

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UTILIZANDO TÍTULOS DE ARTIGOS
CIENTÍFICOS NA CONSTRUÇÃO DE
REDES SEMÂNTICAS PARA
CARACTERIZAR ÁREAS DE PESQUISA

JANSEN CRUZ DE SOUZA

JOÃO PESSOA-PB
AGOSTO-2015

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

**UTILIZANDO TÍTULOS DE ARTIGOS CIENTÍFICOS
NA CONSTRUÇÃO DE REDES SEMÂNTICAS PARA
CARACTERIZAR ÁREAS DE PESQUISA**

JANSEN CRUZ DE SOUZA

JOÃO PESSOA-PB
AGOSTO-2015

Universidade Federal da Paraíba
Centro de Informática
Programa de Pós-Graduação em Informática

Utilizando Títulos de Artigos Científicos na Construção de
Redes Semânticas para Caracterizar Áreas de Pesquisa

Jansen Cruz de Souza

Dissertação submetida à Coordenação do Curso de Pós-Graduação em
Informática da Universidade Federal da Paraíba como parte dos requisi-
tos necessários para obtenção do grau de Mestre em Informática.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Computação Distribuída

Alisson Vasconcelos de Brito

João Pessoa, Paraíba, Brasil

©Jansen Cruz de Souza, 27 de agosto de 2015

S729u Souza, Jansen Cruz de.
Utilizando títulos de artigos científicos na construção de
redes semânticas para caracterizar áreas de pesquisa / Jansen
Cruz de Souza.- João Pessoa, 2015.
79f. : il.
Orientador: Allison Vasconcelos de Brito
Dissertação (Mestrado) - UFPB/CI
1. Informática. 2. Ciência da computação. 3. Computação
distribuída. 4. Análise de redes sociais. 5. Redes semânticas.

UFPB/BC

CDU: 004(043)

Resumo

O conhecimento científico não se resume na descoberta de fatos e leis novas, mas também em sua publicação. Visando compreender o complexo sistema de produção científica, inúmeras técnicas de análise vêm sendo aplicadas em repositórios digitais. Assim sendo, a análise de redes sociais (ARS) tem sido um tópico de interesse de diversos estudos realizados nos últimos anos. As redes sociais podem ser estabelecidas a partir da relação entre pessoas ou da relação de informações que podem ser analisadas a fim de auxiliar a tomada de decisão. A proposta desta dissertação é apresentar uma técnica que possibilite caracterizar áreas de pesquisa científica através da construção de redes semânticas variáveis no tempo. Essas redes são construídas tomando por base a coocorrência de palavras-chave nos títulos de artigos científicos, que servem para representar a evolução dos temas de pesquisa abordados pelos meios de publicação ao longo do tempo.

Palavras-chave: Análise de Redes Sociais. Redes Semânticas.

Abstract

The scientific knowledge is not just the discovery of new facts and laws, but also in its publication. Aiming to understand the complex scientific production system analysis techniques have been applied in many digital repositories. Therefore, the social network analysis (SNA) has been a topic of interest of several studies in recent years. Social networks can be established from the relationship between people or the relationship of information that can be analyzed to assist decision making. The purpose of this thesis is to present a technique that makes it possible to characterize areas of scientific research by building semantic networks vary in time. These networks are built building on the co-occurrence of keywords in the titles of scientific papers, used to represent the evolution of research themes addressed by means of publication over time.

Keywords: Social Network Analysis. Semantic networks

Agradecimentos

Agradeço primeiramente a DEUS pela minha existência, pelas bênçãos que tem concedido ao longo de toda minha vida, me guiando, acompanhando e incentivando diariamente, a fim de concluir, com muito esforço, mais uma etapa acadêmica.

Gostaria de dedicar, com todo meu amor, ao homem mais importante de minha vida, que nunca deixou faltar nada dentro de casa, sempre dando aos seus filhos uma vida com muito conforto e investindo sempre no futuro de cada um. Um homem batalhador, honrado, inteligente, um verdadeiro pai, em quem vou me espelhar o resto de minha vida. Esse vai pra você, painho. Tenho certeza de que um dia te darei o abraço que guardei para esse momento tão especial.

