

Universidade Federal do Rio Grande
Engenharia de Computação

Research.Net: um sistema para análise de redes de colaboração baseado na Plataforma Lattes

Lucas Rodrigues de Farias

Rio Grande, 17 de Abril de 2013

Universidade Federal do Rio Grande
Engenharia de Computação

Research.Net: um sistema para análise de redes de colaboração baseado na Plataforma Lattes

Lucas Rodrigues de Farias

Trabalho de Conclusão do Curso de Graduação em Engenharia de Computação submetido à avaliação como requisito parcial à obtenção do título de Engenheiro de Computação.

Orientador(a): Prof. Msc. Eduardo Nunes Borges

Co-orientador(a): Prof. Msc. André Prisco Vargas

Rio Grande, 17 de Abril de 2013

Este trabalho foi analisado e julgado adequado para a obtenção do título de Engenheiro de Computação e aprovado em sua forma final pelo orientador.

Prof. Msc. Eduardo Nunes Borges

Banca Examinadora:

Prof. Msc. Eduardo Nunes Borges

C3 – FURG (Orientador)

Prof. Msc. André Prisco Vargas

C3 – FURG

Profa. Dra. Ana Marilza Pernas

CDTec - UFPel

Profa. Dra. Karina dos Santos Machado

C3 – FURG

Aos meus pais, que me apoiaram incondicionalmente e, nos momentos difíceis, não me deixaram desistir.

Agradecimentos

Aos meus pais, Ciro e Rosangela, e minhas irmãs, Cibele e Cinara, que sempre estiveram comigo nos bons e maus momentos.

Aos meus orientadores, Eduardo e André, por todo o ensinamento, paciência e incentivo ao longo dos quase dois anos de convivência.

Aos colegas de classe que fizeram essa trajetória comigo, especialmente ao Rafael e ao Cássio, companheiros de inúmeros trabalhos.

Aos amigos, em especial ao Lânderson, ao Tiago e à Francine, que por vezes até me tiraram o foco do trabalho, mas também entenderam os momentos de ausência.

A todos colegas e professores da FURG que, de alguma forma, contribuíram para a minha graduação.

Conteúdo

Lista de Figuras	iii
Lista de Tabelas	vi
Lista de Abreviaturas	vii
Resumo	viii
Abstract	ix
1 Introdução	1
2 Trabalhos Relacionados	5
2.1 ArnetMiner	5
2.2 CiênciaBrasil	6
2.3 ScriptLattes	7
2.4 Análise dos trabalhos relacionados	8
3 Arquitetura do Sistema	11
3.1 Identificação de Rélicas	13
3.2 Banco de Dados	19
3.3 Cálculo das Métricas	21
3.4 Processamento da Rede	22
4 Implementação	25
5 Protótipo	30

6 Considerações Finais	40
Referências	44
Apêndices	47
A Diagrama de Classes do Cliente	47
B Diagrama de Classes do Servidor	49

Listas de Figuras

1.1	Múltiplas representações de uma mesma referência bibliográfica.	2
2.1	Rede de colaboração gerada pelo sistema ArnetMiner. Obtida em 4 de junho de 2012.	6
2.2	Rede de colaboração obtida no portal CiênciaBrasil. Obtida em 7 de março de 2013.	7
2.3	Rede de colaboração obtida com a ferramenta scriptLattes. Obtida em 7 de março de 2013.	8
2.4	Relatório da produção científica de um grupo de pesquisadores obtido com a ferramenta scriptLattes. Obtido em 7 de março de 2013.	9
3.1	Diagrama da arquitetura do sistema.	11
3.2	Algoritmo de identificação de réplicas usado no módulo <i>Nova Coleta</i>	13
3.3	Criação das listas de publicações de cada pesquisador divididas por ano, e das listas de referências únicas e de pares.	14
3.4	Análise das publicações do ano de 2011.	15
3.5	Análise das publicações do ano de 2012.	16
3.6	Estado final das listas de referências únicas e de pares, após a análise de todas as publicações.	16
3.7	Análise das referências bibliográficas de um terceiro pesquisador.	17
3.8	Comparação da referência do autor com as referências únicas do mesmo ano. .	17
3.9	Par de referências identificadas como réplicas é armazenado na Lista de Pares.	18

3.10	Duas referências que são réplicas da mesma referência de BORGES, mas não foram identificadas como réplicas.	18
3.11	Estado final das listas de referências únicas e de pares, e o <i>cluster</i> formado pelo algoritmo.	19
3.12	Diagrama relacional do esquema <i>Coleta</i>	21
3.13	Diagrama relacional do esquema <i>Aplicação</i>	22
3.14	Consulta SQL que cria uma tabela temporária no banco de dados con- tendo o conjunto de adjacências da rede a ser processada. No exemplo, é processada a rede de <i>id</i> 1, ano inicial 2008 e ano final 2012	23
4.1	Diagrama de classes simplificado do cliente.	26
4.2	Diagrama de classes do servidor.	28
5.1	Tela de <i>login</i> . Usuário pode acessar a ferramenta ou criar um novo usuário.	31
5.2	Tela de redes.	31
5.3	Edição das propriedades da rede.	33
5.4	Inclusão de um novo membro na rede.	33
5.5	Membro clicado destacado em vermelho, com os seus coautores destacados em amarelo.	34
5.6	Listagem das publicações do pesquisador selecionado.	34
5.7	Dois membros com publicações em coautoria selecionados.	35
5.8	Listagem das publicações em comum dos pesquisadores selecionados. . . .	35
5.9	Visualização das diferentes representações da referência bibliográfica sele- cionada na lista de publicações em comum, destacada em cinza.	36
5.10	Visualização do número de publicações do pesquisador no período.	36
5.11	Visualização das colaborações entre pesquisadores, representada pelos pe- sos das arestas.	37
5.12	Valores da métrica <i>Grau</i> de cada pesquisador.	37
5.13	Valores da métrica <i>PageRank</i> de cada pesquisador.	38
5.14	Valores da métrica <i>ClusterCoefficient</i> de cada pesquisador.	38
5.15	Valores da métrica <i>Closeness</i> de cada pesquisador.	39
5.16	Destacadas em vermelho, as arestas que formam um dos <i>Diâmetros</i> da rede.	39

6.1	Rede de colaboração dos pesquisadores do Centro de Ciências Computacionais (C3), considerando publicações entre 2008 e 2012, apresentando os resultados da métrica <i>PageRank</i> para cada um dos pesquisadores.	42
A.1	Diagrama de classes do cliente.	48
B.1	Diagrama de classes do servidor.	49

Lista de Tabelas

2.1	Comparativo entre os trabalhos relacionados	10
4.1	<i>Strings</i> de mensagens definidas pelo protocolo criado.	29
6.1	Comparativo das características dos trabalhos relacionados e do Research.Net	41

Listas de Abreviaturas

C3 Centro de Ciências Computacionais

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

CRUD Create, Read, Update and Delete

FURG Universidade Federal do Rio Grande

IES Instituição de Ensino Superior

INCT Institutos Nacionais de Ciência e Tecnologia

SGBD Sistema de Gerenciamento de Banco de Dados

Resumo

Este trabalho descreve um sistema de informação denominado Research.Net que tem por objetivo construir e analisar redes de colaboração acadêmica baseadas na produção científica extraída dos currículos dos pesquisadores. O sistema proposto analisa os dados extraídos em busca de referências bibliográficas redundantes e gera um relacionamento de coautoria entre os autores para cada referência replicada. A rede de colaboração gerada pode ser visualizada graficamente e de forma interativa. É possível visualizar a lista de publicações em comum a dois pesquisadores quaisquer e avaliar a qualidade da identificação de réplicas. Também foram implementadas várias métricas de redes sociais que permitem entender melhor como se comportam as interações entre pesquisadores.

Abstract

This paper describes an information system called Research.Net that aims to build and analyze academic collaboration networks based on scientific production extracted from researchers curricula. The proposed system analyzes the extracted data searching for redundant bibliographic references and generates co-author relationships between the authors for each replicated reference. The collaboration network may be graphically and interactively visualized. It is possible to visualize the list of co-authored publications of any two authors and to evaluate the replica identification quality. Also, several social networks metrics were implemented, which help to understand the authors interactions behavior.

Capítulo 1

Introdução

Atualmente, analisar a produção científica de uma instituição de ensino como um todo é uma tarefa bastante difícil. Além do elevado número de publicações dos servidores da instituição, as conferências e revistas científicas que contêm estas publicações, nacionais ou internacionais, podem não estar indexadas por ferramentas de busca como *SciELO*¹, *Scopus*² e *Google Scholar Citations*³.

Além disso, podem existir múltiplas fontes de dados contendo citações para as mesmas publicações, mas que diferem na forma com que foram escritas e representadas, o que poderia ser identificado como publicações distintas. A figura 1.1, apresenta um exemplo contendo representações diferentes da mesma referência bibliográfica. Existem várias diferenças entre as representações: quantidade de autores citados, forma de escrita dos nomes, diferenças nos títulos do artigo e do veículo de publicação, entre outras.

A identificação de réplicas, ou deduplicação, é a tarefa de identificar dois ou mais registros em um repositório de dados que se referem à mesma entidade no mundo real. A identificação de registros duplicados pode ser difícil por vários motivos, como variações na grafia, estilo de escrita, padrão de metadados utilizado, ou até mesmo erros ortográficos (Borges et al., 2011).

A Plataforma Lattes do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) é uma base de dados que contém, entre outras informações, os currículos da

¹<http://www.scielo.org>

²<http://www.scopus.com>

³<http://scholar.google.com/citations>

Elmagarmid, A. et al., 2007. Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 1, pp. 1-16.

Ahmed K. Elmagarmid, Vassilios S. Verykios. Duplicate Record Detection: A Survey, IEEE TKDE, 2007.

Elmagarmid, A.K.; Ipeirotis, P.G.; Verykios, V.S. Duplicate record detection. *Knowledge and Data Engineering, IEEE Trans.*, 19(1), 2007.

Figura 1.1: Múltiplas representações de uma mesma referência bibliográfica.

maior parte dos pesquisadores que atuam no Brasil. Grande parte dos editais de financiamento de projetos feitos por instituições de amparo à pesquisa, como o próprio CNPq, utilizam os currículos Lattes dos pesquisadores como uma das formas de avaliação das propostas. Este fato motiva os pesquisadores a manter seus currículos com informações corretas e atualizadas, tornando a Plataforma Lattes uma fonte adequada para análise da produção científica brasileira.

Para analisar a produção científica de uma Instituição de Ensino Superior (IES) usando a Plataforma Lattes, é necessário extrair informações coletivas de dados individuais dos pesquisadores. A partir destas informações, é possível inferir relacionamentos de coautoria entre os pesquisadores, que podem ser usados para gerar redes de colaboração acadêmica, uma especialização de rede social.

Uma rede social é um conjunto de pessoas ou grupos de pessoas com algum padrão de relacionamento ou interações entre eles (Newman, 2003). Esses relacionamentos podem ser de vários tipos. Por exemplo, relações de amizade (*Facebook*⁴), profissionais (*LinkedIn*⁵), educacionais (*Moodle*⁶) ou proporcionadas pela necessidade de algum prestador de serviço (*Betterfly*⁷). Quando uma rede social é formada por pesquisadores e o relacionamento entre eles é acadêmico, ou seja, por participação nos mesmos grupos de pesquisa, por coautoria de publicações ou até mesmo por uma relação de orientador e orientando, a

⁴<http://www.facebook.com>

⁵<http://www.linkedin.com/>

⁶<http://moodle.org/>

⁷<http://www.betterfly.com/>

rede social é denominada rede de colaboração acadêmica. Neste trabalho usamos o termo simplificado “rede de colaboração” como sinônimo de rede de colaboração acadêmica.

Sendo assim, o objetivo deste trabalho consiste em desenvolver um sistema de informação que facilite a análise das redes de colaboração acadêmica entre membros de uma ou mais IES utilizando dados individuais dos pesquisadores extraídos dos currículos Lattes. A aplicação chamada Research.Net também remove a redundância da produção bibliográfica, estabelece os relacionamentos de coautoria a partir das redundâncias, e produz uma visualização gráfica da rede de colaboração, além de implementar um conjunto de métricas de redes que permitem ao usuário entender melhor como as colaborações entre os pesquisadores se comportam.

Algumas ferramentas como *Pajek* (Batagelj and Mrvar, 2002), *Gephi* (Bastian et al., 2009) e *GraphViz* (Ellson et al., 2003) focam apenas na visualização de dados e recebem como entrada uma rede de colaboração previamente processada e modelada por meio de um grafo. *ArnetMiner*⁸ (Tang et al., 2008) e *Microsoft Academic Search*⁹ integram dados de múltiplas fontes e possuem interface *web*. Outras ferramentas como *SocNetV*¹⁰ e *UCINET*¹¹ (Borgatti et al., 2002) permitem análises baseadas em métricas de redes sociais, mas não incorporam os resultados das métricas na visualização da rede, tornando a interpretação dos dados mais difícil.

Diferentemente das ferramentas anteriores, este trabalho envolve todo o processo, desde a geração da rede de colaboração com dados extraídos da *web* até a análise de métricas de redes sociais. Esse processo inclui a visualização de redes, a exploração visual das referências bibliográficas originais nos currículos de cada pesquisador, a identificação de réplicas e sua avaliação pelos usuários, e a apresentação dos resultados das métricas embutido na visualização da rede.

O restante do texto está organizado da seguinte forma. No capítulo 2, são discutidos os trabalhos relacionados. No capítulo 3, é apresentada a arquitetura do sistema desenvolvido. No capítulo 4, são apresentados aspectos técnicos da implementação do sistema. No capítulo 5, é apresentado um protótipo funcional do sistema proposto, com um exem-

⁸<http://www.arnetminer.com>

⁹<http://academic.research.microsoft.com>

¹⁰<http://socnetv.sourceforge.net>

¹¹<http://www.analytictech.com/ucinet>

plo de uso da ferramenta por um usuário. Por fim, no capítulo 6, são apresentadas as conclusões, sintetizando as principais contribuições do sistema proposto. Além disso, são apresentadas as publicações resultantes do trabalho, assim como os trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

Neste capítulo são apresentados diversos trabalhos existentes na área de redes de colaboração acadêmica. Ao final do capítulo, é feita uma análise dos trabalhos relacionados apresentados.

2.1 ArnetMiner

O sistema ArnetMiner (Tang et al., 2008) fornece serviços de busca *online* e mineração de dados para redes sociais de pesquisadores. Ele constrói perfis para os pesquisadores integrando informações acadêmicas e bibliográficas de múltiplas fontes de dados. Os relacionamentos entre os pesquisadores são baseados na coautoria da produção bibliográfica. Entre os principais recursos destacam-se:

- busca por especialistas em um determinado tema de pesquisa;
- *rankings* acadêmicos baseados em diversas métricas.

A figura 2.1 apresenta um exemplo de rede de colaboração gerado pelo sistema ArnetMiner. O elemento central representa a pesquisadora Sílvia Botelho do C3 da Universidade Federal do Rio Grande (FURG). Os outros elementos ligados a ela representam orientadores (em vermelho), orientandos (em amarelo) e outros coautores (em verde). É possível aumentar o nível de profundidade da rede clicando nos seus pesquisadores, possibilitando outras análises, como, por exemplo, quem orienta um ex-aluno atualmente, ou com quem um colega de trabalho costuma interagir.

