SENAI CIMATEC

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL E TECNOLOGIA INDUSTRIAL

Mestrado em Modelagem Computacional e Tecnologia Industrial

Dissertação de Mestrado

Um Modelo Computacional para Extração Textual e Construção de Redes Sociais e Complexas

Apresentada por: Patrícia Freitas Braga
Orientador: Hernane Borges de Barros Pereira
Co-orientador: Marcelo A. Moret

Setembro de 2010

Patrícia Freitas Braga

Um Modelo Computacional para Extração Textual e Construção de Redes Sociais e Complexas

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial, Curso de Mestrado em Modelagem Computacional e Tecnologia Industrial do SENAI CIMATEC, como requisito parcial para a obtenção do título de **Mestre em Modelagem Computacional e Tecnologia Industrial**.

Área de conhecimento: Interdisciplinar

Orientador: Hernane Borges de Barros Pereira
SENAI CIMATEC

Co-orientador: Marcelo A. Moret
SENAI CIMATEC

Salvador
SENAI CIMATEC
2010

Nota sobre o estilo do PPGMCTI

Esta dissertação de mestrado foi elaborada considerando as normas de estilo (i.e. estéticas e estruturais) propostas aprovadas pelo colegiado do Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial e estão disponíveis em formato eletrônico (*download* na Página Web http://ead.fieb.org.br/portal_faculdades/dissertacoes-e-teses-mcti.html ou solicitação via e-mail à secretaria do programa) e em formato impresso somente para consulta.

Ressalta-se que o formato proposto considera diversos itens das normas da Associação Brasileira de Normas Técnicas (ABNT), entretanto opta-se, em alguns aspectos, seguir um estilo próprio elaborado e amadurecido pelos professores do programa de pós-graduação supracitado.

SENAI CIMATEC

Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial

Mestrado em Modelagem Computacional e Tecnologia Industrial

A Banca Examinadora, constituída pelos professores abaixo listados, leram e recomendam a aprovação [com distinção] da Dissertação de Mestrado, intitulada "Um Modelo Computacional para Extração Textual e Construção de Redes Sociais e Complexas", apresentada no dia 28 de Setembro de 2010, como requisito parcial para a obtenção do título de **Mestre em Modelagem Computacional e Tecnologia Industrial**.

Orientador:

Prof. Dr. Hernane Borges de Barros Pereira
SENAI CIMATEC

Co-orientador:

Prof. Dr. Marcelo A. Moret
SENAI CIMATEC

Membro externo da Banca:

Prof. Dr. Roberto C. S. Pacheco
EGC/UFSC

Membro interno da Banca:

Prof. Dr. Nome completo
Instituição do membro da banca

Dedico este trabalho a Deus que me manteve firme até o fim e aos meus pais, sem os quais não teria chegado tão longe.

Agradecimentos

Agradeço em primeiro lugar a Deus, por ter me mantido persistente a alcançar meus objetivos e aos meus pais amados, pelo apoio e carinho constantes durante esta jornada.

Ao meu noivo, que foi bastante paciente e companheiro, com minhas ausências justificadas pelos momentos de trabalho árduo na elaboração desta dissertação.

Ao meu orientador, Prof. Dr. Hernane Borges de Barros Pereira, pelo incentivo e confiança no desenvolvimento deste trabalho.

Ao Prof. Dr. Marcelo Moret, e Doutoranda Teresinha Tamanini, pelo suporte e apoio necessários para a elaboração deste trabalho.

Aos meus amigos e colegas do mestrado, pelo apoio e companherismo.

Salvador, Brasil — Patrícia Freitas Braga
28 de Setembro de 2010

Resumo

As redes complexas estão presentes em diversos níveis, desde redes biológicas até redes sociais, sendo as redes de colaboração científica o foco de estudo desta dissertação. Visando subsidiar o estudo acerca da dinâmica da difusão do conhecimento, este trabalho de mestrado objetivou apresentar uma modelagem computacional para extração de dados de textos para a construção de redes sociais e complexas a partir da detecção destas estruturas implícitas nos textos, sendo estas redes construídas com base em relações de colaboração científica em autoria de publicações. Considerando que boa parte das informações se encontram em repositórios textuais digitais, o modelo provê métodos para otimização na obtenção destes dados de textos e geração de redes a partir destes. Atualmente, há poucos trabalhos que realizem mineração de textos para coleta de dados específicos e que construa redes complexas destes dados minerados. Este trabalho apresenta as etapas processuais do modelo, que envolve mineração dos textos e obtenção das redes, e apresenta os resultados obtidos a partir da utilização do modelo proposto, utilizando como procedimento experimental, a construção das redes de colaboração em produções científicas no contexto de programas de pós-graduação. Na avaliação dos resultados obtidos a partir do trabalho experimental, no aspecto de criticidade do modelo, constatou-se que a dependência de conhecimento mínimo para minerar textos está associado com a precisão dos resultados obtidos da extração dos dados. Quanto a construção das redes, todas foram geradas corretamente e a partir destas, propriedades de redes sociais e complexas puderam ser analisadas.

Palavras-chave: Mineração de Textos, Expressões Regulares, Redes Complexas, Redes Sociais

Abstract

Complex networks are present in various levels, from biological networks to social networks, being the scientific collaboration networks the focus of study in this thesis. Aiming to support the study of the dynamics of diffusion of knowledge, this master's thesis aims to present a computational model for extracting data from texts to build social and complex networks from the detection of these structures implicit in the texts, which are networks built based on relations of scientific collaboration in authorship of publications.Whereas much of the information are in textual digital repositories, the model provides methods for optimization in obtaining these data and generating texts from these networks. Currently, there are few jobs that perform text mining to collect specific data and build complex networks from data. This work presents the procedural steps of the model, which involves text mining and obtaining the network and will present the results obtained by the use of the proposed model, using the experimental procedure, the construction of networks of collaboration in scientific production in the context of post-graduation. The evaluation of the results obtained of the experiemental work, int the critical aspect of the model, found that the dependence of minimum knowledge to text mining is associated with the accuracy of the results of data extraction. As the construction of networks, all corretamente were generated and from these, properties of complicated networks were analyzed.

Keywords: Text Mining, Regular Expressions, Complex Networks, Social Networks

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Siglas

CAPES	Coordenação de Aperfeiçoamento de Pessoal Nível Superior
PPGMCTI ..	Programa de Pós-graduação em Modelagem Computacional e Tecnologia Industrial
PLN	Processamento de Linguagem Natural
PPG	Programa de Pós-Graduação
RI	Recuperação de Informações
EI	Extração de Informações
GNU	General Public Licence
KDD	Knowledge Discovery in Databases
KDT	Knowledge Discovery in Texts
WWW	World Wide Web
PDF	Portable Document Format
HD	Hard Disk
QED	QED Text Editor
HTML	Hypertext Markup Language

Capítulo Um

Introdução

O interesse pelo estudo das redes sociais e complexas é cada vez mais comum na sociedade. Isto pode ser justificado pela importância que a compreensão das estruturas formadas pelas redes tem na análise de um contexto específico, tais como, redes de amigos, redes biológicas, redes de links entre páginas da web, redes científicas entre outros. De forma simplificada, uma rede pode ser definida como um conjunto de vértices que são interconectados por arestas. Uma outra definição para redes, segundo Boccaletti (2006), é que redes podem ser entendidas como entidades definidas em um espaço abstrato, tais como redes de conhecimentos e colaborações entre indivíduos.

Uma rede complexa é uma rede com características topológicas não triviais. Uma rede social também se caracteriza como rede, entretanto o foco são relações sociais em grupos de indivíduos. Redes de colaboração científica podem ser consideradas como redes sociais, já que considera as interações entre os pesquisadores em tarefas acadêmicas, sendo este o foco desta pesquisa. O entendimento da natureza das redes e sua dinâmica de conectividade conduzem a uma clara visão acerca do comportamento das redes, como se formam, quais são os elos fracos, os vértices de maior centralidade, vulnerabilidades na estrutura, capacidade de expansão, presença de agrupamentos, nível de coesão entre os elementos de sua composição, entre outras características que configuram as redes, sejam sociais e/ou complexas.

Dentre os benefícios obtidos com os estudos das características topológicas de redes sociais e complexas, está a capacidade de previsibilidade da dinâmica da rede a partir da inserção de novos eventos neste contexto da estrutura rede. O entendimento desta dinâmica também auxilia a tomada de decisão em resposta a esses eventos atuantes na rede, podendo-se citar como exemplo de possível evento, a introdução ou remoção de vértices desta rede. Uma boa representação deste fato é a identificação de pontos de vulnerabilidade em um emaranhado de conexões entre pontos. Em um único ponto da rede onde haja alta conectividade com outros pontos, se uma falha ocorrer, pode rapidamente afetar outros pontos desse sistema, provocando o efeito falha em cascata. Um exemplo bastante comum de uma rede onde se observa esse comportamento é o sistema de rede elétrica. Se há um problema na central de energia, todos os pontos de luz conectados a ela serão derrubados, causando um blecaute. Outro exemplo que retrata bem o efeito que um evento pode produzir em um vértice muito conectado é a internet.

O primeiro foco dado às redes e suas propriedades veio com a teoria dos grafos, que surgiu a partir de um experimento para resolução do problema das pontes de Konigsberg,

realizado por Leonhard Euler, e proporcionalmente, criou-se um ambiente propício para o desenvolvimento do campo ou área das redes complexas. Isto estimulou trabalhos posteriores de outros pesquisadores como Paul Erdós e Alfred Renyi e o estudo de redes randômicas, Duncan Wattz e seu orientador Steven Strogatz e a teoria de mundos pequenos, e Réka Albert e Albert-László Barabási e suas pesquisas em redes livres de escala. A partir daí, formam-se marcos para estabelecer topologias distintas de redes. Tendo em vista a necessidade de análise das características topológicas das redes formadas a partir das associações entre autores em tarefas acadêmicas, o foco de estudo desta pesquisa são as redes de colaboração científica em programas de pós-graduação.

Vivemos uma realidade onde se discute exaustivamente a questão da importância na difusão e compartilhamento da informação para socialização do conhecimento. Considerando essa realidade, a participação coletiva em produções científicas é incentivada pela comunidade científica, em decorrência do potencial de qualidade obtido em publicações científicas, resultado do esforço coletivo na união de conhecimentos e competências diversas e interdisciplinares. As características topológicas das redes de co-autoria podem revelar aspectos importantes na dinâmica das associações entre os pesquisadores, não apenas a mera quantidade de conexões e agrupamentos, mas a evolução comportamental e temporal dessas redes. Pode-se revelar, por exemplo, a existência de preferências e privilégios em grupos de pesquisadores, agrupamentos preferenciais formados por aproximação profissional ou por relações de amizade, entre outros.

Uma dificuldade encontrada para análise de redes sociais reside no mapeamento, extração e quantificação dos dados necessários à construção das redes, devido à sua complexidade e composição, que nem sempre têm proporções modestas. O processo de busca e filtragem de dados para construção das redes não é um trabalho trivial, e se apresenta demasiadamente custoso em termos de tempo, esforço e orçamento se realizado inteiramente sob intervenção humana. Em virtude desta dificuldade apresentada para análise das redes de colaboração, a pesquisa visa desenvolver um modelo que auxilie neste processo de obtenção dos dados e construção das redes, que por fim subsidiarão o estudo da dinâmica destas redes.

1.1 Definição do problema

Em decorrência da falta de acesso à base de dados da CAPES [1], ou pelo menos uma cópia dela, o modelo proposto para este trabalho foi idealizado no sentido de prover uma forma de obtenção dos dados necessários para construção das redes. As redes a serem criadas deverão ser compostas por dados da CAPES. Entretanto, como os dados não podem ser obtidos diretamente da base de dados desta instituição, a fonte de informações mais

[1] www.capes.org.br, último acesso em 03/08/2010

completa acerca das tarefas acadêmicas em PPGs são oriundas dos cadernos indicadores da CAPES, que são textos em formato digital. Por esta razão, o principal problema é como desenvolver um modelo computacional que realize mineração de textos, gestão de dados e construção de redes sociais e complexas para estudar as redes de colaboração científica em programas de pós-graduação.

No estudo das redes sociais são identificadas duas etapas essenciais: a primeira consiste na obtenção e classificação das informações e a segunda consiste na construção das redes. São procedimentos independentes que, apesar de trabalhoso, resultarão nos dados necessários para o estudo das redes. Como pontos a serem resolvidos no desenvolvimento do modelo destacam-se a criação de um mecanismo de extração e filtragem dos dados, por meio de mineração dos textos e reconhecimento de padrões, e o desenvolvimento de um algoritmo que gere a construção das redes.

1.2 Objetivo

Com base na argumentação do problema identificado, o objetivo da pesquisa é propor um modelo que possibilite a realização de mineração de textos, gerência de dados e construção das redes. Para consolidar a modelagem proposta, será desenvolvida uma ferramenta que realize a extração de informações específicas de textos digitais, armazene os dados no banco de dados, recupere os dados e gere posteriormente as redes de colaboração científica. Os objetivos específicos, são definidos a seguir:

1. Estudar os modelos de mineração de dados e textos: Analisar alternativas de técnicas mineração de textos para facilitar a tarefa de extração de dados dos textos;
2. Selecionar o modelo mais adequado para o estudo das redes de colaboração científica: Identificar técnicas ou métodos de mineração de textos mais adequados para extração de textos;
3. Desenvolver uma ferramenta que realize a extração de informações específicas de textos digitais, armazene os dados no banco de dados, recupere os dados e gere posteriormente as redes de colaboração científica;
4. Validar os resultados obtidos com base em documentos oficiais (Cadernos CAPES): Realizar a verificação das redes construídas comparando-se os documentos utilizados na mineração de textos;

1.3 Importância da pesquisa

O modelo proposto contribui com o processo de instituição de políticas nacionais de incentivo e fomento à pesquisa, uma vez que se tornam possíveis a detecção de padrões de comportamento entre os pesquisadores, a identificação de onde se concentra maior participação em co-autorias intra e extra instituições, a detecção das instituições que possuem mais pesquisadores ativos, entre outros fatores. Por exemplo, a partir das análises conseqüentes aos artefatos gerados pelo modelo, do ponto de vista econômico, novos investimentos em pesquisas poderão acontecer na visão de estimular o aumento e a circulação de produções científicas no Brasil. Neste sentido, o modelo computacional proposto tem como fator de motivação, facilitar o estudo do comportamento da dinâmica dessas estruturas formadas pelas redes de colaboração científica, por meio da utilização de ferramentas existentes no modelo que auxiliam na coleta de informações, necessárias para análise destas redes presentes nos programas de pós-graduação.

Tendo em vista o peso da importância no estudo de redes, é interessante adotar métodos não apenas qualitativos e empíricos, mas também quantitativos, que validem o diagnóstico resultante das análises apresentadas para a rede estudada. Considerando que as informações serão retiradas de textos, a modelagem proposta provê ferramentas para auxílio na execução de tarefas que precedem à análise dessas redes. Durante as etapas do processo de utilização do modelo, espera-se minimizar a intervenção humana na coleta das informações e construção das redes. A partir da obtenção das redes de colaboração científica dos textos, o pesquisador poderá ter a compreensão de como se estabelece as relações de colaboração científica nos programas de pós-graduação, identificar padrões de comportamento nestas associações, observar a existência de interações preferenciais e agrupamentos entre os autores, estabelecer metas a partir do diagnóstico gerado pela análise das características destas redes, entre outros.

O pensamento de centralizar os processos necessários ao estudo das redes sociais em um ambiente é interessante, porém há pontos críticos a serem analisados. De que forma se dará a coleta de dados? Como serão compostas as redes para análise? A fonte de informações para coleta dos dados serão textos digitais, e nesse aspecto, faz-se necessário a utilização de técnicas e conceitos encontrados em mineração de textos e expressões regulares. Os textos selecionados para análise contêm informações de autorias e produções científicas, que deverão ser extraídas para posterior geração das redes. As informações contidas nestes textos serão extraídas por meio de reconhecimento de padrões textuais, assim, será utilizado o conceito de expressões regulares na mineração dos textos.

A compreensão da dinâmica social das estruturas das redes de co-autorias, está correlacionada ao conhecimento das propriedades topológicas de redes sociais e complexas, que são intrísecas a elas. Dimensionar e perceber a evolução destas redes, só é possível se houver

entendimento da mecânica de redes complexas. A partir das redes obtidas pelo modelo, será possível identificar essas propriedades topológicas de redes sociais e complexas, e fazer análises do comportamento destas redes de colaboração científica em programas de pós-graduação.

1.4 Limites e limitações

O fato de os dados das redes e suas estruturas interativas estarem implícitas nos documentos demonstrou ser um aspecto a ser investigado. Como extrair, não apenas dados explicitos nos textos, mas as relações implícitas nos documentos? As estruturas que compõem as redes a serem construídas são baseadas no contexto das informações, não nas estruturas semânticas. Estruturas semânticas se baseiam em relações existentes entre as palavras (signos) de um texto que dão significado a um conjunto maior de palavras. No caso previsto nesta pesquisa de mestrado, as redes construídas estão no contexto das redes de co-participação em programas de pós-graduação, recomendados pela CAPES. O modelo deve detectar e construir as redes implícitas nos textos das relações entre os pesquisadores, considerando os seguintes tipos de publicações: artigos em periódicos, trabalhos em anais de eventos, capítulos de livro e livros, além das participações em bancas e projetos.

Considerando que as relações estruturais das redes estão em âmbito contextual, ou seja, no contexto do texto e não na estrutura formada pelas ligações entre as palavras, passa-se a usar expressões regulares para coleta destes dados e relacionamento destes. Expressões regulares se baseiam em reconhecimento de padrões em textos. Como os dados relevantes para as redes são muito específicos, técnicas de mineração de textos comuns, a exemplo do uso de técnicas para sumarização de conteúdo de textos, não desempenhariam de forma eficiente a garimpagem de dados e criação de estruturas das redes sociais e complexas que se deseja encontrar. As técnicas mais utilizadas para mineração de textos tratam apenas de coletar dados mais frequentes em textos, entretanto, não englobam todos os dados de interesse e não relacionam de forma direcionada estes dados.

1.5 Aspectos metodológicos

Partindo-se da idéia de que seria necessária a aplicação de técnicas específicas para extração dos dados dos textos, a utilização de alguns conceitos de mineração de textos foi necessária, haja vista que a coleta de dados proposta neste modelo computacional se baseia em identificação de dados relevantes dos textos para construção das redes. A Mineração de Textos, também conhecida como Descoberta de Conhecimento em Textos (Knowledge

Discovery in Text -KDT) (MONTEIRO, 2006), consiste basicamente em um processo de extração e classificação de conhecimento significante em fontes de dados textuais não estruturadas ou semi-estruturadas. O procedimento adotado neste processo compreende as etapas de varredura do texto, pré-processamento do texto, que envolve a eliminação de palavras insignificantes denominadas de *stopwords*, correção ortográfica, redução a radicais da palavra, análise dos dados extraídos e for fim indexação dos termos resultantes. A mineração de textos explora técnicas e metodologias da área de recuperação de informações ou RI (Information Retrieval) (FELDMAN; SANGER, 2007), e utiliza processamento de linguagem natural no tratamento dos dados, o que não é tarefa simples.

Uma das etapas da modelagem proposta nesta pesquisa, visa encontrar dados específicos e de importância para a construção das redes. A aplicação de técnicas mais comuns de mineração de textos, tais como a sumarização de conteúdo de textos, se baseia em identificar dados mais frequentes nos textos. No caso desta pesquisa, os dados de interesse não podem ser considerados pelos seus números de ocorrências nos textos, assim sendo, deve ser considerado todo e qualquer dado que seja pertinente a contrução das redes. Considerando estes fatos, minerar textos para a contrução de redes, conforme a proposta desta modelagem, exige maior especificidade quanto a coleta de informações dos textos, para que os dados coletados sejam significantes para as redes.

A mineração de textos, na identificação de termos relevantes nos documentos de uma forma geral, considera e seleciona apenas palavras mais frequentes nos textos. Entretanto, no modelo desta pesquisa, a utilização desta forma de coleta de dados, traria muitos dados irrelevantes, que não seriam utilizados na contrução das redes. Desta forma, o esforço empreendido na limpeza dos dados, teria um aumento significativo se comparado a técnica de mineração de textos por reconhecimento de padrões, a qual foi aplicada nesta modelagem. O reconhecimento de padrões, se apresentou como a forma mais adequada de mineração de textos para coleta de dados específicos, haja vista que nos documentos selecionados, nomes de autores quando citados em publicações, estão contidos em um padrão de formatação das palavras, que são facilmente identificados.

Em vias de agilizar e viabilizar essa coleta dos dados, a utilização de expressões regulares na captura das informações se mostrou mais eficiente que minerar textos de forma global. As expressões regulares são compostas por uma linguagem formal, que descreve um padrão a ser identificado dentro de uma cadeia de caracteres. Segundo (FRIEDL, 1997), as expressões regulares constituem um poderoso, flexível e eficiente processador de textos. Baseadas em notações de padrões, elas podem encontrar e transformar textos dentro de textos, examinando em cada sequência de caracteres a igualdade com o padrão pré-determinado.

Como procedimento metodológico adotado para verificação da usabilidade do modelo

abordado nesta pesquisa, será realizado um experimento cujo objetivo é estudar as redes de colaboração científica em programas de pós-graduação, a partir de um corpus documental em formato PDF, de onde serão retiradas as informações para construção das redes. Estes documentos digitais são os cadernos indicadores de produções científicas da CAPES.

O experimento realizado para aplicação do modelo compreende três etapas básicas: a extração dos dados dos textos, a inserção deste dados em uma base de dados e construção das redes e suas estatísticas descritivas, que são relativas a percentagem de participação de pesquisadores por vínculo e por tipo de publicação. A primeira etapa do modelo compreende a coleta das informações, que abrange conceitos de reconhecimento de padrões e expressões regulares, que serão utilizados para obter os dados textuais.

Esta primeira etapa engloba algumas atividades fundamentais para o processo de obtenção dos dados, tais como a submissão dos documentos selecionados ao modelo, onde serão convertidos para extensão TXT, a execução da varredura textual na busca por padrões reconhecíveis e designados pelo pesquisador e a limpeza dos resultados encontrados, para filtragem e armazenamento dos dados filtrados em listas de dados.

Na segunda etapa, já com dados obtidos na etapa anterior e devidamente armazenados em listas, os dados serão registrados no banco de dados. São dois tipos de inserção a serem executados nesta etapa: A inserção dos dados primários e a inserção das relações destes dados. A primeira se refere às unidades formativas das redes, aquelas que representarão os vértices da rede. A segunda está vinculada à relação existente entre essas unidades formativas, são representadas pelas conexões presentes na rede (eg.: arestas). Então se observa que nesta fase do experimento já se tem os componentes básicos de uma estrutura de rede: os vértices e suas arestas.

A terceira etapa do processo consiste na recuperação dessas informações registradas. Essas informações serão recuperadas do banco por meio de filtros de seleção, que delimitarão um contexto para geração das redes. É nesta etapa que será possível guardar as informações da rede para utiliza-las em outros softwares, a exemplo do Pajek [2], onde serão calculados os valores dos índices ou propriedades da rede criada, que servirão de subsídio para análise desta rede.

Na realização do experimento serão necessários, considerando o aspecto técnico, um software para desenvolvimento (Visual Studio 2008), um Banco de Dados, neste caso o MySQL versão 5.01, e conhecimentos razoáveis em CSharp (linguagem de programação). No aspecto teórico será imprescindível conhecer os conceitos de mineração de textos e reconhecimento de padrões para a criação do módulo de geração de expressão regular e características e conceitos de redes sociais e complexas.

[2] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

O experimento, pretende demonstrar que o modelo auxilia de fato a coleta das informações e construção das redes, considerando-se que o tempo para realização de todas as tarefas necessárias para composição das redes (e.g. coletar dados, relacionar dados e construir as redes), de forma puramente manual seria maior, e estaria mais vulnerável a erros humanos. Desta forma, este modelo dará suporte a construção das redes de modo que o estudo de suas propriedades permita compreender a dinâmica da formação das redes sociais na colaboração científica.

1.6 Organização da Dissertação de Mestrado

Esta dissertação é composta por 6 capítulos e está estruturada da seguinte forma:

- **Capítulo 1 - Introdução**: Contextualiza o âmbito, no qual a pesquisa proposta está inserida. Apresenta, a definição do problema, objetivos e justificativas da pesquisa e como esta dissertação de mestrado está estruturada;
- **Capítulo 2 - Mineração de Textos**: Aborda conceitos gerais de mineração de textos, processamento de linguagem natural e apresenta aspectos processuais, além de fazer um comparativo em relação ao processo de mineração de dados. Apresenta uma dicussão sobre as expressões regulares, sintaxe padrão e criação de notações de padrões;
- **Capítulo 3 - Redes Sociais e Complexas**: Apresenta fundamentos teóricos acerca das redes sociais e complexas, abordando os modelos teóricos de redes aleatórias, redes mundo-pequeno e redes livres de escala e suas características topológicas;
- **Capítulo 4 - Modelo para Extração de Dados Textuais e Geraçaõ de Redes**: Neste capítulo é apresentado o modelo computacional para extração textual de dados para construir redes sociais e complexas, onde se analisa a modelagem e implementação do modelo;
- **Capítulo 5 - Trabalho Experimental**: Demonstra o experimento realizado, assim como apresenta as etapas do processo, redes construídas pelo modelo e análise de suas redes;
- **Capítulo 6 - Considerações Finais**: Apresenta as conclusões, contribuições da pesquisa e algumas sugestões de atividades de pesquisa a serem desenvolvidas no futuro.

Capítulo Dois

Mineração de Textos

2.1 *Mineração de Textos - Conceitos e Fundamentos*

A tendência para armazenagem de textos em meios digitais vem aumentando nos últimos anos, em decorrência da facilidade de acesso e redução de espaço físico. Na internet pode-se observar diversos tipos de arquivos textuais, disponibilizados como documentos em extensão PDF, DOC, páginas de web, emails entre outros. Entretanto, diferente de um armazenamento estruturado, como é o caso de um banco de dados, os textos não apresentam estruturas bem definidas das suas informações. Isto torna o processo de busca de dados mais complexo em um grande volume de dados textuais, se comparado a busca de informações em um banco de dados. Em decorrência dessa complexidade, foram criadas técnicas para realização do tratamento destes dados e extração de conteúdo relevante.

Mineração de Textos (Text Mining) é também conhecida por Descoberta de Conhecimento em Textos ou KDT (Knowledge Discovery in Texts) (MONTEIRO, 2006) e consiste basicamente em técnicas extrativas de dados relevantes de uma coleção de textos. Segundo Aranha e Passos (2006), a mineração de textos procura extrair conhecimentos úteis de dados não estruturados ou semi-estruturados. Na visão de Feldman e Sanger (2007), a mineração de texto procura extrair informação útil de fontes de dados textuais, por meio de identificação e exploração de padrões de interesse. Em ambos os casos, tratam-se de fontes de dados provenientes de textos, que necessitam de processamento de linguaguem natural (PLN) para descoberta de informação de interesse, já que textos não seguem um padrão de armazenagem de dados.

As técnicas de mineração de texto têm como base conceitual as técnicas de mineração de dados e processamento de linguaguem natural. Assim como a mineração de textos, a mineração de dados busca extração de conhecimento implícito por meio do reconhecimento de padrões em fontes de dados. Porém, a fonte de informação a ser explorada na mineração de dados se baseia em base de dados estruturados, ao passo que na mineração de textos a fonte de dados são textuais não estruturados ou semi-estruturados. A mineração de texto também está associada a técnicas da Recuperação de Informações ou RI (Information Retrieval). A RI envolve conceitualmente o processo de representar, armazenar e procurar informação relevante para um ideal específico (INGWERSEN, 1999).

O aspecto da linguaguem na mineração de textos tem grande importância, porque um

dos princípios básicos para extração de conteúdo relevante de textos é o tratamento das informações, por meio de processamento de linguagem natural. Isto porque os textos, além de não apresentarem estrutura de dados ordenada, estão sujeitos a fatores humanos de forma escrita das palavras. Em linhas gerais, a mineração de textos compreende as seguintes etapas (MONTEIRO, 2006):

1. Obtenção da fonte de dados;
2. Pré-processamento dos dados;
3. Análise dos dados e Extração de Conhecimento;
4. Avaliação das Descobertas.

A obtenção de dados é a fase de seleção do corpus documental a ser utilizado no processo de busca da informação requerida. As fontes de dados podem ser arquivos de texto em extensão PDF, DOC ou páginas HTML, por exemplo. Uma vez obtida a fonte de dados, esses dados são pré-processados para posteriormente serem analisados e daí se obter conhecimento. Nesta etapa de pré-processamento, é onde os textos são tratados por meio de algoritmos para minimizar dados sem importância incluídos nos resultados obtidos para análise.

O processamento da linguagem natural (PLN) é o primeiro processo para mineração dos textos e se refere ao tratamento aplicado aos textos. O PNL é composto pelas tarefas de correção ortográfica, eliminação de palavras sem relevância na compreensão do texto denominadas *stopwords* (e.g. preposições), e redução a radicais da palavra, cujo processo é denominado *stemming*. Uma vez processado e refinado o texto, este então é submetido a etapa de análise dos dados, que é de onde serão extraidos os conhecimentos. Nesta fase, o documento é submetido ao processo de sumarização, que é composto por quatro ações básicas: a separação do texto por sentenças, o pré-processador, os analisadores estatísticos (que realizam os cálculos de frequências de termos nos documentos e indexação dos termos em uma matriz de termos e frequências) e a formatação do sumário (VEIGA, 2009).

O processo de sumarização do texto se inicia pela separação do documento com base nas sentenças encontradas, sendo estas armazenadas em um documento pós-processado. Após processamento, este documento é submetido a um algoritmo que fará a leitura dos termos encontrados (i.e. os *stems*) e realizará a análise estatística do texto, que consiste em calcular o número de ocorrências de cada termo e atribuir um grau de relevância a cada um. Esses resultados encontrados pela análise estatística, são armazernados em um vetor de termos com seu grau de relevância. A partir deste vetor formado, será composto o sumário de conteúdo do texto minerado.

2.1.1 Processamento de Linguagem Natural

Dois conceitos são amplamente estudados quanto se trata de minerar textos: processamento de informações e recuperação de informações. Quando se trata de extrair conhecimento de fontes de dados não estruturados, que é o caso de textos, é necessária uma preparação desta fonte de dados para que o processo de descoberta do conhecimento seja realizado. O Processamento de Linguagem Natural (PLN), é o processo que trata os dados utilizando-se de técnicas específicas para preparar o documento de modo a ser possível sua submissão ao processo de extração das informações. O PLN é multidisciplinar e está associado a conhecimentos de áreas como Inteligência Artificial, Cognição, Computação e Estatística. Segundo Liddy (2005), a linguagem pode ser analisada em diversos níveis linguísticos: semântico, morfológico, sintático, etc. Considerando os diversos níveis de compreensão linguística, para a realização de tratamento de textos, é necessário se definir em quais destes níveis o texto será processado, porque cada nível exige um processamento e análise específicos. Sendo assim, a tarefa de tratamento de textos não é uma tarefa trivial.

Todo texto deve ter sua forma original representada de maneira reduzida para extração de conhecimento, isto porque a linguagem humana não assume um formato organizado de representação, ela se constitui em agrupamentos de palavras que adquirem significado no geral. Em vista desta disposição extensa e não estruturada de dados, a representação do conteúdo do documento deve ser a mais relevante e precisa possível. Alguns dos níveis para analisar a linguagem são: a fonética, a morfologia, o léxico, o sintático, o semântico, o discursivo e o pragmático (LIDDY, 2005). Nas técnicas de mineração de textos, na etapa do pré-processamento não necessariamente são feitas as análises em todos os níveis linguísticos supra citados, mas é utilizada especialmente a análise morfológica. Em alguns casos, a análise semântica é realizada quando são criadas ontologias de conteúdos.

As técnicas de PLN mais conhecidas aplicadas na mineração de textos são correções ortográficas, remoção de *stowords* e *stemming*. Na correção ortográfica, um dicionário contendo vocabulário de linguagem específica é comparado ao texto, já transformado em vetor de palavras, para comparação dos termos, e recriação de novo vetor contendo o texto já corrigido. Na remoção de *stowords* o vetor do texto é novamente lido e remontado, eliminados termos específicos sem relevância para preservação do conteúdo do texto. E por último o *stemming* reduz as palavras a seus radicais para evitar repetições e reduzir dimensão de termos na matriz gerada no processo de indexação dos termos dos documentos.

Segundo Ingwersen (1999), a Recuperação de Informações se preocupa com o processo de representar, armazenar, procurar e encontrar informações que são relevantes a uma requisição de informação desejada por um usuário. Basicamente a RI se fundamenta sobre

três aspectos: representação do conteúdo, representação de consultas e busca (LOH, 1999). A busca por informações em textos, refere-se a forma como estes textos serão organizados e representados, de modo a acelerar as buscas por conteúdo de interesse. Para realização desta tarefa, é utilizado um indexador do vocabulário relevante, o qual cria interpretações e representações unificadas de significados de conteúdo (INGWERSEN, 1999). A mineração de textos se utiliza de ferramentas para criação de vocábulos unificados, como o corretor ortográfico e o *stemming*. O terceiro aspecto da RI é a busca, que trabalhará com o grau de relevância do documento para a consulta realizada.

Existem três modelos clássicos para aplicação da RI em textos:

- Modelo Booleano:

  Se baseia na teoria dos conjuntos e na álgebra booleana. Os documentos são representados por conceitos e características em um conjunto finito, segundo (LOH, 1999). As consultas realizadas por este modelo, são contruídas em forma de expressões booleanas, utilizando como operandos, estas características dos textos, e como operadores, a lógica booleana representada por AND, OR, NOT. Critérios de decisão binária são aplicadas para seleção de documentos de interesse, a partir desta expressão booleana. A Figura 2.1 mostra um exemplo do funcionamento da consulta em documentos baseada em lógica booleana.

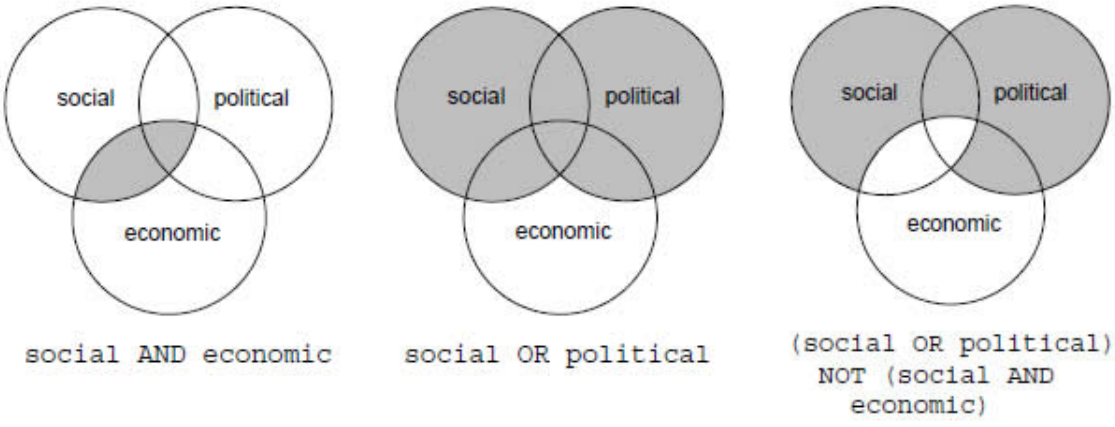


Figura 2.1: Exemplo de consultas que utilizam a lógica booleana. Fonte: (HIEMSTRA, 2008).

O primeiro conjunto mostrado na Figura 2.1, é uma combinação booleana que na consulta aceita apenas documentos que contenham as palavras "social" e "economic" no texto. É o caso da utilização do operador "AND" (i.e. "social" AND "economic"), ou seja o documento deve conter os dois termos. No segundo conjunto mostrado, os documentos aceitos pela combinação booleana devem necessáriamente conter ou o termo "social" ou o termo "political". Utiliza-se do operador "OR" (i.e. "social" OR "political"), ou seja basta que um dos termos apareça no texto

de um documento. O terceiro conjunto é uma combinação booleana mais complexa. Na consulta que utiliza essa lógica booleana, apenas documentos que contenham os termos "social" ou "political", e não contenham ao mesmo tempo os termos "social" e "economic" serão aceitos. Neste caso, há ainda a utilização do operador "NOT", que representa a negação de um termo (i.e. ("social" OR "political") NOT ("social" AND "economic")).

Na lógica booleana, ou o documento tem ou não tem determinado termo da consulta. Por essa razão esse modelo utiliza critério binário, ou seja 0 (se o documento não tem o termo) ou 1 (se o documento tem o termo). Apesar de ser o modelo mais utilizado de forma comercial, e possuir um formalismo semântico preciso, a dificuldade encontrada para aplicação deste modelo está em se formular expressões booleanas. Um exemplo de extração de informações que utiliza esse modelo são as expressões regulares, as quais se fundamentam na teoria algébrica de kleene e álgebras booleanas.

- Modelo Vetorial: Este modelo admite valores não binários como parte do conjunto de pesos dos índices, sendo estes pesos associados ao grau de similaridade entre os documentos. Desta forma, é possível estabelecer um ranking por relevância dos documentos considerando esses valores intermediários que estão compreendidos em um intervalo entre 0 e 1. Neste modelo, uma consulta é representada por um vetor de termos, onde cada termo tem um peso, sendo esta representada por:

$$\vec{q} = (\omega_{1q}, \omega_{2q}, ..., \omega_{tq}) \tag{2.1}$$

onde $\vec{q}$ é o vetor de termos da consulta, ω_{tq} é o peso de cada termo incluido na consulta. O vetor de termos de um documento $\vec{d_j}$ representado por:

$$\vec{d_j} = (\omega_{1d1}, \omega_{2d2}, ..., \omega_{tdt}) \tag{2.2}$$

onde o $\vec{d_j}$ é o vetor de termos de um documento e o ω_{td} é o peso para cada termo indexado do documento. A partir dos pesos dos termos indexados do documento $\vec{d_j}$, será estabelecido um grau de similaridade para ordenar o documento com base no termos incluídos na consulta $\vec{q}$. O cálculo de similaridade dos termos é dado pela equação2.3:

$$sim(d_j, q) = \frac{\sum_t \omega_{tq} \times \omega_{td}}{\sqrt{\sum_t \omega_{tq}^2} \times \sqrt{\sum_t \omega_{td}^2}} \tag{2.3}$$

onde ω_{tq} representa o peso dos termos na consulta, e ω_{td} representa o peso dos termos nos documentos.

Esse modelo permite aglomerização de documentos para associação por similaridade com base no ordenamento dos pesos. Desta forma admite-se resultados parciais de documentos dentro de um intervalo.

- Modelo Probabilístico: Este modelo atua independemente do fator binário como critério de decisão. O seu critério de consulta depende da probabilidade em se selecionar documentos que satisfaçam a necessidade do usuário na consulta.

  Diferente do modelo vetorial, este modelo não depente do ordenamento dos pesos para selecionar documentos, como foi feito em outros modelos. Este modelo se baseia na probabilidade que documentos possam ser considerados relevantes para uma dada consulta, sendo que estes documentos possuem características. Estas características podem ser frases ou palavras correntes em um documento. Documentos são classificados ordenadamente por grau de relevância em relação a consulta, considerando a probabilidade de similaridade.