Dedico também a minha maravilhosa e amada mãe, que sempre zelou por mim, me incentivando, dando seu apoio nos momentos difíceis pelos quais passamos, principalmente nos últimos dois anos. Mainha, sem você com certeza eu não conseguiria.

Aos meus irmãos, Faber e Kilmer, e suas respectivas esposas, que sempre me deram apoio e torceram por mim. Enfim, a toda essa minha família maravilhosa, pelo privilégio de tê-la sempre ao meu lado.

A minha amada namorada, Anny Carollyne, que sempre me deu forças pra continuar, suportando meus desabafos, meu mau humor nos momentos de pressão. Sua compreensão e seu apoio, nos momentos que mais precisei, foram fundamentais para eu vencer os obstáculos nessa longa jornada. Aos professores, Alexandre Duarte e Thais Gaudêncio, pela paciência e competência na transmissão dos conhecimentos. Ao meu orientador e amigo Alisson Vasconcelos de Brito, pelo exemplo de profissionalismo, pela paciência, apoio e contribuição para realização deste trabalho.

Muito Obrigado!

Conteúdo

1	Introdução	1
1.1	Motivação	3
1.2	Metodologia	4
1.3	Objetivos	4
1.4	Contribuição	5
1.5	Estrutura da Dissertação	6
2	Fundamentação Teórica	7
2.1	Linguagem e Sistemas Complexos	7
2.1.1	Palavras: Gramaticais e Lexicais	8
2.1.2	Redes Complexas	9
2.1.3	Modelos de Redes	13
2.2	Redes Sociais	19
2.2.1	Definições e Termos Gerais	19
2.2.2	Análise de Redes Sociais	21
2.2.3	Componente Gigante	22
2.3	Métricas para Análise das Redes Sociais	23
2.3.1	Centralidade de Grau	23
2.3.2	Centralidade de Intermediação	24
2.3.3	Centralidade de Proximidade	24
2.3.4	Modularidade	25
2.4	Análise de Tendências	27
2.4.1	Regressão Linear Simples	27

3	Trabalhos relacionados	29
3.1	Análise de redes complexas	29
3.1.1	<i>The network of concepts in written texts</i>	29
3.2	Análise de redes semânticas baseada em títulos de artigos científicos	30
3.2.1	Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática	30
3.2.2	Redes de títulos de artigos científicos variáveis no tempo	31
3.2.3	<i>Semantic networks based on titles of scientific papers</i>	31
3.2.4	<i>A network approach based on cliques</i>	32
3.3	Considerações	32
4	Procedimentos Metodológicos	34
4.1	Processo para criação e análise da rede semântica	34
4.2	Criação do suporte automatizado	38
4.3	Ferramentas e tecnologias	40
4.3.1	Java	40
4.3.2	Gephi	40
5	Estudos de caso	42
5.1	Análise de redes de palavras baseada em títulos extraídos de um sistema de atendimento	43
5.1.1	Análise quantitativa	43
5.1.2	Análise da Rede Semântica dos Títulos das Solicitações	44
5.1.3	Considerações	46
5.2	WebMedia	46
5.2.1	Análise das Redes Semânticas de Títulos	47
5.2.2	Considerações	50
5.3	Sistemas Embarcados	51
5.3.1	Análise da Rede Semântica de Títulos	51
5.3.2	Classificação das Comunidades	54
5.3.3	Análise temporal das palavras-chave	56

5.3.4	Considerações	59
5.4	Publicações	59
6	Considerações Finais	61
6.1	Ameaças à validade	61
6.2	Discussão	62
6.3	Trabalhos Futuros	64
	Bibliografia	68

