org/10.1016/j.is.2009.05.002"
target="_blank" style="padding-right: 4px;">
          <img src="images/curriculo/ico_doi.gif">
        </a>
        Dorneles, Carina F. ; Nunes, Marcos Freitas ; Carlos A. Heuser
; Moreira, Viviane P. ; da Silva, Altigran S. ; de Moura, Edleno S. .
A strategy for allowing meaningful and comparable scores in
approximate matching. Information Systems (Oxford), v. 34, p. 673-689,
2009.
      </td>
    </tr>
    ...
  </tbody>
</table>
```

Analisando esse trecho de código HTML, pode-se notar que sempre, antes do inicio da listagem dos trabalhos bibliográficos, existe uma tag "*a*", que serve de ancora para essa sessão especifica. A partir dessa constatação o *parser* procura pela *string ""* como marcador de início para obtenção dos trabalhos publicados.

Continuando a análise deste trecho de código HTML, pode-se notar que a produção bibliográfica está agrupada de acordo com sua classificação. O título da classificação dos trabalhos sempre está inserido em um elemento "*td*", que utiliza uma classe CSS de nome "*agrupadorsub"*. Partindo desse pressuposto, o *parser* considera que todos os artigos subseqüentes pertençam a essa categoria, até a ocorrência de um novo elemento "*td*" que utilize a classe "*agrupadorsub*".

Seguindo adiante com a análise, percebe-se que os dados relativos a uma publicação também estão inseridos em um elemento do tipo "*td*", mas que desta vez utiliza uma classe CSS chamada "*textoProducao*". A partir disso, o *parser* considera que o texto desse elemento, retirando todas as demais tags HTML, representa os dados de uma publicação.

Conforme é possível notar pelo exemplo dado, não existem marcações no texto que o separem em campos específicos, como autor, título e veiculo de publicação, mas isso pode ser feito a partir de padrões no texto que se repetem ao longo da lista de publicações.

Os nomes dos autores da publicação sempre se encontram no inicio do texto e prosseguem até a string " . ", que os separa do titulo da publicação.

Publicações das categorias "*Trabalhos completos publicados em anais de congressos*" e "*Capítulos de livros publicados*" possuem como separador entre titulo e veículo de publicação, a string "*In:*". Para os trabalhos das demais categorias, é necessário considerar que o titulo é uma frase, sendo, portanto, terminado com um ponto final, o que acaba sendo utilizado como um delimitador do titulo.

4.5 Pesquisas realizadas no Google Scholar

As consultas submetidas ao Google Scholar utilizam o título do trabalho e os nomes utilizados pelo autor em citações bibliográficas para limitar os resultados. Basicamente dois tipos de consultas são utilizados.

O primeiro tipo busca publicações que possuam as mesmas palavras em seu título que as encontradas no título do trabalho obtido no Currículo Lattes, além de pertencerem ao autor desejado. Para isso utiliza as palavras reservadas do Google Scholar "allintitle", "OR" e "author".

Tabela 4.1: Palavras reservadas do Google Scholar utilizadas na primeira consulta.

Palavra	**Descrição**
allintitle	Força que todas palavras se encontrem no título da publicação.
OR	Operador de ou lógico em buscas do Google Scholar.
author	Realiza uma busca utilizando o nome do autor.

Por exemplo, a pesquisa por um trabalho intitulado "Uma Ferramenta Para Análise Automatizada de Pesquisadores" e cujo autor tivesse como nomes utilizados em citações bibliográficas os nomes "MP Frosi" e "Frosi, MP", geraria a seguinte consulta:

allintitle:Uma Ferramenta Para Análise Automatizada de Pesquisadores author:"MP Frosi" OR author:"Frosi, MP"

O problema desta consulta é que ela é muito restritiva. Basta que uma das palavras do título do trabalho pesquisado esteja diferente da encontrada em seu nome real para que a publicação não seja encontrada.

Caso essa pesquisa não retorne resultados, é realizada uma segunda tentativa, desta vez, com uma pesquisa menos restritiva.

Essa nova pesquisa é baseada no posicionamento de algumas palavras do título, além do seu número de palavras. Para isso utiliza os caracteres reservados aspas duplas (") e asterisco (*), além das palavras reservadas "*intitle*", "*OR*" e "*author*".

Tabela 4.2: Palavras reservadas do Google Scholar utilizadas na segunda consulta.

Palavra ou Caráter	**Descrição**
"	Delimita uma string literal, ou seja, que a frase dos resultados sejam exatamente como a apresentada.
*	Dentro de uma string literal representa uma, e apenas uma, palavra desconhecida.
intitle	Força que a palavra se encontre no título da publicação.