  Dada uma situação em que q a *query* de consulta, $\vec{d_j}$ o vetor de termos de um documento, R representa a relevância de documentos e $\bar{R}$ a não relevância de documentos, a similaridade deste documento d_j com a query apresentada é dada por:

$$sim(d_j, q) = \frac{P(R \mid q, \vec{d_j})}{P(\bar{R} \mid q, \vec{d_j})} = \frac{P(R \mid q)}{P(\bar{R} \mid q)} \cdot \frac{P(\vec{d_j} \mid R, q)}{P(\vec{d_j} \mid \bar{R}, q)} \tag{2.4}$$

Cada aplicação para extração de informação pode utilizar um modelo que melhor se adapte ao objetivo da consulta.

Por exemplo, se a consulta deve ser abrangente, busca por conceitos e grupos de conteúdo associados, o modelo probabilístico trará resultados mais refinados. Entretanto, se a mineração exige uma precisão maior na filtragem quanto ao nível de comparação por conteúdo, o modelo booleano realiza melhor essa tarefa.

2.1.2 Aspectos Metodológicos

A mineração de textos por PLN de forma ampla segue as etapas demonstradas na Figura 2.2.

Primeiramente, é feita a seleção da coleção de documentos que será a fonte de dados a ser minerada. Esses documentos podem ser de diversos tipos, tais como PDF, DOC,

Figura 2.2: Etapas do Processo de Mineração de Textos. Fonte: Autor.

documentos web, entre outros. Porém a grande dificudade encontrada na coleta do material é encontrá-los em formato adequado para tratamento. Eles podem estar amazenados em um diretório do HD ou mesmo na Internet. Uma vez localizados, esses arquivos são submetidos ao pré-processamento, que consiste em preparar os dados para armazenar em vetores e posterior indexação dos termos.

A fase de pré-processamento utiliza técnicas de PLN e engloba outros sub-processos: correção ortográfica, remoção de *stopwords* e *stemming*. Esses sub-processos não necessariamente devem ocorrer, porém vale ressaltar que é no pré-processamento que a limpeza dos dados é feita e isto facilitará as análises posteriores.

A correção ortográfica é dependente de domínio de linguaguem, porque cada palavra do texto é toda verificada com base em um dicionário (MONTEIRO, 2006). Um dicionário bastante utilizado para textos em português é o br.inspell que está sob licença da GNU (*General Public Licence*). Quando uma palavra é negada na verificação ortográfica, o corretor ortográfico substitui a palavra por outra sugerida pelo dicionário. Apesar da correção ortográfica ser muito eficiente na eliminação de possíveis termos escritos errados, ainda sim a tarefa de conferir cada palavra do texto considera apenas o aspecto morfológico da palavra, o que do ponto de vista conceitual pode limitar a aceitação de certos termos utilizados. Algumas palavras por exemplo, podem apresentar a mesma grafia, porém com significância diferente, e o corretor ortográfico não resolverá essa questão, porque do ponto de vista sintático, está correta.

Nesse aspecto, as técnicas que são utilizadas na RI seriam úteis na correção do texto, uma delas é a utilização de vocabulário conceitual, ou *thesaurus* (GONZALEZ; LIMA, 2009). Os *thesauri* trabalham com o conceito de ontologias para realizar por mapeamento de termos. Essas ontologias se constituem em uma abstração de dominios do conhecimento por meio de hierarquias e relacionamentos entre os objetos das hierarquias, neste caso de relacionamentos lexicais. A aplicação dos *thesauri* complementaria a correção ortográfica no aspecto semântico das palavras.

Após o texto ter sido processado pelo analisador léxico para correção de erros de grafia das palavras, este documento é submetido a um algoritmo para remoção de *stopwords*. Estas

stopwords são termos sem relevância para o conteúdo a ser extraído, ou de demasiada repetição (MONTEIRO, 2006) palavras como por exemplo preposições, artigos e verbos auxiliares. O algoritmo lê o texto linha por linha, fazendo verificação de existência de stopwords e remonta em um novo vetor, compondo uma representação reduzida do texto. Após essa reconstrução do texto, este é submetido ao processo de *stemming*, que trata de reduzir palavras a seus radicais, os chamados *stems*.

Uma palavra pode ter diversas flexões gramaticais, tais como os sufixos, prefixos, plural, gerúndio, entre outros. Sendo assim, uma palavra de mesmo significado pode se repetir no texto inúmeras vezes, porém com formações gramaticais diferentes (e.g. "casa" e "casinha"). Visando evitar a repetição de uma palavra na indexação de termos relevantes, são aplicados alguns algoritmos para tratamento destas palavras e redução destas a seus radicais (e.g. "casa" e "casinha" ambas possuem o radical "casa", exclui-se o sufixo diminutivo). Existem alguns algoritmos que executam essa tarefa de *stemming*, no caso da língua inglesa, o mais famoso é o Stemmer, e para língua portuguesa existem alguns como o Orengo, bastante divulgado, o PegaStemming, criado em 2003, porém com menor aplicação prática, o Porter, o PortugueseStemmer, entre outros.

O que diferencia esses algoritmos é o número de passos necessários para o tratamento do texto. Porém, dois erros comums ocorridos no processo de *stemming* são o *overstemming* e o *understemming* (CHAVES, 2003). O *overstemming* ocorre quando o há redução da palavra até atingir a raiz do termo, o *stem*. Exemplo, a palavra "maluquice", quando reduzida e sofre o *overstemming*, ela é transformada em "malu", ou invés de "maluc". O understemming ocorre quando o termo reduzido não tem todo seu sufixo ou prefixo. É o caso de por exemplo, do termo "caixote", quando sofre o *understemming* ela é transformada em "caixo" em vez de "caix". O PLN não é um processo simples, e está sujeito a falhas por conta de algoritmos que não refinam o tratamento do texto, que pode afetar os resultados dos termos anexados ao vetor.

Após o tratamento da linguagem de todo o texto, esse documento passará pelo processo de indexação dos termos pós-processados. Cada documento é representado em um vetor $Vd = P1, P2, ...Pn$, onde P_i representa cada uma das n palavras pós-processadas que compõem o vetor, assim é formada matriz da relação dos vetores de documentos e palavras, denominada Bow (bag of words), mostrada na Tabela 2.1.

Tabela 2.1: Representação da matriz do Bag of words.

Documentos (D)/ Termos(T)	T_j	...	T_n
D_i	$f(D_i, T_j)$	...	$f(D_i, T_n)$
...	...	...	...
D_n	$f(D_n, T_j)$	...	$f(D_n, T_n)$

Fonte: (PAES, 2008)

O peso de cada termo presente no conjunto vetorial de termos gerado é calculado para ser inserido no arquivo de indexação, que é a representação reduzida do documento no processo de busca. O cálculo dos pesos referentes a cada palavra tem importância para a expressão do conteúdo relevante na consulta, e leva em consiredação uma série de parâmetros, entre eles, a frequência e localização dos termos no documento na coleção (GONZALEZ; LIMA, 2009). Assumimos que a definição da significância do termo para a sumarização considera os seguintes parâmetros: TF (frequência do termo no documento) e IDF (frequência inversa do termo no documento). A Equação 2.5 representa uma matriz de palavras (Bow) utilizando estes parâmetros citados (PAES, 2008):

$$BoW(D_i, T_j) = TF(D_i, T_j) * IDF(T_j) \tag{2.5}$$

onde, $TF(D_i, T_j)$ é a frequência do termo T_j no documento D_i e $IDF(T_j)$ é o logarítmo do inverso da frequência do termo T_j. O cálculo da frequência do termo T_j no documento D_i é definida por 2.6:

$$TF(D_i, T_j) = \frac{f_{ij}}{\sum_{h=1}^{N} f_{ih}} \tag{2.6}$$

onde f_{ij} é o número de ocorrências do termo T_j no documento D_i e f_{ih} é o número de ocorrências do termo em toda coleção de documentos.

O logarítimo do inverso da frequência do termo $IDF(T_j)$ é representada pela Equação 4.2:

$$IDF(T_j) = \log \frac{N}{N_j} = \log N - \log N_j$$
$$N_j = \sum_{i=1}^{N} a_{ij} \quad a_{ij} = \begin{cases} 1 & \text{if } a_{ij} \neq 0 \\ 0 & \text{if } a_{ij} = 0 \end{cases} \tag{2.7}$$

onde N é o número total de documentos na coleção e N_j é o número total de documentos onde termo T_j ocorre.

Após a normalização das frequências do termos, regras podem ser criadas para filtragem dos termos mais relevantes. São duas as estratégias básicas para a tarefa de redução da dimensionalidade da matriz de palavras: por seleção do termo e por extração do termo com base no peso deste dentro de conjunto de regras lógicas de classificação .

Existem algumas técnicas para criação dos arquivos de indexação, sendo o mais conhecido a arquivo invertido. Este arquivo é constituido de dois componentes básicos: O dicionário de termos e suas localizações. Cada termo presente na coleção de documentos tem um grau de frequência e uma localização específica dentro de cada documento. Ao se indexar a palavra no arquivo invertido, ela conterá listas com a localização daquele termo em cada documento, facilitando assim a busca porque somente conterá palavras de relevância na representação do conteúdo. Considerando uma matriz palavra-documento e sua similaridade e presença na coleção de documentos, o quadro 2.2 mostra um exemplo de um elemento ij, seu grau de similaridade e seu peso representado pelo intervalo [0,1], sendo S_{ij} é o indice de similaridade contido nos documentos D_i e D_j.

Tabela 2.2: Representação da matriz de similaridade de um termo nos documentos.

Documentos	D_i	...	D_j
D_i	Sim_{ii}	...	Sim_{ji}
...	...	...	...
D_j	Sim_{ij}	...	Sim_{jj}

Fonte: Autor.

É possivel, a partir dos termos indexados e seus respectivos pesos, inferir relacionamentos entre palavras, com base em suas frequências, o que possibilita extrair conceitos e relacionamentos hierárquicos entre eles. Na descoberta de conhecimento, existem diversos modos de extração da informação que utilizam técnicas diferentes: baseado em regras, em redes

neurais, em k-NN (nearest neighbours, ou vizinhos mais próximos) que classifica termos por proximidade de similaridade, em métodos probabilisticos (e.g. teorema de Bayes), entre outros. Assim sendo, ao submeter o arquivo de indexação do texto processado a um algoritmo que analise os dados consultados, esse algoritmo deverá comparar o vetor de termos com os parâmetros estabelecidos para classificação.

A abrangência, precisão, acurácia e curva de erro obtidas nos resultados da consulta, estão associadas ao uso das técnicas e algoritmos escolhidos no processamento do texto. A Figura 2.3 retrata o processo geral para sumarização dos textos.

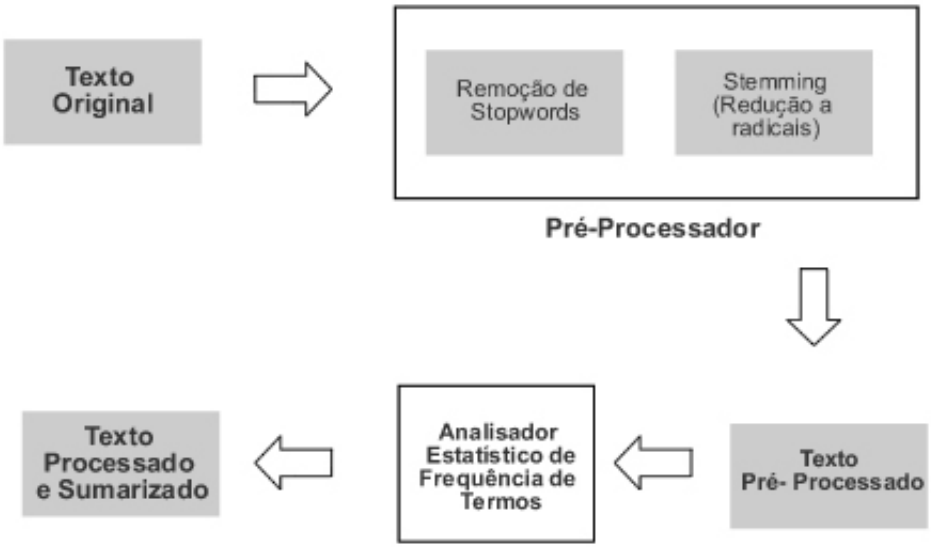


Figura 2.3: Processo de sumarização de textos. Fonte: Autor.

2.1.3 Aplicações de Mineração de Textos

A mineração de textos no Brasil ainda é pouco utilizada, apesar de ser um processo de grande potencial na descoberta de conhecimento em um grande volume de dados não estruturados, como é o caso da maior parte das informações disponibilizadas atualmente, em formato digital. São muitas as áreas que se beneficiariam com a utilização das técnicas de mineração de textos para automatização da extração de informação. A abrangência de áreas para mineração de textos é grande, e engloba negócios, medicina, direito entre outras.

Relatando como exemplos para aplicação de mineração de textos, pode-se citar análise de resultados nas pesquisas de opinião e auxílio a tomada de decisão por vantagem com-

petitiva. No primeiro caso, é muito comum se observar que pesquisas com resposta pré-definidas não possuam alta complexidade na quantificação dos dados e análise posterior com base nas estatísticas geradas. Porém, estas estatísticas revelam apenas quantificação de dados direcionados e pré-definidos com opção, não demonstra outras características mais detalhadas nas opiniões relatadas, como tendências pessoais expostas em respostas subjetivas.

Considerando pesquisas de respostas abertas, a avaliação e análise estatísticas das respostas geradas demandam um esforço maior em relação a pesquisas objetivas, e neste caso, a mineração de textos possibilita a extração de tendências com base nas respostas, se é favorável ou desfavorável em relação a um tema. Essa área de estudo é conhecido por *Sentiment Analysis*, que é uma área da mineração de textos que trata específicamente da extração e análise de conhecimento para tendências de opinião em pesquisas.

No caso de vantagem competitiva, trata-se do apoio à tomada de decisão estratégica. Muitas empresas devem estar atentas às tendências da concorrência. Agregar informações da concorrência, além de tendências do mercado, auxilia no planejamento estratégico, e neste aspecto, a mineração de textos permite rastrear informações, principalmente em ambiente *web*, do setor de mercado que possam ser de importância para a empresa.

2.1.4 Cometários da Mineração de Dados

A análise estatística de dados é um processo que pode ser aplicado a qualquer área do conhecimento onde haja grandes quantidades de dados a serem tratados. Muitas organizações aplicam técnicas de mineração de dados para geração de conhecimento útil a partir de grandes volumes de dados, como ferramenta de apoio à tomada de decisão ou mesmo para detecção de tendências.

De forma resumida, minerar dados significa extrair conhecimento útil de um grande volume de dados. Mineração de dados compõe uma das etapas realizadas no processo de descoberta de conhecimentos implicítos em uma coleção de dados estruturados, sendo este processo também conhecido como KDD (*Knowledge Discovery from Data*). Porém, diferente da mineração de textos, a mineração de dados não utiliza como repositório de dados conjuntos de textos. Neste caso, os dados estão contidos em bancos de dados ou outros repositórios, como a World Wide Web ou DataWare Houses. A definição de mineração de dados, segundo Giudici (2003) é:

> "Mineração de Dados é o processo de seleção, exploração e modelagem de grandes quantidades de dados para descoberta de regularidades ou relações que a principio são desconhecidos, com o objetivo de obter resultados úteis e limpos para o proprietário do banco de dados." (GIUDICI, 2003, p.2)

Segundo Han e Kamber (2003), o termo KDD não pode ser aplicado ao processo de mineração de dados, porque este faz parte de um processo maior, com etapas bem definidas para extração do conhecimento. As etapas consistem em: Limpeza dos Dados, Integração dos Dados, Seleção dos Dados, Transformação dos Dados, Mineração dos Dados, Identificação de Padrões, Representação do Conhecimento. A mineração de dados é inserida como uma parte da transformação de dados em informação relevante. É nesta etapa que são aplicados métodos inteligentes para extração de padrões dos dados, portanto é a etapa de mineração de dados propriamente dita. Pode-se observar a arquitetura geral para o KDD na Figura 2.4.

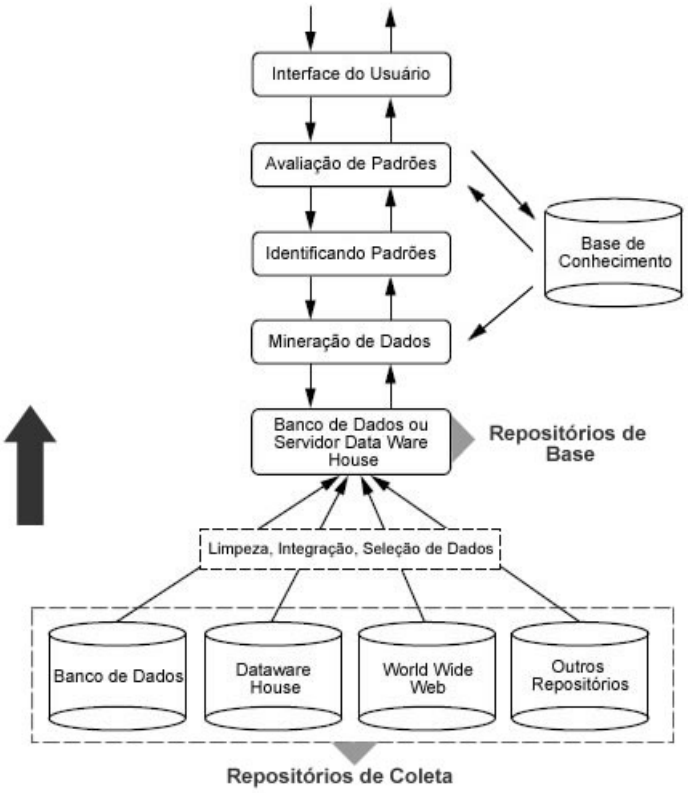


Figura 2.4: Arquitetura do Processo de Descoberta de Conhecimento (KDD).Fonte: (HAN; KAMBER, 2003)

Esta arquitetura demonstra os componentes elementais na extração do conhecimento que se contituem em:

1. **Repositórios de Coleta**: São as fontes de dados de onde serão coletados os dados

para análise, e podem ser bancos de dados, data ware houses, a World Wide Web ou qualquer outro repositório. Em uma primeira instância, esse dados passam pelo tratamento de limpesa, integração de repositórios e seleção dos dados;

2. **Repositórios de Base**: Os dados resultantes do processo de limpeza e seleção dos dados, são novamente armazenados em outro repositório, que também poderá ser um banco de dados ou um dataware house, onde só conterá dado relevante para análise;

3. **Motor de Mineração de Dados**: São funções que executam tarefas para extrair conhecimento dos dados, por meio de classificação e análises dos dados;

4. **Identificação de Padrões**: Consiste basicamente na observação de padrões nas informações encontadas;

5. **Avaliação de Padrões**: A partir da observação dos padrões encontrados, direciona-se a busca por padrões de interesse, para filtragem de informações úteis;

6. **Base de Conhecimento**: É a fonte de domínio de conhecimento que direciona a procura ou avalia os padrões;

7. **Interface com o Usuário**: É a comunicação entre o usuário e o sistema minerador, onde o usuário entra com a query desejada, que guiará a busca na mineração dos dados.

Segundo Giudici (2003), muitas das metodologias aplicadas a mineração de dados estão associadas a dois campos de pesquisa: a aprendizagem de máquina e a estatística computacional. Por um lado, a aprendizagem de máquina permite a generalização de conceitos a partir de dados observáveis. Por outro lado, os métodos estatisticos possibilitam a análise dos dados observados com base no paradigma conceitual. A análise estatística dos dados realizada pela engine mineradora pode ser distinta por três classes (GIUDICI, 2003):

1. **Métodos Descritivos**: visam descrever grupos de dados. As variáveis são analisadas no mesmo nível sem relação de causalidade;

2. **Métodos Preditivos**: visam descrever uma ou mais variáveis em relação as demais. A predição acontece pela regras de classificação dos dados;

3. **Métodos Locais**: Procuram analisar características particulares relacionadas ao subconjunto de interesse do repositório;

O método estatístico escolhido deve ser traduzido para um algoritmo que realize as análises dos resultados. Uma das grandes vantagens na aplicação de mineração de dados é que,

além de analisar dados, ela possibilita a utilização do conhecimento extraído deste processo como ferramenta de apoio à tomada de decisão. No ambiente corporativo, a mineração dos dados pode ser direcionada a diferentes áreas de competências, para extração de informações específicas de interesse.

2.1.5 Mineração de Dados x Mineração de Textos

Relativo ao objetivo primário, tanto a mineração de textos quanto a mineração de dados visam obter conhecimento de grandes columes de informação. A diferença reside em dois aspectos básicos: algumas técnicas aplicadas e os repositórios utilizados. A mineração de dados utiliza repositórios de dados estruturados, oriundos de um banco de dados ou outro repositório em que os dados estejam organizados em tabelas e campos, basicamente como um vetor ordenado. As informações neste caso são explícitas e estruturadas, com campos bem definidos, por isso, deve haver uma procupação na armazenagem, manutenção e integridade desses dados para facilitar a posterior recuperação.

Na mineração de textos, os dados que são extraídos para análise estão implicitos no texto, não há padronização na organização dos dados. São textos de arquivos digitais, arquivos de e-mails, páginas html, entre outros. As informações neste caso são mais fáceis de preservação e o risco de haver falhas na armazenagem ou integridade dos dados é baixa, haja visto que os dados do texto são fixos e não passíveis de alterações. Além da origem da fonte dos dados diferir em ambos os casos de mineração, a forma de tratamento dos dados também é um fator de relevância no processo.

Os dados a serem tratados na mineração de dados, são submetidos à limpeza e seleção para evitar ruidos na filtagem das informações necessárias na análise, enquanto na mineração de textos, o texto passa pelo processamento da linguaguem natural que é a normalização dos termos encontrados, objetivando evitar repetições e reduzindo a escala da matriz de dados gerada na indexação de termos. No primeiro caso o custo em termos de desempenho é menor considerando-se o fato que as informações já estão organizadas, e no segundo caso, o PLN é demanda uma complexidade maior e exige algoritmos mais elaborados para tratamento desses dados. Entretanto, uma vez tratados esses dados, o processo de reconhecimento de padrões se assemelha em ambos os casos, dado que "minerar" informação implica procurar, reconhecer e avaliar padrões em um grande volume de dados.

Há outro ponto de divergêngia que diz respeito a como as informações são consideradas relevantes para a consulta. Na mineração de dados há uma base de conhecimento que é a base para avaliação dos padrões reconhecidos pelo motor minerador. Em se tratando de textos, a base para análise de relevância está na frequência dos termos encontrados,

que é o ponto chave para o ordenamento de documentos em uma consulta. A Tabela 2.3 mostra um comparativo entre os processos:

Tabela 2.3: Comparação entre a Mineração de Dados e a Mineração de Textos.

Características	**Mineração de Dados**	**Mineração de Textos**
Repositórios	Bancos de Dados, Data Ware Houses, World Wide Web, ou qualquer outro repositório estruturado	Arquivos textuais em formatos digitais, arquivos de e-mails, páginas HTML ou outros repositórios textuais.
Tratamento dos Dados	Limpeza dos Dados, Integração dos Repositórios e Seleção dos dados relevantes. Fácil manipulação dos dados. Armazenagem, manutenção e integridade mais complexa.	Transformação do repositório cadeias de strings. Sumissão ao PLN (três etapas). Os dados são fixos e não passíveis de manipulação. Não há preocupação com manipulação ou integridade dos dados.
Dados Pós-Processados	Dados submetidos ao motor minerador. Métodos estatisticos complexos.	Indexação dos termos com base nas frequências encontradas. Matriz de documentos vs. palavras.
Análise dos Padrões	Base de Conhecimento como elemento norteador para avaliação dos padrões encontrados.	A padronagem se baseia nas frequências encontradas. Rankeamento de relevância por frequência de termos.

Fonte: Autor.

Como no trabalho desta dissertação as fontes de dados são originadas de coleções de textos digitais, o foco conceitual da pesquisa é mineração de textos. Assim, o uso de expressões regulares para extração de informação se torna relevante.

2.2 Expressões Regulares

Na mineração de textos formal, o processo de tratamento do texto transforma todo o conteúdo textual de um documento inicialmente em um vetor de sentenças e depois em vetor de palavras, para posteriormente serem aplicadas as técnicas de identificação de padrões de repetição e limpeza dos dados para extração do conhecimento. No modelo desta pesquisa, a técnica utilizada na extração e descoberta de informação útil dos textos

será a criação de expressões regulares.

A diferença que existe entre a aplicação de apenas expressões regulares e a aplicação de algoritmos para o pré-processamento do texto, tais como a correção ortográfica, remoção de *stopwords* e o processo de *steaming*, é que no primeiro caso, a expressão regular identifica padrões textuais com base da representação escrita das palavras (e.g. A palavra "casa" é diferente de "CASA"). No segundo caso, algoritmos de pré-processamento utilizam analisadores gramaticais para remoção de palavras sem relevância (e.g. preposições) e palavras repetidas que serão posteriormente utilizadas na sumarização dos textos.

A mineração de textos envolve mais trabalho no tratamento das informações para analisar o conteúdo, por utilizar artifícios mais complexos que criar um padrão para reconhecer palavras específicas. Porém, é mais abrangente em relação ao que se filtra do conteúdo. No uso de expressões regulares, a forma escrita da palavra faz diferença na seleção dos dados extraídos (e.g. A palavra "casa" é diferente de "CASA"). Algoritmos usuais de mineração de textos, não necessitam de representação exata das palavras na coleta de informação relevante, porque eles se baseiam em análises estatísticas de ocorrências dos termos e não na identificação de semelhança gráfica das palavras.

As expressões regulares apóiam-se na idéia da existência de padrões formais de formatação das palavras, em como estão dispostas as sequências de caracteres, em quais tipos de caracteres estão sendo empregados em certo tipo de sentenças, etc. Sendo assim, o grande problema para minerar textos é que nem todo texto possui formatações expressas bem definidas, pois é justamente nesse ponto que aparece a grande complexidade na mineração dos textos: a ausência de estruturas dos dados bem definidas.

Um texto pode ser construído por centenas de sentenças sem uma padronização qualquer sobre as palavras que o compõe (e.g. Um texto de um documento web pode não ter nenhum padrão de formatação das palavras, pode apenas ser constituído de palavras desordenadas, sem separações, quebra de linha, etc.). Diferentemente de um texto sem padrão de forma, um livro é composto por páginas que contém padrões específicos de formatações, tais como numeração de páginas, citações de autores entre aspas duplas, título todo escrito em maiúculo, entre outros. A aplicação de expressões regulares no primeiro caso será menos eficiente em termos de extração do conteúdo de relevância devido à inexistência de padrões bem definidos de formatações de sentenças e palavras. No caso de um livro, a busca de resultados será mais precisa se comparada a busca em um texto sem padrões de formatação, já que existem alguns padrões na formatação das informações (e.g. Geralmente datas são representadas na forma "xx/xx/xxxx" e isso se constitui em um padrão de formatação).

2.2.1 Conceitos e Fundamentos

Históricamente, a base das expressões regulares deriva da teoria de autômatos finitos e teoria das linguagens formais. O matemático Stephen Cole Kleene, a partir 1950, publicou artigos que descreviam modelos matemáticos que explicavam a lógica recursiva de predicados, que é a fundação da mecânica de expressões regulares, a exemplo do artigo de Erdös e Rényi (1959). Surgiu como resultado destas pesquisas uma notação que foi designada de conjuntos regulares, ou, álgebra de Kleene. Posteriormente, Ken Thompson utilizou a notação criada por Kleene, para reconhecer padrões em arquivos de textos, sendo primeiramente utilizado em um editor de texto conhecido como QED. Depois desta experiência surgiram outras aplicações para embutir esses conceitos de padrões reconhecíveis para outros editores de textos e bibliotecas para diversas linguagens de programação.

As expressões regulares, também conhecidas como Regex, são notações de padrões que permitem o reconhecimento de sequências de caracteres dentro de um texto. Segundo Good (2005), as expressões regulares são como expressões matemáticas que operam em sequência de caracteres ou *strings*, em vez de números. Estas notações de padrões podem ser compostas apenas por caracteres simples ou podem agregar meta-caracteres. Os caracteres simples são símbolos formais que caracterizam de forma literal as letras do alfabeto e numéricos. Os meta-caracteres se caracterizam por serem caracteres simples, porém com um significado especial diferente do sentido literal (GOOD, 2005).

Diversas linguagens de programação utilizam expressões regulares para descobrimento de padrões em textos, como o C# , VB script, Java, Perl e outras. Essas linguagens trazem esse recurso como uma classe embutida na linguagem própria, que deve ser importada para o código fonte, onde será montada a notação para a busca do padrão. Isto nos permite ter mais opções ao escolher uma forma de desenvolver um buscador de padrão que melhor se adapte a necessidade. Por exemplo, o C# é uma linguagem menos complexa se comparada a Java, então no desenvolvimento de um modelo que gerencie a criação de expressões regulares, a produtividade em termos de esforço e tempo é maior em C# que em Java.

Uma notação pode identificar palavras como unidades reconhecíveis, ou ainda, agrupamentos de palavras reconhecíveis por padrão. As sintaxes das expressões regulares permitem que se criem padrões do mais abrangente ao mais específico por meio de utilização de meta-caracteres, que podem quantificar, qualificar, restringir e agrupar caracteres.

Por exemplo, em uma situação onde se deseja buscar por um padrão de palavras onde todos os caracteres estão em caixa alta (e.g. “BRAGA”), a expressão regular deve ser criada seguindo o padrão de formatação desejado (neste exemplo, todas as letras em caixa

alta). Desta forma, ao buscar resultados desta expressão, somente serão buscadas palavras que apresentam forma escrita, conforme padrão descrito na expressão regular. A notação neste caso seria descrita, de maneira geral, desta forma:

[A-Z]+

A expressão [A-Z]+ indica que palavras que contenham sequências de caracteres somente em caixa alta serão selecionados como resultado de busca. O "A-Z" representa um intervalo de letras que vai de A a Z, sendo todas maiúsculas. O símbolo + indica um quantificador que concatena os caracteres seguintes na montagem da palavra encontrada. Na próxima seção serão detalhadas as premissas de expressões regulares. Pode-se observar que a forma como foi descrita a notação que fará a filtragem a depender do objetivo empreendido, se apresenta bem diferente. A escrita de uma notação não é trivial, a sintaxe que constitui a expressão regular pode ser algo bem simplificado, porém pode se tornar complexo com o maior refinamento do padrão.

Existe atualmente uma série de ferramentas disponíveis para testes de expressões regulares, o que facilita na criação dos padrões. Geralmente, estas ferramentas estão disponíveis on-line, tais como o RegexMagic [1] e o Txt2Re [2]. Porém é mais difícil encontrar ferramentas que criem automaticamente uma notação, portanto o refinamento de um padrão está associado com o nível de conhecimento que se tem das sintaxes de expressões regulares.

2.2.2 Descoberta de padrões e Extração e Filtragem dos dados: Modelos Matemáticos

A dinâmica de reconhecimento de padrão das expressões regulares se baseia em testes contínuos de strings. De forma resumida, a mecânica da busca por padrões se traduz da seguinte forma: o documento é interpretado como um vetor de caracteres, onde uma expressão regular representará a condição de busca das cadeias de caracteres. O comportamento desta mecânica é fundamentada na álgebra de Kleene (1956) e na teoria dos autômatos finitos.

Segundo (COHEN, 1996), a álgebra de Kleene é definida como uma classe de estruturas algébricas que são utilizadas nas mais diversas áreas da ciência da computação: programas lógicos, design e análise aritmética. Essa álgebra, de propriedades idempotentes e comutativas, é definida por

[1] http://www.regexmagic.com/, data de acesso 20/08/2010
[2] http://www.txt2re.com/index-csharp.php3, data de acesso 20/08/2010

$$(k, +, ., *, 0, 1)$$

onde k representa o conjunto de caracteres guardados ao longo de alfabetos finitos e os simbolos $+, ., 0$ e 1 representam os operadores que atuam sobre essas cadeias, que se referem respectivamente à escolha, composição, falhar ou pular elemento do alfabeto. A algebra de Kleene é um conjunto A juntamente com duas operações binárias $+: k \times k \to k$ e $.: k \times k \to k$ e uma função $* : k \to k$, escrito como a + b, ab e *, respectivamente, que deve satisfazer os seguintes axiomas:

- Associatividade de + e .: $a + (b + c) = (a + b) + c$ e $a(bc) = (ab)c$ para todo a, b, c em k;
- Comutatividade de +: $a + b = b + a$ para todo a, b em k;
- Elementos de Identidade para + e .: Existe um elemento 0 em k, tal que para todo a em $k : a + 0 = 0 + a = a$, assim como existe um elemento 1 em k, tal que para todo a em $A : a.1 = 1.a = a$.

Os axiomas apresentados tratam das propriedades de adição e multiplicação na álgebra de Kleene, e resume o k da expressão de Kleene como uma estrutura algébrica idempotente, não assumindo valores negativos. Além dos operadores + e ., há a propriedade do operador * que define o comportamento transitivo de fechamento de relações binárias em um conjunto de strings.

- $1 + a(a*) \leq a*$ para todo a em k;
- $1 + (a*)a \leq a*$ para todo a em k;
- se a e x estão em A tal que $ax \leq x$, então $a * x \leq x$;
- se a e x estão em A tal que $xa \leq x$, então $x(a*) \leq x$.

De forma simplificada, em expressões regulares, o operador $*$ funciona como iteração (e.g. a* = 1 + a + aa + aaa...), e o + como união (e.g. a+b = ab, abab, ...) e o . como sequenciamento (e.g. a.b = ab, aaab, aaaaaaaaab, ...). sobre a cadeia de caracteres.

Com base nesta teoria, as expressões regulares utilizam nas suas operações alguns aspectos utilizados na álgebra de Kleene, como os operadores *, . e + na leitura e identificação de sequências de caracteres equivalentes. Dentro da lógica booleana da álgebra de Kleene, o fecho de Kleene (Kleene star) é a base matemática para a identificação desta equivalência.

Kleene utilizou a a teoria da álgebra booleana do seu teorema para caracterizar certos autômatos, criando assim o fecho de Kleene. O fecho de Kleene é uma operação unária realizada sobre uma cadeia de caracteres ou símbolos. Considere k um conjunto de strings. $k*$ é o menor sub-conjunto de k, e contém 0 ou mais strings e se encerra pela operação de concatenação entre elas, conforme mostrado no modelo a seguir:

{“a”, “c”}* = {ε, “a”, “c”, “abab”, “abc”, ...}

onde ε é o conjunto vazio.

Na construção da expressão, qualquer símbolo pode representar uma expressão regular na álgebra de Kleene. Por exemplo, uma expressão regular definida por “a.*”, quando aplicada ao longo de um alfabeto, testa todas as cadeiras de caracteres encontradas nele, porém, só é guardado como resultado, as cadeias de caracteres que começam com a letra “a”. A medida que encontra um sequência de caracteres que começa com “a”, a cadeira de caracteres resultante da busca é contruída a partir da concatenação dos outros elementos do alfabeto (e.g. A palavra “anão” dentro de um texto, definido como o alfabeto, é uma sequência de caracteres que começa com a letra “a”, e como atende ao que foi definido na expressão algébrica, ele além de guardar como resultado a própria letra “a”, concatena todo o resto da palavras “não”).

Conforme citado, um dos fundamentos para a criação de expressões regulares é a teoria do autômatos finitos. Um autômato é um modelo matemático que define uma máquina de estados finitos. Essa máquina é um sistema de entradas e saídas discretas (MENESES, 2000) determinadas estados, os quais são definidos pela condição de transição que compõe a expressão matemática da máquina, conforme mostrado na Figura 2.5.

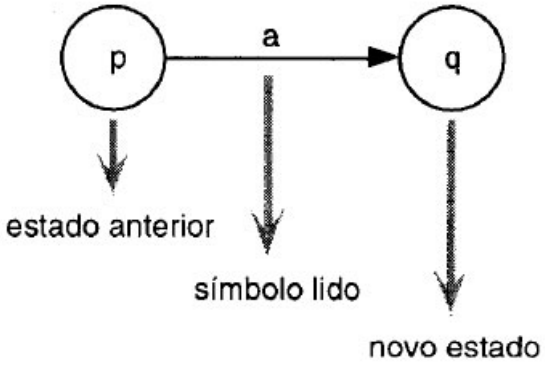


Figura 2.5: Ilustração da transição dos estados em autômato. Fonte: (MENESES, 2000, p.34)

Segundo (MENESES, 2000), Analisadores Léxicos e Processadores de textos, que são comumente utilzados na etapa de pré-processamento de textos na mineração de textos, são

exemplos de sistemas de estados finitos, onde cada estado memoriza apenas a estrutura do prefixo da palavra em análise. Por exemplo, o processo de *stemming* que acontece na etapa de pré-proocessamento de textos, onde as palavras são reduzidas a radicais e armazenadas para classificação posterior de outras palavras semelhantes, que possuam o mesmo radical da palavra, como em "casa" e "casinha", ambas possuem mesmo radical.

Por definição, autômatos finitos são estruturas matemáticas constituídas de 3 partes: Um conjunto de estados, um alfabeto e um conjunto de transições (VIEIRA, 2000).Os estados são compostos pelo estado inicial, estados intermediários, e os estados finais. O autômato finito é determinado pela quintupla $(E, \Sigma, \sigma, i, F)$, onde:

- E é um conjunto finito de estados;
- Σ é um alfabeto;
- σ é a função de transição, onde $\sigma : E \times \Sigma \rightarrow E$;
- i é o estado inicial dentro de E;
- F - é o subconjunto de E estados finais.

A Figura 2.6 representa o ciclo de execução do autômato segundo regras de expressões regulares. Para Meneses (2000), uma linguagem é regular, se e somente se, é possível construir um autômato finito que reconheça a linguagem.

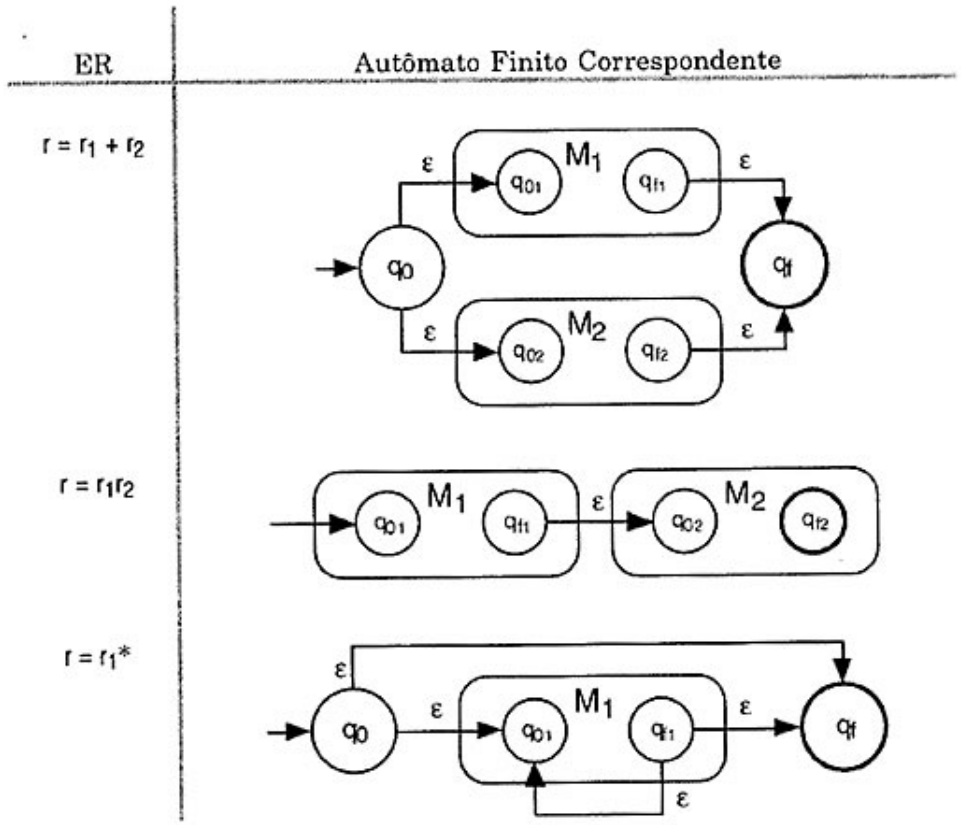


Figura 2.6: Representação de autômatos finitos correspondentes às expressões regulares. Fonte: (MENESES, 2000, p.52)

Nas expressões regulares, cada caractere de texto corresponde a um caractere do alfabeto a ser lido pela regra da função de transição, que determinará quais sequências de strings serão apresentadas como resultados. A função de transição é determinada pela notação descrita na expressão regular, os estados são compostos pelos caracteres coletados e incluídos nas *strings* finais resultantes.