Lista de Figuras

2.1	Esquema simplificado de uma rede	9
2.2	Representação semiabstrata das 7 pontes da Prússia (a); Grafo representando as 7 pontes da Prússia (b)	10
2.3	Exemplo dos dois tipos de grafos	12
2.4	Rede aleatória formada a partir do modelo $G(n,p)$, com $n = 200$ e $p = 0,17$.	13
2.5	Distribuição de Graus de uma rede aleatória formada a partir do modelo $G(n,p)$, com $n = 200$ e $p = 0,17$	14
2.6	Ilustração dinâmica da realização do fenômeno small-world	15
2.7	Exemplo simplificado de distribuição de graus da diferença entre a rede aleatória e a livre de escala	16
2.8	Rede semântica formada por cliques	17
2.9	Sentenças de um discurso em forma de cliques (a); Após a junção das cliques, formação da rede semântica (b)	18
2.10	Estado inicial de cliques isoladas (a) e uma possível configuração para redes de cliques (b)	20
2.11	Componente Gigante	22
2.12	Equação da Centralidade de Grau	23
2.13	Nós com maior grau de intermediação	24
2.14	Equação da Centralidade de Intermediação	24
2.15	Equação da Centralidade de Proximidade	25
2.16	Grafo representando 3 comunidades determinadas pela modularidade . . .	26
2.17	Equação da regressão linear.	27
2.18	Exemplo de Regressão Linear	28
2.19	Cálculo do coeficiente linear (a) e coeficiente angular (b).	28

4.1	Processo Metodológico Referencial	35
4.2	Processo Metodológico Proposto	36
4.3	Módulo de Criação da Rede de Termos	39
5.1	Sub-rede formada contendo os pesos das arestas	45
5.2	Distribuição do grau para os vértices das redes semânticas	49
5.3	Rede Semântica	52
5.4	Evolução da rede semântica dos títulos na área de sistemas embarcados . .	54
5.5	Tendência do termo <i>Application</i>	57
5.6	Tendência do termo <i>Architecture</i>	57
5.7	Tendência do termo <i>Design</i>	57
5.8	Tendência do termo <i>Implementation</i>	58
5.9	Tendência do termo <i>Realtime</i>	58

Lista de Tabelas

5.1	Quantitativos por setor e seus respectivos índices de centralidade	44
5.2	Relação dos pesos em relação às arestas que conectam os vértices	46
5.3	Estatísticas referentes à rede semântica de títulos em português	47
5.4	Estatísticas referentes à rede semântica de títulos em inglês	48
5.5	Estatísticas referentes à distribuição do grau das redes semânticas.	50
5.6	Estatísticas referentes ao grau dos vértices da rede semântica dos títulos . .	52
5.7	Estatísticas referentes à rede semântica de títulos	53
5.8	Cinco maiores comunidades	55
5.9	Palavras-chave mais relevantes para comunidade A	55
5.10	Palavras-chave mais relevantes para comunidade B	55
5.11	Palavras-chave mais relevantes para comunidade C	55
5.12	Palavras-chave mais relevantes para comunidade D	56
5.13	Palavras-chave mais relevantes para comunidade E	56
5.14	Estatísticas referentes aos índices do coeficiente angular	58

Lista de Siglas

ARS Análise de Redes Sociais

CMA Coeficiente Médio de Aglomeração

CMM Caminho Mínimo Médio

DAM Diâmetro

GML Graph Modeling Language

TI Tecnologia da Informação

Capítulo 1

Introdução

Nas últimas décadas têm aumentado os estudos sobre o periódico científico pois, para os cientistas, a linguagem humana vem sendo cada vez mais caracterizada por uma perspectiva interdisciplinar. A pesquisa científica é definida como um conjunto de atividades que têm por objetivo a descoberta de novos conhecimentos de interesse para a solução de problemas existentes (LUNA, 1999). Ela é crucial para que ocorra o desenvolvimento, oferecendo novos subsídios que irão permitir indicar soluções adequadas para os problemas detectados pela utilização de novas técnicas desenvolvidas.

Embora a realização da pesquisa científica seja muito importante, igualmente necessárias é a sistematização e a divulgação destes conhecimentos produzidos. Assim, o conhecimento científico não se resume na descoberta de fatos e leis novas, mas também em sua publicação. Trata-se de obter e comunicar resultados através de artigos científicos, relatórios técnicos ou outras formas de produção científica. Essa literatura, hoje conhecida como periódico científico, foi criada em 1665 e, desde então, transformou-se, de um veículo cuja finalidade consistia em publicar notícias científicas, em um meio de divulgação do conhecimento, que se origina a partir das atividades de pesquisa de cientistas e colaboradores da ciência (MIRANDA; PEREIRA, 1996).