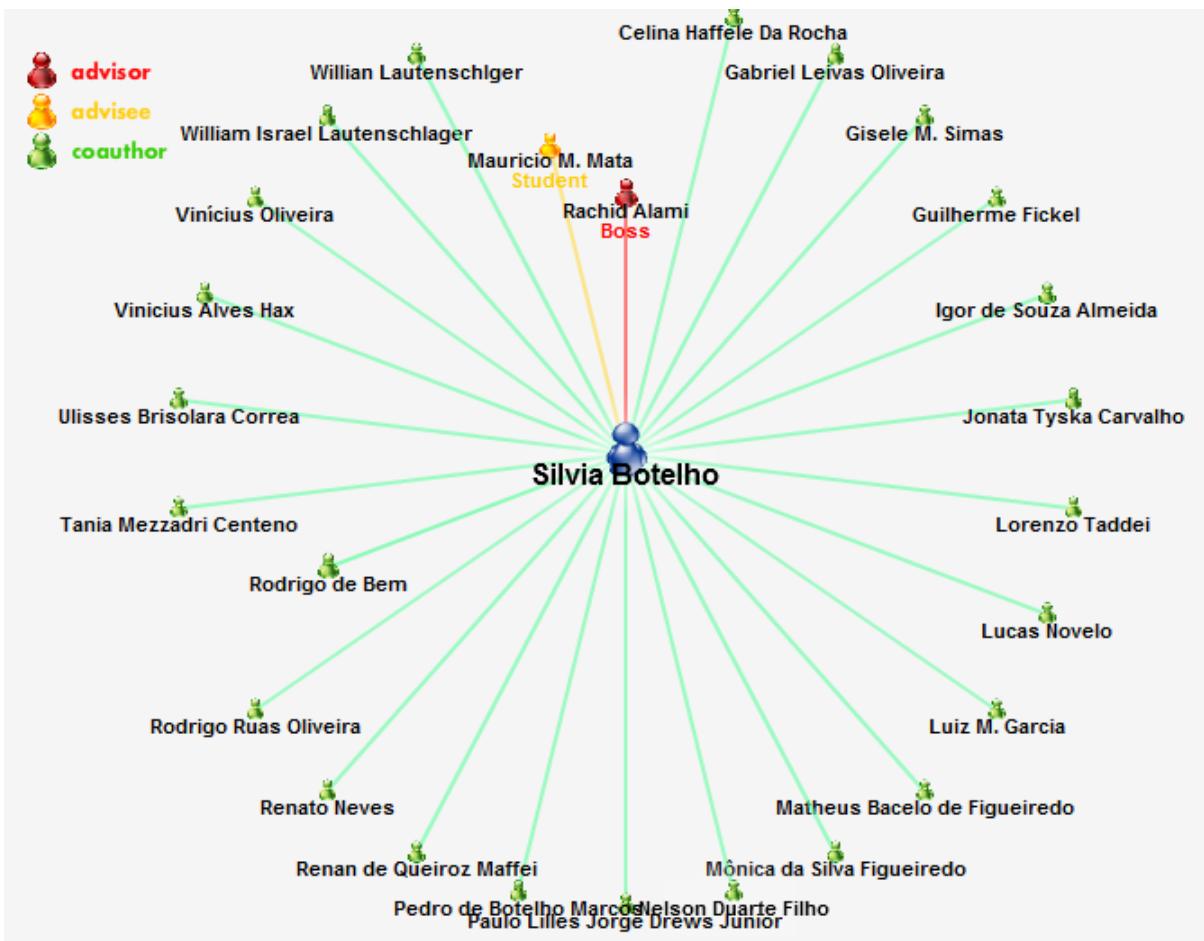


Figura 2.1: Rede de colaboração gerada pelo sistema ArnetMiner. Obtida em 4 de junho de 2012.

2.2 CiênciaBrasil

O portal CiênciaBrasil (Laender et al., 2011a) permite analisar e visualizar informação sobre pesquisadores brasileiros que participam dos Institutos Nacionais de Ciência e Tecnologia (INCT). Entre os principais recursos oferecidos pelo portal estão as visualizações das redes de colaboração entre os pesquisadores de cada instituto. A figura 2.2 apresenta um exemplo de rede de colaboração formada pelos pesquisadores do INCT da web. O sistema destaca as interações criadas e intensificadas após a formação de um INCT. As redes sociais são construídas de forma automática a partir de dados coletados da Plataforma Lattes e das páginas web dos INCT (Laender et al., 2011b). CiênciaBrasil identifica os relacionamentos de coautoria através de um algoritmo específico para deduplicação de citações bibliográficas (Borges et al., 2011).

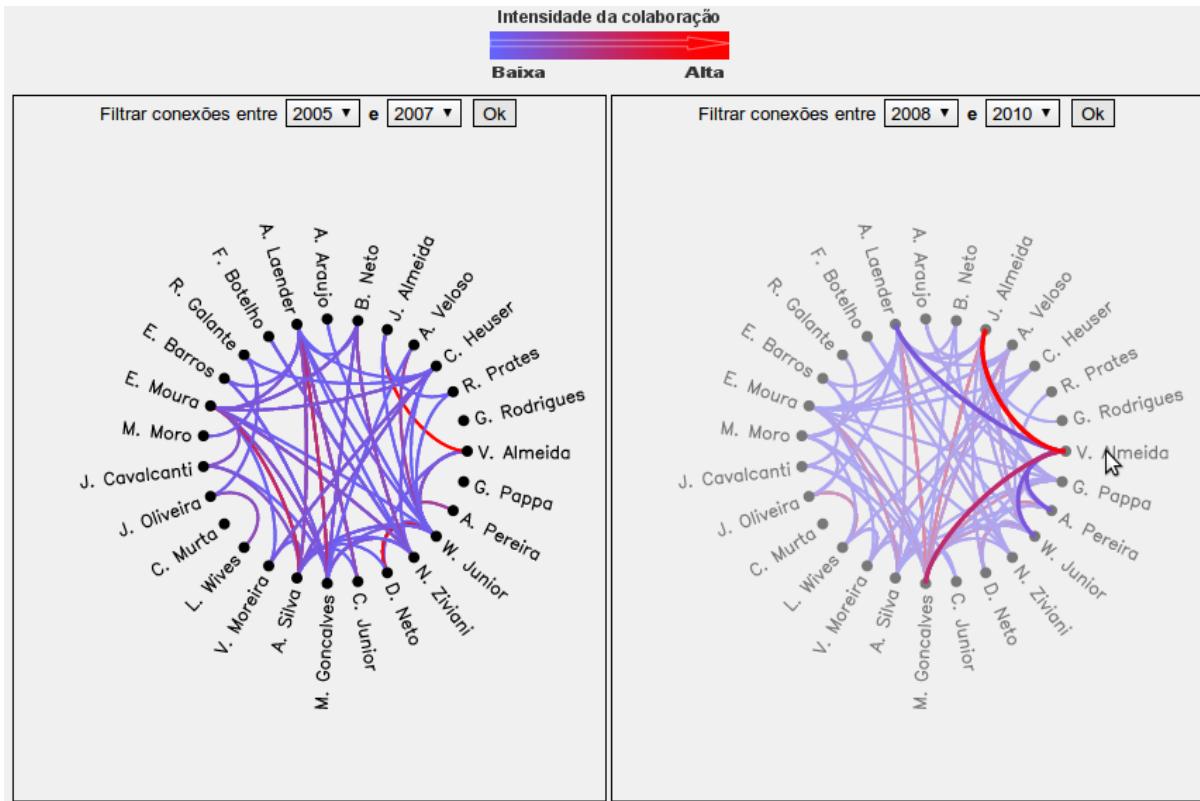


Figura 2.2: Rede de colaboração obtida no portal CiênciaBrasil. Obtida em 7 de março de 2013.

2.3 ScriptLattes

O scriptLattes (MenaChalco and Cesar Junior, 2009) é uma ferramenta de código-fonte aberto que extrai dados de um conjunto de currículos Lattes e gera relatórios a partir deles. Estes relatórios contêm uma lista de todas as publicações do conjunto, com tratamento das publicações replicadas, gráficos da produção científica e um grafo das redes de colaboração entre os pesquisadores. A figura 2.3 apresenta um exemplo de rede de colaboração formada por alguns professores do C3 da FURG obtido com a ferramenta scriptLattes.

Na figura 2.4 temos um exemplo de um relatório gerado pelo scriptLattes. São descritas informações dos artigos completos publicados em periódicos produzidos pelos pesquisadores da rede de colaboração da figura 2.3 entre 2008 e 2012. A ferramenta gera um gráfico anual de publicações, além de apresentar a referência bibliográfica de cada uma delas, com *links* para outras ferramentas, como Google Scholar e Microsoft Academic Search.

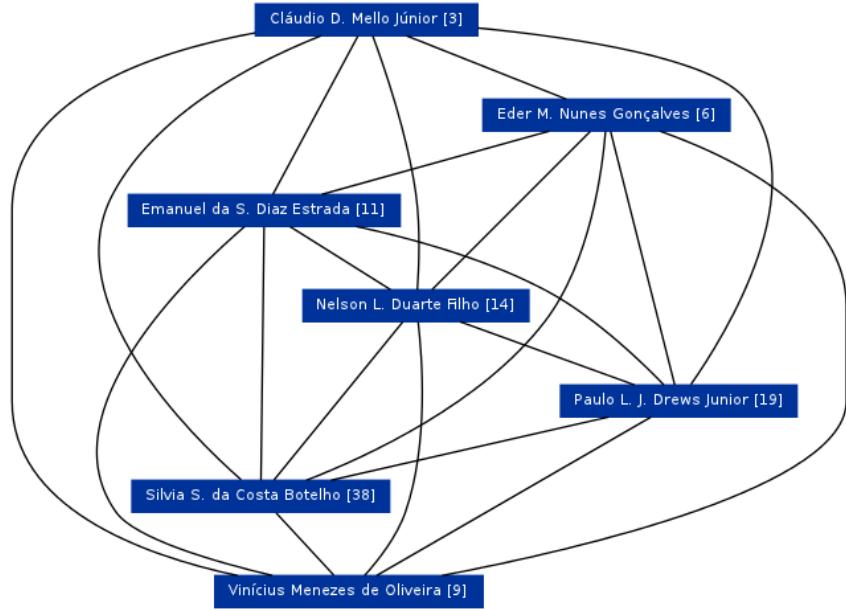


Figura 2.3: Rede de colaboração obtida com a ferramenta scriptLattes. Obtida em 7 de março de 2013.

2.4 Análise dos trabalhos relacionados

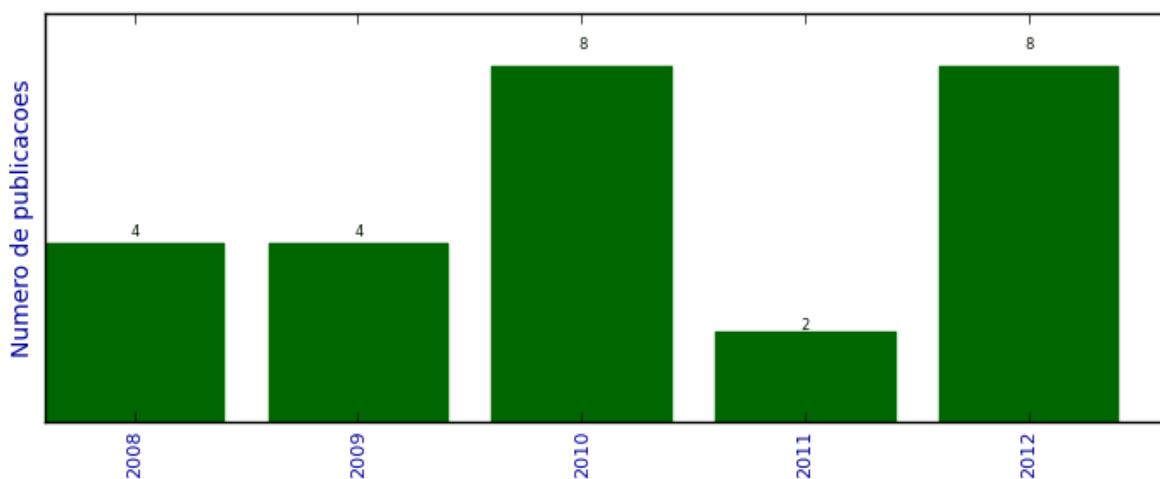
Os grafos gerados pelo portal CiênciaBrasil e pela ferramenta scriptLattes possuem alguns problemas. Primeiramente, as imagens geradas são estáticas, não sendo possível reposicionar vértices e arestas do grafo. Em redes complexas, com um número elevado de vértices e arestas, isto pode ser um grande empecilho para a visualização e entendimento das colaborações entre os pesquisadores.

Nenhum dos sistemas estudados permite visualizar quais são as produções bibliográficas que dois pesquisadores publicaram em conjunto, ou seja, que artigos/livros/trabalhos geraram a aresta entre esses dois pesquisadores. Também não é possível visualizar como cada referência bibliográfica está representada em sua forma original em cada fonte de dados de onde o relacionamento foi extraído. O scriptLattes não possui nenhuma funcionalidade temporal, não permitindo analisar o comportamento da rede de colaboração ao longo do tempo.

A tabela 2.1 resume o capítulo apresentando um conjunto de características importantes, algumas identificadas nos trabalhos relacionadas e outras desejáveis. As características apresentadas são:

Teste do scriptLattes

Artigos completos publicados em periódicos



2012

1. GONÇALVES, E. M. N.. **Specifying Knowledge in Cognitive Multiagent Systems Using a Class of Hierarchical Petri Nets.** *Journal of Software.* v. 7, p. 2405-2414, 2012. [doi](#)
[\[citações Google Scholar\]](#) [\[citações Microsoft Acadêmico\]](#) [\[busca Google\]](#)
2. JACOBI, E. ; POPOLEK, P. ; HAX, V. ; Tyska Carvalho, Jonata ; Duarte Filho, Nelson L. ; MENDIZABAL, Odorico. **Modelo de Computação em Nuvem e sua Aplicabilidade.** *Revista Junior de Iniciação Científica em Ciências Exatas e Engenharia.* v. I, p. 1-6, 2012.
[\[citações Google Scholar\]](#) [\[citações Microsoft Acadêmico\]](#) [\[busca Google\]](#)
3. M, Santin ; SILVA, J. ; Botelho, Silvia S. C.. **TOPOBO: Aspectos motivacionais do uso da robótica com crianças.** *RENOTE. Revista Novas Tecnologias na Educação.* v. 10, p. 1-11, 2012.
[\[citações Google Scholar\]](#) [\[citações Microsoft Acadêmico\]](#) [\[busca Google\]](#)
4. SANTIN, M. M. ; SILVA, J. A. ; BOTELHO, S. S. C.. **Artefatos educacionais com memória cinética Topobo: uma abordagem para o currículo dos anos iniciais.** *RENOTE. Revista Novas Tecnologias na Educação.* v. 10, p. 1-11, 2012.
[\[citações Google Scholar\]](#) [\[citações Microsoft Acadêmico\]](#) [\[busca Google\]](#)
5. SANTIN, M. M. ; SILVA, J. A. ; BOTELHO, S. S. C.. **Aspectos motivacionais do uso da robótica com crianças.** *RENOTE. Revista Novas Tecnologias na Educação.* v. 10, p. 10-10, 2012.
[\[citações Google Scholar\]](#) [\[citações Microsoft Acadêmico\]](#) [\[busca Google\]](#)

Figura 2.4: Relatório da produção científica de um grupo de pesquisadores obtido com a ferramenta scriptLattes. Obtido em 7 de março de 2013.

Tabela 2.1: Comparativo entre os trabalhos relacionados

Característica	ArnetMiner	CiênciaBrasil	scriptLattes
Busca por Especialistas	✓		
Rankings acadêmicos	✓		✓
Geração de relatórios da produção		✓	✓
Visualizações dinâmicas	✓		
Listagem de publicações em comum			
Visualização de múltiplas representações das referências bibliográficas			
Temporalidade das redes de colaboração		✓	
Avaliação da identificação de referências bibliográficas duplicadas			

- Busca por especialistas – determinando uma área do conhecimento, o sistema encontra especialistas nesta área;
- *Rankings* acadêmicos – pesquisadores classificados por diversas métricas, como H -index ou número de citações;
- Geração de relatórios – relatórios textuais com número de publicações por ano, referências destas publicações, etc;
- Visualizações dinâmicas – visualizações nas quais o usuário possa interagir, rearranjando vértices e arestas;
- Listagem de publicações em comum – a partir de dois pesquisadores, ver em quais publicações eles colaboraram entre si;
- Visualização de múltiplas representações das referências bibliográficas – visualizar como as referências bibliográficas estão descritas em diferentes locais, como nos currículos de dois pesquisadores;
- Temporalidade das redes de colaboração – variação da rede de colaboração ao longo do tempo;
- Avaliação da identificação de referências bibliográficas duplicadas – usuário indicar ao sistema se a identificação de duplicatas está sendo feita corretamente ou não.

Capítulo 3

Arquitetura do Sistema

A arquitetura do Research.Net está representada graficamente na figura 3.1. O sistema foi desenvolvido em um modelo de cliente-servidor, no qual a ferramenta cliente é responsável apenas pela descrição e visualização das redes, enquanto o servidor é responsável pelo restante do processo.