2.2.3 Símbolos e Notações de Expressões Regulares

As notações possuem características que delimitam e refinam os padrões de busca. Segundo Good (2005), dentre essas características estão:

- Qualificadores (ou qualifiers): restringem o número de vezes que uma expressão precedente vai aparecer nos resultados de busca. Os símbolos que caracterizam qualificadores são ?,+ e *.
  - ? Significa um ou zero resultados da expressão encontrados;

  - + Significa um ou mais resultados da expressão encontrados;
  - * Significa zero ou mais resultados da expressão encontrados.

  O uso de qualificadores tendem a maximizar os resultados encontrados, o que pode se inferir que dentre os resultados encontrados poderão ser achados outros resultados a mais que não estejam dentro do esperado pela busca. Por exemplo, a expressão dada por [abc]+ trará como resultados de busca todas as ocorrências a sequência identificada por abc, não importa a posição em que elas ocorram nas strings, no mínimo a sequência abc deve aparecer uma vez. No exemplo abc[d]?, os resultados obtidos poderiam conter palavras como abc, abcd, abcde, isto porque no mínimo a sequência abc deve aparecer em qualquer resultado, entretanto, a letra d, deve aparecer nenhuma ou pelo menos uma vez. No exemplo [abc]*, as sequências passíveis de serem encontradas seriam: a, aaaa, abb, abccc, accc, cccc, etc. Isto porque no mínimo a sequência encontrada deve conter nenhum, um ou mais de um elemento.

- Escalas (ou ranges): assim como os qualificadores, ele também define quantas vezes uma determinada expressão pode ocorrer, com a diferença que neste caso, a quantidade pode ser especificada previamente. Ela é determinada pelos símbolos { }.

  - { n} O número máximo de ocorrências para determinada expressão é de n vezes.
  - { n, } O número mínimo de ocorrências para uma expressão é de n vezes.
  - { n,m} O número mínimo de ocorrências de uma expressão é de n vezes, e não podem ocorrer mais de m vezes.

  Exemplo: Dada a sequência [abc]{2}, serão filtrados resultados onde ocorram extamanente duas vezes a sequência abc dentro de uma string. Neste exemplo, as sequências encontradas só poderiam ser: abc e abcabc. Entretanto, se a expressão fosse [abc]{2,}, a amplitude dos resultados aumentariam: abcabc, abcabc, abcabcabc, etc. Isto porque no mínimo, a sequência deve conter duas ocorrências do termo anterior abc.

- Âncoras de linhas (ou line anchor): Essas âncoras pegam o início ou fim de uma string, não necessariamente um caractere. Essas âncoras são representadas pelos símbolos ^ e $.

  - ^ Representa o início de uma string
  - $ Representa o fim de uma string.

  Exemplo: a busca pela expressão ^gato no texto, encontrará ocorrências onde o início do texto for gato. Frases como: gato feliz, gato pula o muro, etc.

- Escapamento (ou escapes): Escapar um caractere significa dar-lhe atribuição literal. Alguns caracteres necessitam de escapamento para obterem significância literal, caso

contrário, seu significado será interpretado como meta-caracter. Caracteres como os qualificadores, por exemplo, se for precedido por um escape, ele não será interpretado como qualificador de caracter, e sim, como caractere literal. O escape é definido por símbolo \ precedendo a expressão.

Exemplo: [abc]+ não é a mesma coisa que [abc]\+ porque no primeiro caso, o simbolo do + quantifica a busca pela sequência, enquanto no segundo caso, ele busca por uma sequência abc seguida do simbol + literalmente. Na expressão [abc]+, as sequências encontradas seriam: abc, abcabc, abcabcabc, etc., enquanto que na expressão [abc]\+, o resultado só poderia ser abc+.

- O simbolo OR: Determina uma condição para dois resultados encontrados, por exemplo, encontrar a expressão à esquerda da barra senão encontre a expressão à direita da barra. O simbolo que define essa condição é —.

  Exemplo : A expressão definida como gr(a—e)y trará como resultados de busca todas a strings que tiverem a sequência gray ou grey em sua composição, porque a condição está nos caracteres “a” ou “e” dentro desta string definida.

- Outras características das expressões: Os simbolos [ ] definem classes de caracteres que devem encontrar resultados que contenham somente os caracteres indicados dentro destes símbolos. Se uma expressão é definida por [az], a busca só trará aquelas strings que contenham somente essas letras. Porém se um caracter delimitador de intervalo por utilizado dentro desta expressão, representado pelo símbolo “-” a busca passa a ser mais abrangente, e pegará todo um intervalo de caracteres. A mesma expressão citada acima, utilizando-se do delimitador de intervalo ficará escrito da seguinte forma [a-z]. Neste caso, a busca filtrará todas as strings onde existas os caracteres especificados pelo intervalo dados: a até z.

O ponto “.” é um simbolo que combina todos os caracteres de uma sequência descrita na notação de uma classe de caracteres. Esse meta-caractere pode ser combinado a outros meta-caracteres para especificar ou ampliar resultados de busca. Espaçamento entre *strings* é outro aspecto que deve ser considerado em uma notação. A sintaxe “\s” representa os espaços em branco, e, se não for incluída na expressão, os resultados não trarão palavras que tenha espaços em sua composição.

Outro aspecto dos meta-caracteres é a utilização de mesmo símbolo com mais de uma significância. É o caso do simbolo ^ já descrito anteriormente como âncora de linha. Este caracter também pode significar negação de outros caracteres, se inserido dentro das classes de caracteres indicadas por [e].

A expressão ^gato e a expressão [^gato] têm significados diferentes. O primeiro caso tem o caracter ^ como indicativo de âncora inicial de linha que contenham a palavra gato.

O segundo caso representa uma negação dos caracteres que estiverem dentro da classe de caracteres, ou seja, não encontrará palavras que contenham na sequência da string os caracteres g, a, t ou o. Segundo Friedl (1997), negar uma classe de caracteres não significa combinar caracteres a menos que tenha o x, senão combinar caracteres que não sejam o caractere indicado na sequência negada. Por exemplo, a expressão dada por [ˆabc]. As sequências encontradas não conterão estes caracteres indicados entre os cochetes. Palavras como "casa" e "bola" não poderiam estar nos resultados, apenas palavras que não tenham as letras "a", "b" e "c".

Expressões também podem agrupar notações utilizando-se dos simbolos (e). Estes caracteres delimitam grupos de notações. Um exemplo já comentado foi a questão do simbolo — para delimitar uma condição de busca por duas expressões. No caso dado, gr(a—e)y há uma delimitação do escopo da condição indicada pelos parêntesis.

Ainda há de se ressaltar outras característica de importância, como os modificadores modo, que alteram a forma como o mecanismo de busca pela expressão regular é interpretado (STUBBLEBINE, 2007). Os principais são os modos multilinha, linha simples e não sensibilidade a caracteres maiúsculos ou minúsculos. O modo multilinha, altera o comportamento do simbolos de âncoras para se ligar ao caractere mais próximo de novas linhas dentro da string de entrada. O modo linha simples, considera as âncoras e caracteres definidos no padrão expresso. O modo ignorar case dos caracteres, desconsidera as formas dos caracteres (i.e. maiúsculas ou minúsculas).

Um último ponto a ser considerado é em relação ao POSIX, que trata de um tipo especial de classes de caracteres. Ela busca pelos mesmos resultados que os caracteres e meta-caracteres comuns fariam, porém de uma forma mais objetiva. Considerando como exemplo a expressão [a-z0-9] tem o mesmo significado que [a-z[:digit:]] e a busca de ambas trarão resultados iguais. A Tabela 2.4 apresenta as características gerais da sintaxe empregada em expressões regulares:

Tabela 2.4: Tabelas de Notações de Expressões Regulares.

Características	**Sintaxe**	**Exemplo**
Âncoras		
	`ˆ começo de linha`	Encontra toda linha que começa com ˆgato
	`\b limite da palavra`	.pega c em abc
	`\A começo de string`	`\A. pega a em abc`
	`\B não é limite da palavra`	`\B.\B pega b em abc`

Continuação na próxima página...

Tabela 2.4 – Continuação

Características	**Sintaxe**	**Exemplo**
	`$ fim de linha`	`.$ pega f em abcndef.`
	`\Z fim de string`	`.\Z pega f em abcndef`
Classes de Caracteres		
	`\s espaço em branco`	`[a\sb] pega qualquer sequência que tenha a,b ou espaço`
	`\d dígito`	qualquer dígito (0-9)
	`\S não é espaço em branco`	`.\S pega abc em abc def`
	`\D não é dígito`	`.\D pega abc em abc2009`
Quantificadores		
	* 0 ou mais	A expressão b[ip]* pega bi, bip, bp, biiip, etc.
	+ 1 ou mais	A expressão [a-c]+ pega qualquer sequência que tenha as letras a, b ou c. Exemplo: aaa, abb, abc, acc, etc.
	`?  0 ou 1`	A expresão fala[r]? pega os resultados fala ou falar
	`{n} exatamente` *n* `vezes`	A expressão [abc]{2} pega a sequência abcabc. Todas as sequências que tenham exatamente duas ocorrências do termo anterior.
	`{n, }` *n* `ou mais vezes`	A expressão [abc]{2,} pega sequências como abcabc, abcabcabc, etc.. Todas as sequências que tenham no mínimo duas ocorrências do termo anterior.
	`{n,m} no mínimo` *n*`, no máximo` *m* `vezes`	A expressão [abc]{2,3} pega as sequências abcabc ou abcabcabc. Todas as sequências que tenham no mínimo duas e no máximo três ocorrências do termo anterior

Continuação na próxima página...

Tabela 2.4 – Continuação

Características	**Sintaxe**	**Exemplo**
Escalas ou intervalos		
	. Qualquer caracter menos quebra de linhas \n	A expressão A.O pega sequências como ANAO, AnjO, ALtO, etc. Entre o A e o O ele aceita qualquer caractere.
	`(a|b) a ou b`	A expressão (gato—lebre) só trará ou gato ou lebre nos resultados.
	`(...) agrupamentos`	([abc]+)[def]+ trará nos resultados abcabcdef, abcdefdef, etc.
	`[abc] intervalo de a ou b ou c`	Só são aceitas sequências que tenham as letras a, b ou c, como aaa, abc,abb, etc.
	`[^abc] que não seja a ou b ou c`	Só são aceitas sequências que não tenham as letras a, b e c.
	`[a-z] intervalo de a a z`	Serão aceitas sequências que tenham qualquer caracter entre a e z e que os caracteres sejam todos minúsculos. Exemplo: [a-z]+ acha resultados como casa.
	`[A-Z] maiusculas de A a Z`	Serão aceitas sequências que tenham qualquer caracter entre A e Z e que os caracteres sejam todos maiúsculos. Exemplo: [A-Z]+ acha resultados como CASA.
	`[0-7] numéricos de 0 a 7`	Serão aceitas sequências que tenham qualquer caracter desde que seja numérico. Exemplo: [0-9]+ acha resultados como 2010.
Caracteres especiais		

Continuação na próxima página...

Tabela 2.4 – Continuação

Características	**Sintaxe**	**Exemplo**
	\ caracter de escape	Exemplo: [a-z]+ é diferente de [a-z]/+. O primeiro acha palavras como casa, gato. o segundo acha palavras minúsculas, porém que sejam seguidas do sinal de +, como casa+.
	\n nova linha	Exemplo: [a-z\n]+ aceita linhas inteiras de palavras deste que sejam todas as letras minúsculas.
Meta Caracteres (devem ser escapados para terem significado literal)		
	^ [ ] ( ) { } . + ? — < > $	Os metacaracteres para serem considerados literalmente deve-se escapa-los. Exemplo: (gato) é diferente de *gato*, porque no primeiro caso os parêntesis representam agrupamento de caracteres e no segundo caso ele representa o próprio caracterer do parêntesis.
Asserções (condições de procura na expressão)		
	?= olhar a frente	Por exemplo, Michael (?=Jackson) só encontrará Michael se for seguido da palavra Jackson.

Continuação na próxima página...

Tabela 2.4 – Continuação

Características	**Sintaxe**	**Exemplo**
	`?! negativa de olhar a frente`	Por exemplo, Michael `(?=Jackson) só encontrará Michael se não for seguido da palavra Jackson, pode encontrar outras combinações de Michael com outras palavras.`
	`?≤ olhar atrás`	Exemplo: `(?≤) Jackson só encontrará Jackson se for antecedido da palavra Michael.`
	`?!≤ negativa de olhar atrás`	Exemplo: `(?≤) Jackson só encontrará Jackson se não for antecedido da palavra Michael.`

Fonte: Autor.

2.2.4 Comparação dos métodos extrativos: Expressões Regulares x Algoritmos de Pré-processamento

Existem diversas formas para tratamento de textos na busca de extração de conhecimento. Técnicas como o PLN (processamento de linguaguem natural), que utiliza e produz resultados independente das características linguísticas do texto, e as técnicas de EI (extração de informações), que é dependente de um domínio de conhecimento e outros modelos. Cada uma das técnicas apresenta um nível de dificuldade de aplicação.

Na mineração de texto por processamento de linguagem natural, as etapas incluem pré-tratamento do texto, extração do conhecimento e análise, e sumarização dos dados. O pré-processamento compreende a aplicação de algoritmos sobre o texto para reduzir ao máximo dados não relevantes, e envolve as fases de correção ortográfica, remoção de *stopwords* e *stemming*.

Esta etapa de pré-processamento do texto é a que demanda maior tempo de execução e desempenho de um processador, porque diferentemente de uma base de dados onde

as informações estão em um vetor ordenado, um texto digital se constitui de um vetor de strings não estruturado. A falta de estrutura bem definida nos textos implica na necessidade de utilização de algoritmos de pré-processamento de textos para limpeza dos dados, para posteriormente extrair conhecimento relevante dos documentos.

Da forma como os dados são gerados por meio das técnicas de processamento natural, as informações extraídas serão sumarizadas com base na frequência de repetição, que representa o grau de relevância daquelas palavras no texto. Com isto pode-se observar que o nível de especificidade da informação gerada nos sumários é mais abrangente se comparado ao resultado gerado pela busca de padrões indicados por expressões regulares, isto porque as expressões procuram por padrões dentro do texto, não necessariamente por informações de conteúdo relevante no texto. Isto não significa dizer que as expressões regulares não sirvam para coleta de dados em textos, apenas depende do propósito da extração da informação.

Quando a busca trata da descoberta de conteúdos em uma coleção de textos, independentemente do quão específica possa ser a informação sumarizada, o PLN é recomendável a esta situação, porque independe de padrões, uma vez que se procura a relevância de informação. Porém, se há a necessidade de descoberta específica de um dado, onde há a existência de padrões nas sentenças do texto, este é o caso de se utlizar padrões na extração da informação, e que se aplica a esta pesquisa. O custo em termos de desempenho comparando-se mineração de texto por meio de PLN e expressões regulares é maior quando se trata de PLN, porque o percursso para se chegar ao conteúdo relevante é maior que a simples busca por padrões.

Porém, ainda que sendo mais simples em termos de etapas processuais de tratamento do texto, o que dificulta o uso de expressões regulares é o conhecimento das regras e sintaxes usadas nas notações. A eficiência na busca dos dados no texto está associada diretamente ao refinamento do padrão definido. O domínio das regras pode ser um fator de impacto, no que se refere a facilidade ou não na criação de expressões regulares.

Por outro lado, a capacidade de se criar e buscar por qualquer padrão reconhecível torna esse método de mineração de textos mais flexível que por meio de PLN, cujo processo é todo automarizado e fixado nos algoritmos já existentes. Para utilização nesta pesquisa foram utilizadas expressões regulares como método de mineração de textos, devido ao propósito específico de se encontrar dados específicos dos documentos analisados. Nos Capítulos 4 e 5 esse processo de mineração de textos por uso de expressões regulares será melhor detalhado.

Como a utilização da mineração do textos surgiu da necessidade da construção de redes de colaboração científica, a partir de dados implicitos em uma coleção de documentos, no

capítulo a seguir, serão apresentados alguns conceitos de redes sociais e complexas para fundamentar teóricamente o produto gerado por esta modelagem.

Capítulo Três

Redes Sociais e Complexas

As redes complexas estão presentes na natureza e podem ser encontradas em vários sistemas, desde sistemas de dimensões microscópicas até os mais complexos agrupamentos sociais. Redes sociais, redes de informação, redes de computadores, rede biológicas são exemplos de tipos de redes observáveis (NEWMAN, 2003). Uma rede é um conjunto de items, os quais são chamados de vértices, algumas vezes e nodos ou nós, com conexões entre eles chamados de arestas, (NEWMAN, 2003). Redes Sociais e Complexas basicamente são redes. Redes Complexas apresentam características topológicas específicas que caracterizam sua conectividade e alta influência na dinâmica de processos executados em uma rede (COSTA, 2007). As Redes Sociais são representadas por estruturas sociais (i.e. relacionamentos sociais) onde os vértices destas redes são compostas por pessoas ou grupos.

Em termos matemáticos, uma rede é representada por um grafo, e se constitui de um par de conjuntos $G = \{V, E\}$, onde V é um conjunto de N vértices $V_1, V_2, ..., V_N$ e E é um conjunto de arestas (ou links ou linhas) que conectam os elementos de P (ALBERT; BARABáSI, 2002). Uma outra definição é a de Gross e Yellen (2004), que diz que qualquer objeto matemático envolvendo pontos e conexões entre eles pode ser chamado de grafo. A Figura 3.1 configura um grafo com um conjunto de vértices $V = \{1, 2, 3, 4, 5\}$ e um conjunto de arestas $E = \{\{1, 2\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{4, 5\}\}$.

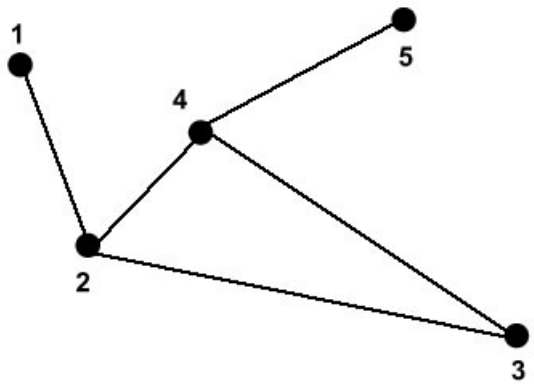
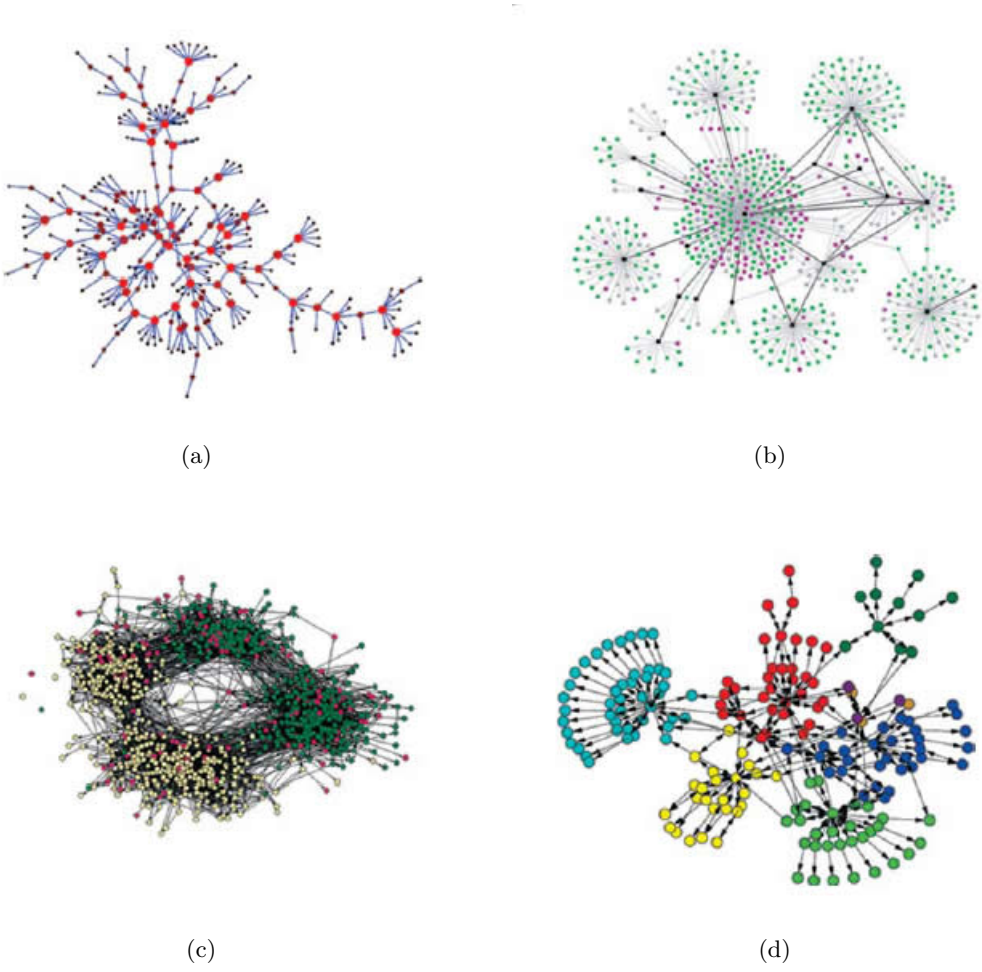


Figura 3.1: Ilustração de um grafo com $N = 5$ vértices e $E = 5$. Fonte: Autor.

De forma sucinta uma rede se constitui em um conjunto de elementos, que podem estar

conectados ou não entre si, e a sua topologia possui propriedades diversas que caracterizam a sua estrutura. O interesse sobre o estudo das redes está na constituição e comportamento de suas estruturas, uma vez que o entendimento desses aspectos auxiliam na compreensão da dinâmica de sistemas complexos.

A internet é o exemplo mais frequentemente lembrado quando se fala em conexões e redes, isto porque, ela se constitui de uma teia de computadores interconectados. Neste tipo de rede, informações são compartilhadas a partir das ligações entre os pontos que constituem esta rede. A Figura 3.2 mostra algumas redes encontradas na natureza.



(a) (b)

(c) (d)

Figura 3.2: a) Rede de contatos sexuais entre indivíduos (POTTERAT J. J., 2002); b) Rede de contágios entre pessoas (KREBS, 2007); c) Rede dos amigos em uma escola dos Estados Unidos (MOODY, 2001); d) Documentos em um sítio da Web e ligações entre eles (NEWMAN; GIRVAN, 2004).

Como se pode notar na Figura 3.2, as redes não são necessáriamente imutáveis e nem sempre são regulares ou uniformes. Cada rede tem sua dinâmica e características próprias, em alguns casos, pode ter uma relação de causa-efeito que emerge como consequência do seu comportamento. Por exemplo, um ponto central de energia, como um poste, que tiver uma queda, todos os pontos que estiverem conectados a eles poderão ter sua energia

interrompida.

Na internet, se houver uma parada no servidor, todos os computadores conectados a ele, sofrerão a perda da conexão. Observando-se os elementos de uma rede e a relação entre eles, pode-se realizar predições com base no comportamento observado. Um exemplo de aplicação da análise da dinâmica das redes é sua utilização ferramenta de apoio a decisões estratégicas, por exemplo, das organizações. Uma empresa não atua sozinha, ela tem ligações com fornecedores, com distribuidores, consumidores, e estes por sua vez têm suas conexões.

3.1 Características Topológicas

Para compreensão do comportamento de uma rede, é interessante analisar suas propriedades a partir dos dados estatísticos significativos. Estes dados podem revelar características do seu processo evolutivo de modo que o entendimento das leis que governam esse sistema seja facilitada. Algumas das características reconhecidas em uma rede são (MENDES, 2006; NEWMAN, 2003; BOCCALETTI, 2006; COSTA, 2006; ALBERT; BARABáSI, 2002; METZ, 2007; RODRIGUES, 2007):

- Os **Vértices** (ou nós), são as unidades que compõem a rede;
- As **Arestas**, são as relações entre os vértices, e podem ou não ser direcionadas ou mesmo ter pesos diferentes. Em uma rede neural por exemplo, as ligações sinápticas têm pesos diferentes que influenciam no comportamento da rede (e.g.: Figura 3.3).

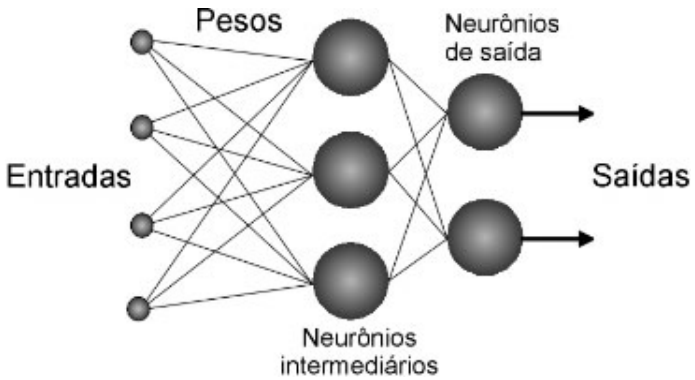


Figura 3.3: Arquitetura de uma rede neural. Fonte: Antiqueira (2007).

- A **Conectividade** ou grau, é o número de arestas incidentes em um vértice e é definida pela expressão:

$$k_i = \sum_{j \in N} a_{ij} \tag{3.1}$$

onde k_i é o grau do vértice i e a_{ij} representa as arestas ligadas a este vértice. No caso de um grafo direcionado, o grau de um vértice, segundo Boccaletti (2006), é a soma dos seus links de entrada e saída. A Figura 3.4 mostra dois exemplos de rede. A rede da esquerda possui um vértice, indicado pela letra A, cujo grau é 4. A rede da direita possui um vértice, indicado pela letra B, cujo grau é 2.

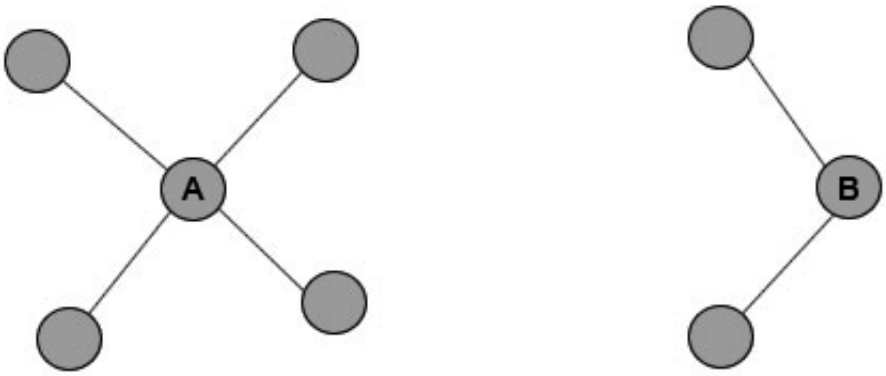


Figura 3.4: Modelos de Redes com graus 5 e 2. Fonte: A AUTORA, 2010.

- A **distribuição de conectividade**, refere-se a forma como estão distribuidas as arestas pelos vértices e a probabilidade de um vértice receber uma nova ligação, e pode ser quantificada por uma função de distribuição cumulativa, dada por:

$$P_k = \sum_{k'=k}^{\infty} p_{k'} \tag{3.2}$$

onde P_k é a função cumulativa de distribuição de probabilidades e p_k é a fração de vértices da rede com grau k. Em redes direcionadas, o cáculo da probabilidade deve

considerar duas variáveis: a fração de vértices de entrada e de saída. A distribuição de graus em uma rede aleatória segue distribuição de Poisson;

- O **caminho geodésico**, ou caminho mínimo, é a menor distância entre dois vértices na rede. A expressão que representa o valor médio desta propriedade é dada por:

$$\ell = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{3.3}$$

onde ℓ é a distância geodésica média, N o número de vértices da rede e d_{ij} a distância entre dois vértices quaisquer.

Entretanto, o problema desta definição matemática é que ela diverge se há pares de vértices desconectados na rede. Para contornar este problema, não se inclui na soma os pares de vértices desconectados (COSTA, 2006). Uma proposta alternativa para esta divergência, proposta por Latora e Marchiori (2001), é considerar a eficiência de um par de vértices, que é dada pelo inverso do caminho mínimo entre estes vértices. A eficiência média, ou eficiência global, da rede é dada pela Equação 3.4

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{d_{ij}} \tag{3.4}$$

onde E é a eficiência global média, N o número de vértices da rede e d_{ij} a distância entre dois vértices quaisquer.

- O **Diâmetro**, é a maior distância entre dois vértices na redes;
- A **transitividade**, ou *clustering*, é a presença de ciclos de três vértices fechados em uma rede. A medida que define a quantidade de ciclos em uma rede é dada pelo seu coeficiente de clusterização. Em termos simples, este índice é a probabilidade média que dois vértices vizinhos de um mesmo vértice têm de também estarem conectados entre si. O coeficiente de clusterização médio de uma rede é definido por (NEWMAN, 2003):

$$C = \frac{1}{\eta} \sum_{i} C_i \tag{3.5}$$

onde é C é coeficiente de clusterização médio da rede (fração de tripas transitivas) e C_i é o coeficiente de clusterização local para o vértice i, que é dada pela expressão (WATTZ; STROGATZ, 1998)

$$C_i = \frac{2E}{k_i(k_i - 1)} \tag{3.6}$$

onde é k_i o grau e E é o número de arestas total entre seus vizinhos.

Nesta pesquisa, considera-se o coeficiente de agregação para observação de clusters nas redes de colaboração científica. A Figura 3.5 mostra um exemplo de formação de um ciclo de vértices em uma rede.

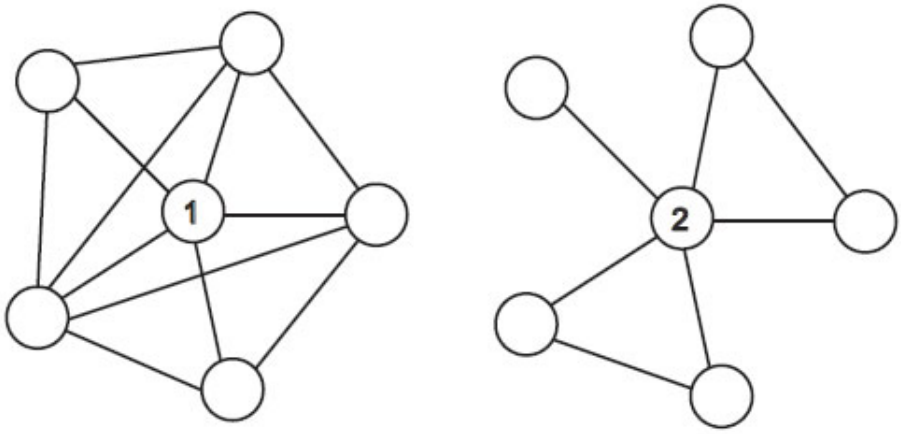


Figura 3.5: Exemplo de rede com um ciclo de vértices fechado. Fonte: Antiqueira (2007).

$$C_{i1} = \frac{2n_i}{k_i(k_i-1)} = \frac{2*7}{5*(5-1)} = \frac{14}{20} = 0,7$$
$$C_{i2} = \frac{2n_i}{k_i(k_i-1)} = \frac{2*2}{5*(5-1)} = \frac{4}{20} = 0,2 \quad (3.7)$$

A Figura 3.5 mostra duas situações de redes. No primeiro caso, o coeficiente de aglomeração é $C_1 = 0,7$ e no segundo caso, $C_2 = 0,2$. Apesar de em ambos os casos, o grau $k_1 = k_2 = 5$, o nível de clusterização é maior na primeira rede.

- Centralidade de Grau: Refere-se a quais vértices são mais conectados aos outros ou quais têm maior influência na rede.
- Centralidade de Intermediação (betweenness centrality): A centralidade de um vértice i se refere ao número de caminhos geodésicos de outros vértices que passam por esse vértice. É definido pela expressão:

$$B_u = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)} \quad (3.8)$$

onde $\sigma(i,u,j)$ é o número de menores caminhos entre os vértices i e j que passam pelo vértice u e o $\sigma(i,j)$ é o número total de menores caminhos entre i e j.

- Densidade: A densidade é referente ao quanto uma rede é *clusterizada*. Uma rede com maior número de links possíveis entre os vértices tem uma densidade maior do que uma rede que possui muitos vértices pouco conectados. Uma expressão que define a densidade média é

$$\rho = \frac{\langle k \rangle}{n-1} \quad (3.9)$$

onde n é o número de vértices de uma rede e $\langle k \rangle$ o grau médio (FERRER; SOLé, 2003).

3.2 Principais Modelos Topológicos de Redes Complexas

Históricamente, a teoria de redes complexas está fundamentada nos conceitos da teoria dos grafos e da mecânica estatística. A primeira tentativa de se explicar os fenômenos ocorridos nas redes encontradas na natureza veio com a teoria dos grafos de Leonhard Euler. Esta teoria surgiu a partir da sua solução proposta para o problema das Setes

Pontes de Königsberg, onde se discutia a possibilidade de se atravessar toda as pontes sem nunca repetir uma ponte. Com os avanços dos estudos neste campo, percebeu-se que a natureza das redes não era sempre estática e uniforme e a partir daí, surgiram outras teorias sobre a dinâmica das redes complexas: modelos de redes aleatórias, de Erdös-Rényi, modelo de redes mundo pequeno de Wattz e Strogatz e redes livres de escala de Barabási.

3.2.1 Redes Aleatórias

Alfred Rényi e Paul Erdös, em 1959, apresentaram em uma publicação (ERDöS; RéNYI, 1959) um modelo matemático de redes aleatórias, que demonstravam que grafos se expandem de forma aleatória. Eles demonstraram que bastava uma conexão entre cada convidado de uma festa para que todos acabassem conectados ao final dela e que, se fossem agregados mais ligações nesta rede, a probabilidade de formação de clusters é maior (GAMEIRO, 2008). Isso representa uma propriedade relevante das redes: o grau de coesão. Quanto maior a clusterização de uma rede mais coesa ela é.

A explicação dada por Erdös-Rényi para a capacidade de um vértice receber novas ligações é que estas ocorriam de forma aleatória. Todo vértice de uma rede tem probabilidade igual de receber novas ligações. Sendo uma rede de N vértices com n arestas, a probabilidade de qualquer vértice deste grafo receber nova ligação é de $N(N-1)/2$ (ALBERT; BARABáSI, 2002). A distribuição de graus neste modelo de rede segue uma distribuição de Poisson, e a probabilidade um vértice ter um grau k é (NEWMAN, 2003):

$$p_k = \binom{n}{k} p^k (1-p)^{n-k} \simeq \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!} \tag{3.10}$$

onde $\langle k \rangle$ é o grau médio, n o número de vértices e k um grau fixo. A conectividade $\langle k \rangle$ média é fixa e a distribuição decai rapidamente em escala natural. Em redes aleatórias é mais difícil encontrar pontos de centralidade, isto porque, segundo esse modelo de rede, todos os vértices da rede possuem a característica de terem igual probabilidade de receber novas ligações. Porém, na natureza encontramos comportamentos que não se explicam pela probabilidade igualitária proposta pelo modelo de redes aleatórias.

3.2.2 Redes Mundo-Pequeno

O primeiro experimento sobre redes mundo pequeno foi realizado por Stanley Milgram em 1967. Milgram se propôs a demonstrar qual a distância média entre duas pessoas qualquer nos EUA. O experimento consistia no envio de algumas cartas a algumas pessoas pedindo-lhes que rementessem a carta a uma determinada pessoa e caso não a conhecesse, enviasse a pessoa que mais provavelmente conhecesse o destinatário (MILGRAM, 1967).

Por meio deste experimento, ele pode comprovar que a distância média entre dois pontos escolhidos ao acaso seriam de aproximadamente 6 pessoas, o chamado 6 graus de separação e a partir daí surgiu o primeiro conceito de "mundos pequenos", pois apesar das redes sociais serem densas a distância média entre dois indivíduos escolhidos ao acaso é de 6 graus.

A ideia de aleatoriedade na formação dos grafos, proposta por Erdös-Rényi, foi revista por Mark Granovetter. Em sua publicação "The Strengh of Weak Ties", ele ressaltou a importância dos elos fracos da rede sobre a estrutura das mesmas, ressaltando que os laços fracos (weak ties) em uma rede têm maior impacto na manutenção da rede que os laços fortes (strong ties) (GRANOVETTER, 1967).

Sob esta ótica, observando-se redes de amigos, amigos com muitos laços fortes tinham a probabilidade de terem os mesmos amigos em comum, enquanto os de laços mais fracos são os que favorecem o aparecimento de clusters na rede, porque conectam grupos diferentes pelo grau de amizade. Isto em uma rede aleatória não seria possível haja visto que a associação de ligações é feita de modo aleatório.

Assim como Granovetter, Ducan Watts e Steven Strogatz perceberam que uma rede social não poderia se expandir sem nenhum tipo de consideração quanto a peso dos laços. Seguindo o modelo de Erdös e Rényi, com base nas teorias de redes,Wattz e Strogatz (1998) observaram que bastava inserir poucos links aleatórios entre os clusters que em pouco tempo a rede se tornaria um grande cluster, o chamado mundo pequeno. O modelo de redes "mundos pequenos" apresentam como uma das propriedades alta transitividade ou alto coeficiente de clusterização quando comparados aos coeficientes de clusterização em redes aleatórias com mesmo N e $\langle k \rangle$.

O caráter aleatório proposto no modelo de redes de Erdös-Rényi, dificulta o aparecimento de pontos de centralidade em uma rede, já que a conectividade média dos vértices é fixa, ou seja qualquer vértice tem a mesma probabilidade de receber ligações. Barabási e Albert (1999) constataram que em sistemas observados na natureza as leis que regem a expansão dessas redes obedecem uma organização, diferentemente das redes segundo modelos de

Figura 3.6: Modelos de Redes. Fonte: Wattz e Strogatz (1998).

Erdös-Rényi e Watts-Strogatz.

3.2.3 Redes Livres de Escala

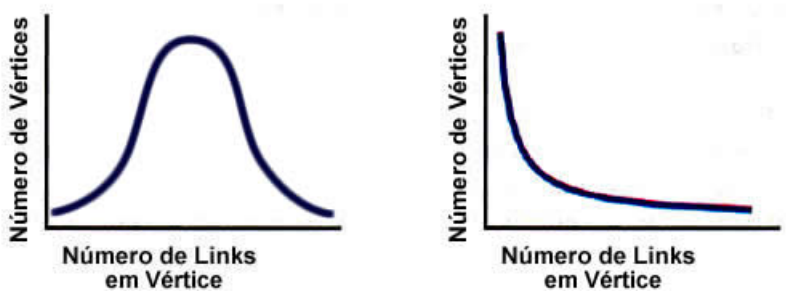
Redes reais não são estáticas, como são assumidas nos modelos aleatórios, elas se expandem segundo características próprias de sua natureza. Barabási e Albert (1999) admitiam que o surgimento de novas ligações nestas redes são desiguais e seguem leis de potências. As leis de potência são expressas pela probabilidade:

$$P(k) \sim k^{-\gamma} \tag{3.11}$$

onde k é o grau dos vértices e γ uma constante de potência.