Muito esforço tem sido realizado pela comunidade científica para criar e manter diversos repositórios digitais para o armazenamento da produção científica, por exemplo, *IEEExplore*, o *ACM Digital Library*, *Science Direct*, entre outros. Visando compreender o complexo sistema de produção científica, inúmeras técnicas de análise vêm sendo aplicadas em repositórios digitais que, por sua vez, armazenam as produções científicas de forma

estruturada, facilitando a aplicação de técnicas semiautomatizadas para recuperação da informação. Essas técnicas objetivam, prioritariamente, a identificação de produções científicas relevantes, baseadas em indicadores quantitativos. Entre elas, podemos citar estudos de produtividade científica como o de Procópio, Laender e Moro (2011) e o de Newman (2004), que analisam uma série de estatísticas e indicadores para verificar a produtividade científica de determinada área ou evento, tais como: a média de artigos por autor, média de artigos por edição, média de coautores por artigo, entre outros.

A ideia de estudar os textos escritos, a partir do ponto de vista estatístico, originou-se no trabalho de Zipf (1972), que consistiu na contagem das palavras, e posteriormente, em outro trabalho, na correlação entre as quantidades das palavras e o *rank* de sua frequência no texto. O último indicou que existe um padrão de comportamento nos textos escritos. Cada ramo da ciência vê as palavras a partir de um ângulo próprio e com interesses bem definidos. O elemento principal deste trabalho é a palavra. A palavra como elemento de construção de um sistema com finalidades bem precisas, os textos escritos. Uma maneira simples de modelar este sistema é utilizando a teoria de redes sociais e complexas. Neste contexto, um periódico científico pode servir de palco para vários tipos de relações sociais (coautoria, citações, vocabulário comum, etc.).

Os títulos possuem um papel fundamental em um documento científico, pois é a primeira parte a ser lida. Ele é composto por palavras selecionadas pelos autores, na busca de uma representação sintética e fidedigna das ideias que serão apresentadas no corpo do trabalho. Através das palavras contidas nesses títulos, pode-se construir redes de palavras a fim de se perceber a relação, por exemplo, de um trabalho com outro. O uso de redes semânticas baseadas em títulos de artigos científicos pode auxiliar no entendimento da integração de coautores de um mesmo periódico. Desta forma, pode-se apontar evidências de como autores decidem os títulos de seus artigos científicos.

Portanto, essa rede de relacionamentos entre palavras é conhecida como rede semântica, que por sua vez, segundo Sternberg (2011), podem ser consideradas como a representação do conhecimento. Para a construção da rede de títulos, de acordo com os estudos realizados por Fadigas et al. (2009) e Pereira et al. (2011), cada título deve ser considerado uma clique¹, onde os vérticos são as palavras que o compõe. Porém, mesmo que as palavras, de um

¹Clique é um conjunto de n vértices mutuamente conectados.

título qualquer, busquem um significado único, a união de mais de uma clique pode possuir, semanticamente, diferenças. Contudo, caberá ao pesquisador, através dos dados fornecidos neste estudo, analisar e identificar a melhor interpretação dos termos encontrados.

Esta dissertação tem como principal foco o estudo da comunicação científica a partir das palavras do vocabulário comum de cientistas, encontradas em títulos de artigos científicos, publicados em um determinado periódico, e visa contribuir também com a identificação e classificação de comunidades nas redes semânticas, a fim de caracterizar determinadas áreas ou subáreas de pesquisa.

1.1 Motivação

Ao longo dos últimos dois anos foi criado um grupo de pesquisa voltado diretamente para a área de Análise de Redes Sociais (ARS), especificamente, em redes de citações, para a qual foi desenvolvida por Júnior (2015), uma ferramenta que coleta informações em um determinado repositório de artigos científicos, com intuito de identificar publicações relevantes em determinadas áreas do conhecimento. Destarte, em busca de fomentar essa pesquisa, resolveu-se adicionar novas funcionalidades ao referido *software*, focadas diretamente na análise de redes semânticas. Portanto, foi realizado um mapeamento sistemático sobre a área de ARS, especificamente, em redes semânticas, com objetivo de identificar métodos e técnicas para a construção dessa.