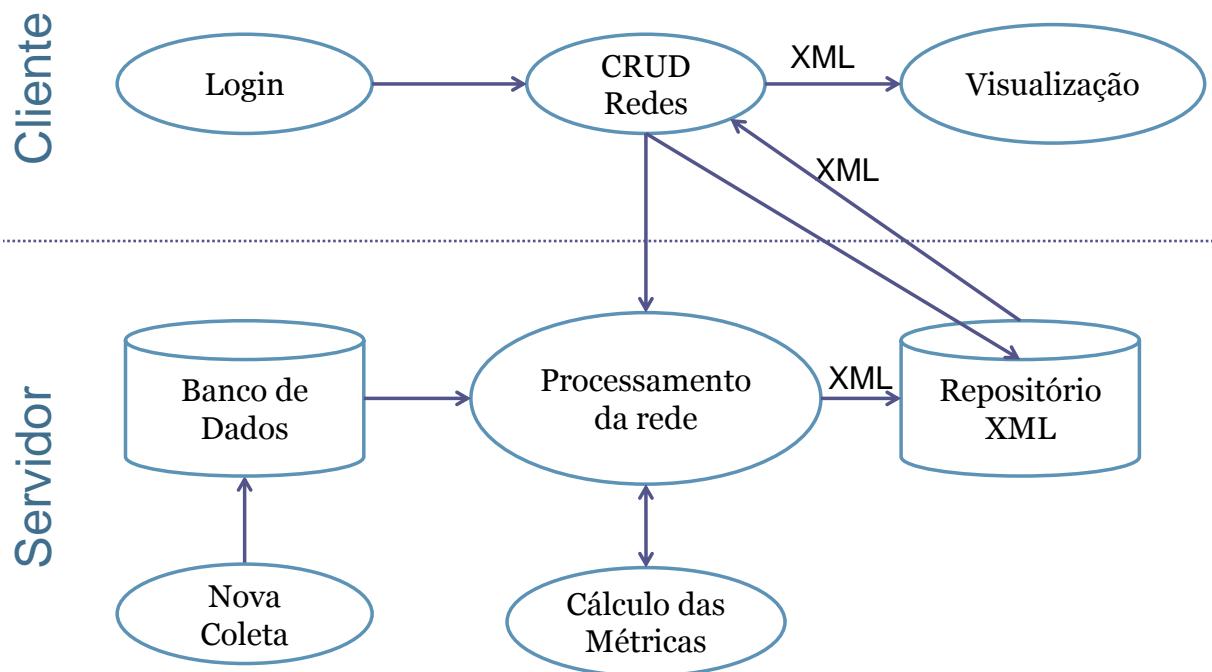


Figura 3.1: Diagrama da arquitetura do sistema.

No cliente, o usuário pode registrar-se na ferramenta no módulo de *Login*, tendo acesso ao módulo de descrição de redes chamado *CRUD Redes*. CRUD é o acrônimo para *Create*,

Read, Update and Delete, ou seja, criar, ler, atualizar e apagar, que são as operações básicas que o usuário tem a disposição. Neste módulo, o usuário pode ver informações de redes públicas criadas por outros usuários ou por si mesmo. Além disso, ele pode buscar pesquisadores disponíveis no servidor e formar uma rede de colaboração a partir deles. O usuário também define nome, ano inicial e final e a visibilidade desta rede para outros usuários. As informações desta nova rede de colaboração são enviadas ao servidor, que ficará encarregado de gerá-la.

O processamento das redes de colaboração é dividido em várias etapas. A primeira etapa é o módulo *Nova Coleta*, responsável por fazer a coleta, extração e análise de dados. A partir de uma lista de currículos Lattes previamente selecionada, é feito o *download* dos dados dos pesquisadores, como informações pessoais e referências bibliográficas. Os dados são analisados em busca de potenciais referências bibliográficas duplicadas. Estas réplicas são usadas para estabelecer as relações de coautoria entre os pesquisadores. Os resultados da análise e os dados extraídos são então armazenados no *Banco de Dados*.

O módulo *Processamento da Rede* é responsável por gerar os arquivos das redes de colaboração. Ele recebe as informações da rede criada pelo cliente e gera uma nova rede de colaboração dentro do banco. Os dados desta rede são enviados ao módulo de *Cálculo de Métricas*, que calcula as métricas de redes sociais, tais como *PageRank*, coeficiente de clusterização e diâmetro. Após, é gerado um arquivo XML no formato GraphML (Brandes et al., 2002), contendo a descrição da rede. O GraphML é um formato de XML criado para a descrição de grafos. Ele facilita a criação de grafos através de estruturas específicas para descrição de vértices e arestas direcionadas ou não direcionadas, e permite a inclusão de novas características nos elementos. O arquivo gerado é armazenado no *Repositório XML*.

Após a criação do arquivo GraphML, a rede fica disponível para *download* pelo usuário. A partir do *download* deste arquivo, o cliente pode acessar o próximo módulo, de *Visualização*. Uma visualização gráfica da rede de colaboração é gerada por este módulo, que desenha um grafo no qual os vértices são os pesquisadores, e as arestas representam as relações de coautoria. Junto ao grafo são apresentados os resultados das métricas previamente calculadas.

3.1 Identificação de Rélicas

O algoritmo usado no processo de identificação de rélicas do módulo *Nova Coleta* é apresentado em pseudocódigo na figura 3.2. Da linha 1 até a linha 11 é feito o processo de identificação de rélicas, utilizando um algoritmo disponível no trabalho relacionado ScriptLattes. O restante do algoritmo foi uma extensão desenvolvida para suprir uma necessidade deste trabalho.

```

1 organizar as referências bibliográficas por autor
2 separá-las em blocos delimitados pelo ano de publicação
3 criar uma lista de referências únicas (deduplicadas)
4 criar uma lista de pares de referências
5 para cada pesquisador:
6     para cada referência bibliográfica:
7         para referências únicas do mesmo ano:
8             se similaridade (referencia_autor, referencia_unica) > limiar:
9                 armazena referências na lista de pares
10            se não encontrou similar
11                armazena referência na lista de referências únicas
12    para cada par de referências:
13        se referência 1 do par não possui cluster
14            cluster da referência 1 = id da referência 1
15        cluster da referência 2 = cluster da referência 1

```

Figura 3.2: Algoritmo de identificação de rélicas usado no módulo *Nova Coleta*

No primeiro passo do algoritmo, as publicações são organizadas como na figura 3.3. Cada pesquisador possui uma lista de suas publicações classificadas por ano, em um processo chamado de blocagem (Elmagarmid et al., 2007), facilitando o trabalho de filtragem dos pares candidatos. Essa estratégia reduz drasticamente o número de pares candidatos ao casamento. O processo de identificação de rélicas é acelerado consideravelmente, já que são comparados apenas esses pares, e não o conjunto de pares formado pelo produto cartesiano do conjunto completo de publicações. Além das listas de cada pesquisador, são criadas também duas listas vazias, a primeira chamada de lista de referências únicas,

que irá armazenar todas as publicações já deduplicadas ou que não possuem réplicas, e a segunda chamada de lista de pares, que irá armazenar todos os pares de referências bibliográficas similares.

Listas de BORGES, Eduardo .N.

Título	Ano
Um sistema para análise de redes de pesquisa...	2012
ARGOsearch: an information retrieval system...	2012
CiênciaBrasil - The Brazilian Portal of Science and...	2011
An Automatic Approach for Duplicate Bibliographic...	2011

2011

Título
CiênciaBrasil - The Brazilian Portal of Science and...
An Automatic Approach for Duplicate Bibliographic...

2012

Título
Um sistema para análise de redes de pesquisa...
ARGOsearch: an information retrieval system...

Listas de FARIAS, Lucas R.

Título	Ano
Um sistema para análise de redes de pesquisa...	2012
Rede de Pesquisa da FURG	2011

2011

Título
Rede de Pesquisa da FURG

2012

Título
Um sistema para análise de redes de pesquisa...

Lista de Referências Únicas

Título	Ano

Lista de Pares

Título 1	Título 2

Figura 3.3: Criação das listas de publicações de cada pesquisador divididas por ano, e das listas de referências únicas e de pares.

Para cada pesquisador, o algoritmo percorre a lista de publicações, comparando as publicações desse autor com as publicações de mesmo ano da lista de referências únicas. Se o algoritmo encontrar alguma publicação na lista de referências únicas que seja similar à publicação analisada, ele guarda o par com as duas publicações na lista de pares, e a publicação analisada não entra na lista de referências únicas. Se percorrendo toda a lista de referências únicas não for encontrada nenhuma produção similar, a publicação

é incluída na lista de referências únicas. Perceba que, na primeira execução do laço de repetição, que começa na linha 5 do algoritmo, a lista de referências únicas será preenchida com todas as publicações do primeiro autor analisado. A similaridade das referências bibliográficas é baseada na distância de edição Levenshtein normalizada (Levenshtein, 1966). Duas referências são identificadas como réplicas quando a distância entre os títulos é menor que um dado limiar.

Nas figuras 3.4, 3.5 e 3.6 é demonstrado um exemplo do processo descrito no parágrafo anterior. A figura 3.4 mostra as listas após a deduplicação de todas as publicações do pesquisador BORGES, Eduardo N. Neste momento, o algoritmo analisa todas as publicações do pesquisador FARIAS, Lucas R. no ano de 2011, como destacado em vermelho na imagem.

Listas de FARIAS, Lucas R.													
2011	2012												
<table border="1"> <thead> <tr> <th>Título</th><th>Ano</th></tr> </thead> <tbody> <tr> <td>Rede de Pesquisa da FURG</td><td></td></tr> </tbody> </table>	Título	Ano	Rede de Pesquisa da FURG		<table border="1"> <thead> <tr> <th>Título</th><th>Ano</th></tr> </thead> <tbody> <tr> <td>Um sistema para análise de redes de pesquisa...</td><td></td></tr> </tbody> </table>	Título	Ano	Um sistema para análise de redes de pesquisa...					
Título	Ano												
Rede de Pesquisa da FURG													
Título	Ano												
Um sistema para análise de redes de pesquisa...													
<hr/>													
<table> <thead> <tr> <th colspan="2">Lista de Referências Únicas</th> </tr> <tr> <th>Título</th><th>Ano</th></tr> </thead> <tbody> <tr> <td>Um sistema para análise de redes de...</td><td>2012</td></tr> <tr> <td>ARGOsearch: an information retrieval...</td><td>2012</td></tr> <tr> <td>CiênciaBrasil - The Brazilian Portal of Science...</td><td>2011</td></tr> <tr> <td>An Automatic Approach for Duplicate...</td><td>2011</td></tr> </tbody> </table>		Lista de Referências Únicas		Título	Ano	Um sistema para análise de redes de...	2012	ARGOsearch: an information retrieval...	2012	CiênciaBrasil - The Brazilian Portal of Science...	2011	An Automatic Approach for Duplicate...	2011
Lista de Referências Únicas													
Título	Ano												
Um sistema para análise de redes de...	2012												
ARGOsearch: an information retrieval...	2012												
CiênciaBrasil - The Brazilian Portal of Science...	2011												
An Automatic Approach for Duplicate...	2011												
<table> <thead> <tr> <th colspan="2">Lista de Pares</th> </tr> <tr> <th>Título 1</th><th>Título 2</th></tr> </thead> <tbody> <tr> <td></td><td></td></tr> <tr> <td></td><td></td></tr> <tr> <td></td><td></td></tr> </tbody> </table>		Lista de Pares		Título 1	Título 2								
Lista de Pares													
Título 1	Título 2												

Figura 3.4: Análise das publicações do ano de 2011.

Como a referência analisada não é similar à nenhuma outra da lista de referências únicas, ela é incluída nessa lista, como visto na figura 3.5. A seguir, o algoritmo analisa as publicações do ano de 2012, como destacado. Como a referência analisada é similar à uma publicação existente na lista de referências únicas, as duas referências são armazenadas na lista de pares, como visto na figura 3.6.

Após o algoritmo iterar sobre todas as publicações de todos os pesquisadores, todas as publicações identificadas como replicadas estão guardadas na lista de pares. Entretanto, podemos ter um caso em que duas publicações são similares mas que não foram compa-

Listas de FARIAS, Lucas R.

2011

Título
Rede de Pesquisa da FURG

2012

Título
Um sistema para análise de redes de pesquisa...

Lista de Referências Únicas

Título	Ano
Um sistema para análise de redes de...	2012
ARGOsearch: an information retrieval...	2012
CiênciaBrasil - The Brazilian Portal of Science...	2011
An Automatic Approach for Duplicate...	2011
Rede de Pesquisa da FURG	2011

Lista de Pares

Título 1	Título 2

Figura 3.5: Análise das publicações do ano de 2012.

Listas de FARIAS, Lucas R.

2011

Título
Rede de Pesquisa da FURG

2012

Título
Um sistema para análise de redes de pesquisa...

Lista de Referências Únicas

Título	Ano
Um sistema para análise de redes de...	2012
ARGOsearch: an information retrieval...	2012
CiênciaBrasil - The Brazilian Portal of Science...	2011
An Automatic Approach for Duplicate...	2011
Rede de Pesquisa da FURG	2011

Lista de Pares

Título 1	Título 2
Um sistema para anál... Um sistema para anál...	Um sistema para anál...

Figura 3.6: Estado final das listas de referências únicas e de pares, após a análise de todas as publicações.

radas. É possível que uma publicação X seja réplica de Y e de Z , mas Y não ter sido identificada como réplica de Z . As figuras 3.7, 3.8, 3.9 e 3.10 apresentam um exemplo deste caso.

Supondo que as publicações de um terceiro pesquisador sejam analisadas, como na figura 3.7. A sua referência será comparada com as referências únicas de mesmo ano e será detectada como similar à uma referência única, como visto na figura 3.8. As duas

Listas de VARGAS, André P.

2012

Título
Um sistema para análise de redes de pesquisa...

Lista de Referências Únicas

Autor	Título	Ano
BORGES	Um sistema para análise de redes de...	2012
BORGES	ARGOsearch: an information retrieval...	2012
BORGES	CiênciaBrasil - The Brazilian Portal of Science...	2011
BORGES	An Automatic Approach for Duplicate...	2011
BORGES	Rede de Pesquisa da FURG	2011

Lista de Pares

Autor 1	Título 1	Autor 2	Título 2
BORGES	Um sistema...	FARIAS	Um sistema...

Figura 3.7: Análise das referências bibliográficas de um terceiro pesquisador.

referências analisadas serão armazenadas na lista de pares, como visto na figura 3.9.

Listas de VARGAS, André P.

2012

Título
Um sistema para análise de redes de pesquisa...

Lista de Referências Únicas

Autor	Título	Ano
BORGES	Um sistema para análise de redes de...	2012
BORGES	ARGOsearch: an information retrieval...	2012
BORGES	CiênciaBrasil - The Brazilian Portal of Science...	2011
BORGES	An Automatic Approach for Duplicate...	2011
BORGES	Rede de Pesquisa da FURG	2011

Lista de Pares

Autor 1	Título 1	Autor 2	Título 2
BORGES	Um sistema...	FARIAS	Um sistema...

Figura 3.8: Comparaçao da referênci do autor com as referências únicas do mesmo ano.

Porém, podemos ver na figura 3.10 que a mesma referência bibliográfica do pesquisador BORGES foi identificada como réplica de uma referência do pesquisador FARIAS e de uma referência do pesquisador VARGAS, mas as duas últimas não foram identificadas pelo algoritmo como réplicas. Para contornar esse problema, são criados os *clusters*.

Todas as referências similares identificadas como réplicas são agrupadas no mesmo *cluster*. As relações de coautoria são geradas extraíndo, para cada pesquisador, a lista de seus colaboradores. Estes colaboradores são os outros autores do conjunto de publicações

Listas de VARGAS, André P.

2012

Título
Um sistema para análise de redes de pesquisa...

Lista de Referências Únicas

Autor	Título	Ano
BORGES	Um sistema para análise de redes de...	2012
BORGES	ARGOsearch: an information retrieval...	2012
BORGES	CiênciaBrasil - The Brazilian Portal of Science...	2011
BORGES	An Automatic Approach for Duplicate...	2011
BORGES	Rede de Pesquisa da FURG	2011

Lista de Pares

Autor 1	Título 1	Autor 2	Título 2
BORGES	Um sistema...	FARIAS	Um sistema...
BORGES	Um sistema...	VARGAS	Um sistema...

Figura 3.9: Par de referências identificadas como réplicas é armazenado na Lista de Pares.

Listas de VARGAS, André P.

2012

Título
Um sistema para análise de redes de pesquisa...

Lista de Referências Únicas

Autor	Título	Ano
BORGES	Um sistema para análise de redes de...	2012
BORGES	ARGOsearch: an information retrieval...	2012
BORGES	CiênciaBrasil - The Brazilian Portal of Science...	2011
BORGES	An Automatic Approach for Duplicate...	2011
BORGES	Rede de Pesquisa da FURG	2011

Lista de Pares

Autor 1	Título 1	Autor 2	Título 2
BORGES	Um sistema...	FARIAS	Um sistema...
BORGES	Um sistema...	VARGAS	Um sistema...