A Figura 3.7 mostra a distribuição de graus em redes aleatórias e em livre escala. Observa-se que no primeiro caso, a distribuição onde a maioria dos nós tem em média o mesmo grau. A distribuição de graus em uma rede livre escala mostra uma queda no número de conexões na rede, à medida que alguns vértices mais conectados recebem cada vez mais ligações, seguindo uma lei de potência. Esses vértices, chamados *hubs*, concentram boa parte das conexões. A lei de potência pode ser representada pela função

$$f(x) = ax^{-k} \tag{3.12}$$



(a) Distribuição de Graus em redes aleatórias

(b) Distribuição de Graus em redes de livre escala

Figura 3.7: Distribuição de Graus em Redes Aleatórias e Livres de Escala.

onde a e k são constantes. Na maioria das rede reais, a constante k está em um valor entre 1 e 3. Segundo Mendes (2006), os vértices são continuamente adicionados a rede, porém não a uma taxa constante, por isso são chamadas redes livres de escala.

Uma característica apontada por Barabási e Albert (1999) para explicar o comportamento expansivo das redes livres de escala é que a probabilidade de um vértice receber novas ligações é proporcional ao grau de conectividade dele dentro da rede, ou seja, os ricos ficam mais ricos. Esta característica foi designada como ligação preferencial (preferencial attachment) e é dada pela probabilidade:

$$\Pi(k_i) \sim \frac{k_i}{\sum_j k_j} \tag{3.13}$$

onde Π é a probabilidade de um vértice receber conexão do vértice i, k_i é o grau do vértice i.

As redes livres de escala apresentam um grau de conectividade baixo, porque há concentração de muitas arestas em poucos vértices, os *hubs*. Na maioria dos vértices a concentração de arestas é baixa, diferente dos modelos aleatórios e de mundos pequenos, onde existe um grau de conectividade médio.

Uma fragilidade destas redes é a dependência dos vértices de menor conectividade dos *hubs* porque se forem removidos poucos *hubs* desta rede, ela praticamente se desfaz, o que não acontece em uma rede altamente clusterizada já que o nivel de coesão é alto. Em redes de sites na internet, pode-se observar este modelo de rede, onde há uma aglomeração de *link*s em torno de sites muito populares, a tendência natural é a preferência pela popularidade

do site.

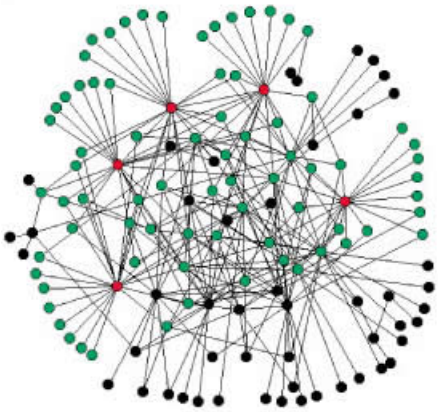


Figura 3.8: Modelo de Rede Livre de Escala. Fonte: Albert (2000).

As propriedades topológicas dos modelos de redes identificados subsidiarão a análise das redes de co-autoria estudadas nesta pesquisa. Aspectos como a identificação de focos de centralidade, pontos de fraqueza, agrupamentos preferenciais, crescimento dentro de um intervalo temporal entre outros, permitirão a análise das redes de colaboração científica, obtidas pelo modelo proposto, a partir de índices das redes sociais e complexas.

3.3 Análise Comparativa entre as topologias de redes

De um modo geral, as redes sociais e complexas apresentam comportamentos dinâmicos, que são observados por suas propriedades topológicas. Por exemplo, em redes sociais, observam-se relacionamentos existentes em contexto cultural, profissional, afetivo, dentre outros. Nas relações sociais, observam-se propriedades que permitem caracteriza-las como redes mundo pequeno. Nestas redes, é evidenciada a presença de *clusters* nos agrupamentos sociais. De uma forma geral, a associação de novos vértices em redes sociais é mais lenta, dado o fator subjetivo envolvido nestas novas conexões, tais como proximidade por grupo de amizade, ambiente profissional, ambiente acadêmico, etc. Por exemplo, um indíviduo tem uma probabilidade maior de se associar a outro indivíduo dentro de um mesmo grupo social do que se associar a qualquer indivíduo de outro grupo social.

A expansão das redes caracterizadas como mundo pequeno ou aleatórias é mais lenta que em redes de livre escala, e a distância geodésica média entre qualquer par de vértices da rede é quase similar com a expansão da rede. Isso ocorre porque no modelo de redes aleatórias, bem como o modelo de mundo pequeno, admite-se a introdução aleatória de novos links, diferentemente do modelo de redes livres de escala. Barabási e Albert (1999) demonstraram que a dinâmica das redes em livre escala seguia algum tipo de ordenação,

as leis de potência.

No modelo de rede proposto por Erdös-Rényi, a probabilidade de conexão de cada vértice da rede é igual. Desta forma, é incomum encontrar grandes diferenças na distribuição de graus e pontos de centralidade na estrutura destas redes e não há grandes distâncias entre os vértices das redes mundo pequeno nem em redes aleatórias.

Considerando que a probabilidade de novos links em redes mundo pequeno é igual para qualquer vértice, este modelo de rede segue uma distribuição de graus média fixa, onde se nota uma expansão lenta na sua estrutura. Porém, essa característica topológica não representa a realidade de diversas redes reais. As conexões em vértices não acontecem sempre com regularidade e aleatoriamente.

Em muitas redes reais, o modelo de redes livres de escala justifica esse comportamento como dependente do fator de potência. A teoria do "rico fica mais rico" é referente à tendência de conexão entre novos vértices e aqueles com alta conectividade, a exemplo das redes de *sites*. *Site* com alto grau de popularidade, como Google, são constatemente associados a outros sites de menor conectividade.

Em redes mais coesas, a resiliência, de uma forma geral, é maior que redes com crescimento sob leis de potência. Se os vértices de maior conectividade em uma rede de livre escala são removidos, a estrutura da rede praticamente se desfaz. Isso se revela como ponto de vulnerabilidade neste modelo de rede. Em uma rede com características de mundo pequeno ou aleatória, a regularidade na distribuição de graus aumenta a capacidade de resiliência da rede. Entretanto, em redes livres de escala se apenas forem retirados vértices pouco conectados, a estrutura destas não sofrerá grande abalo.

Em redes de livre escala, a existência da ligação preferencial com vértices com alta conectividade, dificulta a formação de clusters. Se uma rede é altamente clusterizada, então ela proporcionalmente é coesa, logo os caminhos mínimos para se atingir qualquer vértice da rede são pequenos se comparados aos caminhos necessários para se chegar a qualquer vértice em uma rede de crescimento em livre escala (BARABáSI; ALBERT, 1999). Dessa forma, se um evento ocorre em um vértice de uma rede livre de escala, esse demora mais a atingir outros pontos da estrutura da rede, salvo o caso dos hubs. Eventos ocorridos em vértices muito conectados, afeta rapidamente a estrutura da rede. Em uma redes clusterizadas, qualquer evento ocorrido em um vértice, rapidamente se alastra pela estrutura.

Capítulo Quatro

Modelo para Extração de Dados Textuais e Geração de Redes

4.1 Descrição formal do modelo

O processo de identificação de padrões na mineração do texto é realizado a partir de uma condição booleana de agregação ou restrição de caracteres definida no início da busca. Um padrão criado como expressão regular é um modelo utilizado para verificação das cadeias de caracteres e identificação de resultados que se encaixem dentro deste modelo. Por exemplo, um padrão a ser reconhecido no texto tem a seguinte formatação :

AUTOR, A.B. (Texto)

Analisando do ponto de vista da condição expressa na notação, é importante observar a palavra quanto a forma das letras (maiúsculas ou minúsculas) e separadores entre estas (vírgulas, pontos, dois pontos, parêntesis, espaços em branco, etc.). Neste exemplo anterior, a condição teste deve aceitar somente letras maiúsculas no início da palavra, separar por vírgula, aceitar espaços em branco, pontos e letras maiúsculas até encontrar o próximo separador, o parêntesis, aceitar letras maiúsculas e minúsculas até encontrar o próximo parêntesis.

Cada caractere lido no texto pode ou não fazer parte de um alfabeto definido pela expressão regular. O processo de leitura e teste dos caracteres utilizam a lógica matemática do fecho de Kleene (Seção 2.2, página 27). Considerando Σ alfabeto, o fecho estrela de Σ , denotado por Σ^*, é o conjunto de todas as cadeias (finitas) obtidas concatenando zero ou mais símbolos de Σ. Por exemplo, se Σ ={a,b}, então

$$\Sigma^* = \{\lambda, a, b, aa, ab, ba, bb, aaa, aa, ...\} \tag{4.1}$$

onde λ representa o conjunto vazio. O Σ^* é o menor conjunto de caracteres encontrados e concatenados a partir de Σ.

O conjunto acima é o resultado obtido da busca pelo alfabeto definido Σ. Desta forma ocorre a busca de palavras e sentenças compatíveis ao padrão da expressão regular criado (o alfabeto Σ). Cada caractere da palavra é avaliado e se for considerado como pertencente

ao alfabeto definido pelo padrão, este é concatenado a cadeia de caracteres. Considere o exemplo de string "**AUTOR, A.B. (Texto)**" .Uma notação que poderia descrever o padrão utilizado para aceitação do texto poderia ser representado por:

[A-Z]+\,[A-Z\.\s]+\([A-Za-z\s]+\)

Considerando o alfabeto Σ determinado pela expressão 4.1, a validação das strings ocorre da seguinte forma (Tabela 4.1):

Tabela 4.1: Tabela demonstrativa de leitura do padrão.

Sintaxe	**Execução**	**Exemplo**
[A-Z]+	Aceita todos os caracteres de letras maiúsculas até que seja esgotada a busca por estes caracteres. Enquanto a condição for satisfeita, o caractere é acrescentado no resultado da busca.	Σ^* = {A}, se A ∈ Σ, Σ^* = A+U+T+O+R = AUTOR
\	Após esgotar a busca por caracteres maiúsculos, a sequência "\" indica que devem ser encontrados textos que, além de conter resultados que atendem a primeira condição [A-Z]+, tenha o separador vírgula na sequência.	Σ^* = {,}, se , ∈ Σ, Σ^* = AUTOR + , = AUTOR,
[A-Z\.\s]+	Após executar as as ações anteriores, procura-se por caracteres em letras maiúsculas, pontos, e espaços entre caracteres até que seja encontrado o próximo caractere delimitador, o parêntesis.	Σ^* = {A, ., B}, se A, ., B ∈ Σ, Σ^* = AUTOR, A. B.
\(	o ao final do processo anterior, é procurada na sequência o delimitador parêntesis.	Σ^* = {(}, se (∈ Σ, Σ^* = AUTOR, A. B. (
[A-Za-z\s]+	Aceita letras maiúsculas (A-Z), minúsculas (a-z) e espaços em branco (\s) na sequência, até que seja encontrado o próximo delimitador, o parêntesis.	Σ^* = {T,e,x,t,o}, se T,e,x,t,o ∈ Σ, Σ^* = AUTOR, A. B. (Texto
\)	Define o término desta expressão.	Σ^* = {)}, se) ∈ Σ, Σ^* = AUTOR, A. B. (Texto)

Fonte: Autor.

A validação de caracteres é cumulativa, se a condição verdade de toda a expressão for atendida, este resultado é incluido como padrão reconhecido no texto. Desta maneira, a abrangência ou restrição nos resultados buscados dependem da expressão criada para

minerar o texto. A construção das redes considerou as regras para expressões regulares, descritas a seguir em álgebra relacional (referência de sintaxe na Tabela 4.2) (ELSMARI; NAVATHE, 2004):

$\Rightarrow \quad RESULT1 \leftarrow \pi_{idPesquisador}(pesquisador\_artigo \bowtie_{idPesquisador=id} pesquisador)$

$\Rightarrow \quad RESULT2 \leftarrow \pi_{idPesquisador}(pesquisador\_anais \bowtie_{idPesquisador=id} pesquisador)$

$\Rightarrow \quad RESULT3 \leftarrow \pi_{idPesquisador}(pesquisador\_capitulo \bowtie_{idPesquisador=id} pesquisador)$

$\Rightarrow \quad PESQ\_REDE \leftarrow RESULT1 \cup RESULT2 \cup RESULT3$

$\Rightarrow \quad RESULTREL1 \leftarrow \pi_{idPesquisador,nome}(pesquisador\_artigo *_{idPesquisador=id} pesquisador)$

$\Rightarrow \quad RESULTREL1A \leftarrow RESULTREL1(pesquisador\_artigo \bowtie_{idArtigo=id} artigo)$

$\Rightarrow \quad SUB\_QUERY1 \leftarrow \pi_{id}(\sigma_{pq.id} = RESULT1(RESULTREL1A \bowtie_{idArtigo=id} artigo)$

$\Rightarrow \quad RESULTREL1B \leftarrow \pi_{id,nome}(\sigma_{idinSUB_QUERY1,id=RESULT1}(RESULTREL1A))$

$\Rightarrow \quad RESULTREL2 \leftarrow \pi_{idPesquisador,nome}(pesquisador\_anais *_{idPesquisador=id} pesquisador)$

$\Rightarrow \quad RESULTREL2A \leftarrow RESULTREL2(pesquisador\_anais \bowtie_{idAnais=id} anais)$

$\Rightarrow \quad SUB\_QUERY2 \leftarrow \pi_{id}(\sigma_{pq.id=RESULT2}(RESULTREL2A \bowtie_{idAnais=id} anais)$

$\Rightarrow \quad RESULTREL2B \leftarrow \pi_{id,nome}(\sigma_{idinSUB_QUERY2,id=RESULT1}(RESULTREL2A))$

$\Rightarrow \quad RESULTREL3 \leftarrow \pi_{idPesquisador,nome}(pesquisador\_capitulo *_{idPesquisador=id} pesquisador)$

$\Rightarrow \quad RESULTREL3A \leftarrow RESULTREL3(pesquisador\_capitulo \bowtie_{idCapitulo=idLivro} capitulo)$

$\Rightarrow \quad SUB\_QUERY3 \leftarrow \pi_{id}(\sigma pq.id = RESULT3(RESULTREL3A \bowtie_{idCapitulo=idLivro} capitulo)$

$\Rightarrow \quad RESULTREL3B \leftarrow \pi_{id,nome}(\sigma_{idinSUB_QUERY3,id=RESULT1}(RESULTREL3A))$

$\Rightarrow \quad REDE \leftarrow RESULTREL1B \cup RESULTREL2B \cup RESULTREL3B$

Tabela 4.2: Sintaxe de Álgebra Relacional.

Símbolo	**Descrição**
$\bowtie$	A função JOIN é uma operação binária que é utilizada para combinar tuplas de duas relações dentro de uma tupla única. Ex.: Relacionar as tabelas "pesquisador_artigo" e "pesquisador" = pesquisador_artigo$\bowtie_{idPesquisador=id}$pesquisador
=	A operação EQUIJUNÇÃO é uma variação de JOIN e compara dois atributos de um relacionamento. Ex.: o atributo "idPesquisador"' da tabela "pesquisador_artigo" é igual ao atributo "id" na tabela "pesquisador".

Continuação na próxima página...

Tabela 4.2 – Continuação

Símbolo	**Descrição**
π	A operação PROJEÇÃO seleciona certas colunas da tabela e descarta outras. Ex.: selecionar os campos "nome" e "vinculo" da tabela "pesquisador" = $\pi_{nome,vinculo}$(pesquisador)
$*$	A operação JUNÇÃO NATURAL é equivalente a uma equijunção, entretanto, nesta última ao se comparar e juntar esses atributos, estes aparecem iguais nos resultados. A junção natural realiza a mesma comparação, porém não traz valores duplicados dos atributos nos resultados, traz apenas um dos atributos, já que são equivalentes. Ex.: o atributo "idPesquisador‘ da tabela "pesquisador_artigo" é igual ao atributo "id" na tabela "pesquisador". Neste caso, nos resultados apenas vem o "idPesquisador" ou "id".
σ	A operação SELEÇÃO é utilizada para selecionar um subconjunto de tuplas de uma relação que satisfaça uma condição de seleção. Ex.: Selecionar pesquisadores da tabela "pesquisador" que sejam docentes = $\sigma_{vinculo='docente'}$(pesquisador)

Fonte: (ELSMARI; NAVATHE, 2004)

Para facilitar o entendimento da álgebra relacional envolvida na criação das redes, será mostrado um exemplo de construção de rede de artigos sob a ótica desta álgebra a seguir:

(1) $RESULT1 \leftarrow \pi_{idPesquisador}(pesquisador\_artigo \bowtie_{idPesquisador=id} pesquisador)$

(2) $RESULTREL1 \leftarrow \pi_{idPesquisador,nome}(pesquisador\_artigo *_{idPesquisador=id} pesquisador)$

(3) $RESULTREL1A \leftarrow RESULTREL1(pesquisador\_artigo \bowtie_{idArtigo=id} artigo)$

(4) $SUB\_QUERY1 \leftarrow \pi_{id}(\sigma_{pq.id} = RESULT1(RESULTREL1A \bowtie_{idArtigo=id} artigo)$

(5) $RESULTREL1B \leftarrow \pi_{id,nome}(\sigma_{idinSUB_QUERY1,id=RESULT1}(RESULTREL1A))$

Tabela 4.3: Rede de Artigos em Álgebra Relacional.

Sintaxe	**Execução**	**Exemplo**
(1)	Existem duas tabelas que se relacionam: "Pesquisadores" e "Pesquisador/Artigo". Neste processo, seleciona-se apenas o campo "idPesquisador" da tabela "Pesquisador/Artigo" que são encontrados na tabela "Pesquisador".	Três pesquisadores A,B e C. Apenas os autores A e B estão relacionados na tabela "Pesquisador/Artigo". Neste caso, o autor C será ignorado na seleção, porque não foi encontrado o registro deste autor na tabela de relacionamento "Pesquisador/Artigo", apesar de existir na tabela Pesquisador.
(2)	As tabelas relacionadas são "Pesquisador" e "Pesquisador/Artigo". Neste caso, seleciona-se apenas o campo "idPesquisador" da tabela "Pesquisador/Artigo" e o campo "nome" da tabela "Pesquisador".	Três pesquisadores A,B e C. Apenas os autores A e B estão relacionados na tabela Pesquisador/Artigo. Neste caso, o autor C será ignorado na seleção, porque não foi encontrado o registro deste autor na tabela de relacionamento "Pesquisador/Artigo", apesar de existir na tabela Pesquisador.
(3)	As tabelas relacionadas são "Artigo" e "Pesquisador/Artigo". Neste caso, utiliza-se a seleção resultante do processo anterior, onde foram selecionados os campos "idPesquisador" da tabela "Pesquisador/Artigo" e o campo "nome" da tabela "Pesquisador'. Entretanto, apenas serão filtrados, dentre estes pesquisadores, apenas aqueles que têm relações criadas na tabela "Pesquisador/Artigo", e que os artigos que estão nestas relações existam na tabela "Artigos".	Três pesquisadores A,B e C e dois artigos D e E. Os autores A e B estão relacionados com D na tabela Pesquisador/Artigo. Neste caso, o autor C será ignorado na seleção, porque não foi encontrado o registro deste autor na tabela de relacionamento "Pesquisador/Artigo", e na construção das redes de artigos, o artigo E será ignorado porque não houve relacionamento dele com nenhum autor.

Continuação na próxima página...

Tabela 4.3 – Continuação

Sintaxe	**Execução**	**Exemplo**
(4)	Utiliza-se a seleção de pesquisadores resultante do processo anterior como filtro de seleção a ser utilizado nesta nova execução do processo. Será filtrado pesquisadores da tabela "Pesquisador" que se encontrem dentro dos resultados da seleção realizada no processo anterior, pegando apenas o campo "id" da tabela "Pesquisador".	Todos pesquisadores: A,B e C e pesquisadores encontrados na seleção anterior: A e B. O autor C ficará de fora nesta nova seleção de pesquisadores.
(5)	Utiliza-se a seleção de pesquisadores resultante do processo anterior como novo filtro de seleção a ser utilizado nesta nova execução do processo. Será filtrado pesquisadores da tabela "Pesquisador" que se encontrem dentro dos resultados da seleção realizada no processo anterior, pegando apenas o campo "id" e "nome" da tabela "Pesquisador". Neste caso, esta nova seleção compara o campo "id" resultante do processo anterior, com o campo "id" resultante do processo 3. O resultado desta seleção serão os pesquisadores que comporão as redes.	Todos pesquisadores: A,B e C e pesquisadores encontrados na seleção anterior: A e B. O autor C ficará de fora nesta nova seleção de pesquisadores.

Fonte: Autor.

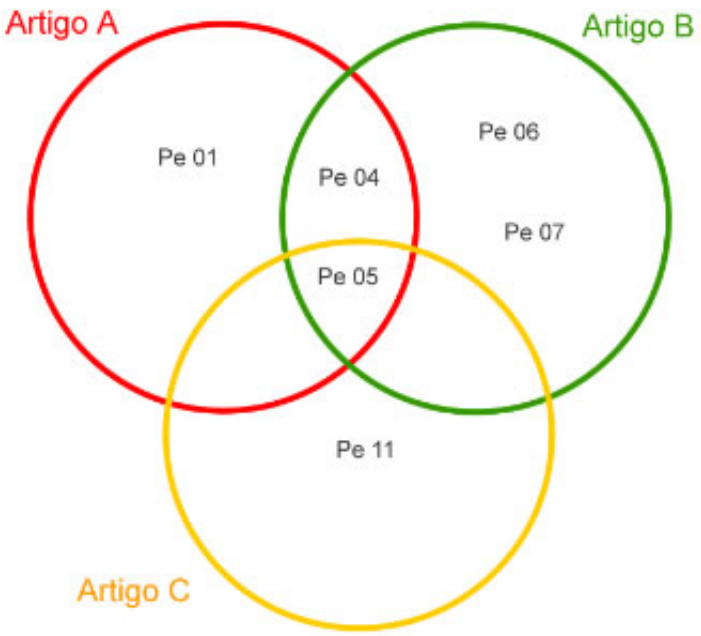


Figura 4.1: Representação da rede de co-autoria em artigos. Neste caso, a representação reflete uma rede de publicações de Artigos, Anais e Capítulos. Fonte: Autor.

No exemplo da Figura 4.1, no artigo A há uma colaboração entre os pesquisadores Pe01, Pe04 e Pe05. Estes dois últimos também colaboram com outros pesquisadores porque estão presentes em outros artigos. O Pe05 colabora com todos os pesquisadores já que está na interseção dos artigos. O Pe04 só não colabora com Pe11, porque não está no conjunto colaborativo do artigo C. Com base neste esquema são montadas as relações de redes construídas pelo modelo proposto.

4.2 Aplicação do modelo formal

São definidas três etapas principais para a realização do processo de construção das redes: Mineração dos Textos, Inserção dos Dados e Construção das Redes. Cada etapa desta contém alguns processos. As etapas do processo em geral estão relacionadas na seguinte sequência:

1. Mineração de Textos

   a. Seleção do conjunto de documentos

   b. Identificação dos padrões

   c. Geração das listas para inserção dos dados

2. Inserção de Dados

   a. Inserção dos dados primários

   b. Inserção de relacionamentos

3. Construção das Redes

   a. Definição de filtros

   b. Construção das redes e gravação da rede para formato .NET

As etapas do processo funcional do modelo proposto está representado na Figura 4.2.

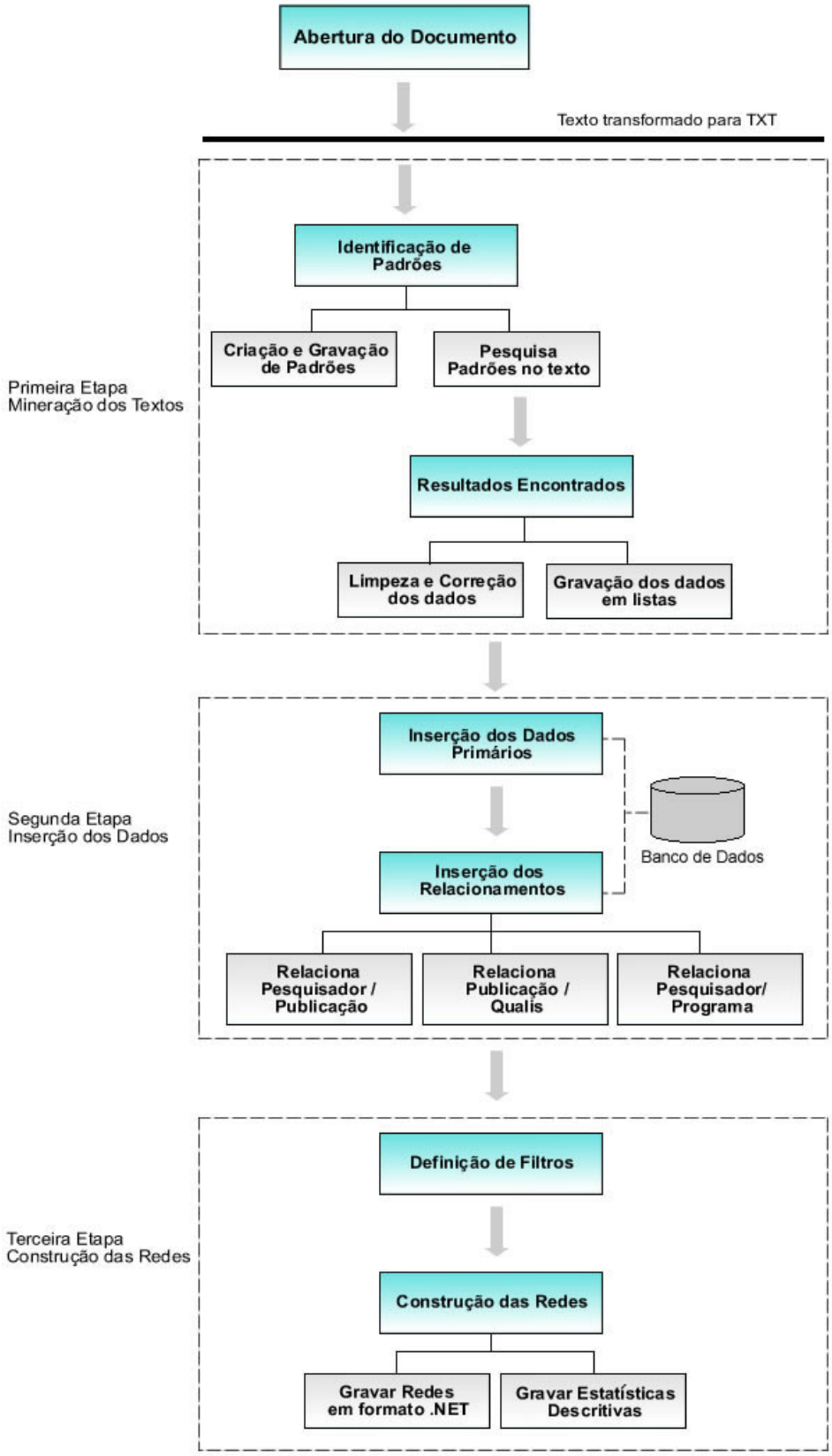


Figura 4.2: Processo Funcional do Modelo Proposto. Fonte: Autor.

1ª etapa : Mineração de Textos

O primeiro processo nesta etapa é selecionar o documento de onde se pretende extrair as informações. Qualquer documento onde haja um padrão identificável no texto (eg.: todo nome de autor possui a formatação "AUTOR, A. B.", ou seja, letras maiúsculas somente e separação por ponto) poderá ser processado pelo modelo, caso contrário, a filtragem de dados será extremamente complexa, isto porque, como já foi dito, a mineração do texto se baseia em reconhecimento de padrão textual. O texto, que deverá estar em formato digital e em extensão PDF, TXT ou HTML, será processado pelo software e transformado em um vetor de palavras de onde serão identificados os padrões.

Um aspecto importante a ressaltar nesta etapa é que, antes da criação das expressões, é interessante estudar os padrões apresentados no texto, porque desta forma a criação do padrão a ser reconhecido será simplificado. O método usado para esta tarefa foi o reconhecimento por uso de expressões regulares. A Figura 4.3 abaixo é um exemplo da expressão criada como padrão.

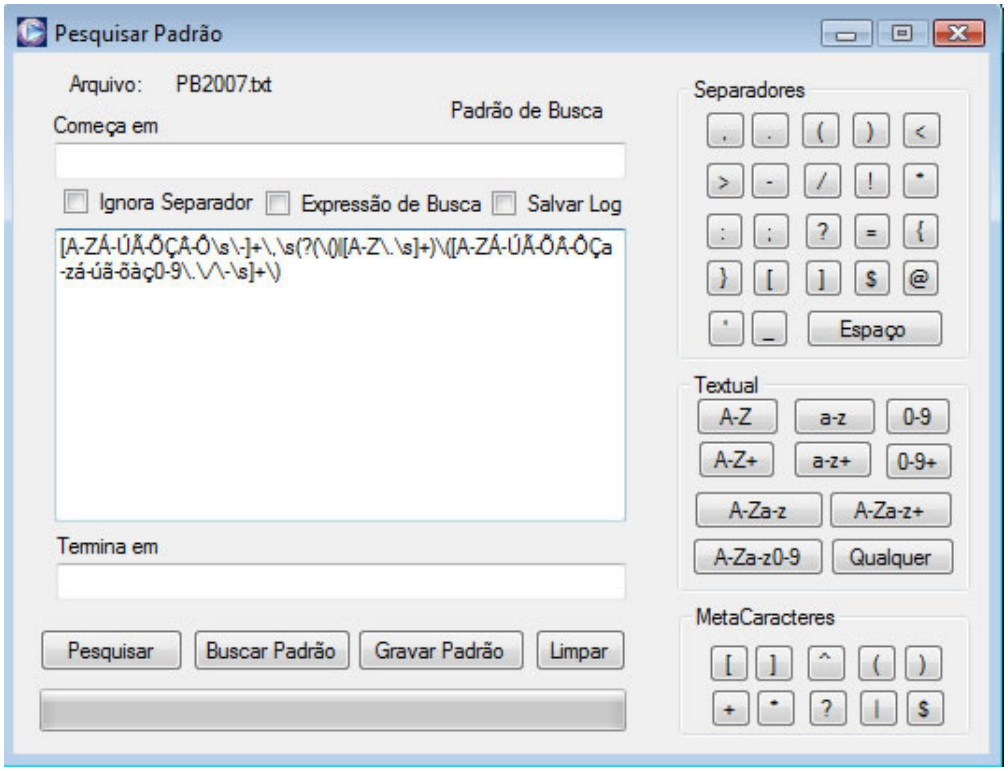


Figura 4.3: Tela para criação de padrões. Fonte: Autor.

A Figura 4.3 mostra uma expressão criada para identificar nomes e vínculos de pesquisadores nos cadernos indicadores da CAPES, que segue a formatação

PATRÍCIA, F. B. (Discente)

e cuja expressão definida para esta busca de padrão foi

[A-ZÁ-ÚÃ-ÕÇÂ-Ô/s]+/,/s(?(/()—[A-Z/./s]+)/([A-ZÁ-ÚÃ-ÕÂ-ÔÇa-zá-úã-õàç0-9/./-/s]+/)

O software provê um módulo (Figura 4.3) para a criação do padrão. Criar padrões por expressão regular não é uma tarefa trivial, e quanto mais refinada for a expressão mais eficiente será a busca dos padrões no texto. Geralmente, as expressões regulares são escritas de forma manual e dependem de um conhecimento mais aprofundado das sintaxes.

No caso deste software, este módulo apresentado facilita a criação da notação para indivíduos que não detenham conhecimento prévio de expressões regulares, porque a tela para pesquisar padrões (Figura 4.3) contém toda a informação necessária para descrição da expressão a partir de botões. As expressões criadas podem ser gravadas para reutilização em outros textos com padrões semelhantes, o que otimiza tempo de trabalho do pesquisador.

As informações que se desejam extrair do texto, devem ser identificadas em padrões e devem ser armazenadas em listas diferentes para posterior inserção no banco de dados. Por exemplo, em um dado documento as informações a serem extraídas são “pesquisador” e “artigos”. Será necessária, neste caso, a criação de duas expressões para extração de dois tipos diferentes de informação , que serão gravadas em duas listas de resultados de padrões identificados diferentes. A Figura 4.4 mostra a tela de resultados encontrados de um padrão.

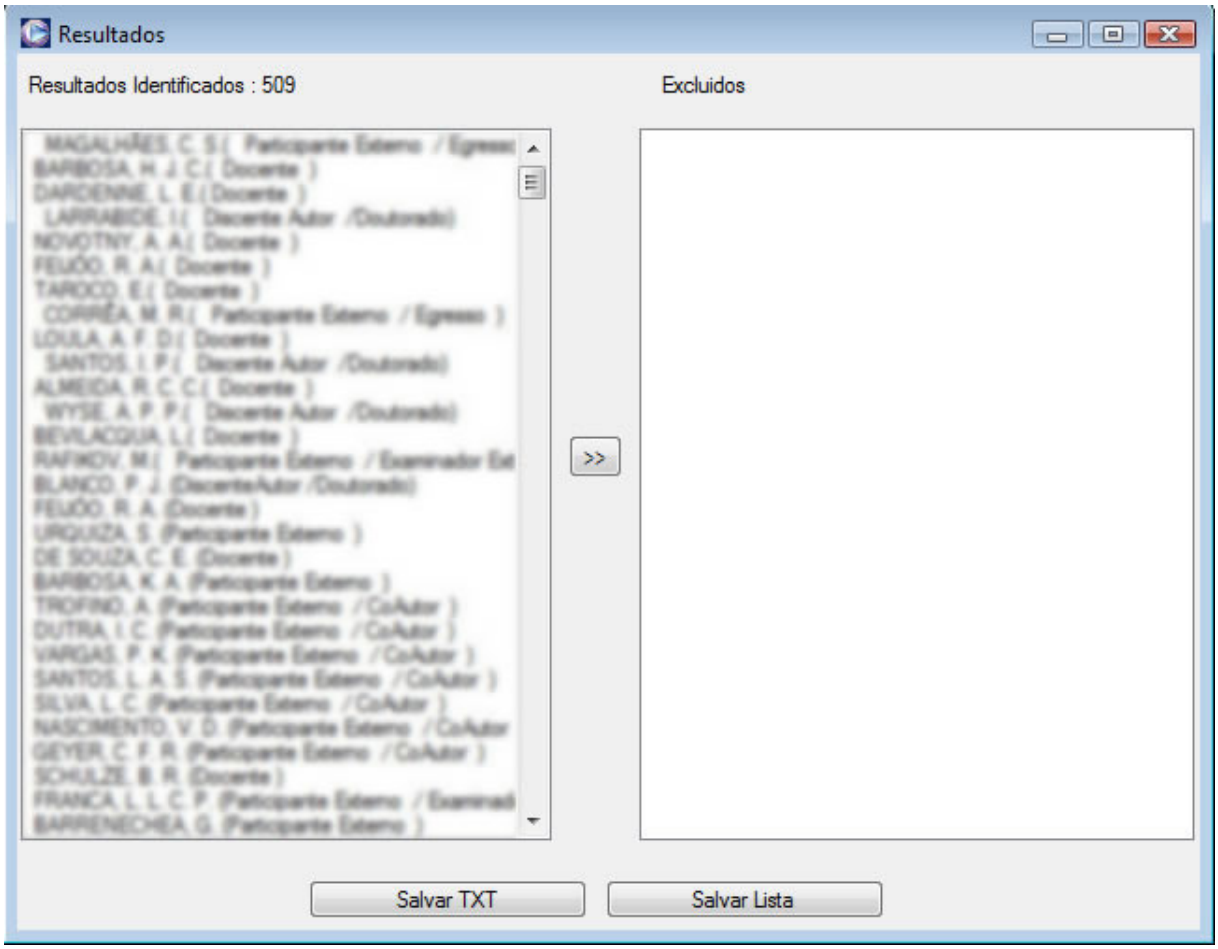


Figura 4.4: Tela de resultados obtidos de um documento a partir da busca por um padrão definido. Fonte: Autor.

Cada lista de resultados deve ser salva para relacionamento destes dados no banco. Isto porque a relação destas listas constituirá as arestas da rede a ser construída. Os resultados obtidos do reconhimento dos padrões podem conter dados irrelevantes, o que pode tornar necessária a limpeza destes dados patra evitar redundâncias, incostistências ou dados errôneos no momento de inserção dos dados e construções das redes.

A limpeza e correção dos dados provenientes dos resultados de busca devem ser feitas com cautela, haja visto que a próxima etapa é a inserção dos dados no banco, e as redes serão compostas com informações contidas no banco. Um exemplo de situação que exije um cuidado maior na limpeza é quando no texto é encontrado um nome de autor "PATRICIA, F.B." e em outro lugar no texto, "PATRICIA, F.", se tratando da mesma pessoa. Neste caso, é interessante verificar se trata-se da mesma pessoa, e se for, excluir um dos resultados para evitar replicação de dados no banco. Uma situação que necessita de correção é quando algum dado aparece incompleto (eg.: o título deveria ser "PATRICIA, F.B.,Titulo do Artigo, 2010." e vem como resultado "PATRICIA, F.B.,Titulo do "). A Figura 4.5 é tela de visualização de listas de dados, contendo nomes de pesquisadores de

um PPG, criadas na mineração do texto.

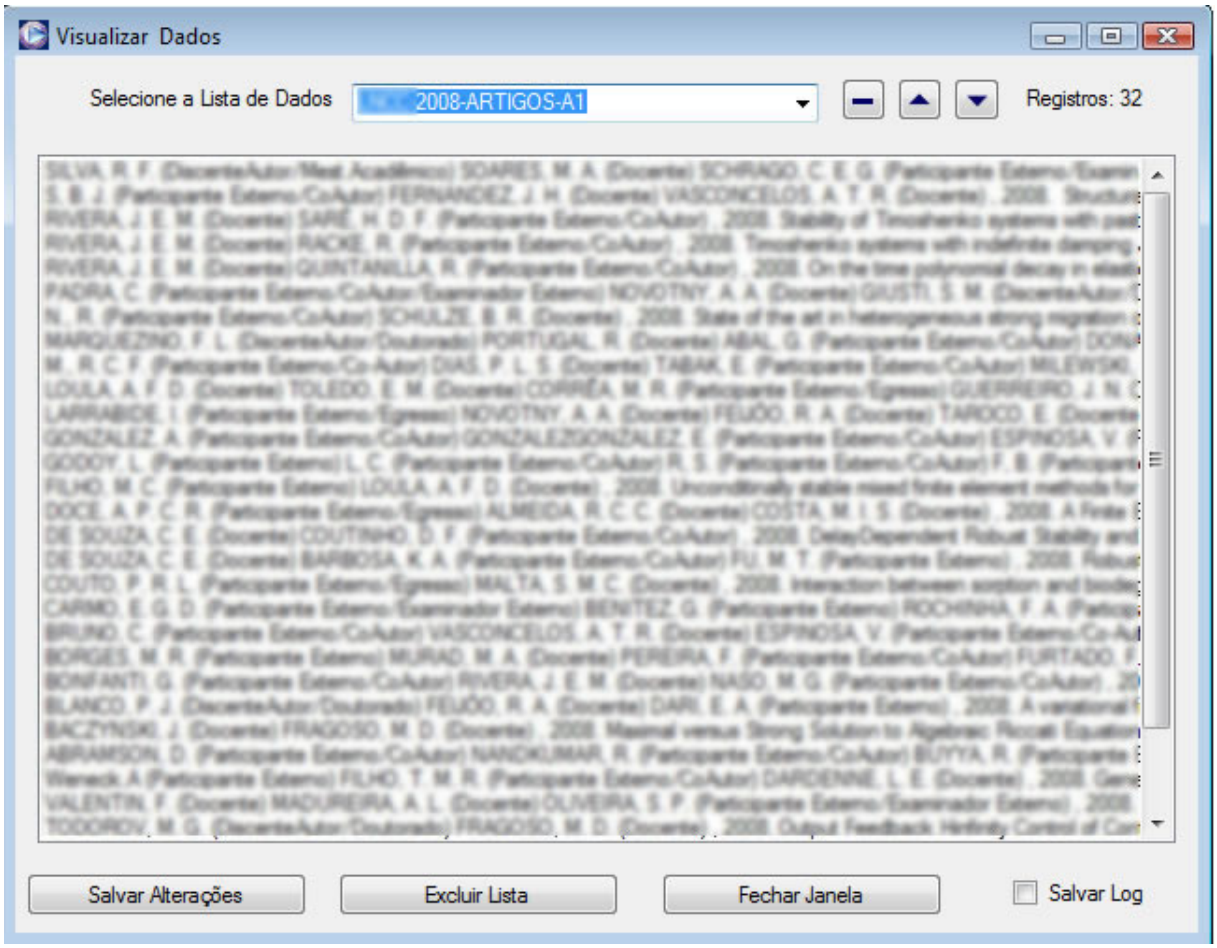


Figura 4.5: Tela de Visualização das listas de dados para limpeza. Fonte: Autor.