A questão de pesquisa que norteou este trabalho foi: “De que forma seria possível caracterizar meios científicos através da análise de palavras-chave, extraídas dos títulos, em artigos científicos?” Em face disso, propõe-se a criação de uma nova metodologia, com suporte semiautomatizado, que permita caracterizar áreas de pesquisa através da análise contextual encontrada nas palavras-chave dos títulos dos artigos científicos, armazenados em repositórios de dados digitais. Contudo, acreditamos que nossa metodologia servirá de suporte a qualquer pesquisador que tenha interesse em obter informações relevantes sobre uma determinada área de pesquisa.

1.2 Metodologia

O interesse dessa pesquisa foi analisar as publicações científicas a partir da criação de uma rede semântica, onde os nós representam as palavras-chave e as ligações são criadas entre pares de palavras que ocorrem em uma mesma sentença que, por sua vez, é vista como a menor unidade de significado de um texto. Assim, cada palavra pode ter um significado diferente a depender das palavras que estejam ao seu redor.

Para suportar a questão de pesquisa, mencionada na seção anterior, alguns procedimentos metodológicos foram devidamente aplicados. Inicialmente, realizamos uma revisão da literatura, com objetivo de identificar métodos e técnicas para a construção de redes semânticas. Em seguida, buscamos, na literatura atual, formas de analisar essas redes, com base em técnicas de ARS. Após a realização desses dois passos, identificamos um modo formal de extração e análise de dados de uma rede semântica. O passo seguinte foi a definição de um suporte semiautomatizado para criação e análise dessa mesma rede. Por fim, nossa metodologia foi avaliada e validada através dos seguintes estudos de caso:

- Análise de redes de palavras baseada em títulos extraídos de um sistema de atendimento;
- Geração e análise da rede semântica a partir dos títulos extraídos dos artigos publicados no evento *WebMedia*;
- Identificação de contexto semântico, ao longo do tempo, para caracterização da área de pesquisa e mapeamento de comunidades referentes ao tema Sistemas Embarcados.

1.3 Objetivos

A presente pesquisa teve como objetivo geral apresentar uma técnica que possibilitasse caracterizar meios científicos através da análise de palavras-chave em artigos científicos, ao longo do tempo, aplicando métricas de ARS. Pretendeu-se caracterizar tais áreas de pesquisa, através do estudo das redes semânticas, construídas a partir das palavras-chave dos títulos de determinados artigos publicados. Para atingirmos o objetivo geral desta pesquisa, algumas foram de fundamental importância:

- Realização de uma revisão da literatura sobre redes sociais e teoria dos grafos, com objetivo de identificar formas de representação e construção das redes semânticas;
- Desenvolvimento de suporte semiautomatizado para criação e análise de uma rede de palavras-chave no contexto de uma rede semântica com aplicação de métricas em ARS;
- Identificação das métricas de análise de redes sociais relevantes, visando à análise dos componentes de uma rede semântica. As métricas de principal interesse dessa pesquisa são aquelas que fornecem suporte a identificação de: (i) Grau de relevância, proximidade e intermediação de uma determinada palavra, (ii) Redes aleatórias, livres de escala e de mundo-pequeno, (iii) Caracterização de uma determinada comunidade científica baseado em algum tema de pesquisa, e por fim, (iv) Classificação das comunidades de palavras-chave;
- Definição de um *framework* de avaliação e validação da metodologia, para análise de redes semânticas dos títulos de artigos científicos, através da realização de estudos de caso feitos, tomando como base repositórios digitais de publicações científicas.

1.4 Contribuição

Este trabalho traz como principais contribuições: (i) definição de um processo metodológico para criação e análise de redes semânticas, (ii) implementação de um novo módulo, na ferramenta desenvolvida por Júnior (2015), voltado para análise de redes semânticas e (iii) através da aplicação da métrica de modularidade, podemos classificar as comunidades que possuem as palavras-chave mais relevantes, a fim de caracterizar possíveis áreas e/ou subáreas de pesquisa.