Figura 3.10: Duas referências que são réplicas da mesma referência de BORGES, mas não foram identificadas como réplicas.

únicas, ou seja, do conjunto de referências bibliográficas em todos os seus *clusters*.

A figura 3.11 apresenta o *cluster* formado pelo algoritmo na situação do exemplo.

O algoritmo apresentado compara apenas referências bibliográficas de mesmo tipo. Podem existir publicações como as *Lectures Notes in Computer Science*, que são descritas em alguns locais como artigos em anais de conferência, e em outros como artigos em periódicos. Se dois pesquisadores descreverem a mesma publicação em tipos diferentes em seus currículos, essas publicações não serão detectadas como réplicas. Entretanto, o

Lista de Referências Únicas			
Autor	Título	Ano	
BORGES	Um sistema para análise de redes de...	2012	
BORGES	ARGOsearch: an information retrieval...	2012	
BORGES	Ciênciabrasil - The Brazilian Portal of Science...	2011	
BORGES	An Automatic Approach for Duplicate...	2011	
BORGES	Rede de Pesquisa da FURG	2011	

Lista de Pares			
Autor 1	Título 1	Autor 2	Título 2
BORGES	Um sistema...	FARIAS	Um sistema...
BORGES	Um sistema...	VARGAS	Um sistema...

Cluster Formado	
Autor	Título
BORGES	Um sistema para análise de redes de pesquisa ...
FARIAS	Um sistema para análise de redes de pesquisa ...
VARGAS	Um sistema para análise de redes de pesquisa ...

Figura 3.11: Estado final das listas de referências únicas e de pares, e o *cluster* formado pelo algoritmo.

sistema pode ser facilmente modificado para considerar estes casos.

3.2 Banco de Dados

O banco de dados desenvolvido no servidor é dividido em dois esquemas: *Coleta* e *Aplicação*. A figura 3.12 mostra o diagrama relacional do esquema *Coleta*. O nome dos pesquisadores que pertencem ao conjunto analisado bem como os identificadores dos currículos Lattes são armazenados na tabela *membros*.

A tabela *publicacao* armazena os principais metadados das citações bibliográficas: título, autores, ano de publicação e número de páginas. Note que estes campos estão presentes em qualquer tipo de publicação analisada. Cada publicação pertence a apenas um membro porque foi extraída de seu currículo. Para cada tipo de publicação foi criada uma tabela auxiliar que armazena os metadados específicos que não se aplicam aos demais tipos. Por exemplo, capítulos de livros publicados (tabela *capitulo*) possuem número de edição, enquanto artigos aceitos para publicação (tabela *revista_aceito*) não possuem. Os relacionamentos entre a tabela *publicacao* e estas tabelas auxiliares podem ser vistos como mapeamentos de herança de tipos.

As instâncias da tabela *publicacao* representam todas as referências bibliográficas ex-

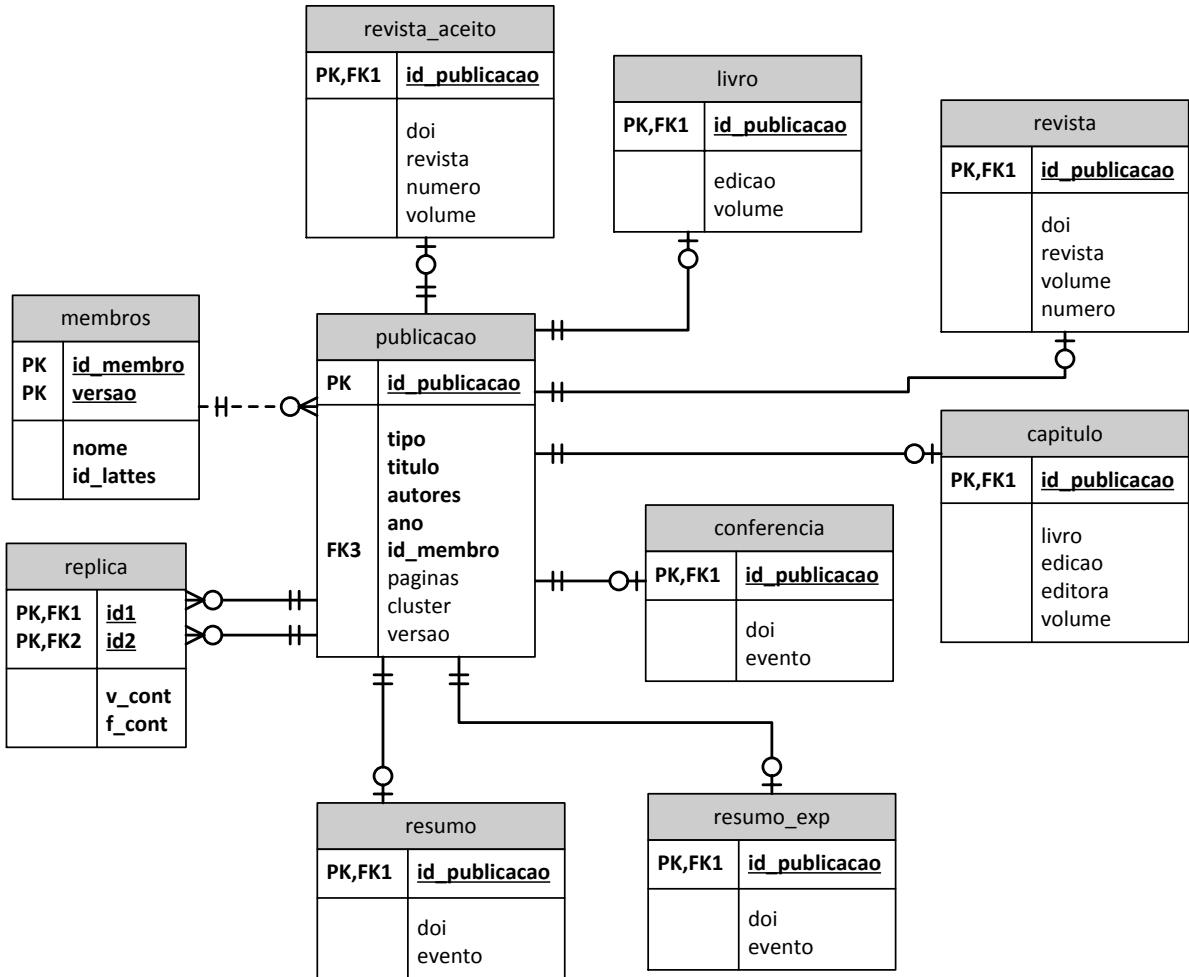
traídas de todos os currículos, sem o processo de identificação de duplicatas. Pares de referências identificadas como réplicas da mesma publicação são armazenados na tabela *replica*. Além do par de identificadores das publicações duplicadas, esta tabela contém os campos *v_cont* e *f_cont*, responsáveis por controlar o *feedback* da deduplicação. Cada vez que um usuário da aplicação avalia a detecção como correta, *v_cont* é incrementado. Caso contrário, *f_cont* é incrementado. A diferença entre os valores destes dois campos pode ser usada como uma boa estimativa da qualidade, independente do processo de deduplicação utilizado. Esses dados não são analisados atualmente, mas deverão ser utilizados futuramente para avaliar a qualidade da identificação de réplicas.

As tabelas *publicacao* e *membros* deste esquema possuem um campo chamado *versao*, que identifica à qual versão da coleta pertence cada tupla. Isto permite que novas coletas sejam feitas sem interferir nas anteriores. Como as outras tabelas do esquema sempre referenciam um registro da tabela *publicacao*, elas não possuem o campo *versao*, pois esta informação já está armazenada no registro correspondente da tabela *publicacao*.

O diagrama relacional do esquema *Aplicação* é descrito na figura 3.13. A tabela *usuarios* armazena os usuários registrados no servidor, utilizando o *email* como nome de usuário. A senha do usuário não é guardada, sendo armazenado no campo *md5* apenas um *hash* da concatenação do *email* com a senha, impossibilitando o acesso de terceiros às senhas dos usuários.

A tabela *rede* armazena as informações das redes de colaboração criadas pelos usuários. O campo *autor* referencia a tabela usuários, determinando o usuário que criou a rede. Além disso, são guardados os outros dados: *nome*; *visibilidade*, ou seja, se essa rede pode ser visualizada por outros usuários do sistema ou não; *xml*, que contém o endereço do arquivo XML da rede no servidor; *anoinicio* e *anofim*, que determinam o período cujas publicações devem ser consideradas. A tabela *membrosrede* contém a informação de quais pesquisadores pertencem a quais redes, utilizando para isso o campo *idlattes*, que corresponde ao identificador do currículo Lattes de cada pesquisador, e o campo *idrede*, que referencia a tabela *rede*.

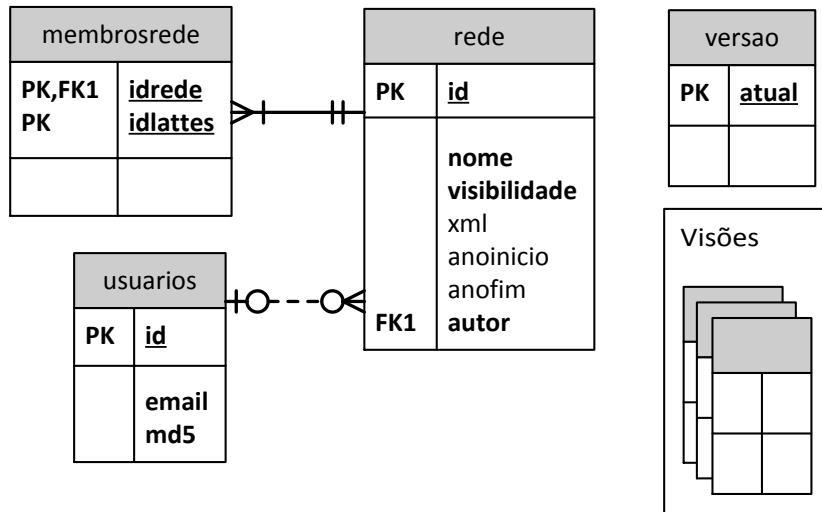
Neste esquema, além das tabelas mencionadas, existem dez visões, uma para cada tabela do esquema *Coleta*: *capitulo*, *conferencia*, *livro*, *membros*, *publicacao*, *replica*, *resumo*, *resumo_exp*, *revista*, e *revista_aceito*. Cada uma dessas visões retorna apenas os

Figura 3.12: Diagrama relacional do esquema *Coleta*

dados da versão atual da coleta, definida na tabela *versao*. Desta forma, o módulo de *Processamento de Rede* acessa apenas o esquema *Aplicação*, garantindo que os dados acessados são os mais atuais. Isto também permite ao administrador do sistema, em caso de algum problema na coleta, voltar a uma versão anterior com um esforço mínimo, apenas alterando o valor na tabela *versao*.

3.3 Cálculo das Métricas

O módulo de *Cálculo das Métricas* foi desenvolvido à parte, como Trabalho de Conclusão de Curso pelo aluno do curso de Sistemas de Informação da FURG Giancarlo Lucca

Figura 3.13: Diagrama relacional do esquema *Aplicação*

(Lucca, 2013). É um componente independente especificado e implementado fora do escopo deste trabalho.

Esse módulo fornece as seguintes métricas de grafos e redes sociais (Easley and Kleinberg, 2010):

- diâmetro do grafo;
- componente gigante;
- densidade;
- *PageRank*;
- grau;
- coeficiente de clusterização;
- centralidade *closeness*.

3.4 Processamento da Rede

O módulo de *Processamento da Rede* é o responsável por criar o arquivo XML das redes de colaboração criadas pelo usuário. Para isso, o componente usa as informações contidas na tabela *rede* do banco de dados.

A consulta apresentada na figura 3.14 cria uma tabela temporária no SGBD responsável por armazenar o conjunto de relacionamentos de coautoria entre os pesquisadores, ou seja, o conjunto de adjacências entre os vértices do grafo que representa a rede de colaboração. São selecionados pares de pesquisadores e o número de colaborações entre eles (linhas 2 e 3) através da junção entre as duas instâncias das tabelas *publicacao* e *membros* (linhas 4-7). Este número é calculado contando a quantidade de *clusters* compartilhados pelos dois pesquisadores (linha 8). O intervalo temporal de interesse do usuário, passado como parâmetro na criação da rede, filtra as colaborações resultantes (linhas 9 e 10). Por fim, são filtrados apenas os pesquisadores que formam a rede e os resultados são agrupados e ordenados em relação ao par de pesquisadores (linhas 11-14).

```

1 create temp table adj as
2 select p1.id_membro as membro1,
3 p2.id_membro as membro2, count(p1.cluster) as colaboracao
4 from publicacao as p1, publicacao as p2, membros as m1, membros as m2
5 where p1.id_membro < p2.id_membro
6 and m1.id_membro = p1.id_membro
7 and m2.id_membro = p2.id_membro
8 and p1.cluster = p2.cluster
9 and p1.ano >= 2008
10 and p1.ano <= 2012
11 and m1.id_lattes in (select idlattes from membrosrede where idrede = 1)
12 and m2.id_lattes in (select idlattes from membrosrede where idrede = 1)
13 group by p1.id_membro, p2.id_membro
14 order by p1.id_membro, p2.id_membro

```

Figura 3.14: Consulta SQL que cria uma tabela temporária no banco de dados contendo o conjunto de adjacências da rede a ser processada. No exemplo, é processada a rede de *id* 1, ano inicial 2008 e ano final 2012 .

A partir do conjunto de adjacências criado, o módulo gera a matriz de adjacência do grafo. Essa matriz é armazenada em um arquivo texto, que será o parâmetro passado ao módulo de cálculo de métricas. Após, o módulo gera o arquivo GraphML. Nesse arquivo,

são armazenados:

- valores das métricas relativas ao grafo;
- os membros da rede, com os seguintes dados:
 - nome;
 - endereço do currículo Lattes;
 - número de publicações no período;
 - valores das métricas relativas aos membros.
- as arestas, ou seja, os relacionamentos entre dois membros, e o número de colaborações entre esses membros.

Capítulo 4

Implementação

O módulo *Nova Coleta* é dividido em três etapas: coleta de dados, extração de informação, e análise. As duas primeiras etapas foram implementadas usando o componente de seleção e pré-processamento de dados do *software* scriptLattes (MenaChalco and Cesar Junior, 2009), uma ferramenta de código aberto desenvolvida para extrair e compilar automaticamente as produções acadêmicas de um conjunto de pesquisadores registrados na plataforma Lattes. Esta plataforma contém, entre outras informações, os currículos de maior parte dos pesquisadores que trabalham no Brasil. O Lattes é mantido pelo CNPq. Esta fonte de dados foi escolhida porque a maior parte das agências de fomento à pesquisa do Brasil utilizam os currículos Lattes como um meio de avaliação dos pesquisadores que solicitam recursos para projetos.

O código fonte do scriptLattes foi adaptado para comunicar com um banco de dados PostgreSQL e armazenar informações sobre pesquisadores e suas publicações, as relações de coautoria e as informações sobre referências bibliográficas replicadas, o que não era possível no *software* original. Foram selecionados para análise dos dados apenas informações sobre artigos em revistas científicas, livros, capítulos de livros, artigos completos e resumos publicados em anais de conferências.

O módulo de *Visualização* foi implementado em linguagem Java, usando a biblioteca *Prefuse* (Heer et al., 2005), que permite a criação de grafos interativos.

O diagrama de classes do cliente é apresentado na figura 4.1. Neste diagrama, apenas as classes mais importantes são mostradas. O diagrama com todos os atributos e métodos está no apêndice A.