Uma vez que todos os dados necessários a construção da rede estejam devidamente gravados em listas, então pode-se inserir estes dados no banco de dados.

2ª etapa : Inserção de Dados

Esta etapa contém dois subprocessos: A inserção dos dados primários e inserção dos relacionamentos. O que define os dados primários são as entidades básicas da rede, ou seja, os pesquisadores, publicações, projetos, etc. Estes dados compõem os vértices e contexto da rede a ser construída. Por exemplo, relações entre pesquisadores que escreveram artigos. As listas anteriormente criadas na etapa da mineração de texto serão inseridas nas respectivas tabelas do banco de dados. Seleciona-se a lista que se deseja inserir e a tabela de destino daqueles dados. A Figura 4.6 mostra a tela para inserção das listas no banco.

Ao completar todas a inserções de todas as listas de dados, deve-se inserir neste momento os relacionamentos destes dados. Os relacionamentos são estabelecidos da seguinte forma:

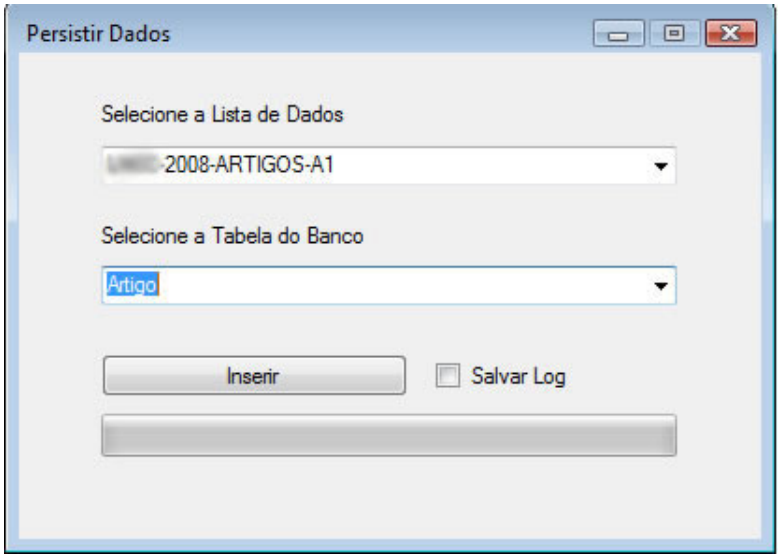


Figura 4.6: Tela de inserção dos dados primários no banco. Fonte: Autor.

o tipo de relação é definido, em seguida duas listas de dados são comparadas para identificação de ocorrências de termos equivalentes em ambas as listas. Por exemplo, a relação "Pesquisador/Artigo" é definida como tipo de relação a ser criada. Em seguida, escolhe-se as listas pertinentes do tipo de relação escolhido, neste caso, uma lista de pesquisadores e uma lista de artigos.

No processamento da relação, o algoritmo varre a primeira lista e a segunda lista procurando por similaridades, exemplo, o pesquisador X é procurado em todos os registros que estão contidos na segunda lista, a lista de artigos. Caso haja identificação de similaridade, insere-se uma relação no banco de dados, contendo a informação do pesquisador X e do artigo onde foi encontrado esse pesquisador.

Diferentemente de uma mineração de dados, a mineração de textos dificulta a identificação de relacionamentos entre as informações, porque os dados não são estruturados nem ordenados. A solução encontrada para o objeto de estudo desta dissertação consistiu em encontrar recorrência de termos entre dados pré-selecionados e assim criar laços de relacionamentos. As informações a serem conectadas são muito específicas, e não poderiam ser criadas apenas pelo processo de PLN e sumarização de conteúdo, como é realizado na maioria dos processos de mineração de textos. A Figura 4.7 mostra a tela para inserção de relacionamentos.

O módulo foi pensado para ser composto de três seções: a seção **Geral**, a seção **Por Qualis** e a seção **Por Programa**. Na seção **Geral**, são feitos todos os relacionamentos entre entidades (eg.: Pesquisador/Artigo, Pesquisador/Anais, etc.), exceto a qualificação das publicações e a associação dos pesquisadores ao programa de atuação, que são realizados

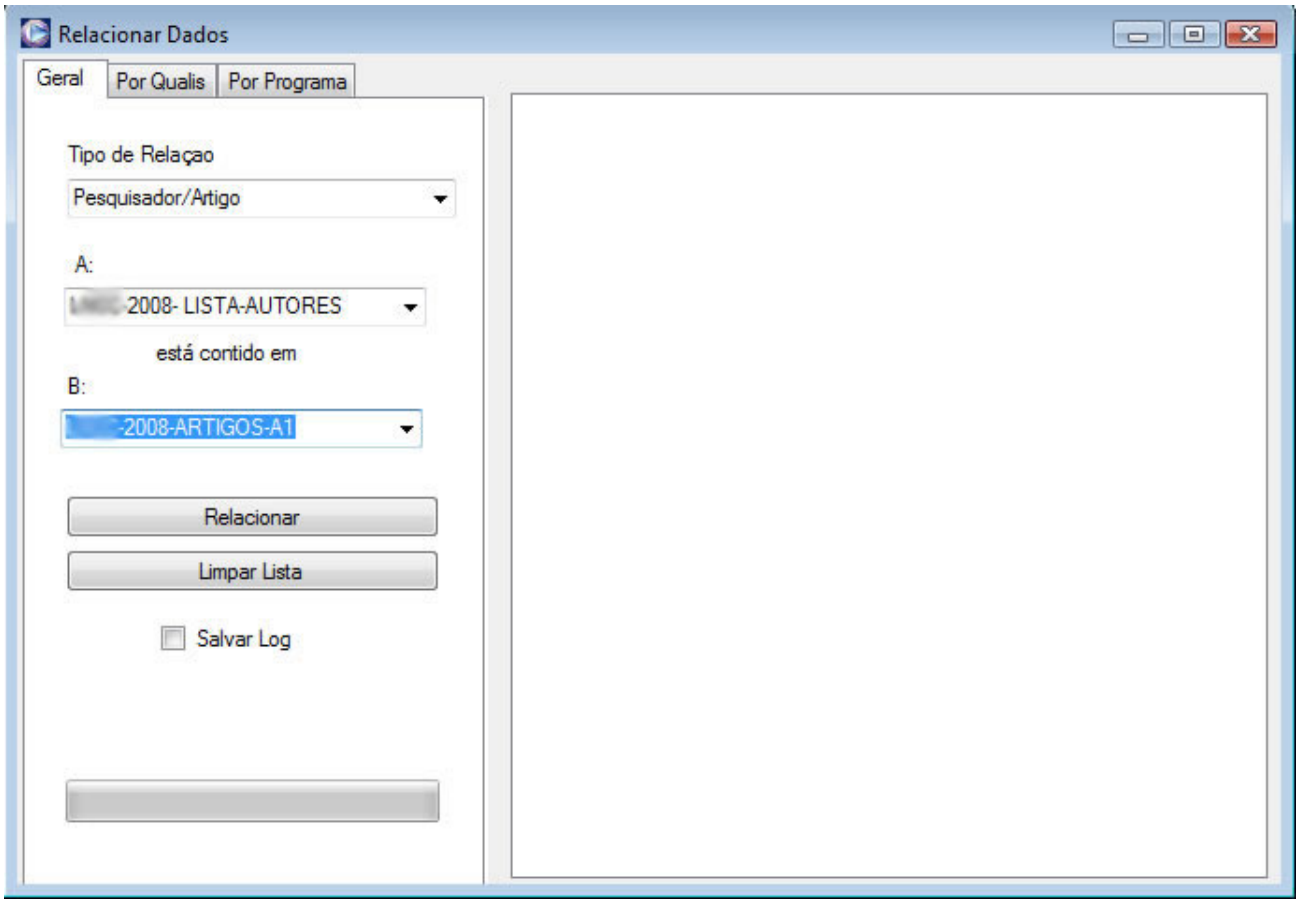


Figura 4.7: Tela para inserção de relacionamentos. Fonte: Autor.

respectivamente nas seções **Por Qualis** e **Por Programa**.

A classificação dos qualis das publicações não pode ser feita por identificação de similaridades encontradas entre os dados, devido ao fato que estas informações não são explícitas nos dados listados. Desta forma, na etapa de mineração de textos, é necessário criar listas separadas por qualis para inserção dos relacionamentos de publicações por qualis (eg.: Artigos/Qualis). O modelo prevê a pesquisa de padrões dentro de intervalos especificados. Por exemplo, buscar artigos dentro do intervalo "Qualis A1" e "Qualis A2". Desta maneira, grava-se esta lista (neste exemplo, lista de artigos qualis A1) para que seja possível classificar estes registros por qualificação.

3ª etapa : Construção das Redes

A construção das redes é composta de dois processos: A definição do escopo das redes por meio dos filtros e a geração e gravação das redes no formato PAJEK[1]. Uma rede pode ser contruída do contexto mais amplo ao contexto mais específico (e.g. pode-se criar redes

[1]http://vlado.fmf.uni-lj.si/pub/networks/pajek/, acessado em 10/08/2010 às 20:30h

amplas contendo todas as publicações de qualquer ano e PPG, como pode ser criadas redes que sejam específicas como redes de artigos qualis A1 de determinados PPG e ano), e a delimitação deste conjunto é definido no módulo de filtros. As redes construídas podem ser gravadas em extensão TXT ou .NET. A extensão .NET é uma extensão para carregar e manipular redes no software PAJEK. A Figura 4.8 mostra a tela para escolha de filtros para gerar redes.

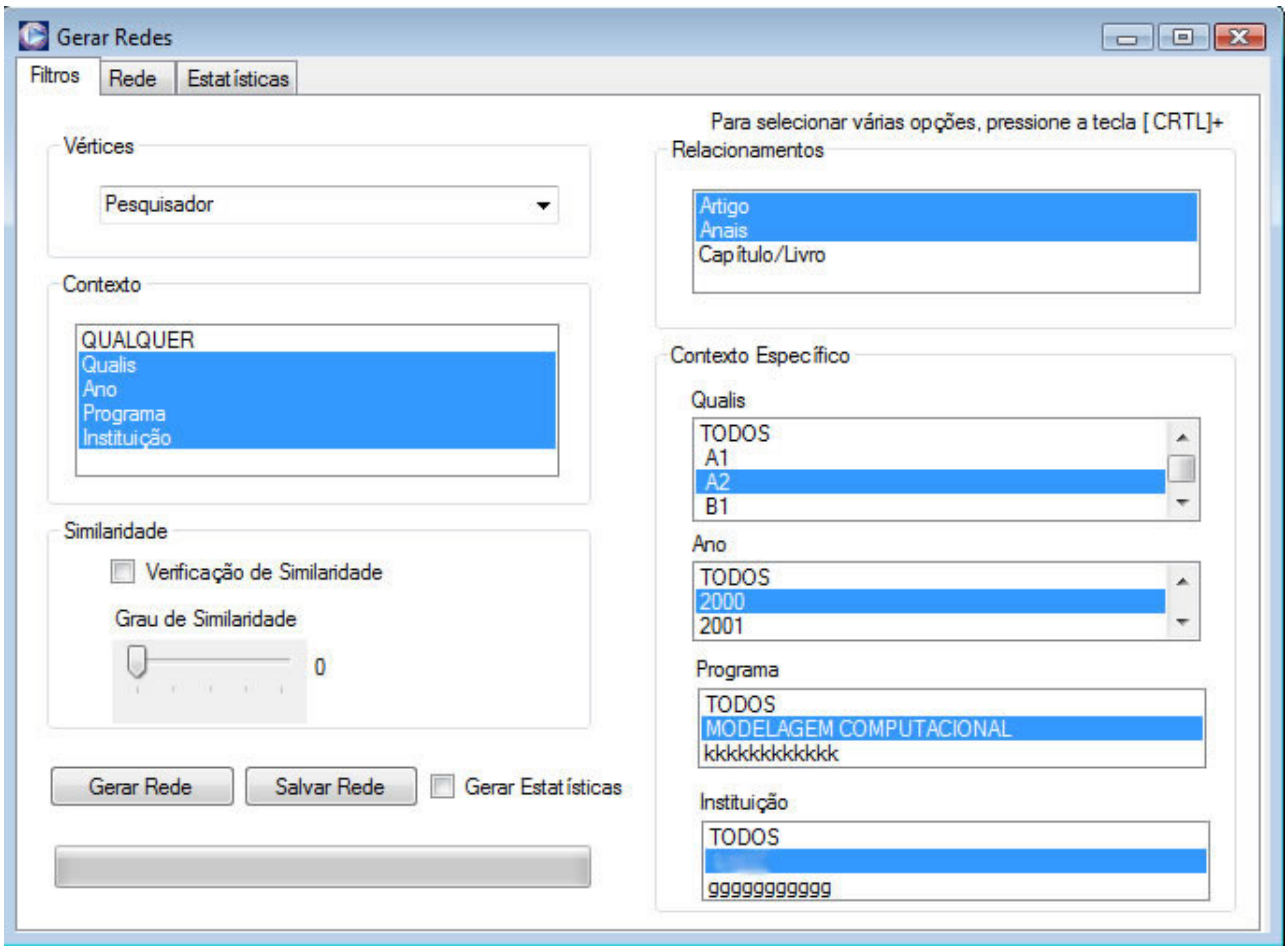


Figura 4.8: Tela para definição de filtros da rede. Fonte: Autor.

Além das etapas principais do modelo proposto, ainda há um módulo para avaliação de um PPG. Neste módulo é possível:

- Visualizar dados dos pesquisadores filtrados por ano, programa e instituição;
- Visualisar dados de quantidades de publicações dos pesquisadores por qualis;
- Calcular índices para avaliação de um PPG, a partir quantidade de publicações por ano e por triênio, com base nos pesos configurados para comissões no módulo de gestão de dados;

A Figura 4.9 mostra a tela que exibe dados de pesquisadores e quantidade de publicações em 2007.

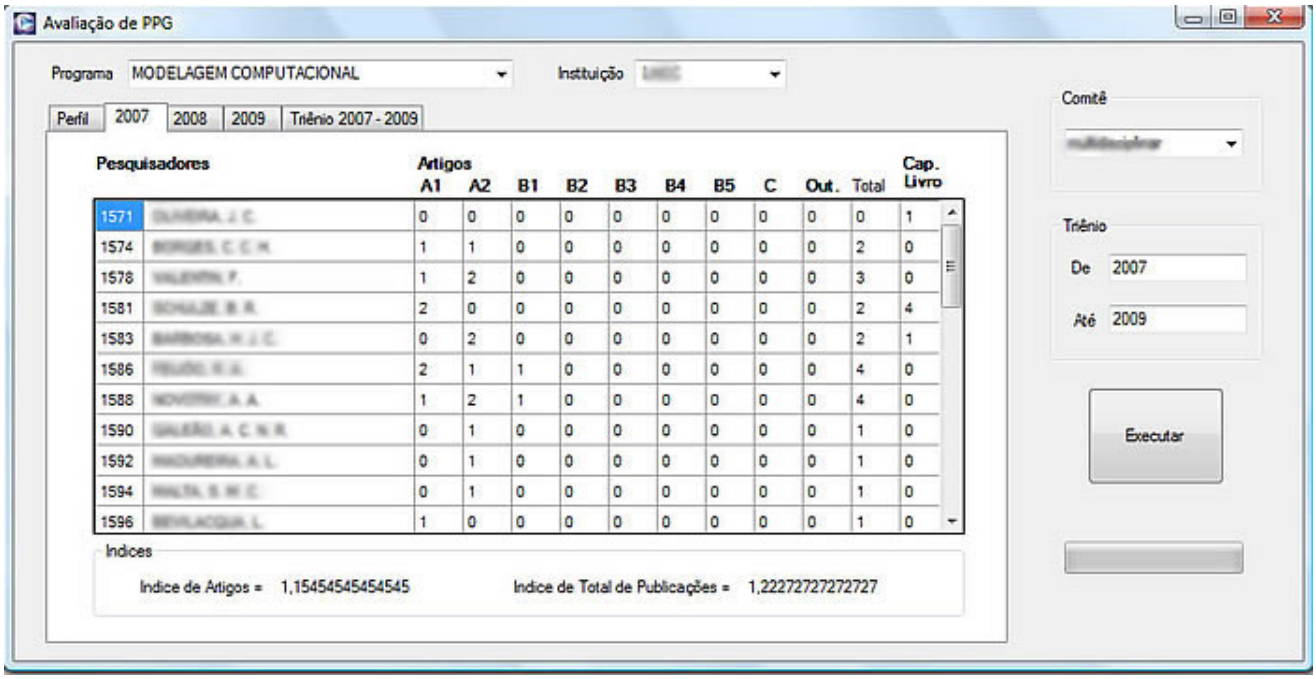


Figura 4.9: Módulo para avaliação de um PPG.

O software contém outras funcionalidades complementares como o cálculo de estatísticas descritivas, onde é possível visualizar o percentual de participação de pesquisadores por tipo de vínculo e tipo de publicações. Permite também visualizar esses dados em um gráfico, sendo para isso necessário habilitar a opção "Gerar Estatísticas".

Além desta função, há também a verificação dos pesquisadores por similaridade. Esta funcionalidade se justifica pelo fato de que em algumas publicações, o mesmo pesquisador é escrito de forma diferente, sendo assim, poderá ocorrer redundâncias nas informações contidas no banco de dados. Por exemplo, "**ALMEIDA, R.C.C.**" e "**ALMEIDA, R.C.**".

Na mineração de texto, o algoritmo interpreta essas informações como dois valores diferentes. Na construção das redes entretanto, habilitando o verificador de similaridade, é possível identificar nomes similares com base no grau de similaridade que se deseja verificar. Esta verificação funciona conforme testes de caracteres em ordem de vetor. A Figura 4.10 demonstra o resultado de uma construção de rede. Na lista ao lado, está a rede de colaboração científica onde os vértices são os autores e as arestas são as publicações em co-autoria.

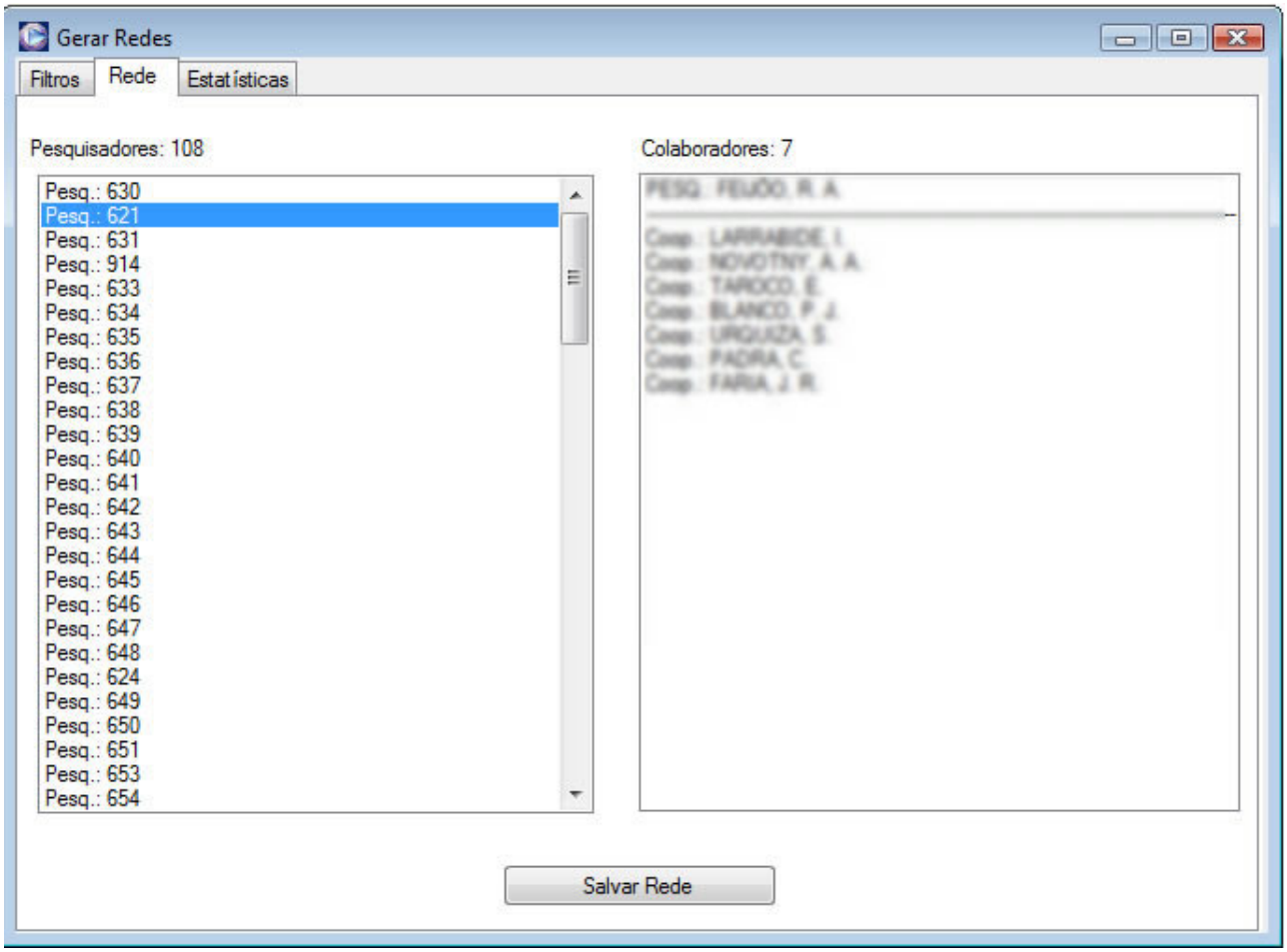


Figura 4.10: Tela de exibição dos componentes (vértices) das redes geradas. Fonte: Autor.

Conforme mostra a Figura 4.10, ao selecionar um pesquisador à esquerda, a lista de autores que colaboraram com aquele pesquisador(neste exemplo, rede de artigos) é exibida no quadro branco a direita. Cada artigo tem vinculação a um ou mais pesquisadores e a interseção de vários artigos compreende as redes individuais de co-autores. Desta forma são constituidas as redes.

Após construir a rede, salvá-la no formato .NET, é possível executar este arquivo gerado no PAJEK para visualização da rede, e a partir daí, extrair informações das características topológicas úteis da mesma. As Figuras 4.11 e 4.12 apresentam exemplos de redes construídas a partir do modelo proposto.

Conforme mostrado nas Figuras 4.11 e 4.12, as redes construídas podem conter uma amplitude maior ou menor a depender do que foi definido pelos filtros.

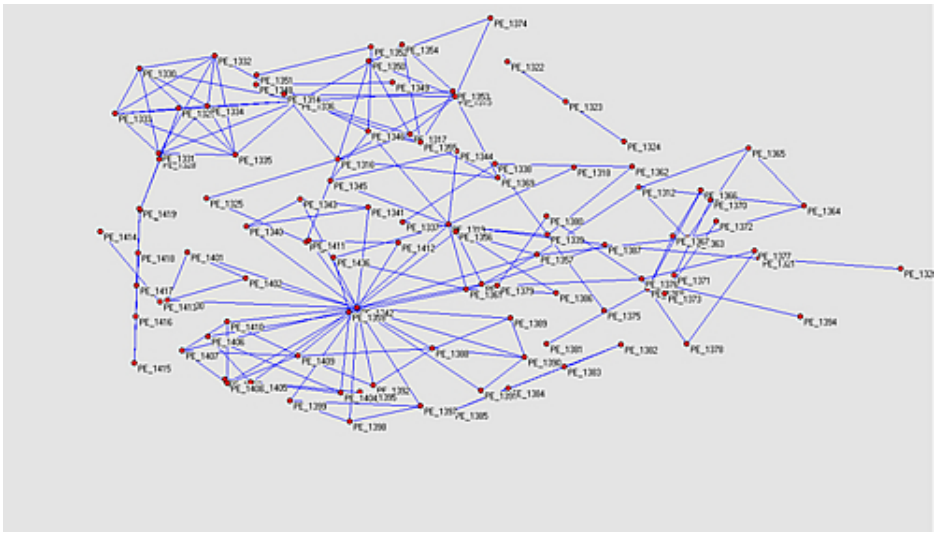


Figura 4.11: Rede de co-autoria entre pesquisadores de um PPG e a partir da publicação de artigos. Fonte: Autor.

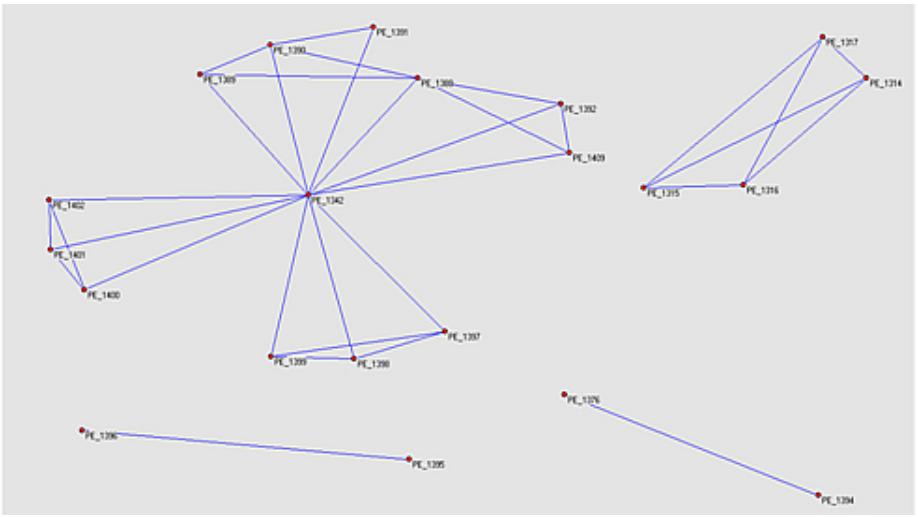


Figura 4.12: Rede de co-autoria entre pesquisadores de um PPG e a partir da publicação de artigos qualis B1. Fonte: Autor.

4.3 Arquitetura do Modelo

A arquitetura prevista para o modelo proposto se baseou em dados pertinentes ao objeto de estudo, identificados em documentos selecionados previamente, que serão a base primária de fontes de textos a serem processados: os cadernos indicadores da CAPES de publicações científicas, bancas de defesas e projetos. Foram identificadas as seguintes entidades principais:

- **Pesquisador**: Esta entidade apresenta os vértices da rede, considerando que o interesse de estudo do modelo proposto se refere ao relacionamento entre pesquisadores

em três atividades em um PPG: publicações, defesas de teses e dissertações e projetos de pesquisa;

- **Publicações (Artigos, Anais e Capítulo/Livro)**: Estas entidades compõem as conexões existentes entre os vértices da rede (Pesquisador). É referente a existência de relacões de co-autoria entre pesquisadores em publicações de artigos, anais ou capítulos.;
- **Defesa**: Além de publicações científicas, outra entidade identificada foi a defesa, de dissertação de mestrado e de tese de doutorado. Neste caso, a relação construída se baseia nos membros participantes da banca examinadora, já que os mesmos também são pesquisadores da comunidade científica;
- **Projeto de Pesquisa**: Esta entidade apresenta outro tipo de vínculo entre pesquisadores, que são os projetos. Projetos em geral têm muitas associações de pesquisadores, normalmente em competências multidisciplinares. Pesquisadores podem estar vinculados a um ou mais Projetos de Pesquisa;
- **Programa**: Esta entidade é referente ao PPG a que os pesquisadores estão vinculados. Cada entidade é representada por um programa, instituição e ano. Pesquisadores podem estar associados a um ou mais Programas.

Além das entidades selecionadas, existem as tabelas de relacionamentos existentes entre as entidades que constituirão as arestas das redes geradas. A Figura 4.13 representa a arquitetura conceitual do modelo proposto:

Conforme representado na Figura 4.13 um pesquisador pode estar associado a um ou mais artigos, anais, capítulos e livros, programas, defesas e projetos. As publicações (artigos, anais, capítulos e livros), projetos e defesas relacionam os pesquisadores participantes em suas autorias. Esses dados compõem a estrutura da rede, sendo os pesquisadores os vértices e as outras entidades (publicações, defesas e projetos) as arestas. Há ainda outras informações adicionais destas entidades, representadas por outros relacionamentos, como anais ligados a eventos, artigos ligados a periódicos, capítulos ligados a livros, que a princípio não têm relevância na construção das redes, porém futuramente poderão representar outros componentes destas.

4.4 Análise e modelagem do software

O modelo proposto de uma forma ampla visa minimar o trabalho manual de coleta de dados (i.e. a complexidade do processo) e maximizar a quantidade de informações extraídas, necessárias para construção das redes a partir dos documentos digitais selecionados. Com

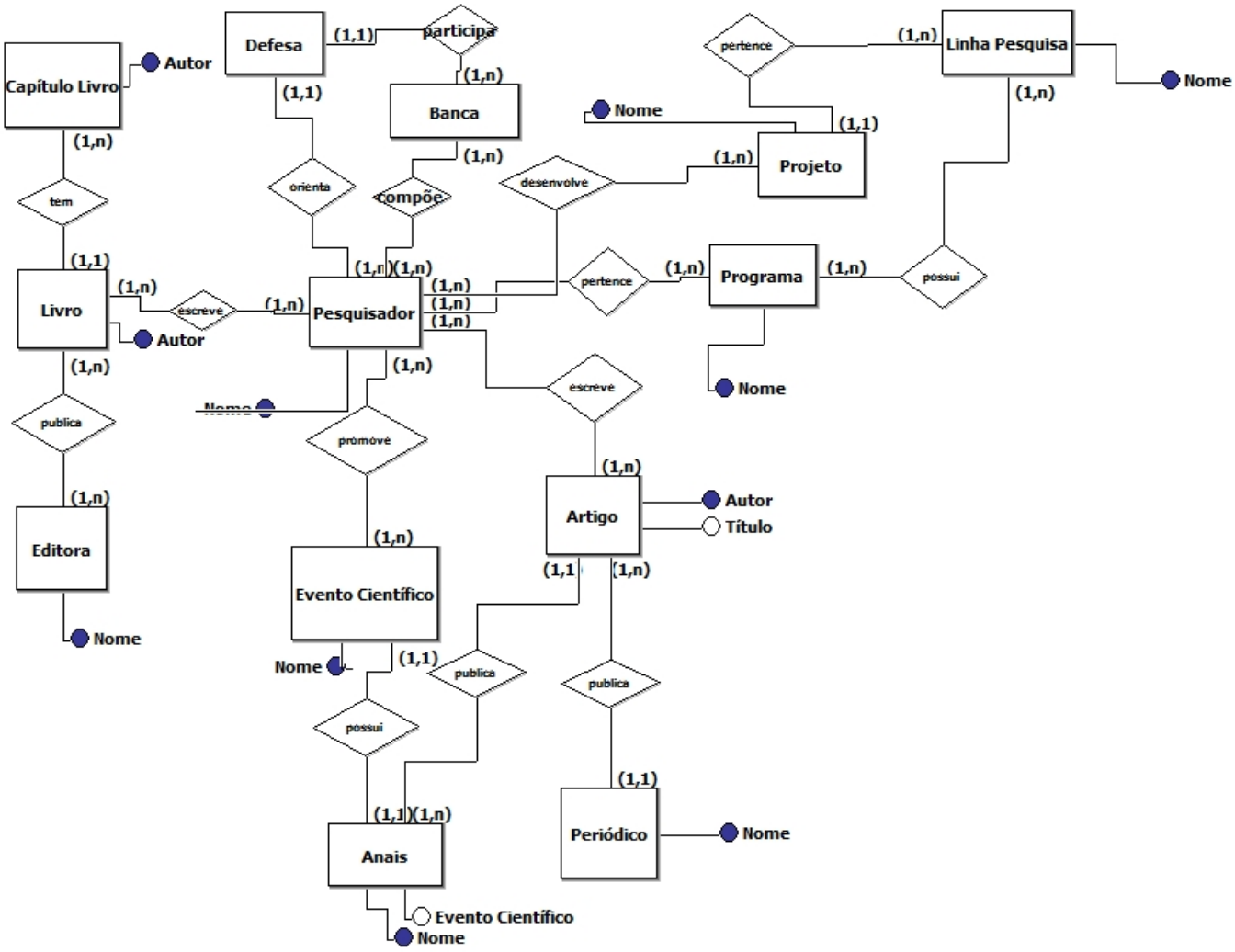


Figura 4.13: Arquitetura Conceitual do Modelo Proposto. Fonte: Autor.

esse intuito foram levantados os seguintes requisitos necessários ao desenvolvimento do software:

- Requisitos funcionais;
  - Prover um método de extração automática de informações específicas sobre pesquisadores (nome e vínculo) além de dados das tarefas acadêmicas (artigos, anais, capítulos e livros, defesas, projetos) de textos digitais, de modo facilitar a manipulação destes dados e evitar a utilização de algoritmos de mineração de textos que, de forma geral, são complexos;
  - Criar um método automático para criação dos relacionamentos dos pesquisadores à suas respectivas tarefas acadêmicas e os programas de pós-graduação que trabalham;

  - Permitir a gravação de padrões para reutilização em outros documentos a serem processados;
  - Prover uma área para correção dos dados filtrados pela mineração do texto, haja vista que a busca por padrões podem trazer ruídos nos resultados alcançados;
  - Construir redes sociais e complexas de um ou dois modos, a partir dos relacionamentos encontrados entre os pesquisadores nas diversas tarefas acadêmicas;
  - Definir filtros para delimitar o escopo da rede a ser construída;
  - Possibilitar gravação da rede em um formato que seja executável no PAJEK;
  - Criar uma área de gestão dos dados já inseridos no banco onde seja possível:
    * Apagar todos os registros, realizar backup e restauração do banco de dados;
    * Cadastrar, Excluir, Consultar, Editar e Imprimir dados das tabelas do banco (pesquisadores, artigos, anais, etc.);
    * Apagar todos os registros, realizar backup e restauração de tabelas específicas do banco de dados (pesquisadores, artigos, anais, etc.);

- Requisitos não-funcionais;

  - Visualizar os dados de cada pesquisador quanto a sua participação em produções, identificadas por qualis da CAPES;
  - Visualizar os dados de relacionamentos de cada pesquisador com outros pesquisadores dentro do contexto de cada rede;
  - Visualizar e gravar dados estatísticos descritivos das redes construídas, a partir dos quais pode se obter a percentagem de participação de pesquisadores por tipo de vínculo (docente, discente, etc.) e por tipo de produção (publicações, defesas, etc.);
  - Permitir que sejam gravados logs de execução;

A seguir, são apresentados três tipos de diagramas para representar o modelo proposto: O diagrama de caso de uso, o diagrama de sequência e o diagrama de classes. Existem dois níveis de usuário: O usuário comum, que esta representado na Figura 4.14 e o usuário administrador, o qual possui login e senha de acesso à area de gestão do software, que está representado pela Figura 4.15.

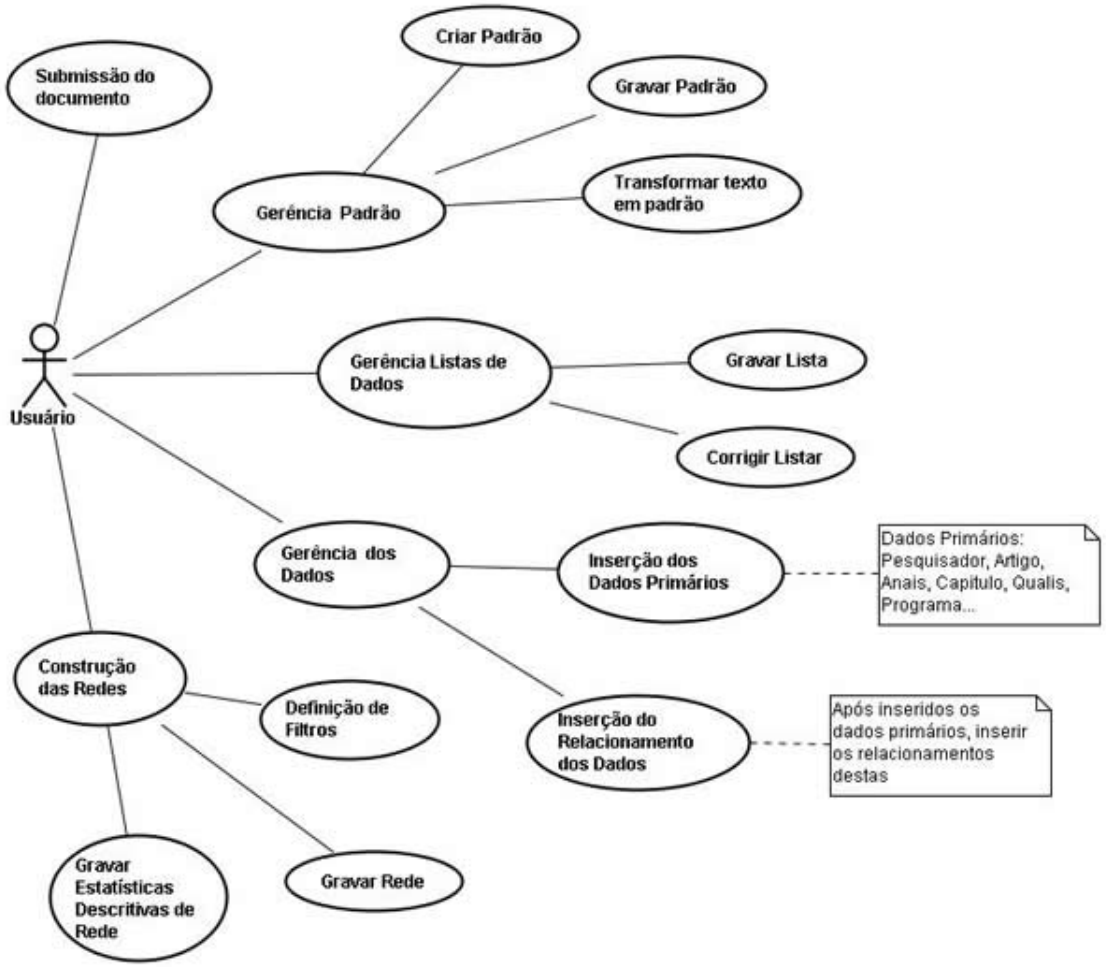


Figura 4.14: Diagrama de Caso de Uso - Usuário. Fonte: Autor.