OR	Operador de ou lógico em buscas do Google Scholar.
author	Realiza uma busca utilizando o nome do autor.

O caractere aspas duplas serve de delimitador para uma string literal, ou seja, os resultados devem possuir a string exatamente como a encontrada entre aspas. Já o asterisco serve como um "coringa", o que significa que em seu lugar existe uma palavra desconhecida.

A consulta montada possui duas partes: uma delas mantém a primeira, a última e a palavra que se encontra exatamente no meio do título do trabalho pesquisado, substituindo todas as outras palavras por asteriscos. A outra parte da pesquisa faz exatamente o inverso, a fim de que, se alguma das palavras mantidas na primeira parte estiver errada ela não interfira negativamente no resultado. Ambas as artes são unidas por um operador "OR".

Para o mesmo caso utilizado como exemplo da primeira consulta, seria gerada a seguinte pesquisa:

*intitle:"Uma * * Análise * * Pesquisadores" OR intitle:"* Ferramenta Para * Automatizada de *" author:"MP Frosi" OR author:"Frosi, MP"*

Embora possa parecer que essa pesquisa retorne resultados indesejados, na prática se mostrou bastante eficiente, indexando o trabalho desejado na grande maioria dos casos.

4.6 *Parser* de uma página de resultados do Google Scholar

Como o Google não disponibiliza uma API para a utilização do Scholar por aplicativos de terceiros, foi necessário implementar um *parser* especializado para extrair as informações relevantes de suas páginas de pesquisa.

Esse *parser* consiste em um algoritmo que percorre o HTML retornado e transforma as informações contidas nesse HTML em objetos tratáveis do sistema, neste caso especifico em objetos da classe PublicaoInfo implementada na camada Model.

A primeira tarefa realizada pelo *parser* é verificar se a página contém algum resultado. Isso é feito através da verificação de que não existe, no HTML analisado, uma determinada *string* que indicaria que a pesquisa não retornou resultados. A *string* utilizada nessa verificação é a seguinte:

* - did not match any articles.
*

Caso o HTML passe dessa validação, o *parser* tenta extrair as publicações retornadas. Analisando o HTML padrão de uma página de resposta do Google Scholar, nota-se que ele pode ser dividido em três partes: cabeçalho, rodapé e o resultado da busca. Tanto o cabeçalho quanto o rodapé utilizam elementos de tabela como base de sua estrutura, enquanto os resultados da busca utilizam elementos de parágrafo como base de sua constituição.

Esse fato facilita o processo de *parsing*: basta remover todas as tabelas ao inicio e final da página, já que tanto o cabeçalho quanto o rodapé da página não possuem informações úteis, e separar os resultados por parágrafos. Isso é feito utilizando como marcadores as tags HTML de tabela: "*<table>*" e "*</table>*" e de parágrafo: "* *" e "* *".

Dentro de cada resultado da pesquisa, duas informações são relevantes: o título da publicação e o seu número de citações. Essas informações são exibidas conforme o exemplo a seguir:

Measuring quality of **similarity functions** in **approximate data matching** - ▶ **ufrgs.br** [PDF]
R da Silva, R Stasiu, VM Orengo, CA Heuser - Journal of Informetrics, 2007 - Elsevier
This paper presents a method for assessing the **quality** of **similarity functions**.
The scenario taken into account is that of **approximate data matching**, in which
it is necessary to determine whether two **data** instances represent the same **...**
Cited by 6 - Related articles - All 3 versions

Figura 4.6: Exemplo de resultado de uma consulta no Google Scholar

Para esse exemplo específico, o código HTML é o seguinte:

```
<p></p>
<h3 class="r">
  <a href="http://linkinghub.elsevier.com/..." class="yC0">
    <b>Measuring quality </b>of <b>similarity functions </b>in <b>
approximate </b>data matching
  </a>
</h3> - <span class="a"> ▶ </span>
<b class="fl">
  <a href="http://www.inf.ufrgs.br/%7Eheuser/..." class="yC1">
    ufrgs.br
  </a>
  <font class="f" size="-2"> [PDF] </font> 
</b>
<font size="-1"><br>
  <span class="a">
```

```
    R da Silva, R Stasiu, VM Orengo, CA Heuser - Journal of
Informetrics, 2007 – Elsevier
  </span><br>
  This paper presents a method for assessing the <b>quality</b> of
<b>similarity functions</b>. <br>
  The scenario taken into account is that of <b>approximate</b> data
matching, in which <br>
  it is necessary to determine whether two data instances represent
the same <b> ...</b> <br>
  <span class="fl">
    <a href="/scholar?cites=...">Cited by 7</a> -
    <a href="/scholar?q=related:...">Related articles</a> -
    <a href="/scholar?cluster=...">All 3 versions</a>
   </span>
</font>
<p></p>
```

Conforme se pode notar, o título da publicação, em um resultado do Google Scholar, está delimitado pelas tags HTML de título tamanho três: "*<h3>*" ao inicio e "*</h3>*" ao final. Basta, portanto, ler o conteúdo destas tags, retirando qualquer outro elemento HTML, como marcadores de link (tags "*<a>*") e negrito(tags "*<b>*"), para obter o titulo da publicação.