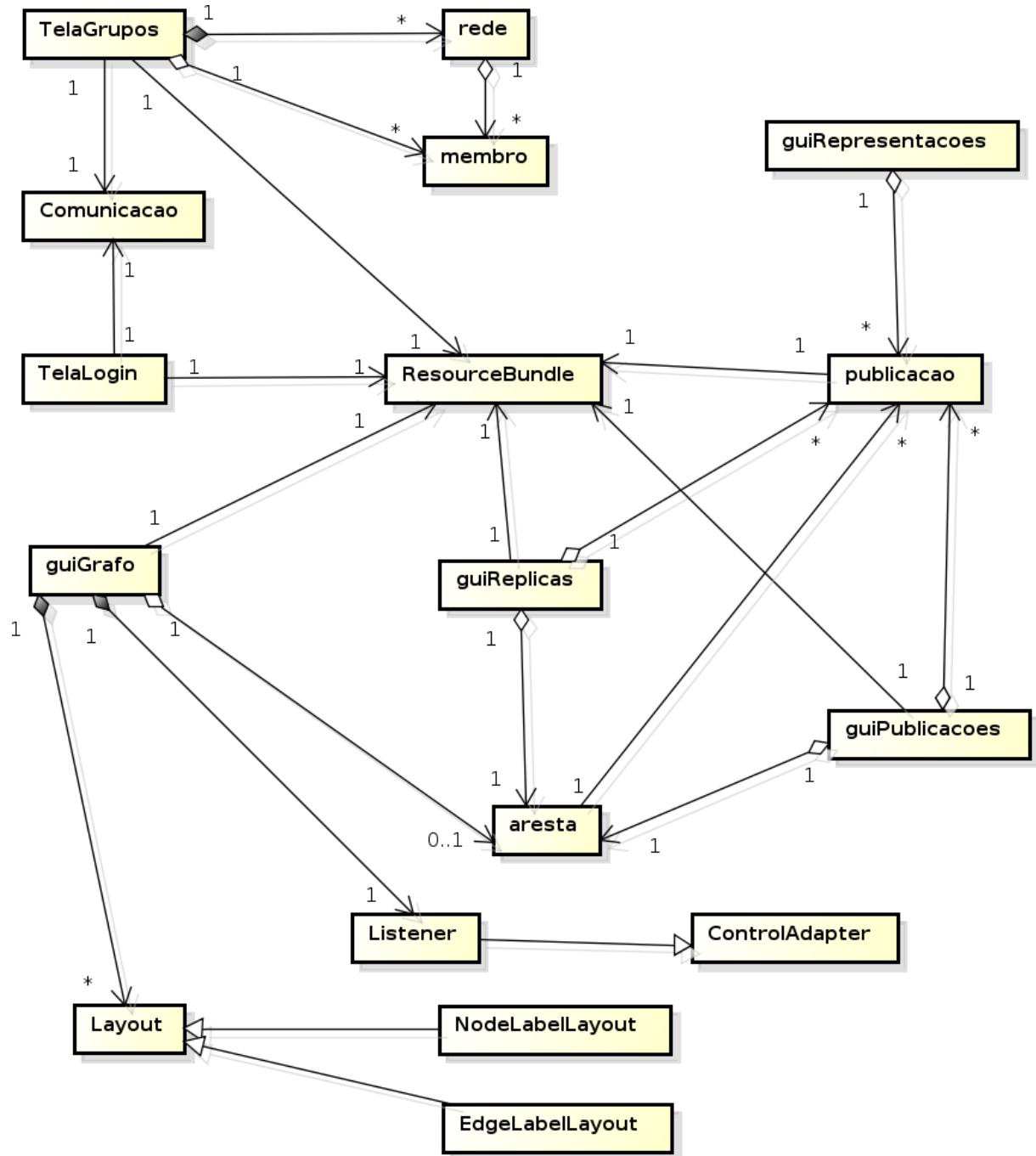


Figura 4.1: Diagrama de classes simplificado do cliente.

A classe *ResourceBundle* é a classe responsável pela internacionalização da ferramenta. Essa classe disponibiliza as *strings* necessárias para a ferramenta de acordo com a linguagem do sistema operacional do usuário, com opção de português brasileiro ou inglês.

A classe *TelaLogin* implementa a tela de login do usuário. Após o usuário iniciar a sessão, essa classe instancia um objeto da classe *TelaGrupos*. Esta classe é a implementação

do *CRUD Redes* explicado no capítulo 3. Para armazenar as informações das redes do usuário, a *TelaGrupos* utiliza uma lista de objetos da classe *rede*, que tem como atributos as informações das redes contidas no banco de dados do servidor, além de uma lista de objetos da classe *membro*, que é a lista de pesquisadores pertencentes à rede. A classe *TelaLogin* também utiliza a classe *membro* para armazenar os dados dos pesquisadores retornados quando o usuário pesquisa por pesquisadores para serem incluídos na rede.

Tanto a classe *TelaLogin* quanto a *TelaGrupos* utilizam a classe *Comunicacao* para se comunicarem com o servidor. A classe *Comunicacao* cria um *socket* com o servidor, e implementa as operações de envio (*send*) e recebimento (*receive*).

O módulo de *Visualização* é implementado pela classe *guiGrafo*. Ela instancia objetos das classes *EdgeLabelLayout* e *NodeLabelLayout*, necessárias para a visualização integrada das métricas no grafo. A classe *Listener* gerencia as ações do usuário na tela do grafo. Ela determina se o clique foi em um vértice, em uma aresta, ou em uma área em branco, e a partir disso, determina qual ação deve ser tomada pela classe *guiGrafo*.

A classe *aresta* é responsável por preparar as listas de publicações de um ou dois pesquisadores selecionados pelo usuário. A lista de publicações guarda objetos da classe *publicacao*. Quando apenas um pesquisador é selecionado, um objeto da classe *guiPublicacoes* é instanciado, gerando uma tela com a lista de publicações do pesquisador selecionado. No caso de dois pesquisadores, a classe *aresta* gera a lista de publicações em coautoria dos dois pesquisadores. Essa lista de publicações em coautoria é utilizada pela classe *guiReplicas*, que é instanciada pela classe *guiGrafo* quando o usuário clica no botão corresponde. A classe *guiReplicas* mostra essa lista de publicações em coautoria ao usuário. Quando uma referência bibliográfica dessa lista é selecionada pelo usuário, um objeto da classe *guiRepresentações* é instanciado, gerando a visualização das diferentes representações da referência.

O diagrama de classes do servidor é mostrado na figura 4.2. As classes *Grafo* e *Componente* implementam o módulo de *Cálculo de Métricas*, descrito na seção 3.3. Como as classes foram implementadas fora do escopo desse trabalho, no diagrama são apresentados apenas as suas associações, sem os seus métodos e atributos. O diagrama com todos os atributos e métodos está no apêndice B.

A classe *Server* é responsável pela criação do *socket* que será utilizado para a comu-

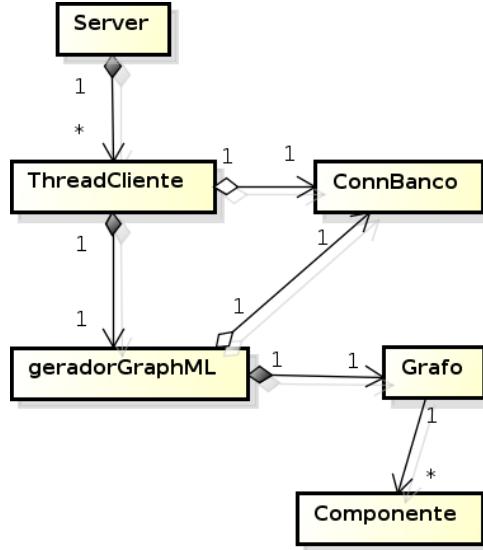


Figura 4.2: Diagrama de classes do servidor.

nicação com o cliente. O *Server* fica ouvindo a porta do sistema pela qual o *socket* faz a ligação e, quando o cliente faz a conexão com o servidor, um objeto da classe *ThreadCliente* é instanciado.

Cada objeto da classe *ThreadCliente* é responsável por toda a comunicação entre o servidor e o objeto da classe *Comunicacao* do cliente. Para a troca de mensagens entre eles, foi criado um protocolo simples baseado em *strings*, descrito na tabela 4.1.

Para atender as requisições do cliente, a *ThreadCliente* acessa o banco de dados através de um objeto da classe *ConnBanco*. Essa classe abstrai a conexão com o banco de dados, e fornece dois métodos para manipulação de dados, um para consultas com retorno e outro para modificações de dados.

A classe *geradorGraphML* é responsável pela criação do GraphML das redes de colaboração, utilizando o processo descrito na seção 3.4.

Tabela 4.1: *Strings* de mensagens definidas pelo protocolo criado.

Remetente	Mensagem	Ação
Cliente	<i>connect</i>	Pedido de conexão ao servidor
	<i>login</i>	Envio das informações de <i>login</i>
	<i>requestnetworks</i>	Solicita as redes do usuário
	<i>createnetwork</i>	Envio das informações de uma rede a ser criada
	<i>createuser</i>	Envio das informações de um usuário a ser criado
	<i>searchmembers</i>	Envio do termo de pesquisa de pesquisadores
	<i>searchnetworks</i>	Envio do termo de pesquisa de redes públicas
	<i>requestpackage</i>	Requisição do arquivo de uma determinada rede
	<i>deletenetwork</i>	Requisição de exclusão de uma determinada rede
	<i>modifynetwork</i>	Envio das informações de uma rede a ser modificada
Servidor	<i>loginaccepted</i>	<i>Login</i> aceito
	<i>loginfailed</i>	<i>Login</i> falho
	<i>package</i>	Envio de arquivo de rede
	<i>usernetworks</i>	Envio das informações das redes do usuário
	<i>members</i>	Resposta da pesquisa de pesquisadores
	<i>publicnetworks</i>	Resposta da pesquisa de redes públicas
	<i>usercreated</i>	Usuário criado
	<i>usedemail</i>	Email já existe no servidor
	<i>networkcreated</i>	Rede criada
	<i>networkdeleted</i>	Rede excluída
	<i>networkmodified</i>	Rede modificada

Capítulo 5

Protótipo

Neste capítulo é mostrado o protótipo da ferramenta cliente do Research.Net. As figuras 5.1 a 5.16 apresentam um exemplo de um usuário utilizando a ferramenta para criar, editar, excluir e visualizar redes de colaboração.

A figura 5.1 mostra a tela de *login* da ferramenta. Nesta tela, o usuário, se já registrado, pode acessar a ferramenta, ou pode criar um usuário. Após esta tela, o usuário tem acesso a tela de redes da figura 5.2, referente ao módulo *CRUD Redes* da arquitetura (capítulo 3). Na parte superior esquerda da tela há uma listagem das redes do usuário, e abaixo um campo pelo qual o usuário pode pesquisar redes públicas criadas por outros usuários.

Selecionando uma de suas redes e clicando em *Edita*, o usuário pode modificar as propriedades da rede e os seus membros, como visto nas figuras 5.3 e 5.4, respectivamente. Na tela da figura 5.3, o usuário pode editar o nome da rede, os anos de início e fim das publicações, e a visibilidade desta rede para os outros usuários. Na tela da figura 5.4, o usuário pode incluir e retirar pesquisadores na rede. Para incluir membros, o usuário procura os pesquisadores disponíveis no banco de dados utilizando o campo *Pesquisar*. Uma lista de pesquisadores cujos nomes são similares ao termo pesquisado será mostrada ao usuário, que pode selecionar um ou mais pesquisadores e incluir na rede.

Após a edição da rede, o usuário deverá salvá-la, utilizando o botão *Salvar* visto na figura 5.3. As informações atualizadas da rede serão enviadas para o servidor, que as armazenará no banco de dados, e irá gerar um novo arquivo XML, que será enviado ao cliente. O processo de criação de uma nova rede é similar ao processo de edição descrito anteriormente.

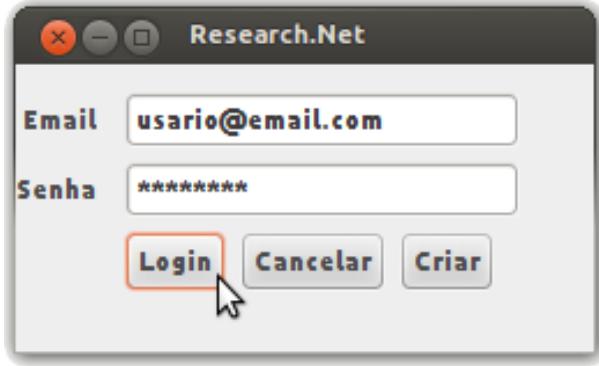


Figura 5.1: Tela de *login*. Usuário pode acessar a ferramenta ou criar um novo usuário.

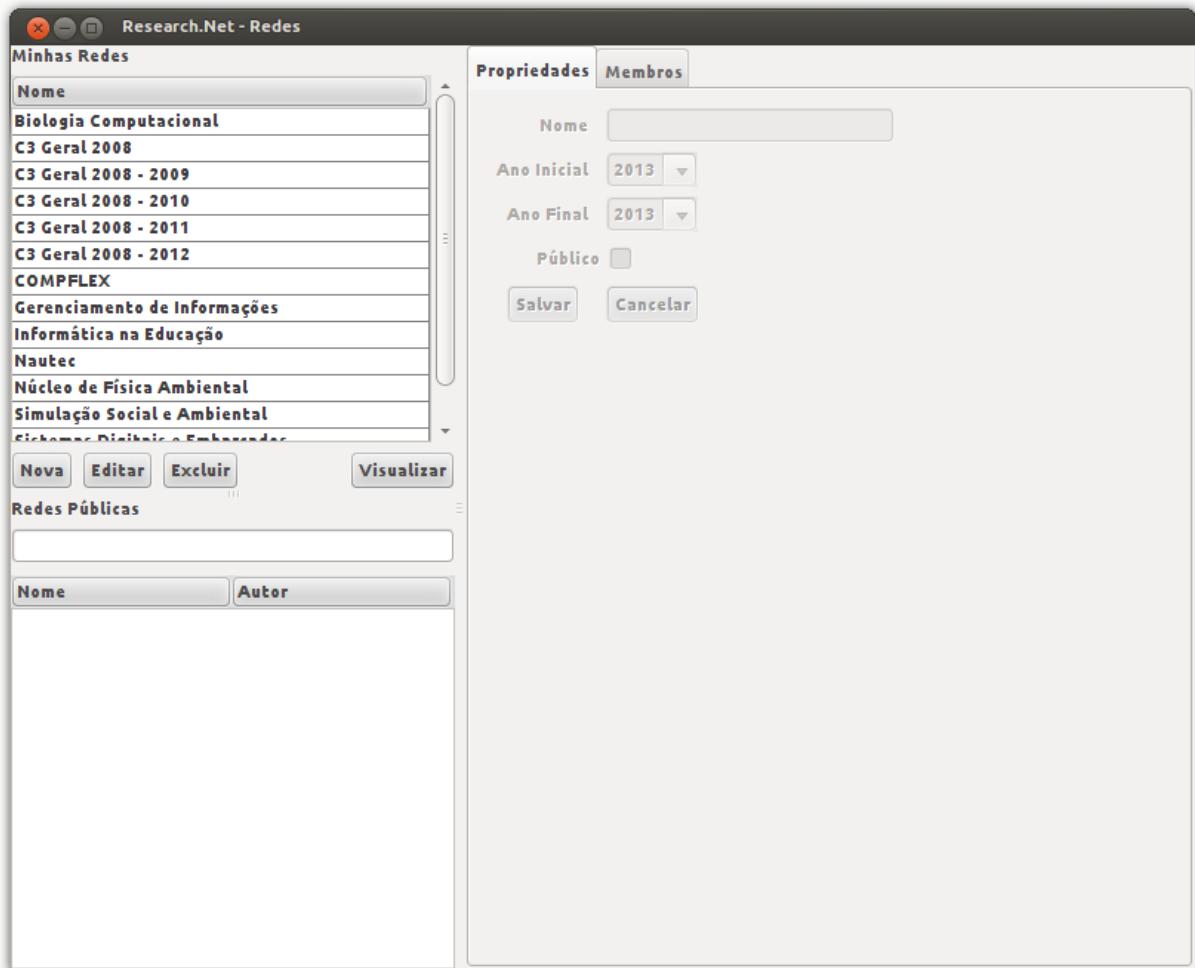


Figura 5.2: Tela de redes.

Clicando em uma rede na lista e no botão *Visualizar*, uma nova tela é exibida com a visualização gráfica da rede de colaboração, como visto na figura 5.5. Nesta tela, o usuário vê os pesquisadores representados por vértices de um grafo, e as colaborações

entre eles representadas pelas arestas. Ao clicar em um membro da rede ele é destacado em vermelho, enquanto os outros membros da rede que possuem relação de coautoria com o membro clicado são destacados em amarelo. É possível visualizar em um navegador *web* o currículo Lattes do pesquisador selecionado, clicando no botão *Curriculum*. O botão *Publicações* abre uma nova tela com a lista das publicações do pesquisador no período, como mostra a figura 5.6.

Ao clicar em outro pesquisador ligado ao primeiro, os dois membros e a ligação entre eles ficam destacados em vermelho, e o botão *Colaborações* fica disponível, como visto na figura 5.7. Este botão ativa uma nova tela que contém a lista das publicações em comum entre os dois pesquisadores identificadas pelo sistema, como mostra a figura 5.8. É possível visualizar as duas representações de uma referência bibliográfica clicando na referência, abrindo assim uma nova janela, mostrada na figura 5.9. A representação da referência bibliográfica no currículo de cada pesquisador é mostrada, sendo possível ao usuário identificar se elas realmente referenciam a mesma publicação e informar ao sistema, utilizando os botões *Sim* ou *Não*.