As tarefas básicas que poderão ser executadas pelo usuário são:

(a) Submissão do documento: O documento que será minerado deverá ser submetido ao modelo proposto (são aceitas as extensões PDF, TXT e HTML);

(b) Gerência de Padrões: Neste processo são criadas as expressões que serão utilizadas na busca dos dados desejados. Uma das funções que podem ser realizadas neste processo é a transformação de pequenos trechos de texto em expressão regular. As expressões poderão ser amazenadas para reutilização em outros textos;

(c) Gerência de Listas de Dados: Os resultados obtidos da pesquisa por padrões deverão ser armazenados em listas de dados (e.g. lista de autores, lista de artigos, etc.). Neste processo, é realizada a limpeza e correção dos dados antes de inseri-los no banco de dados;

(d) Gerência de Dados: Uma vez os dados todos corrigidos e limpos, os mesmos devem ser inseridos no banco. Esta etapa se divide em duas tarefas: inserção dos dados

primários e inserir dados de relacionamentos. A primeira se refere a inserir os dados nas tabelas básicas (e.g. Pesquisador, Artigos, Anais, etc.). A segunda se refere a inserção dos dados que associam esses dados primários nas tabelas relacionais (Pesquisador/Artigo, Pesquisador/Anais, etc.);

(e) Construção das redes: O módulo para construção das redes dispõe de uma tela onde há filtros a serem definidos para delimitação do escopo da rede. Após ser gerada a rede, grava-se o arquivo .NET. Ainda é possível habilitar a opção para gerar estatísticas descritivas da rede, onde um arquivo contento estas informações poderá ser gravado.

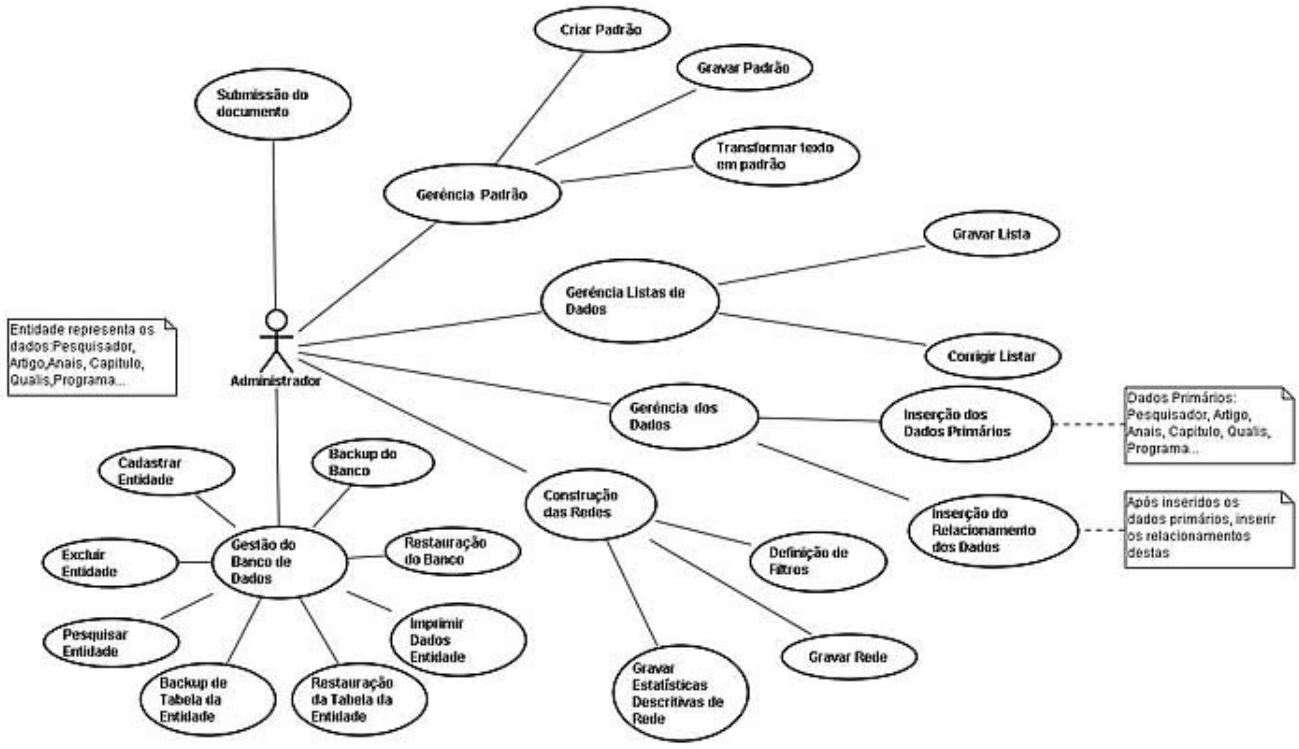


Figura 4.15: Diagrama de Caso de Uso - Administrador. Fonte: Autor.

No caso de o usuário possuir permissão de administardor, ele possuirá alguns privilégios a mais que o usuário comum, conforme mostrado na Figura 4.15. A gestão do banco de dados é uma tarefa própria do administrador. Neste módulo, é possível manipular todos os dados que já foram inseridos no banco, tanto dados primários (Pesquisador, Artigo, Anais, etc.) quanto os dados de relacionamentos (Pesquisador/Artigo, Pesquisador/Anais, etc.). As funções básicas que podem ser executadas são: Cadastro, Edição, Exclusão, Pesquisa e Impressão dos dados de cada entidade. Além das funções básicas de gestão de dados, há ainda a possibilidade de realizar *Backup* e Restauração de todo o banco de dados e de cada tabela deste banco individualmente.

A arquitetura de dados do modelo prosposto está representado no diagrama de classes apresentado na Figura 4.16.

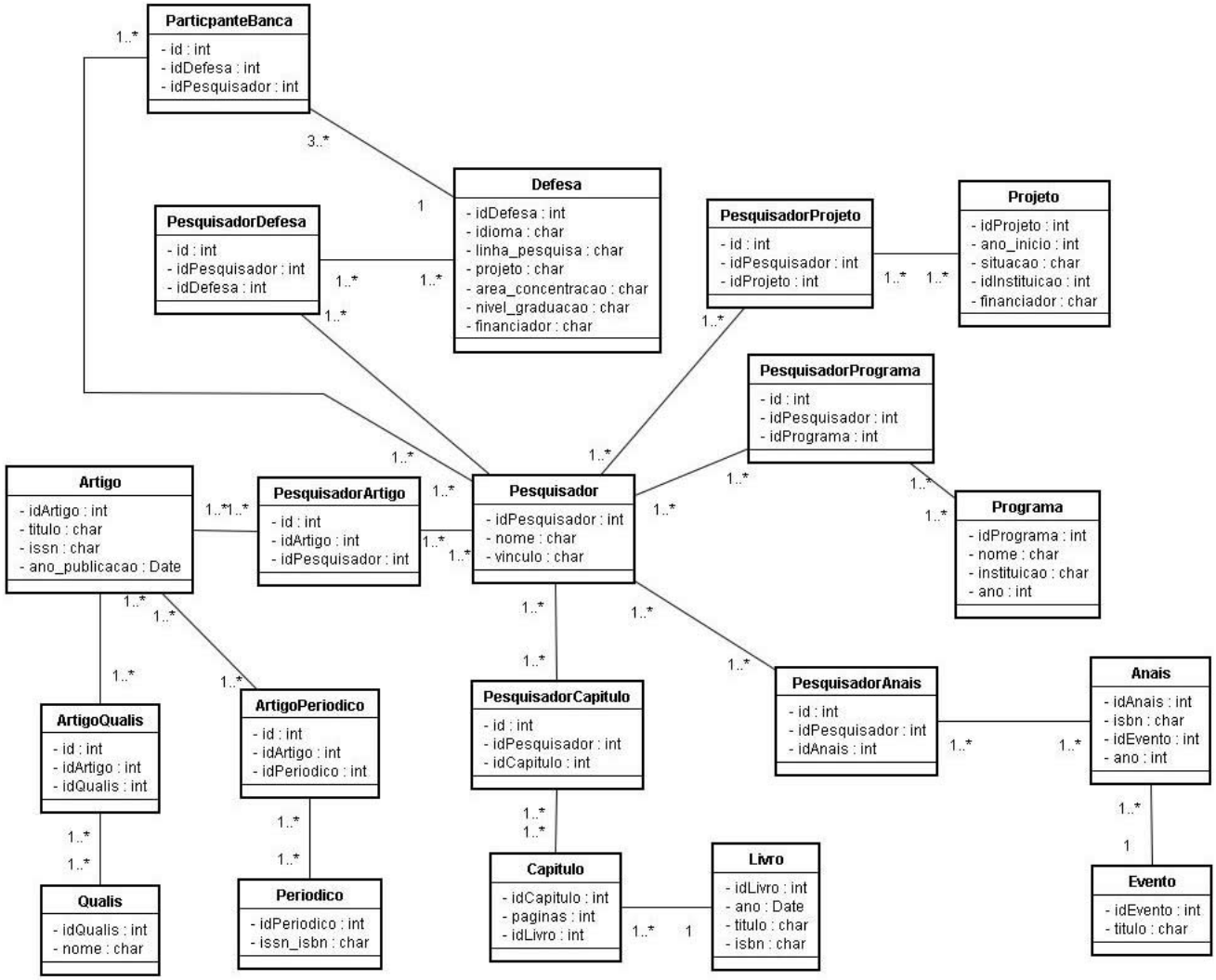


Figura 4.16: Diagrama de Classes. Fonte: Autor.

Conforme mostrado na Figura 4.16, as entidades principais são definidas por: “Pesquisador”, “Artigo”, “Anais”, “Capítulo”, “Livro”, “Defesa”, “Projeto”, “Programa” e “Qualis”. Estas entidades são principais porque são necessárias à constuição da estrutura da rede. As outras entidades (Evento e Periódico) são informações adicionais da rede. Há ainda as entidades que representam os relacionamentos entre as entidades principais, e são definidas por: “Pesquisador/Artigo”,“Pesquisador/Anais”, “Pesquisador/Capítulo”, “Pesquisador/Programa”, “Artigo/Qualis”, “Anais/Qualis, “Pesquisador/Defesa” e “Pesquisador/Projeto”. Na construção das redes, as entidades principais constituirão os vértices e as entidades de relacionamento as arestas existentes entre estas entidades.

A arquitetura dos dados mostrados na Figura 4.16, foi construída com base no objetivo específico desta pesquisa, que são as redes de colaboração científica. Entretanto, será possível criar outras redes, de outros textos além dos cadernos de indicadores da CAPES, que não sejam apenas estas redes de pesquisadores participantes em co-autoria. A estrutura dos dados apenas teria sua nomenclaturas alteradas para nomes mais genéricos, entretanto, as hierarquias relacionais entre as entidades se manteria. Por exemplo, em vez de usar o nome da entidade "Pesquisador", poderia se utilizar um nome genérico "Pessoa".

As ações executadas no processo de mineração do texto, inserção dos dados e geração das redes por um usuário comum são mostradas na Figuras 4.17, 4.18 e 4.19, respectivamente .

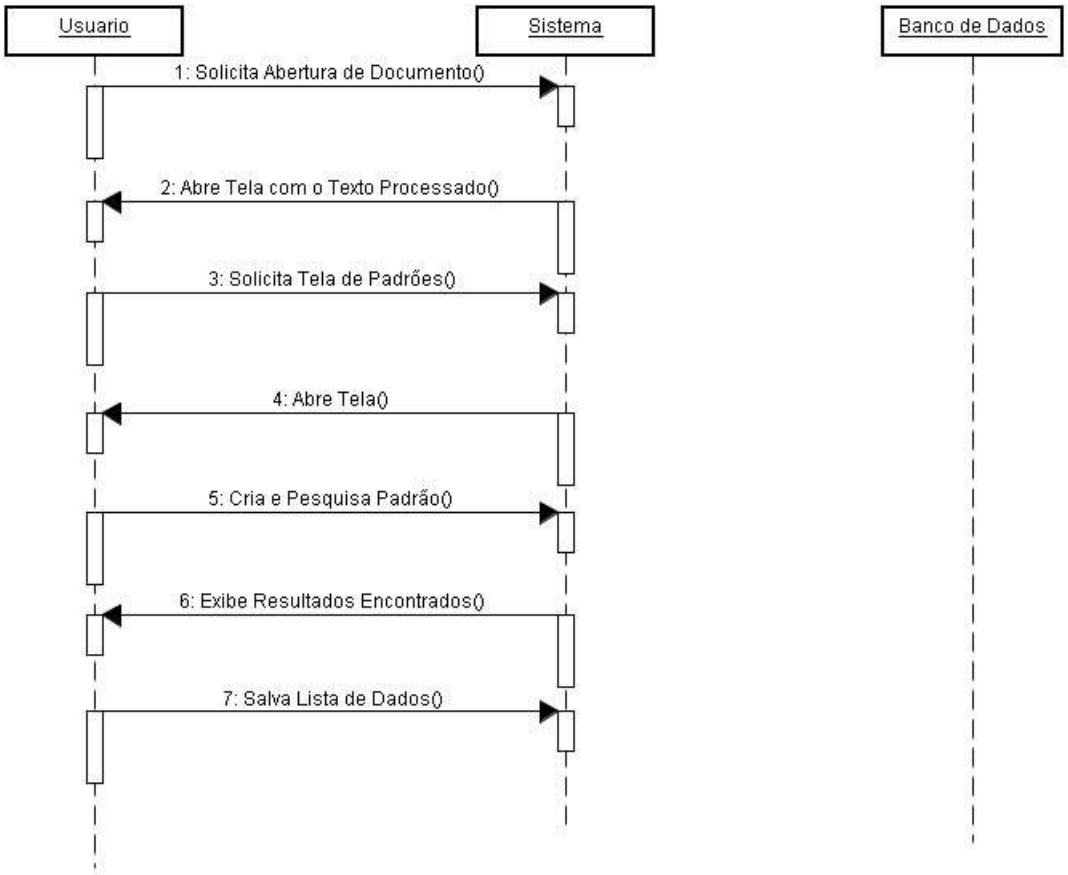


Figura 4.17: Diagrama de Seqüência - Etapa 1 - Mineração de Textos. Fonte: Autor.

A etapa da mineração do texto, representada pela Figura 4.17, consiste em basicamente em três tarefas: abrir o documento desejado, criar padrões para pesquisa no texto e salvar os resultados encontrados como listas de dados. Ao final destas tarefas, serão obtidos os dados que deverão ser inseridos no banco na próxima etapa: a inserção dos dados (Figura 4.18).

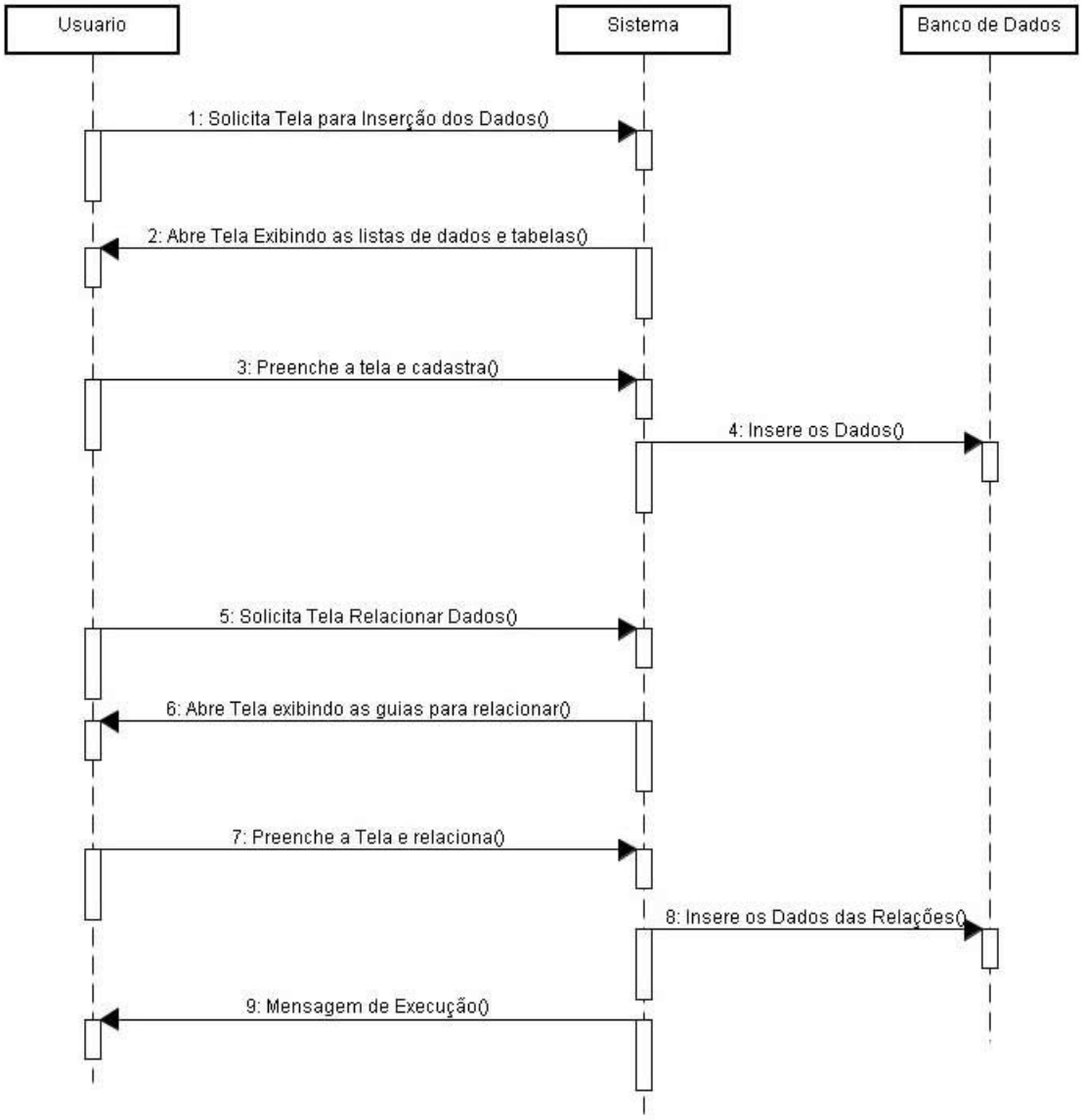


Figura 4.18: Diagrama de Seqüência- Etapa 2 - Inserção de Dados. Fonte: Autor.

A etapa de inserção dos dados, representada pela Figura 4.18, é dividida em duas tarefas: a inserção de dados primários e inserção dos relacionamentos entre os dados. Na primeira tarefa, seleciona-se a lista de dados que se deseja inserir no banco, e a tabela que receberá esses dados (e.g. lista de dados " lista de autores" e a tabela "Pesquisador", onde os dados da lista serão inseridos). Este processo se repete para cada lista de dados criada.

Na segunda tarefa, já com os dados primários inseridos (Pesquisador, Artigos, etc.), serão criadas as relações entre estes dados. Neste módulo de relacionamento de dados, há três tipos de relacionamentos, indicados pelas seções "Geral", "Por Qualis" e "Por Programa". É necessário criar todos os tipos de relações para a construção das redes de forma mais completa. As seções contém informações sobre o tipo de relação a ser criada e dados de listas onde devem ser criados relacionamentos (e.g. lista de autores e lista de artigos). Ao

término desta etapa, todos os dados que foram incluídos no banco de dados constituirão as redes na próxima etapa do processo: construção das redes (Figura 4.19).

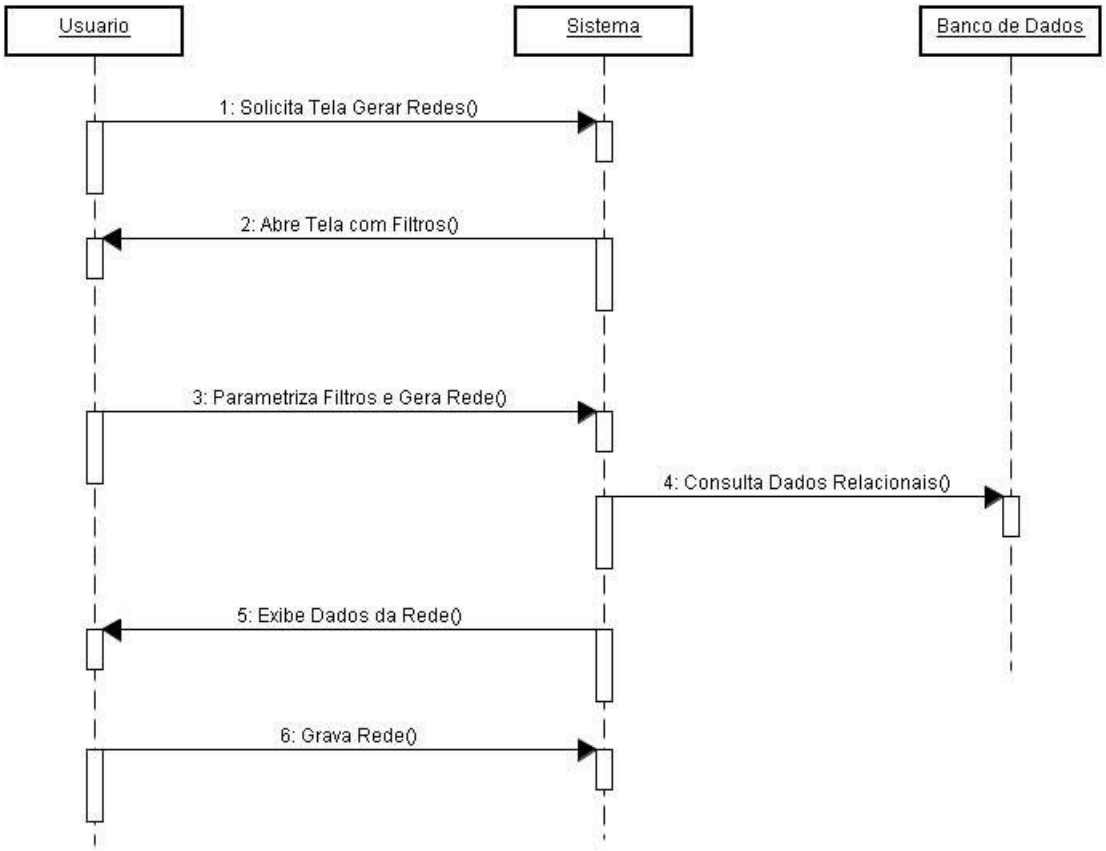


Figura 4.19: Diagrama de Seqüência- Etapa 3 - Construção das Redes. Fonte: Autor.

A terceira e última etapa do modelo proposto, a construção das redes, se resume em duas tarefas: a definição de filtros e gravação dos dados. Ao solicitar este módulo de redes, filtros serão exibidos para delimitar o escopo das redes. A primeira tarefa é parametrizar estes filtros e executar a geração da rede. A segunda tarefa é a gravação da rede resultante deste processo, em formato .NET para estudo posterior das mesmas.

As ações representadas nos diagramas de sequência (Figuras 4.17, 4.18 e 4.19) são realizadas por qualquer usuário do modelo proposto. Entretanto, ainda há outras ações passíveis de serem realizadas por um usuário com maior nível de permissão, os administradores. O diagrama da Figura 4.20 corresponde a ações realizadas apenas por usuários que possuem login e senha administrativa, que dá acesso ao módulo de Gestão do Banco de Dados.

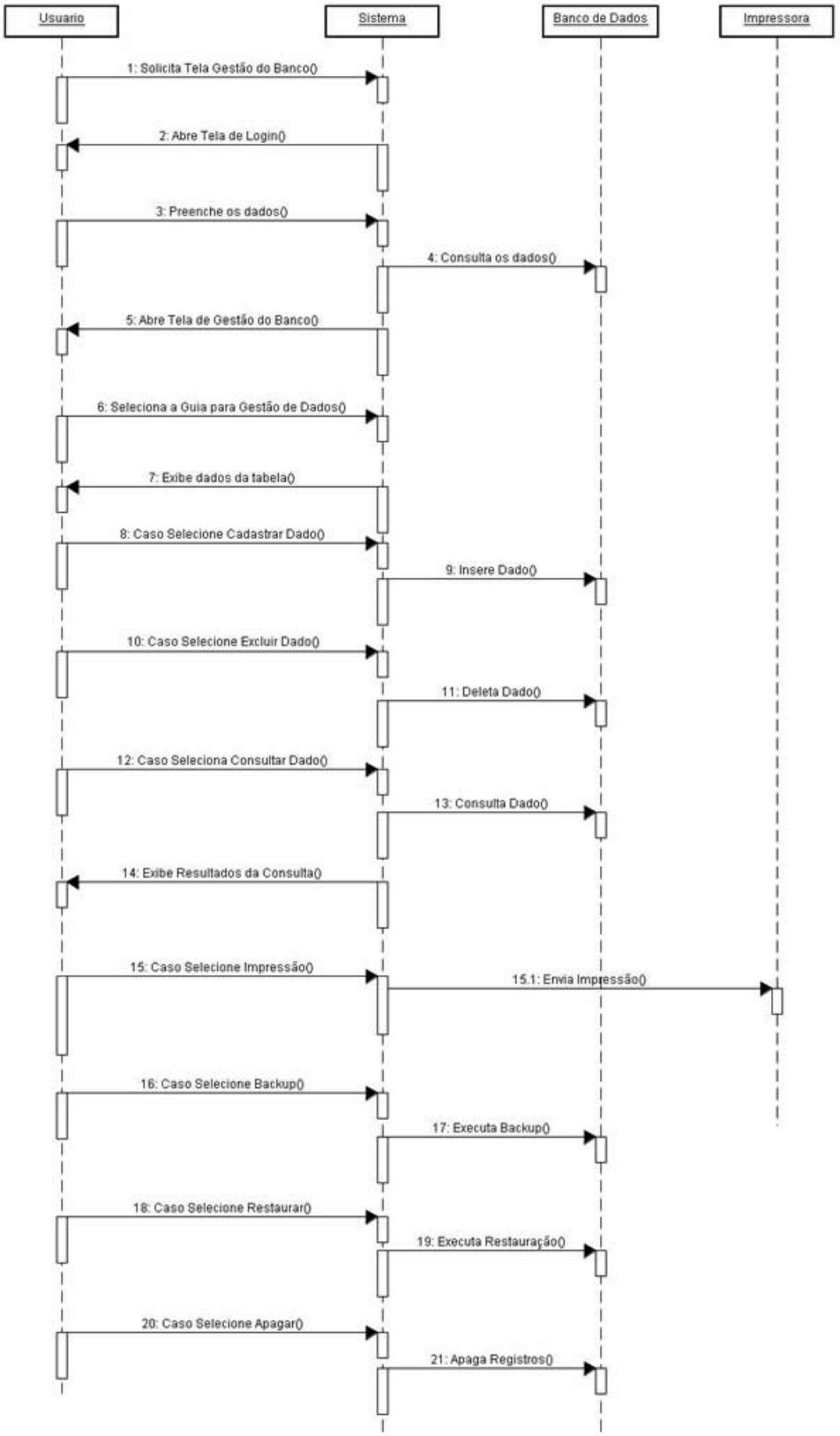


Figura 4.20: Diagrama de Seqüência - Gestão do Banco de Dados. Fonte: Autor.

No módulo de gestão do banco de dados, há dois tipos de manipulação dos dados: a manipulação do banco de dados total e a manipulação dos dados por tabela do banco. O

primeiro tipo se refere a manutenção do repositório inteiro, podendo ser feitos backups e restaurações do banco, verificação e correção da integridade do banco e alteração dos dados de acesso a este módulo.

- *Backup* do Banco: Seleciona-se um diretório onde deve ficar gravado o *backup*, define-se um nome para este arquivo e executa o processo;
- Restauração do Banco: Seleciona-se um arquivo de backup e executa a restauração do banco;
- Corrigir Banco: Verifica-se a integridade do banco e , se houver erros, o banco é corrigido;
- Alterar: Alteração de registro de acesso ao módulo de gestão;

O segundo tipo é referente a manipulação de dados de uma única entidade (eg.: tabela Pesquisador). Neste caso, os usuários podem realizar as seguintes ações:

- Cadastro: Inserção de registros no banco de forma manual, contendo dados necessários a cada entidade;
- Exclusão: Eliminação de registros do banco;
- Pesquisa: Consultar registros no banco, utilizando filtros de consulta;
- Edição: Alteração de registros;
- Impressão: Lista os registros existente em cada tabela, e prepara-os para impressão;
- Backup da tabela: Cria uma cópia de segurança apenas de uma referida tabela;
- Restauração da tabela: Restaura todos os dados apenas de uma referida tabela;
- Apagar todos os registros: Apaga todos os dados apenas de uma referida tabela.

Capítulo Cinco

Trabalho Experimental

Para avaliação do modelo proposto, o experimento é realizado utilizando o conjunto de cadernos indicadores da CAPES relativos às publicações (artigos, anais de eventos, livros, defesas de dissertação e teses e projetos de pesquisa), para alguns programas de pós-graduação em funcionamento. Dos cadernos de indicadores, foram selecionados documentos de 2007 e 2008. Estes cadernos contêm informações de tarefas acadêmicas em PPGs, em disposição de textos organizados. Estes textos seguem uma organização formal de representação das informações sobre as tarefas acadêmicas. Por exemplo, são listadas publicações, separadas por classificação qualis do periódico. Cada publicação listada tem nome de autores (no caso de 2007 e 2008 todas as letras estavam em caixa alta e separados por ponto e vírgula), seguidas pelo ano da publicação, título da publicação, entre outros dados.

A cada triênio, foram observadas poucas alterações ocorridas em relação ao posicionamento dos dados e formatação gráfica das palavras. Por exemplo, em 2003 a formatação do nome dos autores tinha letras maiúsculas e minúsculas, e em 2007 a formatação era composta apenas por letras maiúsculas, como em "Patricia, F. B." e "PATRICIA, F.B.". Isso exigiu adequações das expressões regulares para cada caderno. Para experimentação do modelo entretanto, será utilizado apenas os cadernos de 2007 e 2008.

5.1 Experimento : cadernos indicadores da CAPES

Observando-se a disposição das informações e os padrões textuais, nos anos de 2007 e 2008, não houve alterações, facilitando o trabalho pelo reuso dos padrões utilizados em ambos os cadernos. A Figura 5.1 mostra um exemplo de como as informações estão dispostas nos cadernos de indicadores da CAPES de um PPG em 2007.

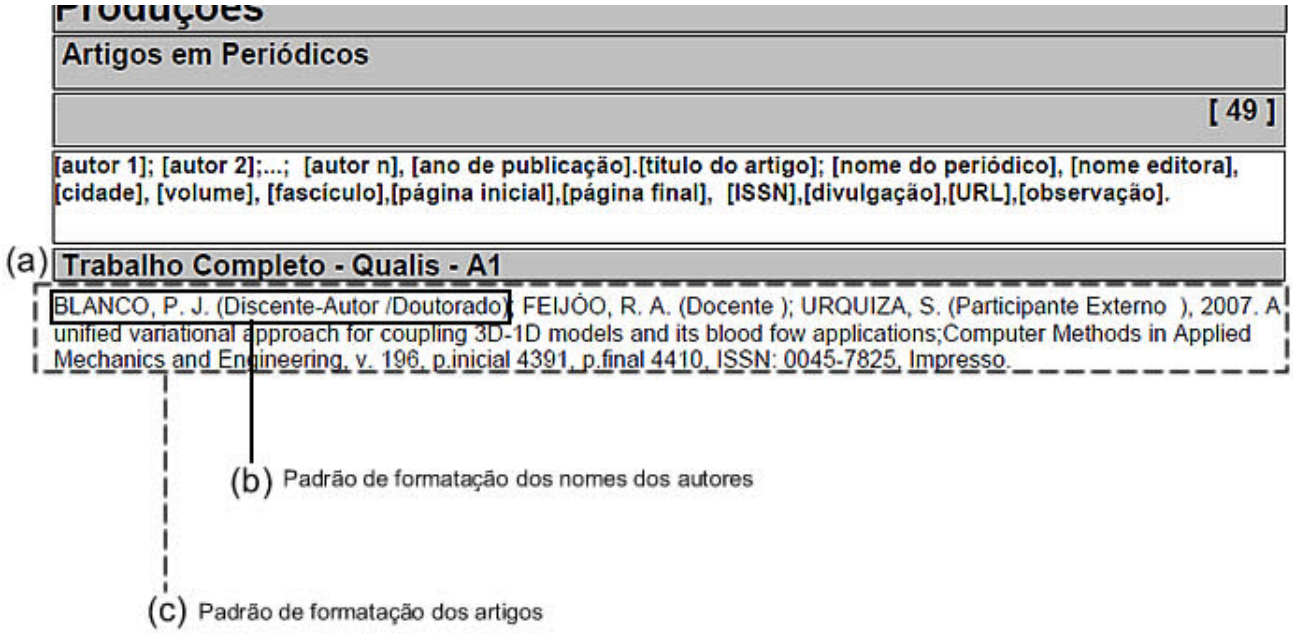


Figura 5.1: Excerto de um dos Cadernos de Indicadores da CAPES de um PPG em 2007.

Na Figura 5.1 podemos ver algumas das informações que serão extraídas por expressões regulares. A indicação do item (a), pode ser considerado como um dos intervalos para seleção de publicações. Todos os títulos que separam as publicações por qualis têm a formatação "Trabalhos Completos - Qualis - ...". Ao se criar as listas de dados de publicações por qualis, deve-se indicar o intervalo de busca da expressão. Por exemplo, todos os artigos qualis A1 estão compreendidos entre o título "Trabalhos Completos - Qualis - A1" e "Trabalhos Completos - Qualis - A2". Neste caso, poderia ser criada uma lista de artigos qualis A1. Indicada pelo item (b), a formatação para nome de autores segue o padrão:

"BLANCO, P. J.(Discente-Autor/Doutorado)".

As letras do nome do autor são todas maiúsculas. O sobrenome é separado por vírgula e espaço antes do nome representado por uma letra e um ponto, seguido de um parêntesis. Dentro do parêntesis se aceita caracteres maiúsculos e minúsculos, além de outros símbolos como o /. Fecha parêntesis. Ao se criar um padrão para formação de lista de dados com os nomes dos autores, deve-se observar estas características das informações. Todo autor que esta listado neste caderno possui esta forma escrita.

Nem toda informação apresentada nos cadernos, será necessária para se construir as redes. Porém, quanto mais informações forem recolhidas do texto, mais dados úteis serão armazenados para aplicações futuras de técnicas de mineração, a exemplo da mineração de

dados, para se descobrir outros conhecimentos implícitos no banco de dados. A princípio, são dados obrigatórios para se construir as redes:

- [autor] - os vértices da rede são identificados e associados entre si pelos nomes dos autores;
- [ano da publicação] - Na geração da rede, pode-se delimitar o escopo da rede de acordo ao ano. Sendo assim, sem essa informação não há como classificar redes segundo o ano de um programa de uma instituição;
- [publicações(artigos, anais e capítulos)] - A verificação da presença de autores em publicações ocorre pela presença do nome deste na publicação;
- [projeto] - A verificação da presença de autores em projetos de pesquisa ocorre pela presença do nome deste no projeto;
- [defesa] - A verificação da presença de autores em defesas de dissertações e teses ocorre pela presença do nome deste na dissertação ou tese de defesa;

Além dos dados citados acima, ainda é necessário identificar o programa, a instituição e ano das publicações, porque as publicações estão associadas aos pesquisadores, que por sua vez estarão associados aos programas.

Outro dado relevante do texto é a classificação Qualis, porque conforme foi observado, em determinadas épocas há mudança na nomenclatura das classificações. O Qualis é um conjunto de procedimentos utilizados pela CAPES para estratificação da qualidade da produção intelectual dos PPGs. Nestes dois cadernos selecionados, não houve mudanças neste aspecto. Por exemplo, até o ano de 2006, as classificações Qualis eram identificadas por: I/A, I/B, I/C, L/C, e <? >. A partir de 2007, essa nomenclatura mudou para: A1, A2, B1, B2, B3, B4, B5, C e NC.

Considerando a forma gráfica dos textos, um padrão identificável para dados necessários a construção das redes foi a forma e separação dos caracteres. Por exemplo, **BRAGA, P. F. (Discente)**, foi um padrão identificado para nome dos pesquisadores nestes dois cadernos, diferente de cadernos de outros anos. A sequência de caracteres pode ser logicamente entendida por: Letras maiúsculas, separadas por virgula, seguido de outros caracteres maiúsculos separados por pontos, seguido de parêntesis, e caracteres maiúsculos e minúsculos, finalizado por parêntesis. A Tabela 5.1 mostra alguns padrões utilizados nos cadernos indicadores da CAPES de um PPG em 2007.

Tabela 5.1: Padrões criados para busca de dados nos cadernos indicadores da CAPES.

Uso	**Padrão Textual**	**Expressão Regular**
Padrão Autores/Vínculo	BLANCO, P. J. (Discente-Autor /Doutorado)	[A-ZÁ-ÚÃ-ÕÇÂ-Ô\s\-]+\,\s(?(\()—[A-Z\.\s]+) \([A-ZÁ-ÚÃ-ÕÂ-ÔÇa-zá-úã-õàç0-9\.\/\-\s]+\)
Padrão Artigos, Anais e Capítulo	BLANCO, P. J. (Discente-Autor /Doutorado); FEIJÓO, R. A. (Docente); URQUIZA, S. (Participante Externo), 2007. A unified variational approach for coupling 3D-1D models and its blood fow applications;Computer Methods in Applied Mechanics and Engineering, v. 196, p.inicial 4391, p.final 4410, ISSN: 0045-7825, Impresso.	[A-ZÁ-ÚÃ-ÕÇÂ-Ô\s]+\,\s(.)+(?(\n)[\n\r]+(.)+ |[\n\r]+(.)+)[\n\r]+(.)+
Padrão Qualis	Trabalho Completo - Qualis - A1	Trabalho Completo - Qualis - [A-Z0-9]+
Padrão Programa, Instituição e Ano	NOME DO PROGRAMA INSTITUIÇÃO - 2007	[A-ZÁ-ÚÃ-ÕÇÂ-Ô]+\s[A-ZÁ-ÚÃ-ÕÇÂ-Ô]+\s\/\s[A-ZÁ-ÚÃ-ÕÇÂ-Ô]+ \s\-\s[0-9]+
Padrão ISSN	ISSN:123456	ISSN\:\s[A-Z0-9\s]+\-[A-Z0-9\s]+\,
Padrão ISBN	ISBN:123456	ISBN [A-ZÁ-ÚÃ-ÕÂ-ÔÇa-zá-úã-õàç0-9\-\/\s]+\,

Fonte: Autor.

Estes padrões acima foram aplicados em ambos os cadernos selecionados e os resultados obtidos foram satisfatórios. Após a aplicação dos padrões criados, as listas de dados geradas foram submetidas ao procedimento de limpeza para garantir dados corretos antes da inserção no banco de dados. As listas criadas foram: lista de pesquisadores, lista de artigos, lista de anais de eventos, lista de capítulos, lista de programa, lista de periódicos e lista de Qualis.

Após a limpeza dos dados, os primeiros dados foram inseridos no banco de dados, a partir das listas de dados já tratadas. O modelo não aceita duplicidades de informação, logo, ele

faz diferença entre uma palavra com a grafia **AUTOR-X** e **AUTORX**. Por essa razão a limpeza e correção das listas gravadas foram as tarefas que demandaram maior tempo em todo o processo realizado pelo modelo proposto.

A inserção dos relacionamentos entre as entidades no banco (Pesquisador/Artigo, Pesquisador/Anais, etc.) não exige uma sequência obrigatória na realização desta tarefa. Por exemplo, não é importante qual relacionamento é feito primeiro, se o "Pesquisador/Programa", ou "Pesquisador/Artigo", desde que todos os relacionamentos possíveis sejam feitos para a construção correta das redes. Neste trabalho experimental as relações foram criadas na sequinte sequência:

1. A relação entre os Pesquisadores e o Programa são criadas;
2. As relações Pesquisadores e Artigos são criadas;
3. As relações Pesquisadores e Anais são criadas;
4. As relações Pesquisadores e Capítulos são criadas;
5. As relações Artigos e Qualis são criadas.