Para extrair o número de citações, é utilizada uma expressão regular. Caso ocorra um match dessa expressão regular, basta ler o valor que causou o match para se encontrar o número de citações. A expressão regular utilizada é a seguinte: "*>Cited by \d**", onde "*\d**" pode assumir o valor de qualquer número. Caso não ocorra *match* dessa expressão regular, é assumido que a publicação não possua citações. Para o exemplo apresentado, a *string* que causa o *match* é a seguinte: *">Cited by 7"* e assim se pode extrair o número de citações facilmente.

4.7 Restrições do Google Scholar

Para evitar pesquisas automatizadas, que podem sobrecarregar sua infra-estrutura, o Google Scholar tenta barrar máquinas que realizem grande número de acessos aos seus servidores. Como a aplicação, por sua natureza, se enquadra nesse perfil, após o processamento de uma certa quantidade de currículos acaba sendo bloqueada.

A maneira de contornar isso, ao menos parcialmente, foi exibir ao usuário a página retornada pelo Google Scholar ao bloquear uma máquina. Essa página possui um código de segurança, exibido como imagem, que permite ao usuário desbloquear temporariamente seu acesso. Ao submeter esse código de segurança, é gerada uma *cookie* que serve como garantia de que o código foi digitado corretamente. Basta ao sistema armazenar esta *cookie* e fornecê-la sempre que uma nova requisição foi enviada aos servidores do Google Scholar.

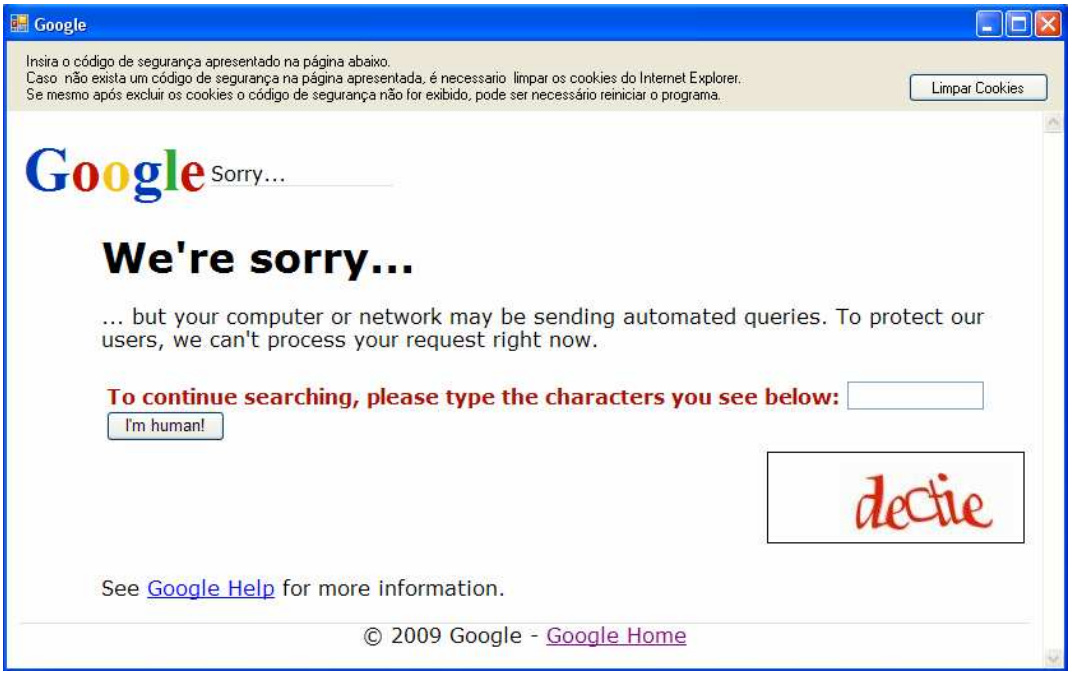

Figura 4.7: Tela de desbloqueio do Google Scholar

Essa maneira de contornar o bloqueio do Google Scholar possui dois problemas. O primeiro é que a *cookie* possui um tempo de validade, em torno de dez minutos. Após esse período o acesso volta a ser bloqueado. O segundo é que, por vezes, após expirar, a *cookie* não pode ser substituída por uma nova. Isso se deve ao fato de que gerenciamento de *cookies*, ao menos na implementação corrente, é feita pelo Internet Explorer, através de uma API. Esse gerenciamento impede a exclusão de *cookies* que estejam sendo utilizadas por alguma instância do Internet Explorer. Esse problema faz com que seja necessário reiniciar a aplicação toda vez que isso acontece.