Voltando à tela de visualização da rede de colaboração, o usuário pode acessar várias informações sobre os pesquisadores e suas colaborações diretamente nas suas representações no grafo, utilizando os botões laterais. Clicando no botão $\#$ *Publicações*, é mostrado em cada um dos membros o número de publicações do pesquisador no período, como visto na figura 5.10. O botão $\#$ *Colaborações* mostra em cada uma das arestas o número de publicações em coautoria entre os membros ligados pela aresta, como mostra a figura 5.11.

Por fim, os botões *Grau*, *PageRank*, *Cluster Coefficient*, *Closeness* e *Diâmetro* mostram os resultados das métricas aplicadas à rede de colaboração, mostradas nas figuras de 5.12 a 5.16.

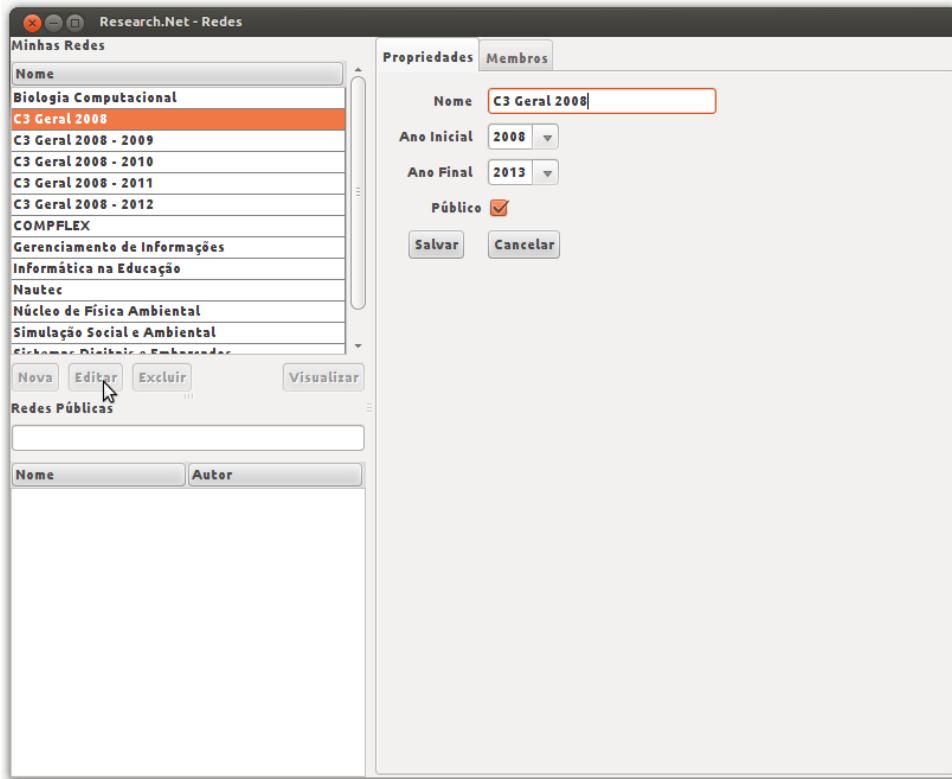


Figura 5.3: Edição das propriedades da rede.

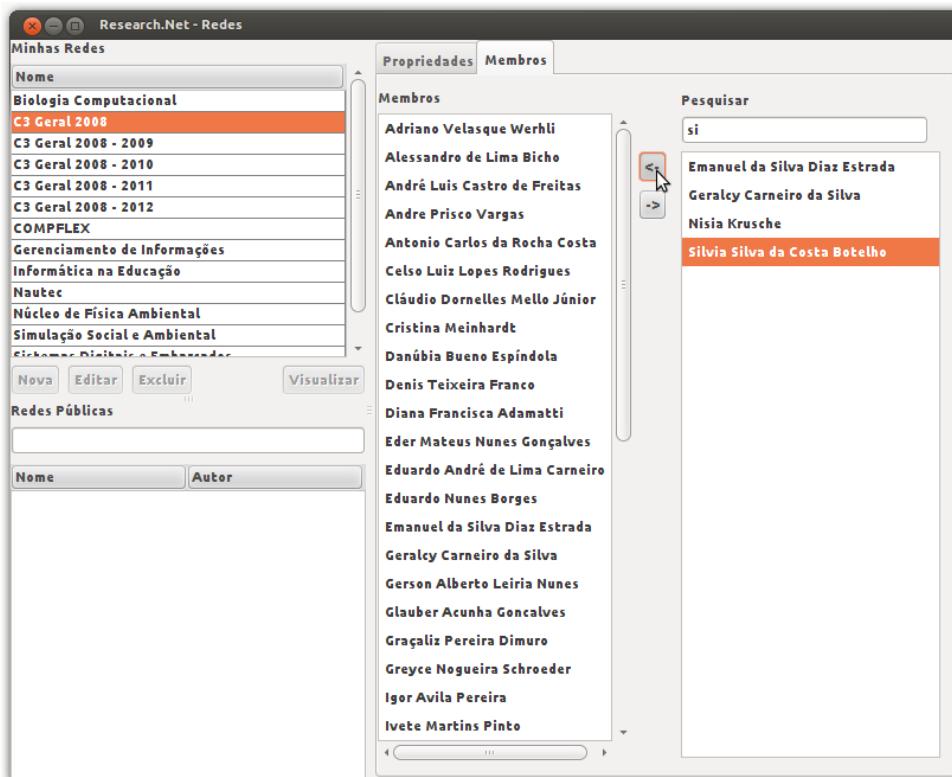


Figura 5.4: Inclusão de um novo membro na rede.

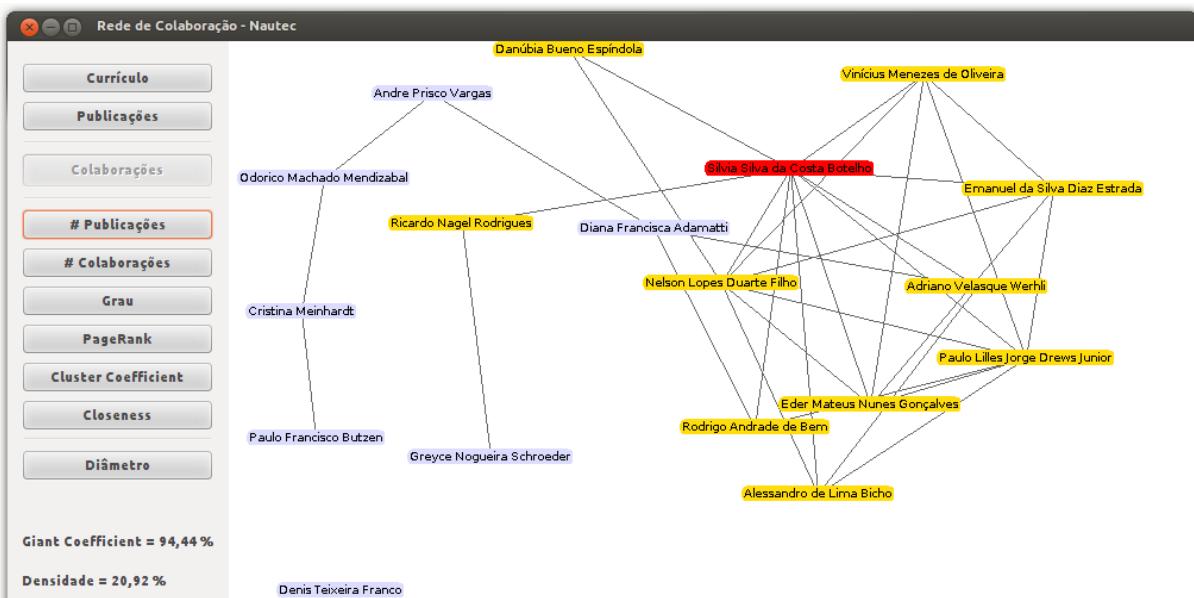


Figura 5.5: Membro clicado destacado em vermelho, com os seus coautores destacados em amarelo.



Figura 5.6: Listagem das publicações do pesquisador selecionado.

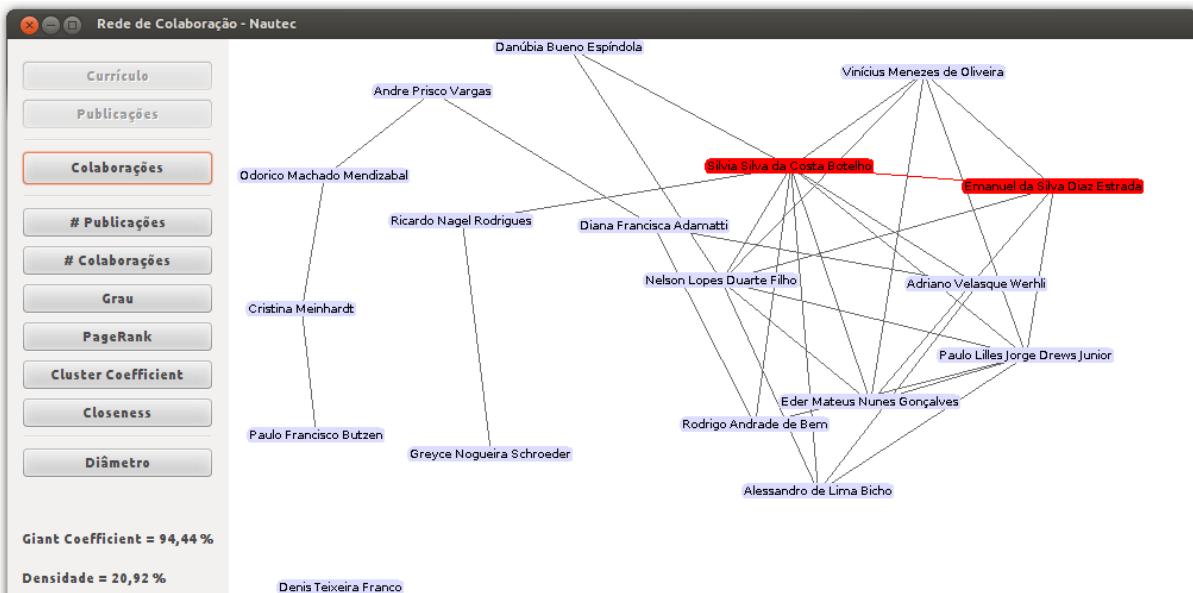


Figura 5.7: Dois membros com publicações em coautoria selecionados.



Figura 5.8: Listagem das publicações em comum dos pesquisadores selecionados.

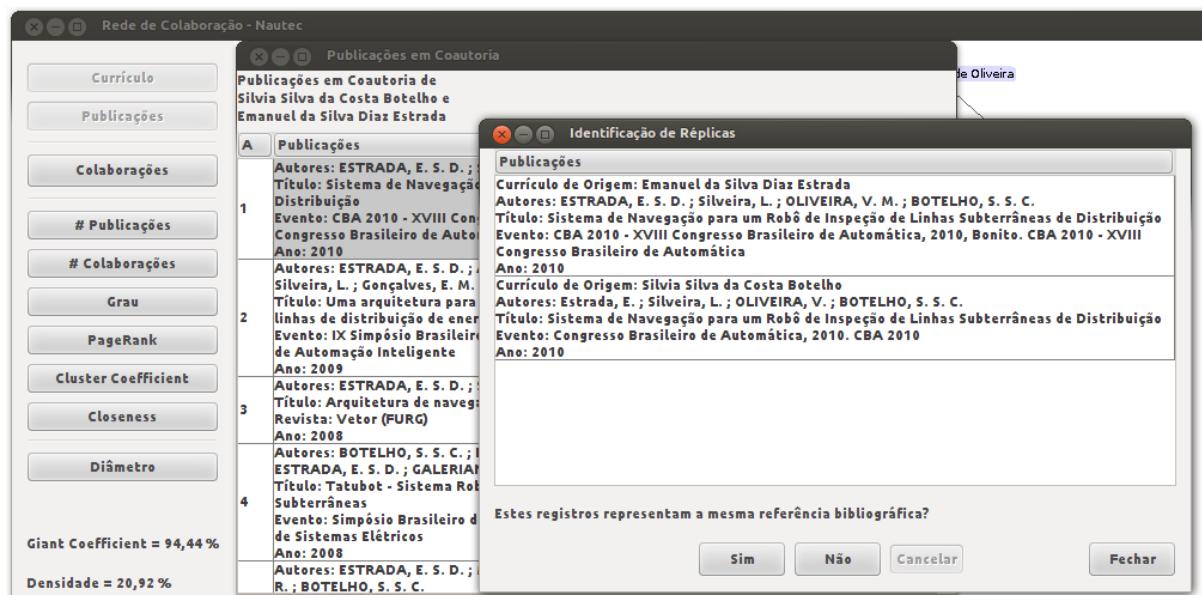


Figura 5.9: Visualização das diferentes representações da referência bibliográfica selecionada na lista de publicações em comum, destacada em cinza.

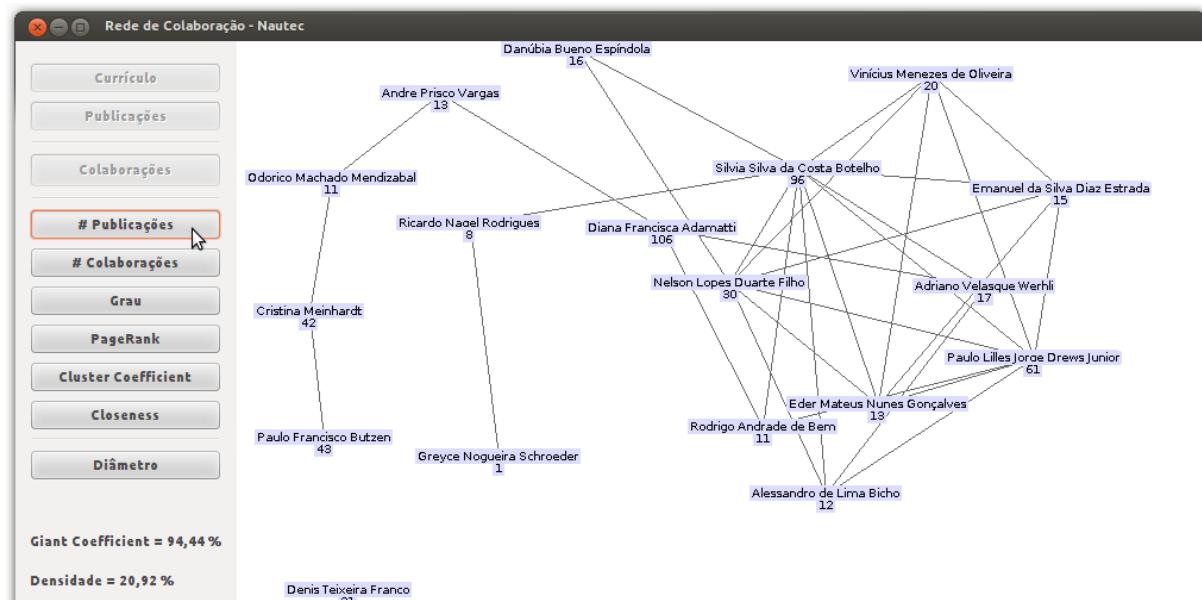


Figura 5.10: Visualização do número de publicações do pesquisador no período.