Todas as relações obrigatórias foram criadas, para geração das redes. As redes construídas foram definidas por filtros de Programa, Instituição, Ano, Tipo de Publicação e Qualis. Abaixo estão listadas as redes resultantes dos filtros aplicados:

Redes Gerais

Filtros : Ano (2007/2008); Instituição(PPG01); Programa (Modelagem Computacional)

- Redes de Artigos;
- Redes de Anais de Eventos;
- Redes de Capítulos de Livros;
- Redes Artigos, Anais e Capítulos.

Redes Específicas

Filtros : Ano (2007/2008); Instituição(PPG01); Programa (Modelagem Computacional); Qualis (A1,A2,B1,B2,B3,B4,B5,NC,?)

- Redes de Artigos Qualis A1;
- Redes de Artigos Qualis A2;
- Redes de Artigos Qualis B1;
- Redes de Artigos Qualis B2;
- Redes de Artigos Qualis B3;
- Redes de Artigos Qualis B4;
- Redes de Artigos Qualis B5;
- Redes de Artigos Qualis NC;
- Redes de Anais de Eventos Qualis ? (Sem classificação definida);
- Redes de Anais de Eventos Qualis NC;

Nas Figuras 5.2, 5.3, 5.4, 5.5, 5.6 e 5.7 estão representadas as redes construídas, visualizadas com auxilio do software PAJEK.

Redes de Colaboração Científica - PPG01 - Modelagem Computacional -2007

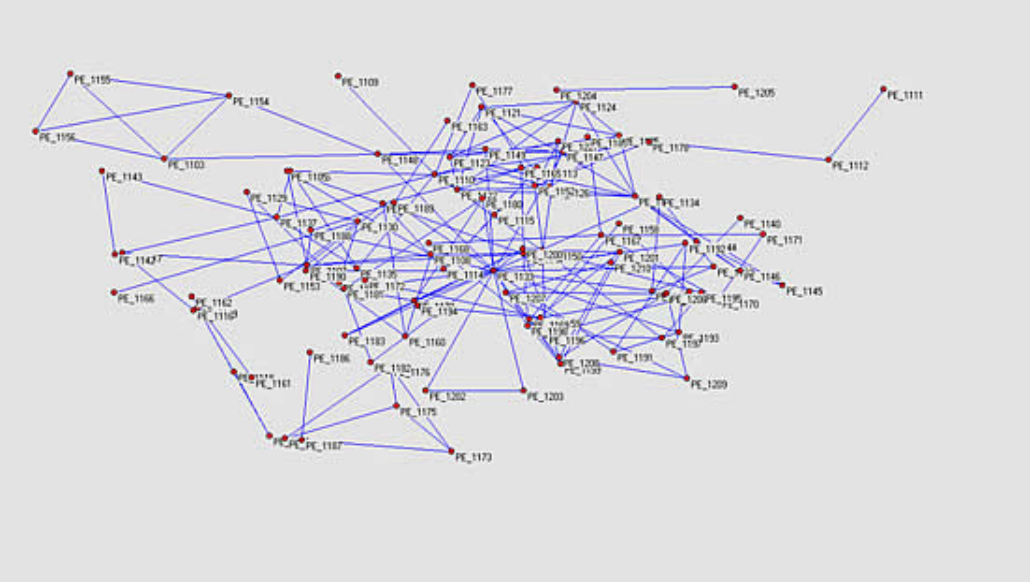


Figura 5.2: Redes de Artigos, todos os Qualis, de 2007. Fonte: Autor.

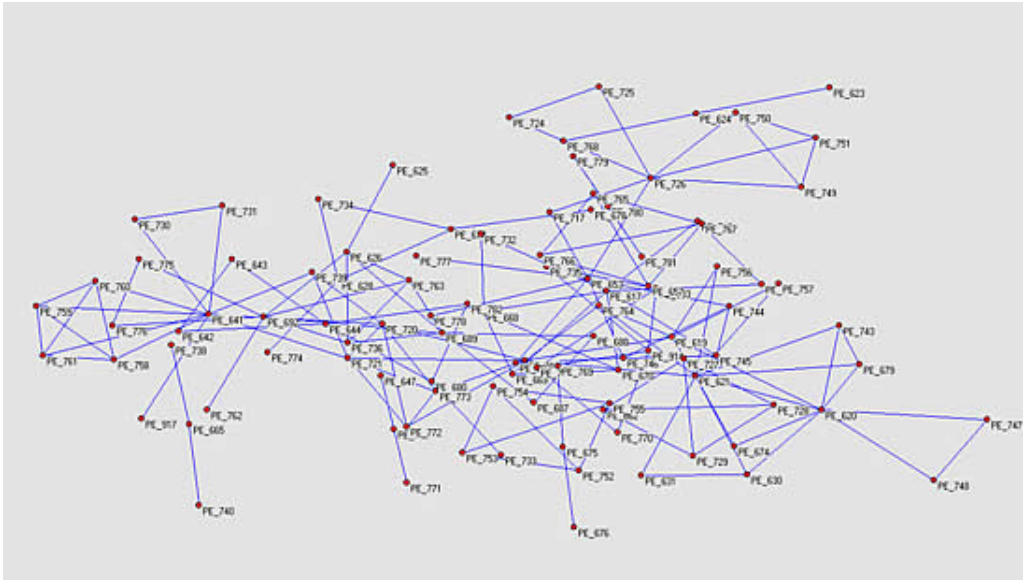


Figura 5.3: Redes de Anais de Eventos de 2007. Fonte: Autor.

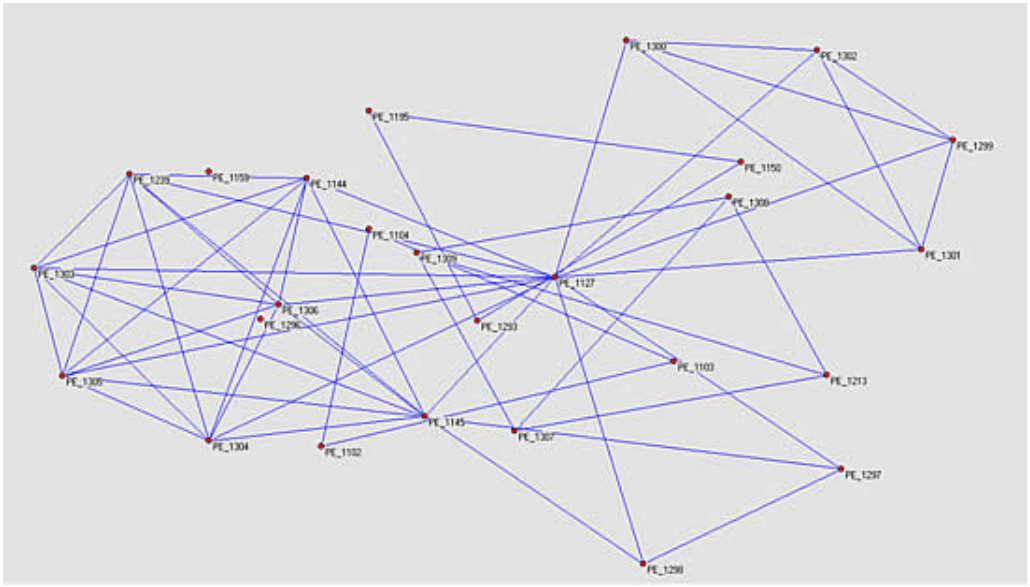


Figura 5.4: Redes de Capítulos em Livros 2007. Fonte: Autor.

Redes de Colaboração Científica - PPG01 - Modelagem Computacional -2008

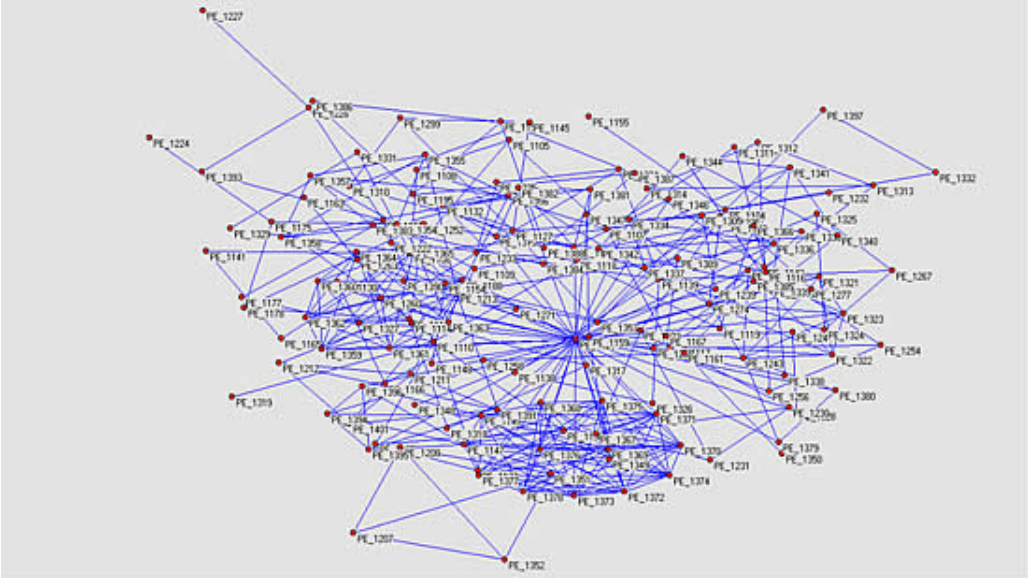


Figura 5.5: Redes de Artigos, todos os Qualis, de 2008. Fonte: Autor.

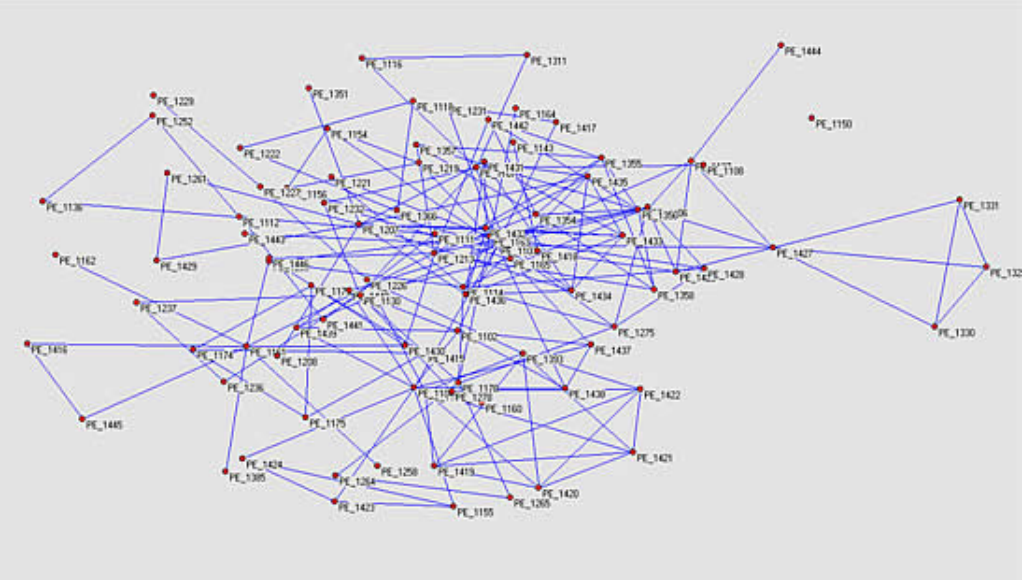


Figura 5.6: Redes de Anais de Eventos de 2008. Fonte: Autor.

As redes construídas por Qualis podem ser vistas no Apêndice A. A partir das redes construídas, dados das características topológicas de redes complexas foram calculadas, afim de subsidiar as análises das relações de colaboração científica. As Tabelas 5.2 e 5.3 mostram as propriedades identificadas nas redes geradas.

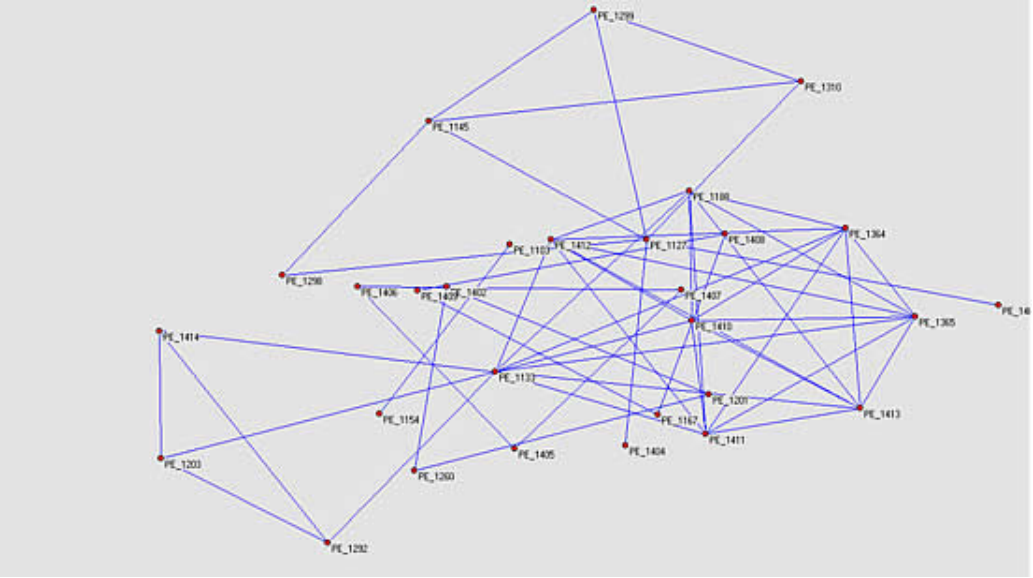


Figura 5.7: Redes de Capítulos de 2008. Fonte: Autor.

Tabela 5.2: Tabela de índices de redes de Artigos, Anais e Capítulos em contexto Geral e por qualis 2007.

Redes	**Vértices**	**Caminhos Mínimos Médio**	**Coeficiente de Aglomeração Médio**	**Densidade**	**Centralização**
Artigos	107	2,0713	0,751	0,0308	0,199
Artigos Qualis A1	46	1,461	0,893	0,073	0,132
Artigos Qualis A2	41	1,483	0,682	0,063	0,117
Artigos Qualis B1	21	1,616	0,736	0,157	0,489
Artigos Qualis B2	5	1,000	0,600	0,400	0,166
Artigos Qualis B3	9	1,000	1,000	0,444	0,071
Artigos Qualis B4	3	1,000	1,000	1,000	0,000
Artigos Qualis B5	6	1,000	0,666	0,466	0,200
Artigos Qualis C	3	1,000	1,000	1,000	0,000
Anais	101	2,496	0,675	0,029	0,0817

Continuação na próxima página...

Tabela 5.2 – Continuação

Redes	**Vértices**	**Caminhos Mínimos Médio**	**Coeficiente de Aglomeração Médio**	**Densidade**	**Centralização**
Anais Qualis ?	83	1,606	0,700	0,034	0,101
Anais Qualis NC	35	1,250	0,718	0,065	0,117
Capítulos	26	1,466	0,443	0,169	0,38

Fonte: Autor.

Tabela 5.3: Tabela de índices de redes de Artigos, Anais e Capítulos em contexto Geral e por qualis 2008.

Redes	**Vértices**	**Caminhos Mínimos Médio**	**Coeficiente de Aglomeração Médio**	**Densidade**	**Centralização**
Artigos	163	1,801	0,399	0,029	0,250
Artigos Qualis A1	52	1,410	0,423	0,061	0,181
Artigos Qualis A2	42	1,431	0,385	0,077	0,123
Artigos Qualis B1	12	1,000	0,333	0,212	0,072
Artigos Qualis B2	4	1,333	0,291	0,666	0,666
Artigos Qualis B3	4	1,000	0,5	1,000	0,000
Artigos Qualis B4	4	1,000	0,000	0,333	0,000
Artigos Qualis B5	9	1,000	0,5	0,5	0,160
Artigos Qualis NC	76	1,636	0,402	0,076	0,359
Anais Qualis ?	97	1.984	0.357	0.038	0.131
Capítulos	29	1,372	0,404	0,133	0,240

Fonte: Autor.

Como informação complementar, após a criação das redes de co-autoria científica, foram calculados os índices para avaliação do PPG analisado.

Tabela 5.4: Tabela de índices de um PPG.

Ano	**Índice de Artigos**	**Índice Total das Publicações**
2007	1,154	1,222
2008	0,982	1,045

Fonte: Autor.

5.2 Análise das Redes Geradas

A maior concentração de pesquisadores em qualis NC indica que a qualidade das publicações em artigos, apesar da entrada de novos pesquisadores neste programa, não cresceu proporcionalmente. A diferença em relação às qualificações obtidas em 2008 comparado a 2007 foi mínima, porém houve um acréscimo em geral indicado pelos números de vértices obtidos das redes em qualis específicos.

Outra propriedade topológica de redes complexas que caracteriza os comportamentos destas redes são os caminhos mínimos médios (CMM). Nas redes em geral de 2007 os CMM foram maior que em 2008, com uma diferença mínima em artigos e capítulos, e uma diferença maior em anais. Quanto maior o caminho mínimo, menos articulada o grupo de pesqusa é, isto é, a rede representa a interação entre pesquisadores de PPGs.

A difusão da informação ocorre mais facilmente onde há minimização de percurso entre o vértice de partida e chegada desta. Caminhos mais curtos indicam uma socialização da informação mais rápida, consequentemente, as redes mais compactas com menor CMM podem apresentar uma melhor articulação dos pesquisadores nas publicações.

As redes colaborativas formadas em 2007 em geral apresentaram uma densidade maior que em 2008, entretanto, a socialização entre os pesquisadores é mais rápida em 2008. Isto pode indicar que os pesquisadores participantes em 2007 estão mais conectados entre si em relação a 2008 na socialização nas produções científicas. Isso pode indicar um caráter de proximidade (a proximidade pode se dar por fatores de amizade, assunto da pesquisa, etc.)

entre os pesquisadores no PPG de 2007. Em 2008, o número de pesquisadores aumentou, entretanto o coeficiente de aglomeração diminuiu em 2008. Desta forma, pode-se inferir que a colaboração em tarefas acadêmicas diminuiu apesar do aumento do número de pesquisadores no PPG de 2008. As redes de colaboração científica e 2008 são menos articuladas que as redes de 2007.

Observou-se uma queda mínima em participações em produções científicas publicadas em anais de eventos de 2007 para 2008. Os índices de 2007 indicam uma concentração um pouco maior de pesquisadores (101 vértices) e mais conectados, com o coeficiente de aglomeração de 0,675, comparado ao coeficiente de aglomeração de 2008 (97 vértices) de 0,357.

Em 2008 foi eliminada a classificação NC de anais de eventos, o que pode explicar a queda nas participações de pesquisadores. Nas redes de participações em capítulos de livros, a quantidade de autores participantes foi quase a mesma, em 2007 participaram 26 pesquisadores e em 2008, 29 pesquisadores. A densidade sofreu uma queda mínima de 0.443 em 2007 para 0.404 em 2008. Percebe-se que em relação a publicação de capítulos em livros, as redes de colaboração de 2007 e 2008 entre os pesquisadores não tiveram grande oscilação na articulação destas.

As redes construídas apresentaram comportamento típico do modelo de redes mundo-pequeno, demonstrados pelos coeficientes de aglomeração entre os pesquisadores, são redes em geral com média a alta clusterização. Comparando-se valores obtidos da criação de uma rede aleatória, com algumas das redes construídas, utilizando os mesmos valores de N, para número de vértices, e $\langle k \rangle$ para grau médio, observou-se que os valores para coeficiente de aglomeração nas redes construídas foram maiores que os obtidos nas redes aleatórias (e.g. redes de artigos : redes aleatórias $C = 0,362092$ e nas redes construídas $C = 0,0456715$) e os CMM (caminhos mínimos médios) nas redes aleatórias foi de 1,248 e nas redes construídas foi de 2,07002.

Entretanto, ainda foi observado um comportamento de ligação tendencioso, onde os pesquisadores mais participativos em publicações de outros autores, tenderam a receber mais cooperação de outros autores. Esse comportamento é caracterizado pelo modelo de redes livres de escala, onde há uma tendência para ligação preferencial entre os pesquisadores. A presença maior de alguns pesquisadores em colaboração científica nas publicações, pode refletir uma preferência em se estabelecer relações de co-autoria com pesquisadores com maior potencial de qualidade científica em tarefas acadêmicas.

5.3 Avaliação do Modelo e Discussão

Os resultados obtidos pelo modelo foram satisfatórios, considerando que as redes obtidas pelo modelo foram construídas corretamente. Os dados e redes geradas foram conferidos manualmente para validação dos dados que compuseram as redes.

Avaliando os resultados parciais obtidos entre as estapas processuais, constatou-se criticidade maior na etapa da mineração de textos, primeiro pela complexidade exigida para criação de uma notação de padrão refinada, que influi diretamente na amplitude dos resultados e sujeira encontrada nestes, segundo pela necessidade da limpeza dos dados que pode demandar um tempo considerável a depender do grau de ruído nos resultados. Como o modelo contém um painel de criação de padrões de fácil utilização, que contém sintaxes de expressões regulares embutidas nos botões, a complexidade na criação da expressão é minimizada.

O processo de inserção de dados e relacionamentos é realizado corretamente. O modelo não permite a inserção de dados redundantes no banco de dados. Apenas um ponto crítico é citado nesta etapa, e está associado a instabilidade do software quando muitas consultas são executadas no banco de dados. A cada inserção de lista de dados, cada dado desta lista é verificado no banco, para validar ou não a existência prévia daquela informação no banco. Caso já exista, o dado não é inserido. Entretanto esse processo de verificação ocorre diversas vezes, até que se esgote as informações da lista de dados. Neste ponto é que pode ocorrer algumas vezes uma instabilidade no modelo, pelo excesso de processos executados. Este problema foi minimizado pelo uso de alguns processos paralelos (threads) na codificação do modelo.

Em relação a última etapa, a contrução das redes, esta é realizada satisfatoriamente, a partir dos dados inseridos no banco. A construção da rede é rápida em termos de processamento, e oferece possibilidades variadas de delimitação do escopo da rede, por meio dos filtros exixtentes no software. Isto permite a contrução de redes de forma mais ampla ou mais restrita (e.g. Redes de Artigos é uma rede mais ampla. Rede de Artigos Qualis A1 do ano de 2007 é uma rede mais restrita.).

5.3.1 Pontos críticos encontrados

O maior impacto em termos de criticidade do modelo se encontra na etapa da mineração dos textos. Em decorrência da especificidade dos dados a serem extraídos e a ausência de padronização exata na grafia em todos documentos, a criação dos padrões para extração dos dados ainda apresenta um nível de complexidade relevante. Apesar do modelo

proposto prover um módulo para criação de padrões facilitada, e funções adicionais para gravação de padrão para reuso do mesmo em outros processos, ainda é relevante que se conheça o mínimo das regras para criação de expressões regulares.

O método de expressão regular utilizada como minerador do texto depende do grau de refinamento da sua expressão para maior eficiência na busca e redução de erros nos resultados encontrados. Isto implica que se deve dominar as regras de expressões regulares. Considerando o baixo domínio das regras, quase todo resultado encontrado necessitará de algum tipo de correção para garantia da integridade dos dados antes da inserção no banco de dados. O procedimento de limpeza dos dados pode demandar um tempo maior ou menor a depender da amplitude dos resultados alcançados e experiência do usuário em usar expressões regulares, e é neste ponto que o tempo no processo geral de mineração de construção das redes é mais crítico.

Outro ponto de criticidade relevante é a execução das inserções e relacionamentos entre os pesquisadores. A execução contínua de consultas ao banco de dados pode causar eventuais paradas no funcionamento do mesmo. Isto porque no processo de verificação da existência dos pesquisadores nas tabelas do banco, para evitar redundância de dados, o banco é consultados diversas vezes. A depender da quantidade de registros verificados, pode haver uma parada no modelo, ou instalibilidades no funcionamento no software (*overhead*).

As listas de pesquisadores geralmente contém um número alto de elementos, e por isso, é necessário um número alto de consultas ao banco. Por exemplo, para uma lista que contém quinhentos pesquisadores, serão executados quinhentas consultas ao banco. Para evitar tantas execuções ao banco e assim limitar ao máximo instabilidades no funcionamento, foram criados no código processamentos paralelos (*threads*). Nesse sentido, uma solução para auxiliar as inserções de dados e relacionamentos, foi a criação do módulo de gestão do banco de dados para inserir informações que não foram registradas no banco em decorrência de algum erro.

Outro tópico crítico na construção das redes é o tratamento das redundâncias nos dados dos textos, onde foi observado que um mesmo autor teve representações textuais diferentes em documentos distintos, em alguns casos até por erro de digitação. Isto exigiu atenção na criação das redes no quesito da similaridade entre os nomes dos pesquisadores. Por exemplo, um autor representado por **BRAGA, P. F.** em um documento, em outro texto era escrito como **BRAGA, P.**.

Considerando a possibilidade de redundância nos dados das redes, foram incluídos no modelo algoritmos para verificação de similaridade dos termos, com base na semelhança de caracteres e o posicionamento de cada um. Cada termo é transformado em vetor de

caracteres e comparado com outros vetores de caracteres a partir do nível de semelhança definido no modelo.

Caso o termo verificado tenha o percentual de similaridade igual ou acima dos outros termos verificados, uma mensagem é lançada confirmando ou não o descarte do dado. Por exemplo, um grau de 90% de similaridade para verificação foi configurado no modelo. Um pesquisador de nome "BRAGA, P. F." é verificado dentre todos os nomes de pesquisadores que tenham grau de similaridade de no mínimo 90% com este nome. Então nomes como "BRAGA, P.", "BRAGA, P.F.G.", ou qualquer outro nome similar com o mínimo de semelhança de 90% será comparado ao nome verificado. Isto assegurou a possibilidade de verificação dos dados que são acrescentados nas redes. Observe a representação da conferência da similaridade na Figura 5.8.

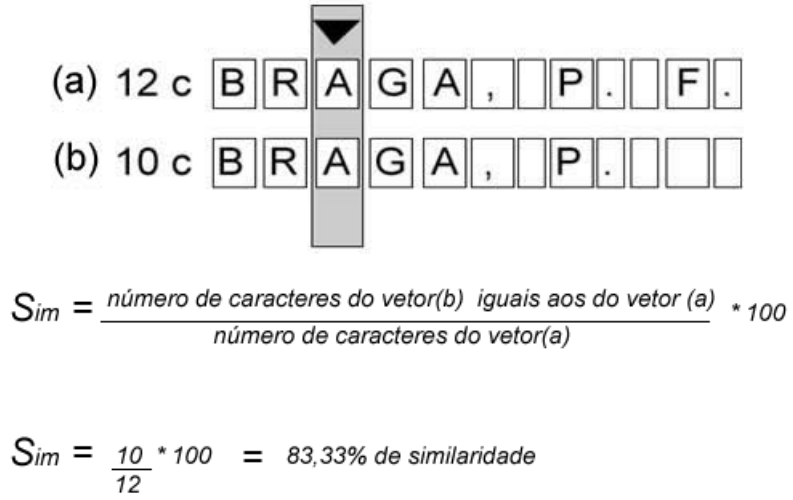


Figura 5.8: Representação de teste de similaridade entre palavras.

onde

- o item (b) representa o vetor de caracteres comparado;
- o item (a) representa o vetor original de caracteres;
- o *Sim* é o cálculo para verificação de similaridade do vetor.

No exemplo da Figura 5.8, apenas dois caracteres não se encaixaram no vetor original o "F" e o ".". A leitura do vetor processa cada posição do vetor original e vetor comparado para validação de igualdade do caracter e passa para a próxima posição. A cada validação verdadeira, é somado mais 01 na contagem dos caracteres. Esta função do modelo garantiu mais uma forma de precisar as informações dos componentes das redes.

5.3.2 Confiabilidade do modelo

Para validação do grau de confiabilidade do modelo, dois indicativos podem ser observados: a verificação manual da rede construída e os *logs*. Conforme citado, durante a fase de experimentação, as informações contidas nos documentos foram verificadas manualmente e validadas nas redes construídas, não sendo encontrados erros. Além desse procedimento, logs de execuções foram gerados ao longo dos processos realizados para dar uma visão da margem de erro e precisão dos resultados obtidos. Por meio destes logs, pôde-se constatar que houve uma margem de erro mínima apenas na etapa de mineração do texto.

Tomando como exemplo os logs gerados na extração dos dados de pesquisadores, artigos, anais e capítulos do programa de Modelagem Computacional do PPG01 2007, são apresentados na Tabela 5.5, um comparativo, entre a quantidade de dados que a expressão criada resultou, e a quantidade resultante da correção e remoção de sujeira.

Tabela 5.5: Tabela de resultados obtidos na mineração dos textos

Elementos	**Antes da Limpeza**	**Depois da Limpeza**	**Erro%**
Pesquisadores	258	258	0
Artigos	57	53	7,1
Anais de Eventos	76	65	14,4
Capítulos de Livros	14	12	14,2

Fonte: Autor.

Esses valores representam uma média encontrada nos logs de processos executados na mineração dos textos. O erro é variante e oscila conforme o resultado decorrente do refinamento das expressões contruidas. Por exemplo, com uma dada expressão, o número de artigos encontrados foi de 15 antes da limpeza. Refinando a mesma expressão, a quantidade de artigos encontrados pode ser de 13. Refinando-se ainda mais a expressão, pode-se encontrar o valor exato correspondente ao real. Logo, neste exemplo citado, o erro médio obtidos nos processos foi de 8.9%Ṅa etapa de inserção e geração das redes não houve erros constatados.

Conforme mostrado na Tabela 5.5, a margem de erro é relativamente baixa, sendo constatada apenas no processo de extração dos dados do texto. Prevendo possíveis lacunas nas listas de resultados, o software dispõe do módulo de gestão do banco de dados que possibilita a inserção deste registro ausente, a fim de evitar possíveis falhas na construção das redes.

Capítulo Seis

Considerações finais

6.1 Conclusões

O modelo proposto nesta pesquisa de dissertação utiliza como idéia básica a mineração de textos, por obtenção de conteúdo relevante nos textos. Entretanto, não utiliza todas as etapas executadas no processo comum, tais como pré-processamento do texto, indexação, aplicação de algoritmos para cálculo de frequências e análise de conteúdo. No pré-processamento comum da mineração de textos, é necessário utilizar algoritmos de tratamento dos textos para se realizar correções ortográficas, redução da quantidade de palavras e indexação dos termos em um vetor de palavras (e.g. remoção de *stopwords* e *stemming* das palavras. ver Capítulo 2). No modelo proposto, a técnica adotada não exige esse pré-processamento do texto, ou seja, as informações extraídas do texto serão originadas diretamente da busca de padrões nos documentos.

O objetivo da mineração de texto, no caso descrito para esta pesquisa, consistiu em apenas minerar conteúdo específico e independente de contexto e frequência. Isto porque as informações não poderiam ser sintetizadas por relevância pois seria necessária a separação deste conteúdo sumarizado em outros dados isolados para relacionamento dos mesmos e construir as redes. Se a idéia principal do projeto fosse construir redes sociais e complexas com base em conteúdo relevante, a aplicação das regras comuns da mineração de texto resolveria de forma eficiente o problema, já que neste caso o dado minerado consideraria a frequência e distância dos termos em relação ao conjunto total de palavras.

No caso deste projeto, o conhecimento necessário a se descobrir está associado às redes sociais e complexas intrísecas nos textos. Como no modelo proposto os dados devem ser extraídos separadamente, o relacionamento entre estes dados foi criado manualmente. Em técnicas comuns de mineração de textos, não há relacionamentos lógicos entre os dados existentes. Por exemplo, não há como saber por meio desses processos de pré-processamento e sumarização de textos, qual a relação entre os autores com suas respectivas tarefas acadêmicas. Isto porque, a mineração de textos usual considera apenas as relações entre as palavras do ponto de vista semântico. No modelo proposto, pode-se separar por dados de interesse as redes que se deseja construir.

O modelo proposto, separa os dados que são de interesse, por exemplo, nos cadernos indicadores das CAPES, há informações sobre publicações e suas autorias. Então para a criação de uma rede de artigos, cria-se uma lista de dados de artigos e uma lista de dados

de pesquisadores separadamente para depois relaciona-las. Para a realização desta tarefa, foi necessário a utilização de técnicas de reconhecimento de padrão para minerar textos, neste projeto, a técnica consistiu na aplicação de expressões regulares.

Em decorrência da inacessibilidade ao banco de dados da CAPES, fez-se necessário encontrar uma forma de retirar informações sobre os PPG para a construção das redes de co-autoria científica. O repositório que comtempla maior quantidade de informações sobre esses PPGs, além dos bancos de dados, são os cadernos de indicadores da CAPES. Considerando que estes cadernos estão sob forma de textos digitais, a forma encontrada para se extrair o máximo de informações desses PPGs foi a mineração de textos.

Devido a dificuldade imposta pela mineração direcionada a conteúdo específico e não a conteúdo relevante, a escolha por utilização de técnica de reconhecimento de padrão demonstrou ser mais eficiente para o problema em questão. A escolha por este tipo de técnica diminuiu o número de etapas do processo de mineração do texto, o que reduz trabalho no processamento do texto em geral. A tarefa principal portanto foi concentrada em outro aspecto, isto é, o entendimento das regras das expressões regulares.

Considerando a complexidade existente na definição de padrões reconheciveis para extração de dados textuais, o modelo foi concebido de forma a minimizar a dificuldade na criação das notações de expressões regulares. De uma forma geral, a técnica empregada atende aos objetivos do modelo sem grandes problemas, concentrando o trabalho árduo apenas na limpeza e correção dos dados. Assim, esta etapa exige um cuidado maior, porque os dados que são inseridos no banco, são oriundos das listas de dados. Haja a vista que as redes são construídas com base nos dados inseridos no banco, caso existam erros nestas listas, as redes também serão compostas de dados errados. As redes criadas montraram a eficiência do modelo proposto, uma vez que as redes construídas foram conferidas e em sua totalidade consideradas corretas, como detalhado no Capítulo 4.

Como a quantidade de dados no banco é considerável, poderão haver instabilidades ou paradas no funcionamento do software em decorrência das consultas ao banco de dados, em especial na etapa da inserção de relacionamentos. As consultas sequenciais ao banco pode causar um *overhead*. A inserção de relacionamentos demanda algumas tarefas internas descritas no algoritmo que podem sobregarregar a memória do computador. No sentido de tentar minimizar esse efeito no modelo proposto, o código foi adaptado para processar múltiplas tarefas em *threads*.

6.2 Contribuições

O objetivo exigiu o uso de técnicas mais específicas de mineração de textos. Existem pesquisas neste campo de geração e caracterização de redes complexas a partir de mineração de textos (e.g. (PAES, 2008)), porém não foi constatado trabalhos com foco em dados específicos como proposto neste projeto de dissertação.

Uma parte considerável das pesquisas existentes se baseiam em sumarização de conteúdo por termos de relevância, como no trabalho de Antiqueira (2007) e Paes (2008). Diferentemente do conceito aplicado nessas pesquisas, o modelo proposto nesta dissertação armazena as informações mineradas em banco de dados, para sua utilização na construção das redes. Diferentemente do modelo de mineração de textos utilizados nos trabalhos anteiormente citados, este projeto,por utilizar dados específicos para composição das redes de colaboração científica, aplicou como técnica de extração de dados dos textos as expressões regulares.

Os trabalhos existentes se baseiam em estruturas semânticas para construção das redes, a exemplo de Antiqueira (2005). As estruturas semântica se baseiam em dados explicitamente interligados, enquanto que no modelo desenvolvido os dados estão implícitos no contexto. Por exemplo, em redes semânticas de textos não há como fazer referência lógica entre o autor e sua participação em artigos. É possível apenas criar as relações do tipo hierarquias de conceitos, como exemplo: o gato(conceito) - come(ação) - o rato(conceito).

No modelo proposto, as relações a serem extraídas não se encontram em nível gramatical ou semântico. São redes implícitas, construídas a partir da extração apenas dos dados de interesse e criação manual de seus relacionamentos. Não é qualquer dado do texto que deve ser inserido nas redes. As redes semânticas construídas de textos, utilizam praticamente todo o texto, verificando as relações semânticas entre as palavras.

Um aspecto de relevância desta pesquisa é a integração de três processos em um único modelo: mineração de textos gestão de dados e construção de redes sociais e complexas. Existem softwares para somente minerar de textos como o AlchemyAPI [1], softwares para análise e visualização de redes complexas (apenas analisam as redes, não constróem as redes), a exemplo do PAJEK[2], etc. De forma geral, os modelos computacionais encontrados realizam esses processos citados, entretanto, dispõem de funções únicas e bem definidas.

[1] http://www.alchemyapi.com/?ga=1021, último acesso em 06/11/2010 às 13:00h
[2] http://vlado.fmf.uni-lj.si/pub/networks/pajek/, acessado em 10/08/2010 às 20:30h

6.3 Atividades Futuras de Pesquisa

O conteúdo coletado dos textos poderá ser utilizado para outros fins além da construção das redes. A informação extraída pode gerar outros conhecimentos que não apenas as estruturas das redes. Por exemplo, pode-se inferir dos dados presentes no banco, conteúdos mais constantes nas publicações, em que redes estão mais presentes, quais programas se concentram maior tipo de publicações e quais assuntos mais pesquisados, entre outras informações. Desta forma a visão que se terá das produções científicas nesses programas de pós-graduação será bem mais ampla, não apenas focando nos pesquisadores como ponto central de análise, mas nos conteúdos de pesquisas.

A proposta futura é a criação de um módulo na área gerencial para realização de mineração de dados no banco. O modelo contará com dois processos distintos, porém complementares: a mineração dos textos em uma etapa inicial, e a mineração dos dados em uma etapa final pós redes geradas. O primeiro processo extrai e armazena os dados de interesse, como já está sendo feito. O segundo processo utilizará os dados já isneridos no banco pelo primeiro processo para descobrir outros conhecimentos implicitos no banco de dados (e.g. pesquisadores que se associam mais). O módulo de mineração de dados será composto por algoritmos para classificação, análise ou predição de dados, a depender do conhecimento que se deseja obter.

Baseado nos resultados a serem obtidos nesse processo, a depender da técnica definida, pode-se obter, por exemplo, classificações por publicações, por qualis, por conteúdo, fazer inferências para predições com base nos conteúdos, anos, instituições e programas e outros conhecimentos.

Do ponto de vista de análise das redes de co-autoria científica, a aplicação de técnicas de mineração de dados identificará padrões de comportamento nas redes construídas, tais como agrupamentos de pesquisadores por participação em produções, classificação dos PPGs quanto a quantidade de produções realizadas, nível das produções por qualis, identificar padrões de associações entre os pesquisadores que mais produzem tarefas acadêmicas, entre outros dados relevantes.