5 CONCLUSÃO

Este trabalho apresentou o desenvolvimento de uma ferramenta cujo objetivo é realizar uma avaliação automatizada da relevância das publicações de um pesquisador. Essa avaliação é realizada através do cruzamento de informações obtidas no Currículo Lattes do pesquisador com informações referentes às suas publicações, obtidas através do Google Schollar. Essa integração de dados possibilita o cálculo de índices, como o *h*-index e o *g*-index, que servem como métricas para a avaliação da relevância de suas pesquisas.

Os resultados obtidos com testes do sistema, utilizando currículos de pesquisadores reais, principalmente do Instituto de Informática da UFGRS, mostram que a ferramenta realizou o cálculo de avaliações de forma bastante apurada, apresentando avaliações bastante próximas de seus valores esperados.

Os tempos de resposta observados variam conforme a quantidade de publicações a serem analisadas, mas pode-se utilizar como métrica de tempo uma média de 1,5 segundo para o processamento de cada publicação, sendo que esse valor pode variar conforme os tempos de resposta do Google Scholar.

O código fonte e os executáveis da ferramenta desenvolvida neste trabalho de conclusão podem ser obtidos através do site http://hlattes.codeplex.com. O código fonte está disponibilizado sob uma licença que permite sua livre alteração e redistribuição, podendo ser utilizado em futuros trabalhos de outros autores sem maiores problemas quanto ao seu licenciamento.

5.1 Melhorias e Trabalhos Futuros

Existem alguns pontos do sistema desenvolvido neste trabalho que poderiam ser melhorados e, de fato, só não o foram, por falta de tempo hábil.

Um desses pontos, que poderia ser melhorado no futuro, é a busca de pesquisador por nome. Essa funcionalidade depende de uma página de busca fornecida pela Plataforma Lattes que exibe apenas dez resultados por página. Atualmente a ferramenta utiliza apenas a primeira página de resultados, exibindo somente os dez primeiros resultados da pesquisa. Isso pode se tornar um problema em pesquisas por nomes muito comuns, que possuam mais de dez resultados. Essa pesquisa poderia ser expandida para as outras paginas de resultados, exibindo todos os pesquisadores encontrados, mas para isso devem ser realizadas alterações no *parser* responsável por esta funcionalidade.

Uma funcionalidade que pode ser interessante de se implementar no futuro é a exclusão de auto-referências. Entende-se como auto-referência uma citação a um trabalho de um determinado autor feita por ele mesmo algum trabalho distinto.

Obviamente esse tipo de citação não é tão relevante quanto às feitas por autores diferentes.

O Google Scholar apresenta recursos que permitem identificar quais publicações referenciam um determinado trabalho. Bastaria, para implementar essa funcionalidade, identificar quantas dessas publicações possuem o mesmo autor do trabalho pesquisado e subtraí-las do total de citações.

REFERÊNCIAS

HIRSCH, J. E. **An index to quantify an individual's scientific research output.** PNAS v. 102 n. 46 p. 16569-16572 November 15, 2005. Disponível em <http://www.pnas.org/content/102/46/16569.full/>. Acesso em novembro de 2009.

SCHOLAR. **Google Scholar.** Disponível em <http://scholar.google.com/>. Acesso em novembro de 2009.

LATTES. **Plataforma Lattes.** Disponível em <http://lattes.cnpq.br/>. Acesso em novembro de 2009.

JCR. **Journal Citation Reports.** Disponível em <http://thomsonreuters.com/>. Acesso em novembro de 2009.

POP. **Publish or Perish.** Disponível em <http://www.harzing.com/pop.htm/>. Acesso em novembro de 2009.

SCRIPTLATTES. **scriptLattes.** Disponível em <http://scriptlattes.sourceforge.net/>. Acesso em novembro de 2009.

WIKIPEDIA **h-index.** Disponível em <http://en.wikipedia.org/wiki/H-index/>. Acesso em novembro de 2009.

EGGHE, L. **Theory and practise of the g-index.** Scientometrics v. 69, n. 1 April, 2006

DOTNET. **Microsoft .NET Framework.** Disponível em <http://www.microsoft.com/NET/>. Acesso em novembro de 2009.

LOG4NET **Apache Log4Net.** Disponível em <http://logging.apache.org/log4net/index.html/>. Acesso em novembro de 2009.