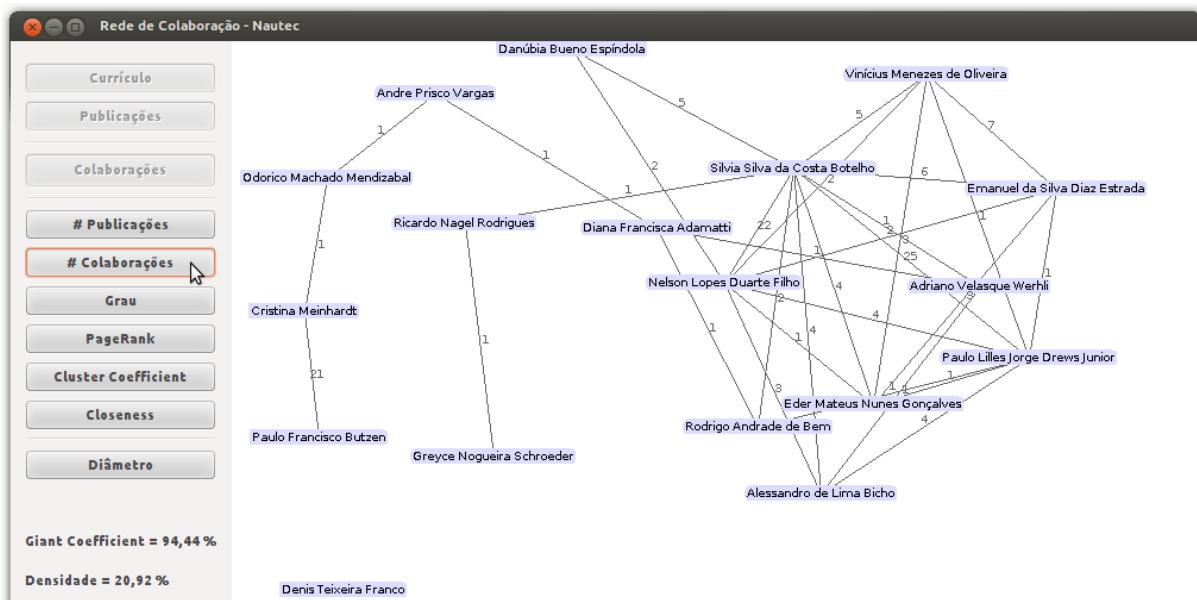


Figura 5.11: Visualização das colaborações entre pesquisadores, representada pelos pesos das arestas.

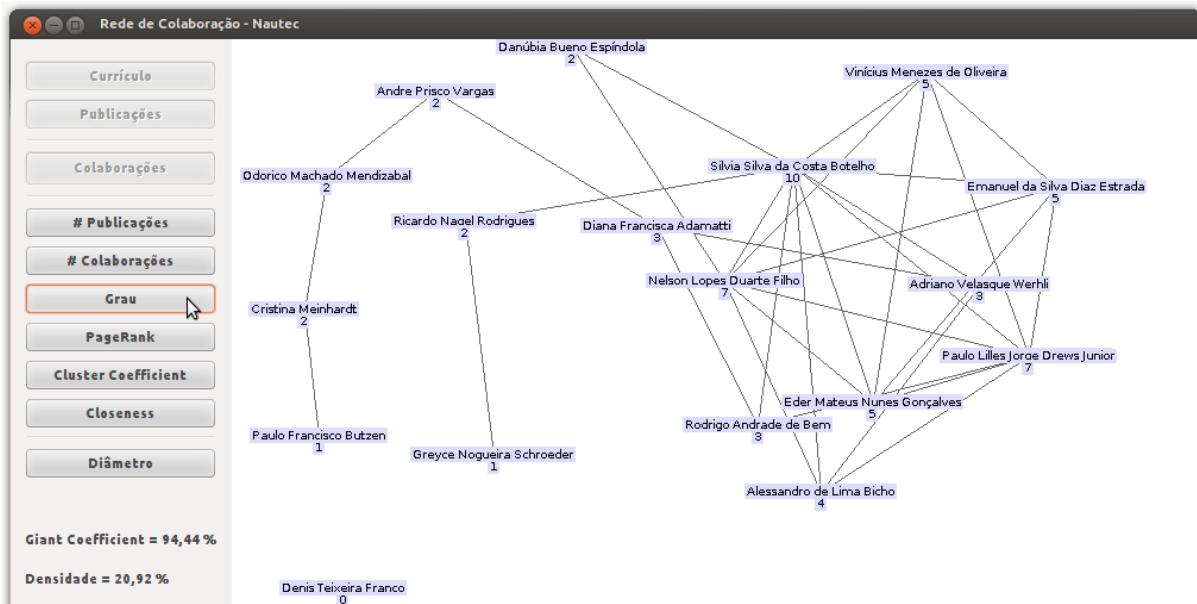
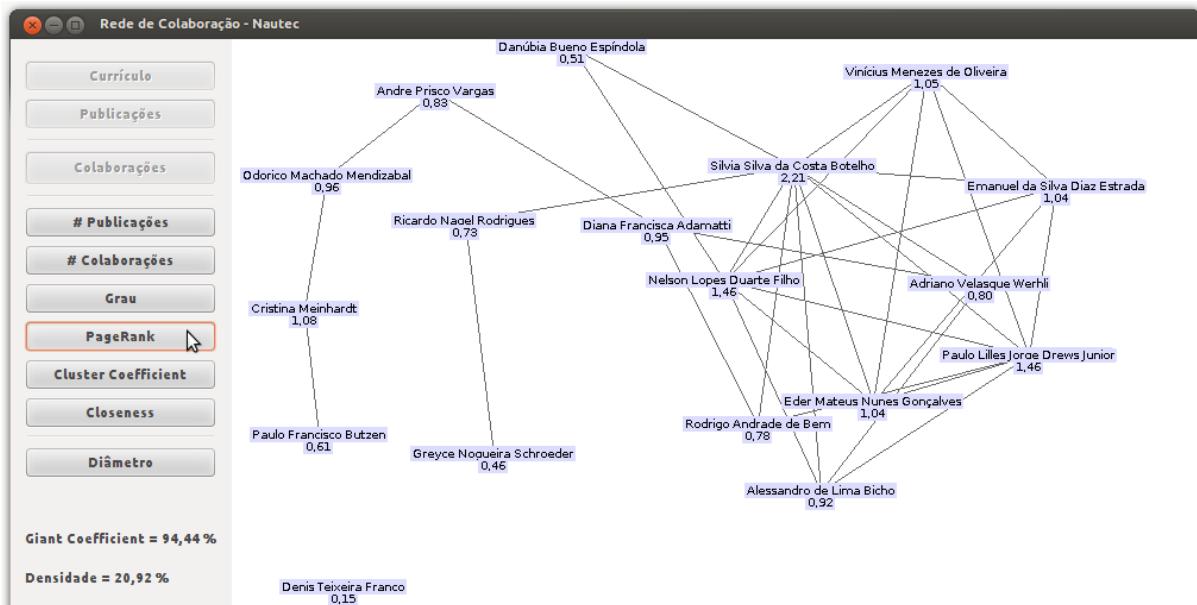
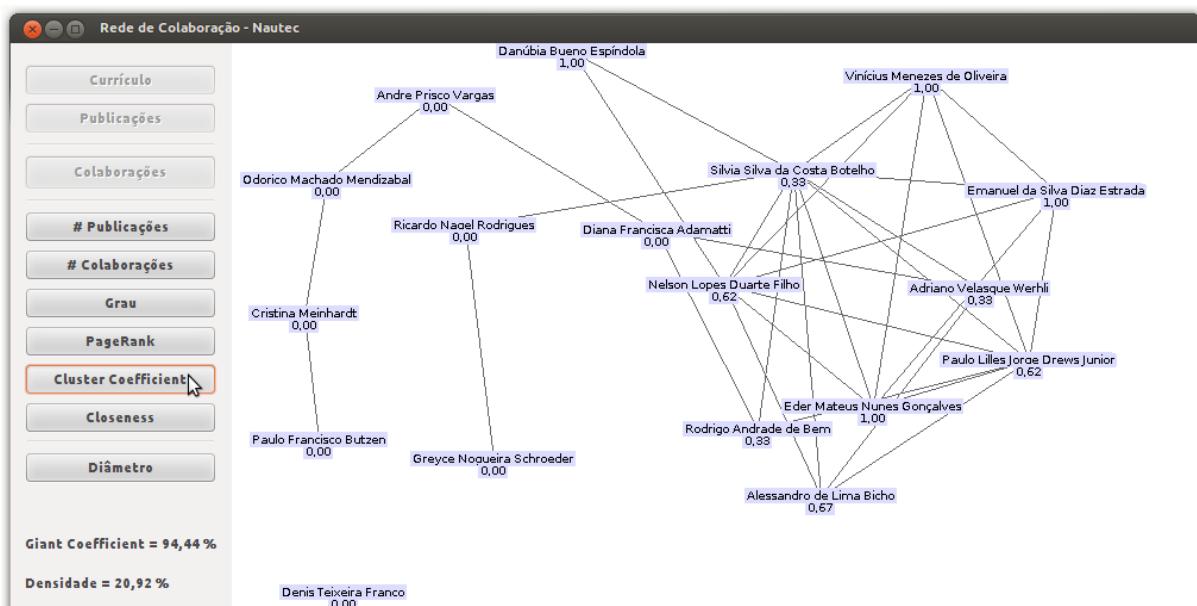
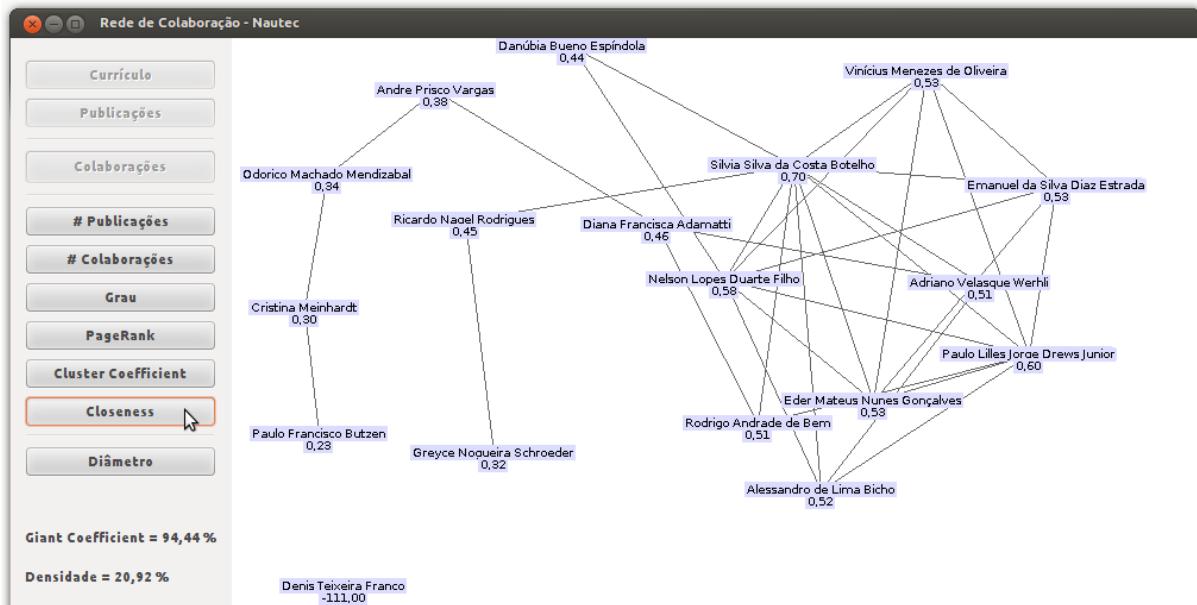
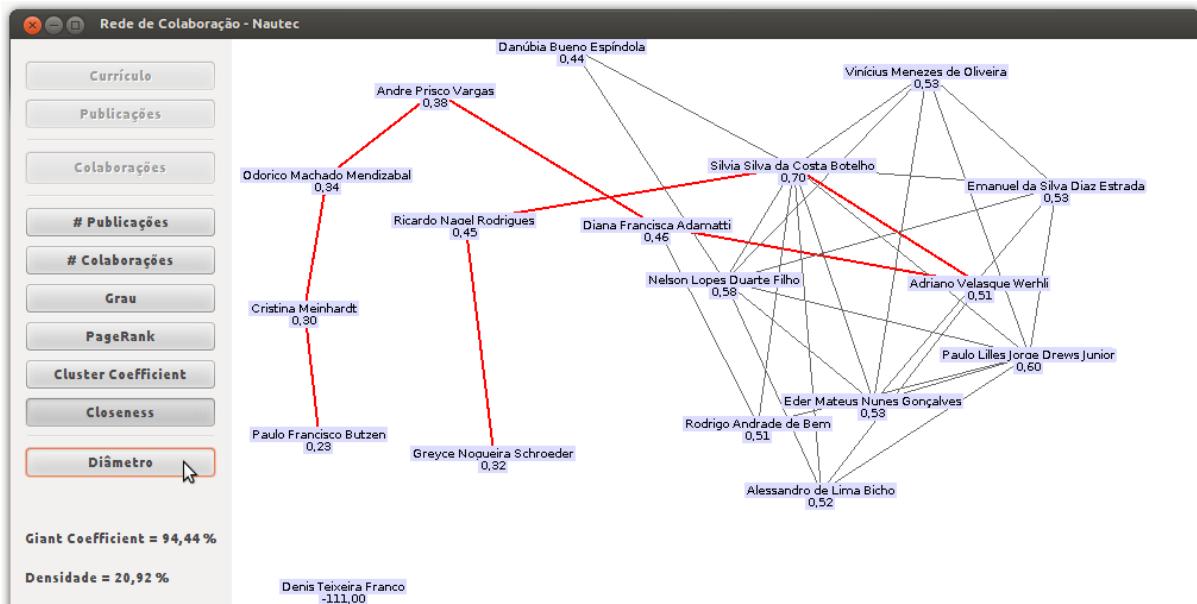


Figura 5.12: Valores da métrica *Grau* de cada pesquisador.

Figura 5.13: Valores da métrica *PageRank* de cada pesquisador.Figura 5.14: Valores da métrica *ClusterCoefficient* de cada pesquisador.

Figura 5.15: Valores da métrica *Closeness* de cada pesquisador.Figura 5.16: Destacadas em vermelho, as arestas que formam um dos *Diâmetros* da rede.

Capítulo 6

Considerações Finais

Este trabalho apresenta o desenvolvimento de um sistema cliente-servidor para análise de redes de colaboração utilizando a plataforma Lattes. A representação gráfica de uma rede de colaboração de pesquisadores permite que uma grande quantidade de informação seja analisada rapidamente.

Ao invés do usuário ter que identificar as interações e os relacionamentos entre os membros do conjunto analisado de forma explícita, como nas redes sociais tradicionais, o sistema extrai estes relacionamentos de forma automática a partir dos relacionamentos de coautoria entre os pesquisadores.

Facilitar o entendimento das interações entre os pesquisadores como uma rede social é uma contribuição significativa. É possível identificar grupos de pesquisa emergentes bem como analisar a produção de grupos bem estabelecidos. O sistema também pode auxiliar o processo de constituição de novos grupos de pesquisa e o gerenciamento de grupos existentes. O uso de métricas de grafos e redes sociais também auxilia as análises da produção científica dos pesquisadores.

A arquitetura modular do sistema facilita modificações e a criação de novas funcionalidades. Por exemplo, a fonte de dados (plataforma Lattes) pode ser substituída, sendo necessários apenas ajustes no módulo *Nova Coleta*, assim como a função de deduplicação, baseada na distância de edição, pode ser substituída por outra função de similaridade.

A criação de arquivos XML para as redes de colaboração pelo servidor permite que o cliente seja modificado, ou que novos clientes sejam especificados, sem que o desenvolvedor precise ter conhecimento do processo de extração e análise de dados.

A tabela 6.1 compara as características dos trabalhos relacionados com as características deste trabalho. Como pode ser visto, o desenvolvimento deste sistema foi focado em funcionalidades inexistentes na maior parte dos trabalhos relacionados.

Tabela 6.1: Comparativo das características dos trabalhos relacionados e do Research.Net

Característica	ArnetMiner	CiênciaBrasil	scriptLattes	Research.Net
Busca por Especialistas	✓			
Rankings acadêmicos	✓		✓	
Geração de relatórios da produção		✓	✓	
Visualizações dinâmicas	✓			✓
Listagem de publicações em comum				✓
Visualização de múltiplas representações das referências bibliográficas				✓
Temporalidade das redes de colaboração		✓		
Avaliação da identificação de referências bibliográficas duplicadas				✓

O principal resultado obtido foi uma análise da pesquisa do Centro de Ciências Computacionais (C3), apresentada no Seminário Anual de Planejamento do C3, produzida utilizando o Research.Net. Esta análise teve como objetivos:

- possibilitar um experimento do trabalho desenvolvido;
- apresentar uma visão geral da pesquisa do C3 desde sua criação, em 2008;
- analisar a integração dos pesquisadores do C3;
- visualizar as colaborações dos integrantes dos grupos de pesquisa cadastrados no CNPq;

A figura 6.1 apresenta a rede de colaboração gerada pelo Research.Net formada por todos os pesquisadores do C3, considerando referências bibliográficas entre os anos de 2008 e 2012. São apresentados também o valor da métrica *PageRank* calculada para cada um dos pesquisadores.