Apêndice A

Resultados das Redes Geradas

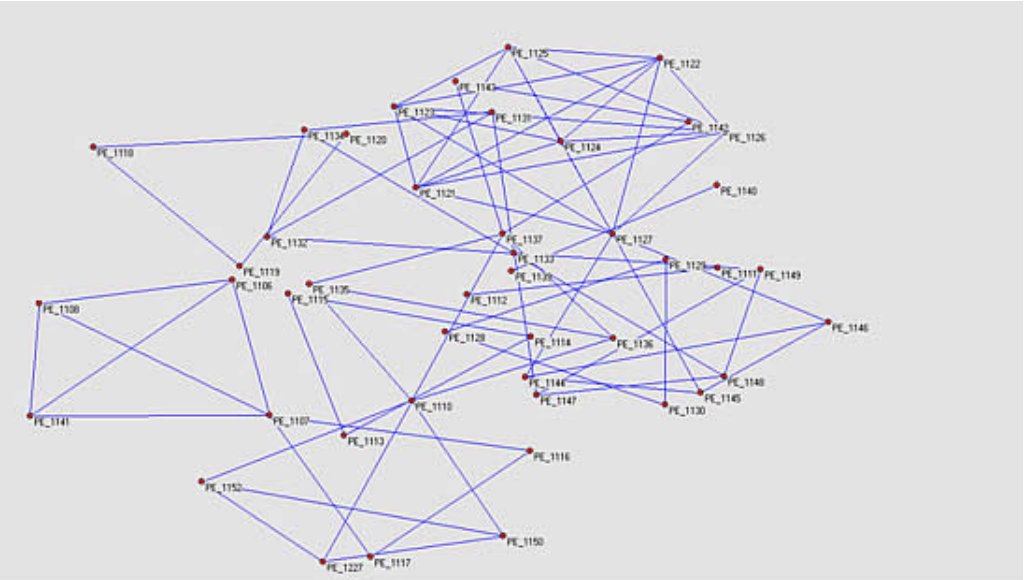


Figura A.1: Redes de Artigos Qualis A1 2007. Fonte: Autor.

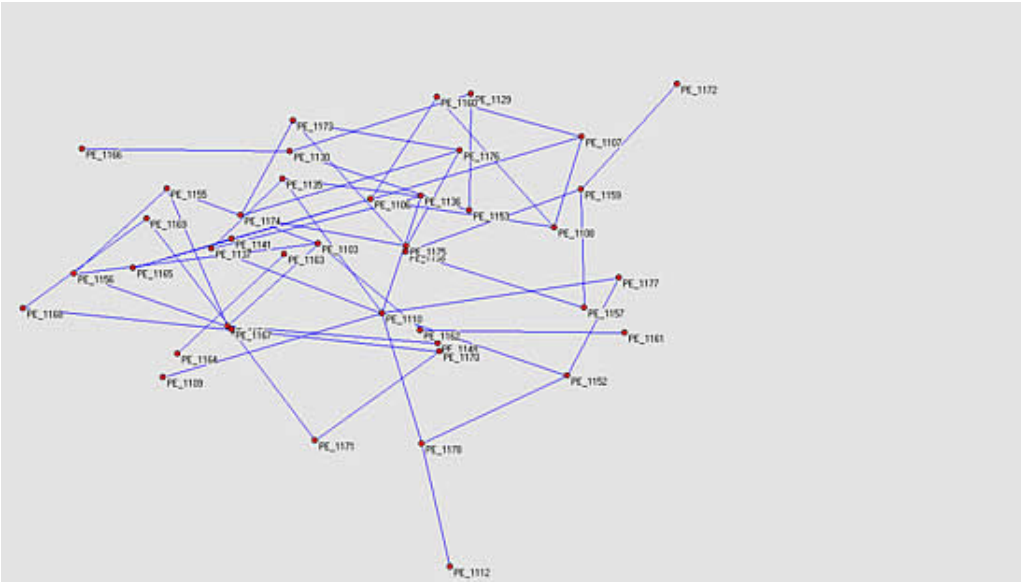


Figura A.2: Redes de Artigos Qualis A2 2007. Fonte: Autor.

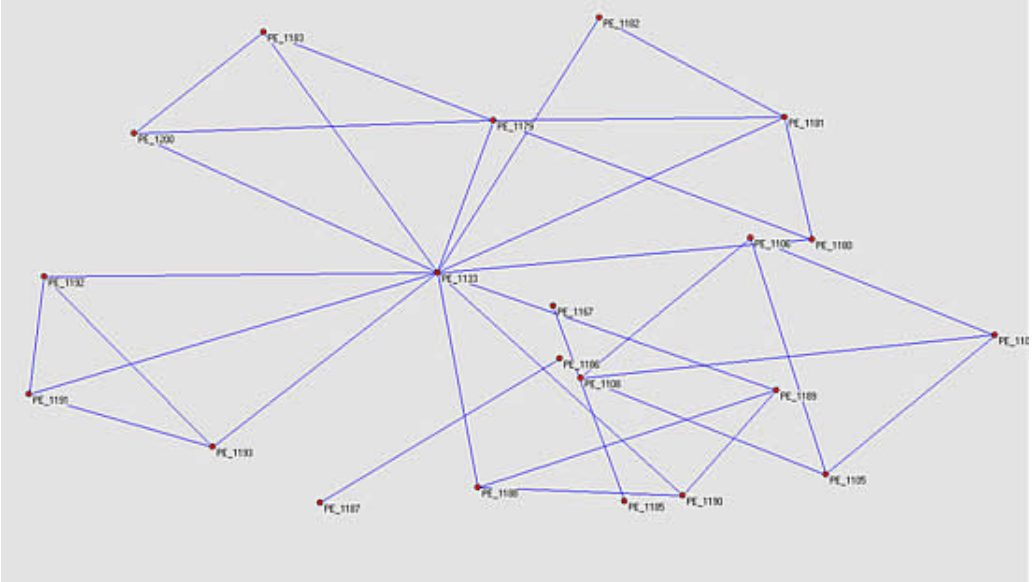


Figura A.3: Redes de Artigos Qualis B1 2007. Fonte: Autor.

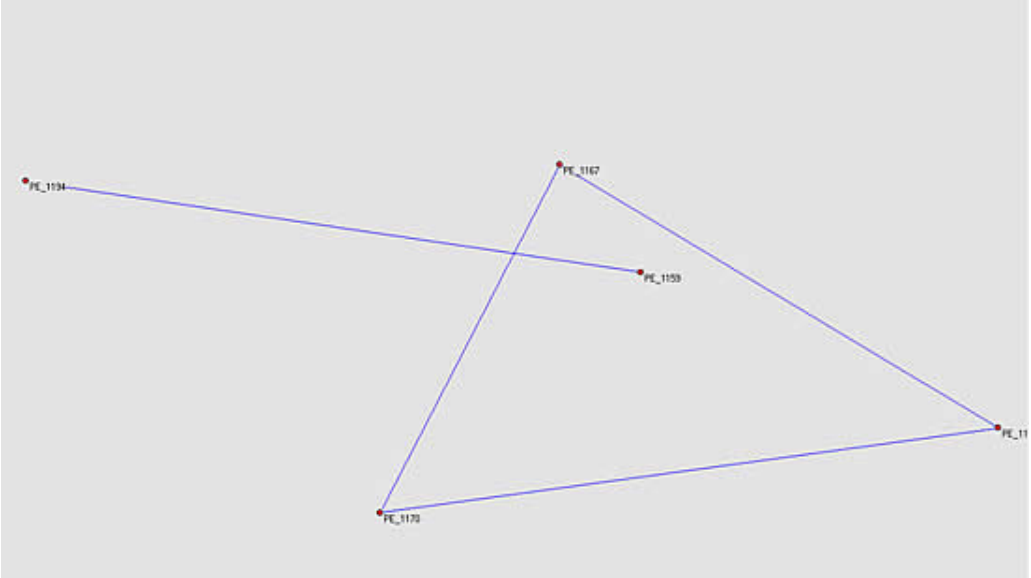


Figura A.4: Redes de Artigos Qualis B2 2007. Fonte: Autor.

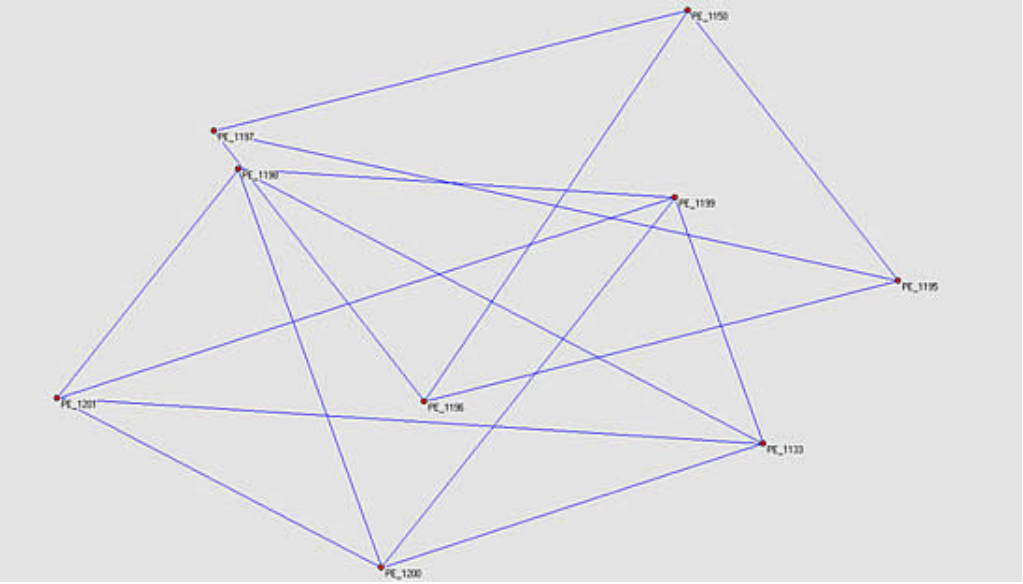


Figura A.5: Redes de Artigos Qualis B3 2007. Fonte: Autor.

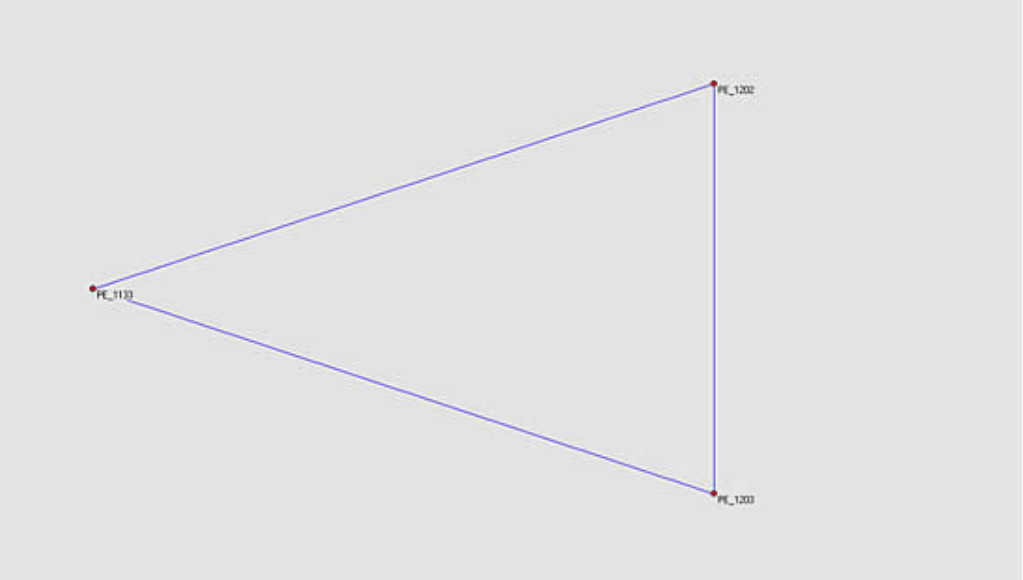


Figura A.6: Redes de Artigos Qualis B4 2007. Fonte: Autor.

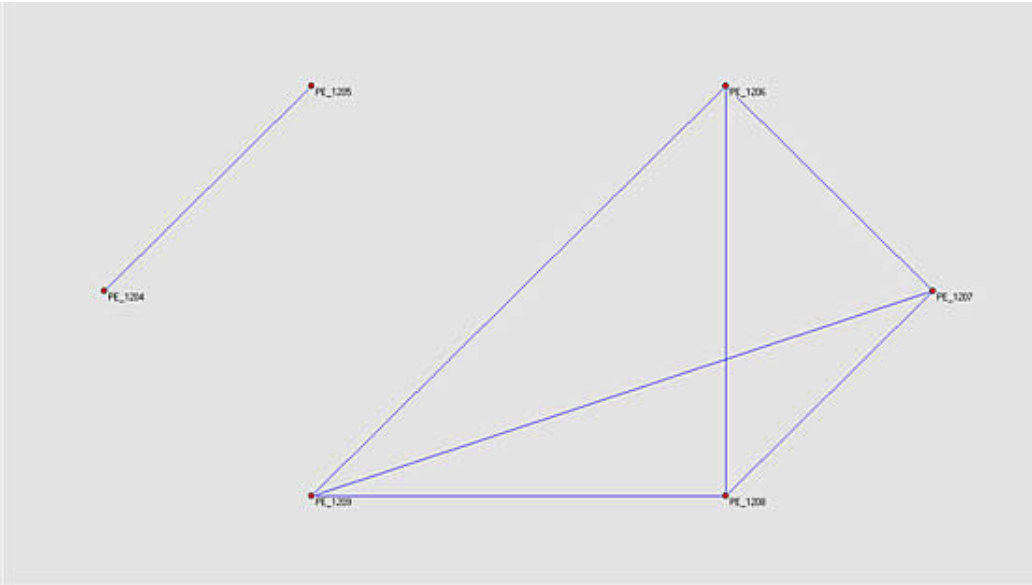


Figura A.7: Redes de Artigos Qualis B5 2007. Fonte: Autor.

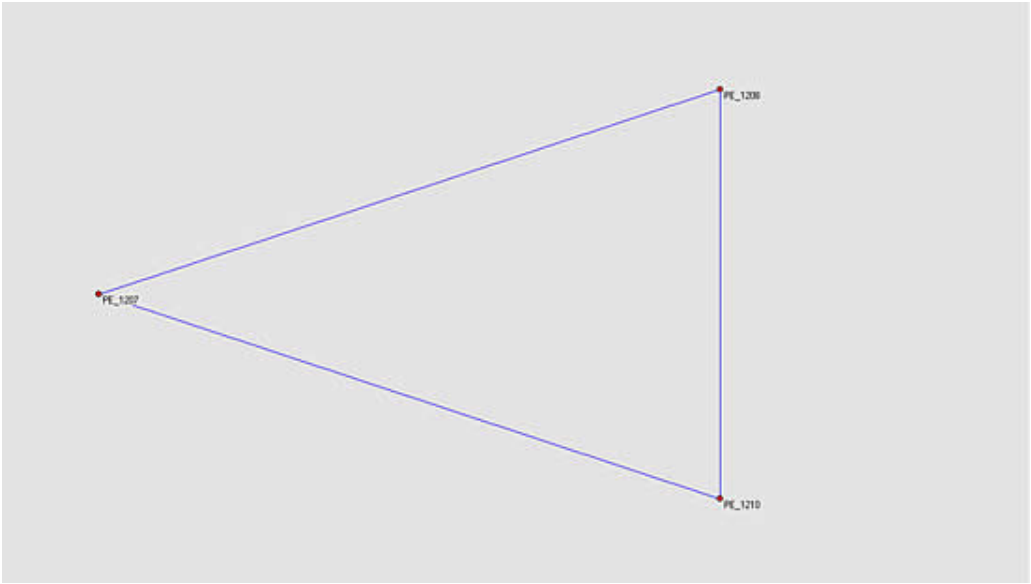


Figura A.8: Redes de Artigos Qualis C 2007. Fonte: Autor.

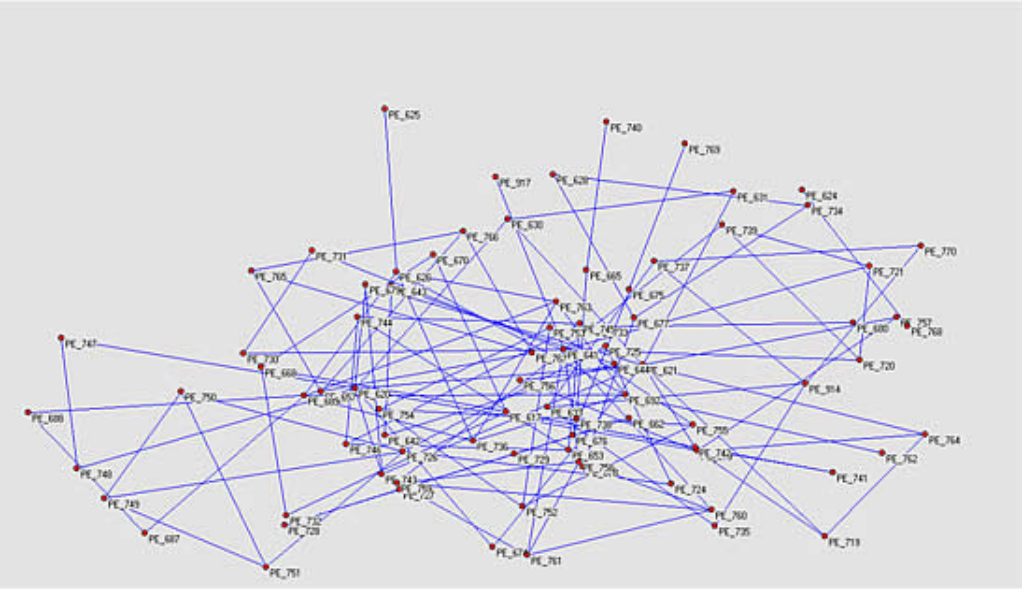


Figura A.9: Redes de Anais Qualis ? 2007. Fonte: Autor.

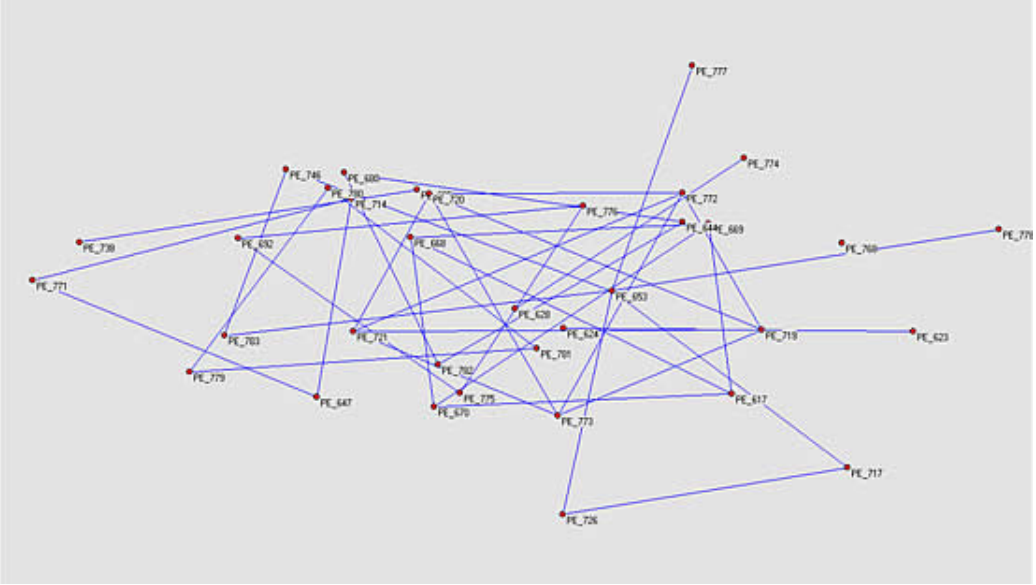


Figura A.10: Redes de Anais Qualis NC 2007. Fonte: Autor.

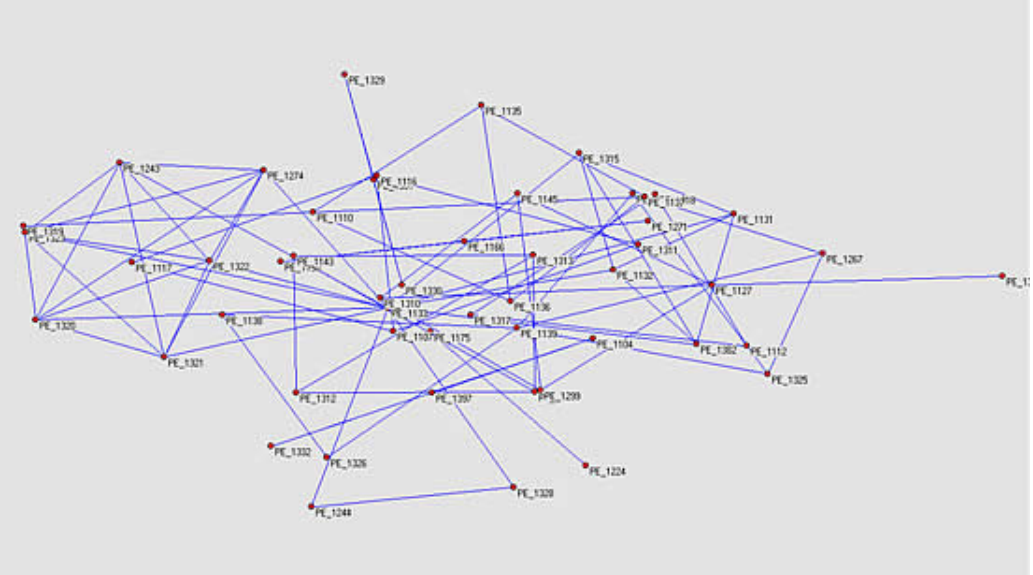


Figura A.11: Redes de Artigos Qualis A1 2008. Fonte: Autor.

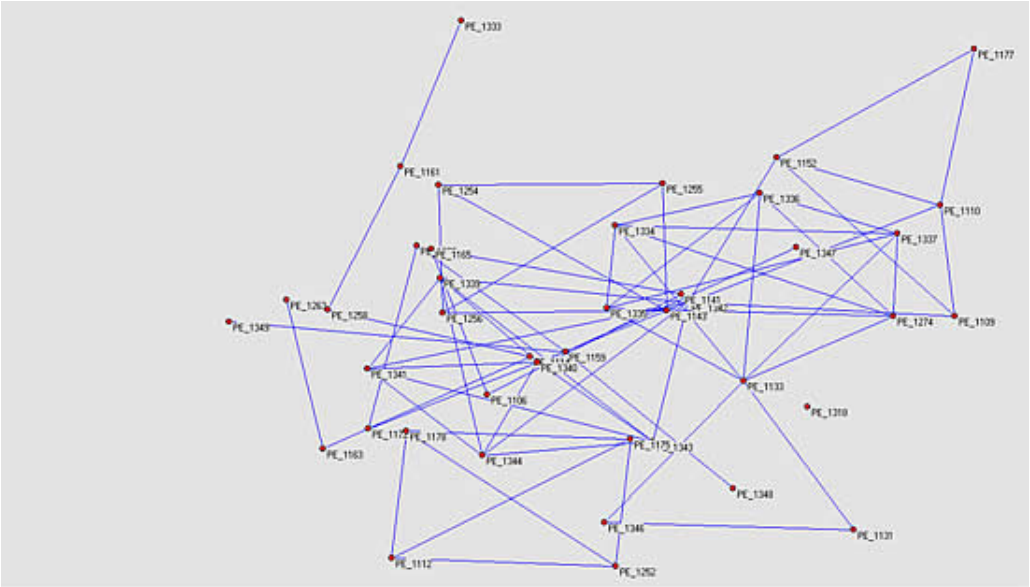


Figura A.12: Redes de Artigos Qualis A2 2008. Fonte: Autor.

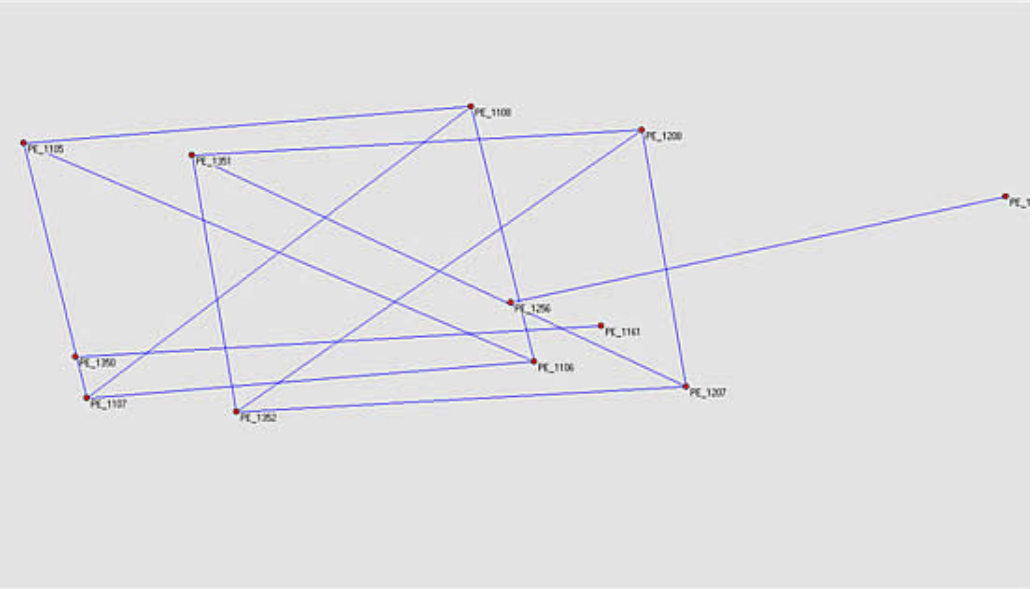


Figura A.13: Redes de Artigos Qualis B1 2008. Fonte: Autor.

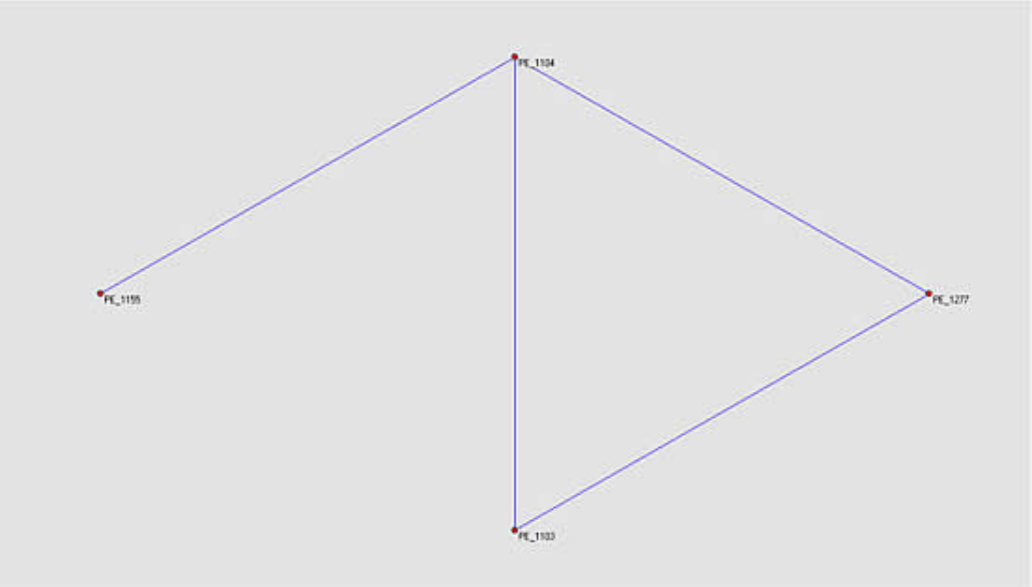


Figura A.14: Redes de Artigos Qualis B2 2008. Fonte: Autor.

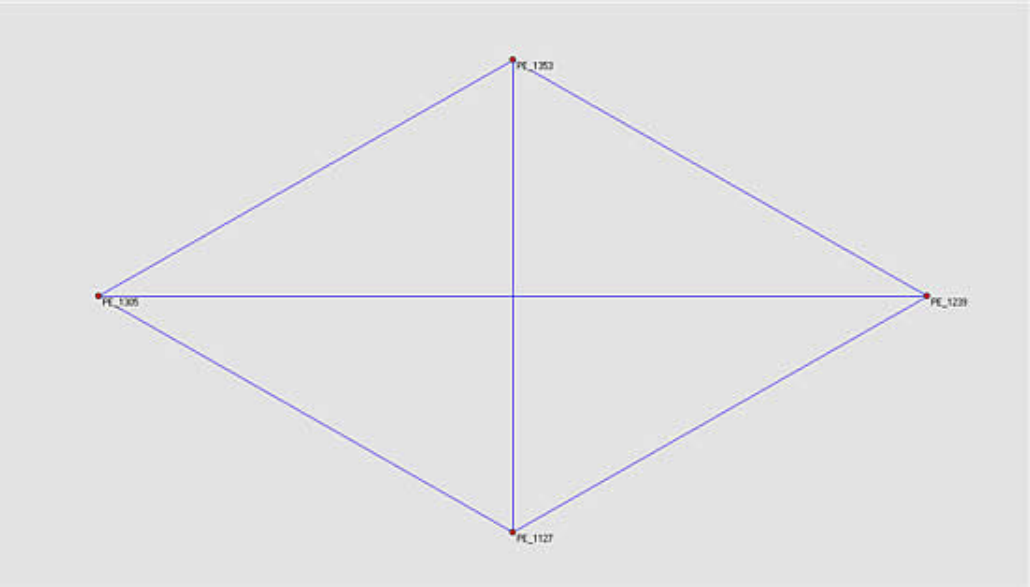


Figura A.15: Redes de Artigos Qualis B3 2008. Fonte: Autor.

Figura A.16: Redes de Artigos Qualis B4 2008. Fonte: Autor.

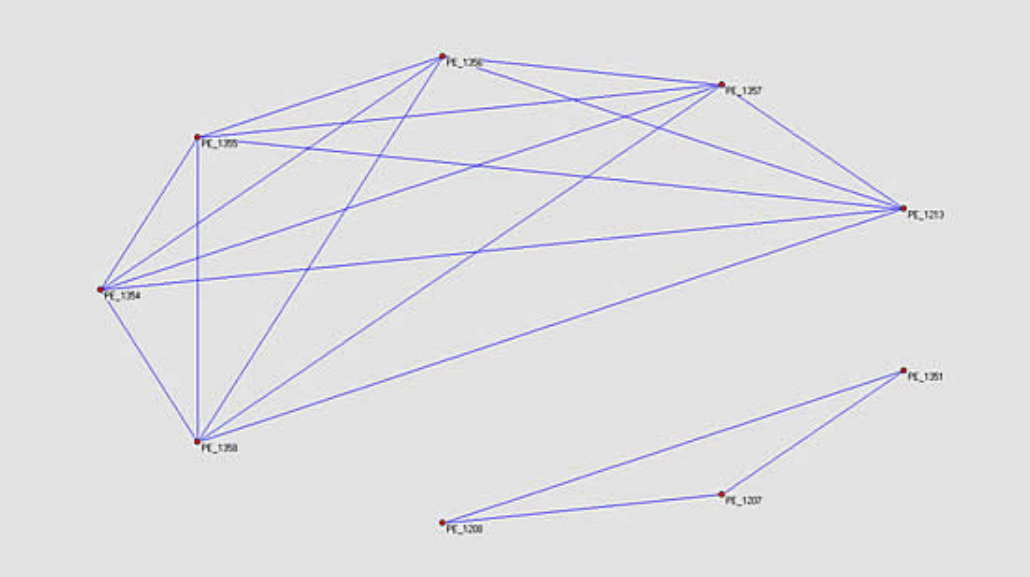


Figura A.17: Redes de Artigos Qualis B5 2008. Fonte: Autor.

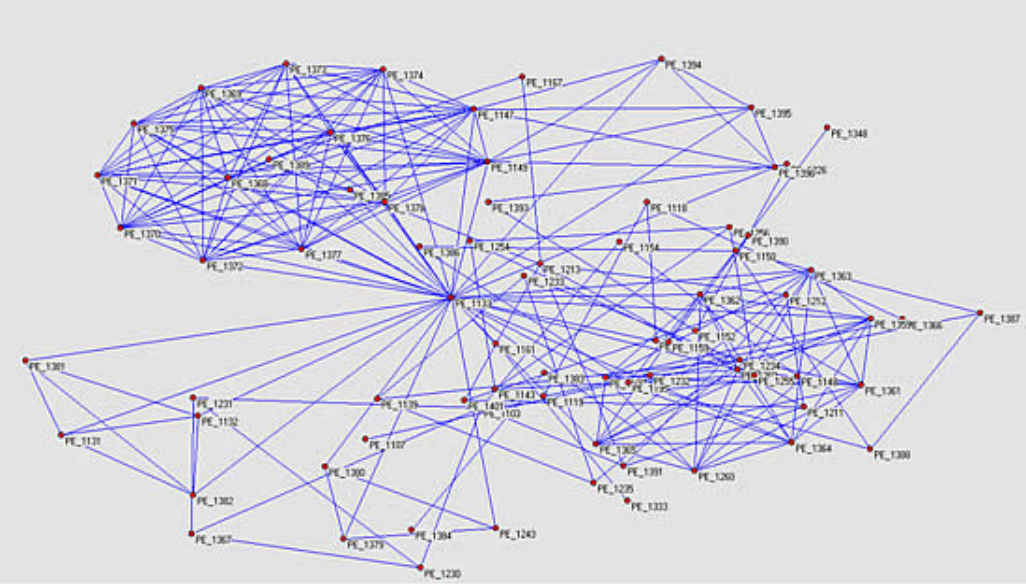


Figura A.18: Redes de Artigos Qualis NC 2008. Fonte: Autor.

Referências Bibliográficas

ALBERT, R.; BARABáSI, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, v. 74, n. 1, 2002.

ALBERT, R. e. a. Error and attack tolerance of complex networks. *Nature*, n. 406, 2000.

ANTIQUEIRA, L. *Desenvolvimento de técnicas baseadas em redes complexas para sumarização extrativa de textos.* Dissertação (Mestrado em Ciência da Computação e Matemática Computacional) — Universidade de São Paulo, São Carlos, São Paulo, 2007.

ANTIQUEIRA, L. e. a. Modelando textos como redes complexas. In: *XXV Congresso da Sociedade Brasileira de Computação*. São Leopoldo, RS, Brasil: [s.n.], 2005.

ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. *RESI-Revista Eletrônica de Sistemas de Informação*, n. 2, 2006.

BARABáSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, 1999.

BOCCALETTI, S. e. a. Complex networks: Structure and dynamics. *Physics Reports*, n. 424, 2006.

CHAVES, M. S. Um estudo e apreciação sobre algoritmos de stemming para lingua portuguesa. *Programa de Pós-Graduação em Ciência da Computação (PPGCC), Porto Alegre*, 2003.

COHEN, E. e. a. The complexity of kleene algebra with tests. *Technical Report TR96-1598*, 1996.

COSTA, L. d. F. e. a. Characterization of complex networks: A survey of measurements. *Instituto de Física de São Carlos, Universidade de São Paulo.*, 2006.

COSTA, L. d. F. e. a. Characterization of complex networks: A survey of measurements. *Advances in Physics*, v. 56, n. 1, p. 167–242, 2007.

ELSMARI, R.; NAVATHE, S. *Sistemas de Bancos de Dados*. [S.l.]: Person Education, 2004. ISBN 85-88639-17-3.

ERDöS, P.; RéNYI, A. On random graphs i. *Publicationes Mathematicae*, n. 6, p. 290–297, 1959.

FELDMAN, R.; SANGER, J. *The Text Mining Handbook - Advanced Approaches in Analysing Unstructured Data*. Cambridge: Cambridge University Press, 2007. ISBN 978-0-521-83657-9.

FERRER, R.; SOLé, R. V. Optimization in complex networks. *Lecture Notes in Physics, Springer (Berlin)*, p. 114–125, 2003.

FRIEDL, J. E. *Mastering Regular Expressions - Powerful Techniques for Pearl and Other Tools.* United States: OReilly, 1997.

GAMEIRO, P. As organizações em rede. *Universidade Lusófona de Humanidades e Tecnologia*, n. 290, 2008.

GIUDICI, P. *Applied Data Mining:Statistical Methods for Business and Industry.* [S.l.]: Ed. Wiley, 2003. ISBN 0 470 84679 8.

GONZALEZ, M.; LIMA, V. L. Recuperação de informação e processamento da linguaguem natural. *Faculdade de Informática (PUCRS), Porto Alegre*, 2009.

GOOD, N. *A Regular Expressions Recipes for Windows Developers: A Problem Solution Approach.* United States: Ed. Apress, 2005. ISBN 1-59059-497-5.

GRANOVETTER, M. The strengh of weak ties. *American Journal of Sociology*, v. 78, p. 1360–1380, 1967.

GROSS, J. L.; YELLEN, J. *Handbook of Graph Theory.* New York: CRC PRESS, 2004. ISBN 1-58488-090-2.

HAN, J.; KAMBER, M. *Data Minning: Concepts and Techniques.* [S.l.]: Ed. Morgan Kauffman, 2003.

HIEMSTRA, D. *Using language models for information retrieval.* Dissertação (PhD) — Uitgeverij voor Lezers en Schrijvers van Talige Boeken., The Netherlands, 2008.

INGWERSEN, P. *Information Retrieval.* United Kingdom: [s.n.], 1999. ISBN 0 94756 854 9.

KLEENE, S. C. Representation of events in nerve nets and finite automata. *Princeton University Press*, p. 3–41, 1956.

KREBS, V. 2007. URL=http://www.orgnet.com/contagion.html.

LATORA, V.; MARCHIORI, M. Efficient behavior of small-world networks. *Physical Review Letters*, v. 87, n. 19, 2001.

LIDDY, E. *Natural Language Processing In Encyclopedia and Information Science.* [S.l.]: Ed. Marcel Decker, 2005.

LOH, S. *Descoberta de Conhecimento em Textos.* Dissertação (Curso de Pós-Graduação em Ciência da Computação) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 1999.

MENDES, J. F. Física das redes complexas. *Gazeta de Física*, 2006.

MENESES, P. B. *Linguagens Formais e Autômatos.* Porto Alegre: Ed. Sagra, 2000.

METZ, J. e. a. Redes complexas:conceitos e aplicações. *Instituto de Ciências Matemáticas e de Computação*, v. 45, n. 290, 2007. ISSN 0103-2569, 2007.

MILGRAM, S. The small world problem. *Psychology Today*, p. 60–67, 1967.

MONTEIRO, L. d. O. e. a. Etapas do processo de mineração de textos - uma abordagem aplicada a textos em português do brasil. In: *Anais do XXVI Congresso da SBC.* Campo Grande-MS, Brasil: [s.n.], 2006.

MOODY, J. Race, school integration, and friendship segregation in america. *American Journal of Sociology*, n. 107, p. 679–716, 2001.

NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Review*, v. 45, n. 2, p. 167–256, 2003.

NEWMAN, M. E. J.; GIRVAN, M. Finding and evaluating community structure in networks. *Physical Review E*, n. 69, 2004.

PAES, C. A. *Caracterização Topológica de Redes Complexas Geradas no processo de Mineração de Documentos.* Dissertação (Mestrado em Engenharia Civil) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2008.

POTTERAT J. J., e. a. Risk network structure in the early epidemic phase of hiv transmission in colorado springs. In: *Sexually Transmitted Infections.* Colorado,EUA: [s.n.], 2002.

RODRIGUES, F. A. *Caracterização, classificação e análise de redes complexas.* Dissertação (Doutorado) — Instituto de Física de São Carlos (USP), São Carlos, São Paulo, 2007.

STUBBLEBINE, T. *Expressões Regulares: Guia de Bolso.* United States: Ed. Altabooks, 2007.

VEIGA, C. e. a. Mineração de textos e bancos de dados textuais. *Departamento de Ciência da Computação (UFBA), Salvador*, 2009.

VIEIRA, N. J. *Linguagens Formais e Autômatos.* Belo Horizonte: Departamento de Ciência da Computação (UFMG), 2000.

WATTZ, D.; STROGATZ, S. Collective dynamics of small-world networks. *Nature*, v. 393, 1998.

Um Modelo Computacional para Extração Textual e Construção de Redes Sociais e Complexas

Patrícia Freitas Braga

Salvador, Setembro de 2010.