Parte dos resultados apresentados nesta monografia foram apresentados em:

- um artigo completo publicado na Escola Regional de Banco de Dados em 2012 (de Farias et al., 2012b);

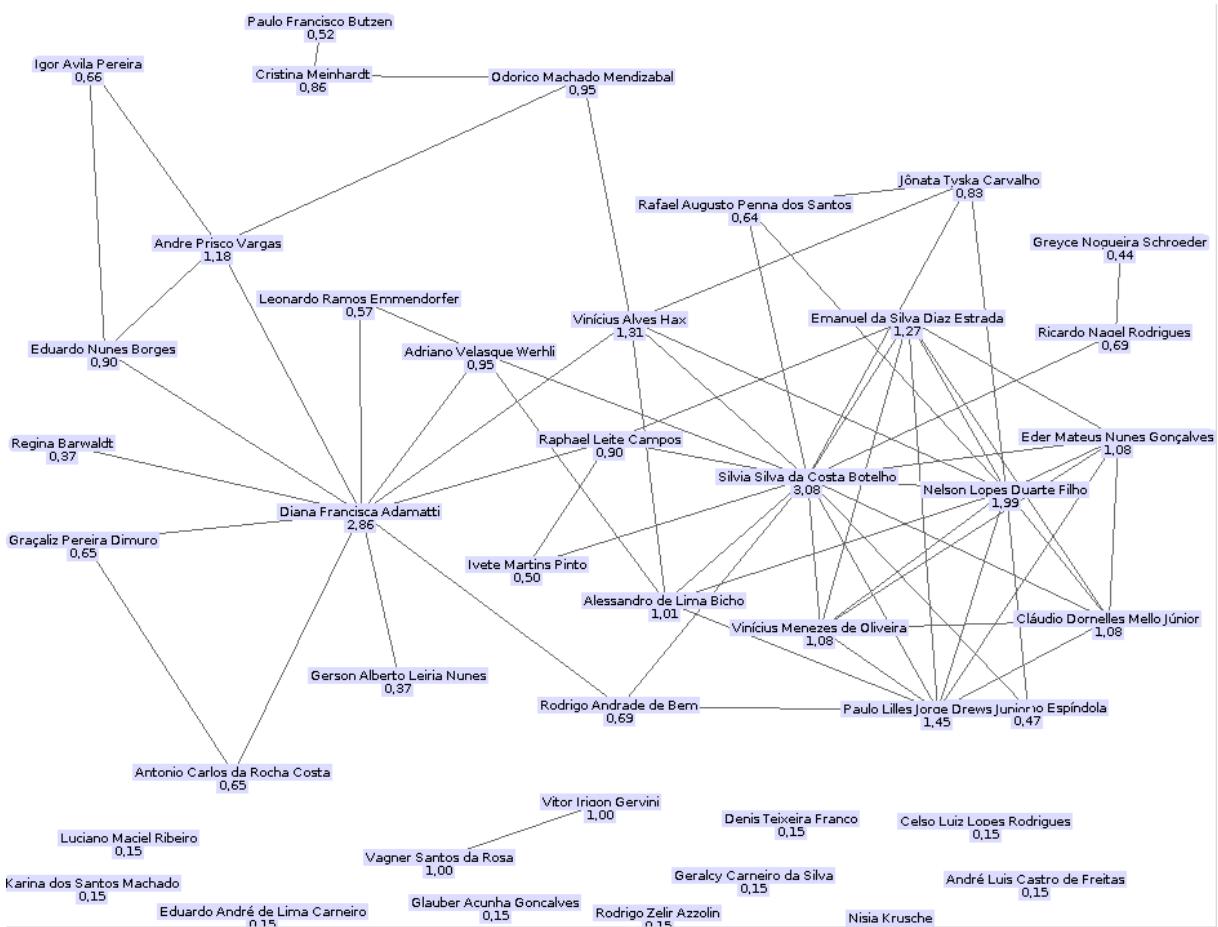


Figura 6.1: Rede de colaboração dos pesquisadores do C3, considerando publicações entre 2008 e 2012, apresentando os resultados da métrica *PageRank* para cada um dos pesquisadores.

- dois resumos publicados na Mostra de Produção Universitária da FURG em 2012 (de Farias et al., 2012a; Lucca et al., 2012);
- um resumo publicado na Mostra de Produção Universitária da FURG em 2011 (de Farias et al., 2011);
- um artigo curto submetido para a International Conference of the Chilean Computer Science Society em 2012.

Apesar do último artigo não ter sido aceito, as revisões apresentadas pelo comitê de programa foram essenciais para o desenvolvimento deste trabalho.

Como trabalhos futuros, destacam-se o desenvolvimento de uma interface *web* para a ferramenta cliente, desenvolvida nas linguagens PHP e JavaScript, e a implementação de

um banco de dados temporal, que permitirá a análise da evolução das redes de colaboração ao longo do tempo. Além disso, novas métricas devem ser implementadas, assim como uma outra função de deduplicação. Também deverão ser analisados os dados obtidos do *feedback* dos usuários para avaliar as funções de deduplicação utilizadas.

Referências

- Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*, pages 361–362.
- Batagelj, V. and Mrvar, A. (2002). Pajek - analysis and visualization of large networks. In Mutzel, P., Junger, M., and Leipert, S., editors, *Graph Drawing*, volume 2265 of *Lecture Notes in Computer Science*, pages 477–478. Springer Berlin Heidelberg.
- Borgatti, S. P., Everett, M. G., and Freeman, L. C. (2002). *Ucinet for Windows: Software for Social Network Analysis*. Analytic Technologies, Harvard, MA.
- Borges, E. N., de Carvalho, M. G., Galante, R., Gonçalves, M. A., and Laender, A. H. F. (2011). An unsupervised heuristic-based approach for bibliographic metadata deduplication. *Information Processing and Management*, 47(5):706–718.
- Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., and Marshall, M. (2002). Graph drawing. In Mutzel, P., Jünger, M., and Leipert, S., editors, *International Symposium on Graph Drawing*, volume 2265 of *Lecture Notes in Computer Science*, chapter GraphML Progress Report Structural Layer Proposal, pages 109–112. Springer.
- de Farias, L. R., Lucca, G., Lopes, G. R., Prisco, A., and Borges, E. N. (2012a). Uma ferramenta para visualização e análise de redes de pesquisa. In *Anais da XI Mostra da Produção Universitária da FURG*, Rio Grande, RS, Brasil.
- de Farias, L. R., Prisco, A., and Borges, E. N. (2011). Rede de pesquisa da furg. In *Anais da X Mostra da Produção Universitária da FURG*, Rio Grande, RS, Brasil.

- de Farias, L. R., Vargas, A. P., and Borges, E. N. (2012b). Um sistema para análise de redes de pesquisa baseado na plataforma lattes. In *Anais da VIII Escola Regional de Banco de Dados*, Curitiba, PR, Brasil.
- Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York.
- Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., and Woodhull, G. (2003). Graphviz and dynagraph - static and dynamic graph drawing tools. In *GRAPH DRAWING SOFTWARE*, pages 127–148. Springer-Verlag.
- Elmagarmid, A., Ipeirotis, P., and Verykios, V. (Jan. 2007). Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1–16.
- Heer, J., Card, S. K., and Landay, J. A. (2005). prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 421–430, New York. ACM.
- Laender, A. H., Moro, M. M., Gonçalves, M. A., Davis, Jr., C. A., da Silva, A. S., Silva, A. J., Bigonha, C. A., Dalip, D. H., Barbosa, E. M., Cortez, E., Procópio Jr., P. S., de Alencar, R. O., Cardoso, T. N., and Salles, T. (2011a). Building a research social network from an individual perspective. In *Proceedings of the International ACM/IEEE Joint Conference on Digital Libraries*, pages 427–428, New York. ACM.
- Laender, A. H. F., Moro, M. M., Silva, A. S., Davis Jr., C. A., Gonçalves, M. A., Galante, R., Silva, A. J. C., Bigonha, C. A. S., Dalip, D. H., Barbosa, E. M., Borges, E. N., Cortez, E., Procópio Jr., P., de Alencar, R. O., Cardoso, T. N. C., and Salles, T. (2011b). Ciênciabrasil - the brazilian portal of science and technology. In *Seminário Integrado de Software e Hardware, Anais do Congresso da Sociedade Brasileira de Computação*, pages 1366–1379.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

- Lucca, G. (2013). Módulo para análise de redes de colaboração acadêmicas do Research.Net. Monografia de Graduação em Sistemas de Informação, Centro de Ciências Computacionais, FURG, Rio Grande.
- Lucca, G., de Farias, L. R., Lopes, G. R., Prisco, A., and Borges, E. N. (2012). Um componente de software para análise de redes de colaboração acadêmica. In *Anais da XI Mostra da Produção Universitária da FURG*, Rio Grande, RS, Brasil.
- MenaChalco, J. P. and Cesar Junior, R. M. (2009). ScriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15:31 – 39.
- Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.
- Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998, New York. ACM.

Apêndice A

Diagrama de Classes do Cliente

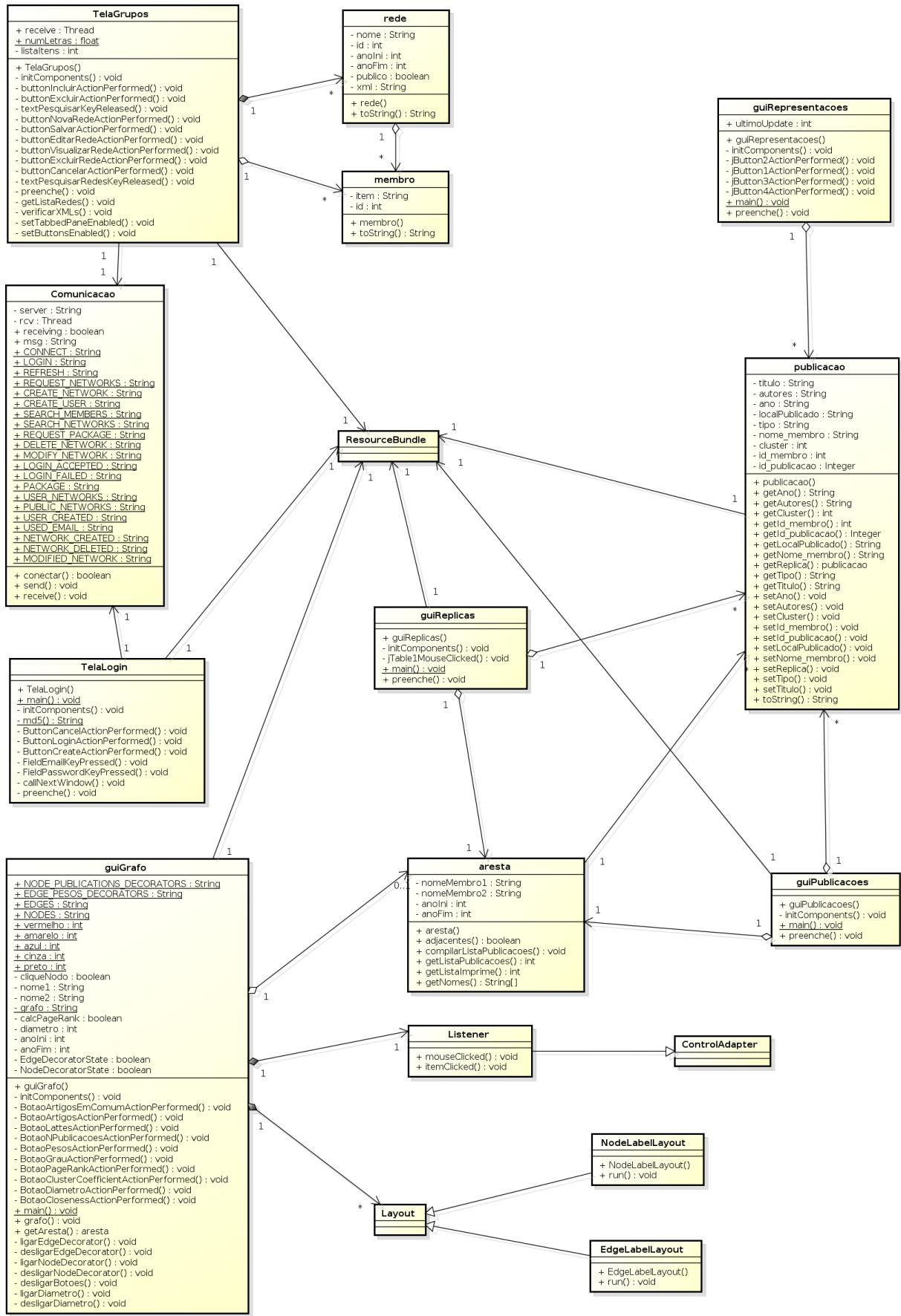


Figura A.1: Diagrama de classes do cliente.

Apêndice B

Diagrama de Classes do Servidor

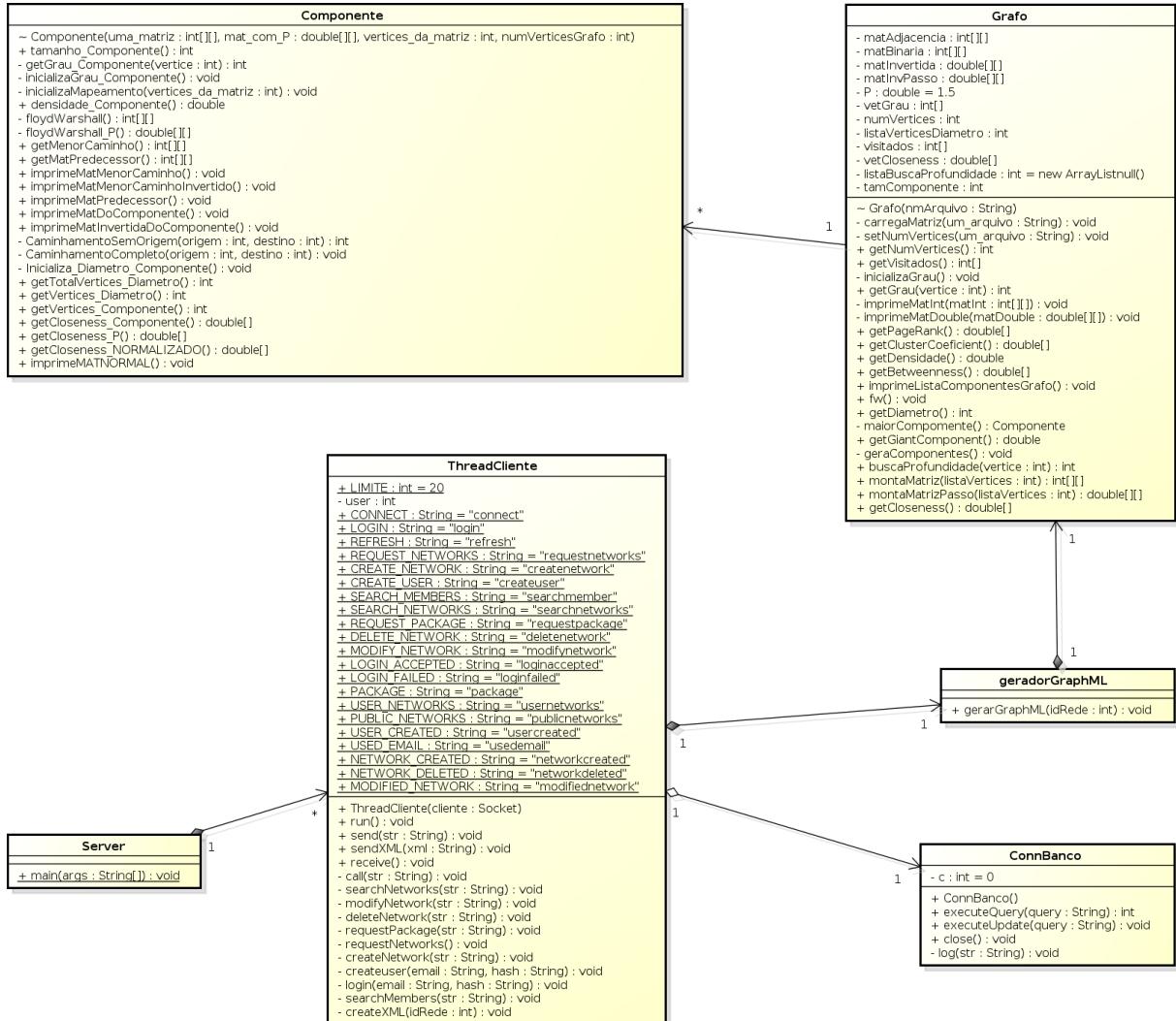


Figura B.1: Diagrama de classes do servidor.