A partir da criação do suporte semiautomatizado para criação e análise da rede semântica, esperamos que os resultados possam ser obtidos em um menor espaço de tempo e com menos esforço, em termos da necessidade da avaliação humana. Com a realização dessa pesquisa, pretendemos avançar o estado da arte relativo à identificação de conteúdo relevante associado a publicações científicas.

Diante do exposto, esperamos prover a comunidade científica meios adequados para avaliação do processo formal de criação da rede semântica, um suporte semiautomatizado; e esperamos que as tecnologias envolvidas na construção da análise de redes semânticas possam estimular pesquisas científicas para o desenvolvimento de novos meios e métodos e, com isso, poder apontar melhorias para trabalhos posteriores.

1.5 Estrutura da Dissertação

Este trabalho encontra-se dividido em seis capítulos. O Capítulo 1 apresenta esta dissertação de maneira geral, descrevendo sua motivação, objetivos, contribuição e sua estrutura. O Capítulo 2, apresenta os conceitos e definições relacionados a publicações científicas, redes sociais, redes complexas, teoria dos grafos, e ARS através de métricas e propriedades da rede. Esses conceitos são empregados para a modelagem da rede semântica. No Capítulo 3, descrevem-se os trabalhos relacionados utilizados para embasar a pesquisa. No Capítulo 4 descreve-se a metodologia empregada para caracterização de determinadas áreas de pesquisa, através da criação de redes semânticas geradas a partir de títulos de artigos científicos. O Capítulo 5 apresenta os estudos de caso realizados (publicações) no contexto da presente pesquisa. Tais estudos, forneceram suporte para responder à principal questão da pesquisa. Finalmente, no Capítulo 6, elencamos as ameaças à validade da pesquisa e as considerações finais do estudo como apontamento para trabalhos futuros na área de pesquisa.

Capítulo 2

Fundamentação Teórica

O presente capítulo pretende modelar o referencial teórico necessário para o entendimento da pesquisa. Por ser um assunto amplo e de domínio de várias áreas de conhecimento, procuramos deixar explícitos apenas os principais conceitos utilizados. Temos particular interesse em definir como criar uma rede semântica a partir dos conceitos de redes sociais. Descreveremos os meios para avaliação dos aspectos quantitativos, associados à topologia da rede semântica, através do uso de métricas aplicadas na área de ARS. Por fim, adicionalmente, abordaremos conceitos relacionados com previsão de dados, a fim de planejar possíveis expansões referentes a contextos utilizados em determinadas áreas do conhecimento.

Iniciamos este capítulo apresentando os principais conceitos sobre o estudo dos significados (semântica), sistemas complexos, teoria dos grafos e seus índices através da estatística (Seção 2.1). Em seguida apresentaremos os conceitos associados a redes sociais, suas propriedades e métricas (Seção 2.2). Na Seção 2.3, serão apresentadas algumas das métricas mais utilizadas na área de ARS. Finalmente, na Seção 2.4, descrevemos um método quantitativo de previsão de dados, que foi utilizado em um dos estudos de caso dessa pesquisa, para análise de tendências das palavras mais relevantes, segundo a aplicação da métrica de centralidade.

2.1 Linguagem e Sistemas Complexos

Pelo fato da linguagem humana ser considerada complexa, definições sobre, especificamente, a palavra e os sistemas complexos foram necessárias para compreender o objeto de

nosso estudo.

Uma propriedade importante de um sistema complexo (ex: linguagem humana) é que o todo é sempre maior que a soma de suas partes, porque, se os elementos do sistema interagem, é comum perceber propriedades que emergem dessas interações, o que seria impossível se analisássemos o elemento sozinho (CUNHA, 2013).

Segundo Caldeira (2005), a unidade linguística é constituída da união de dois termos, e é denominada de signo linguístico que, por sua vez, une não uma coisa e uma palavra, mas um conceito (significado) e uma imagem acústica (significante). Portanto, diante do exposto, e baseado no dicionário ¹, consideraremos a seguinte definição sobre palavra: “unidade da língua escrita, situada entre espaço em branco”.

2.1.1 Palavras: Gramaticais e Lexicais

A reunião dos vocábulos de uma língua (léxico) compreende dois tipos de palavras, gramaticais e lexicais (MARTINS, 2003). Segundo Caldeira (2005) (apud Martins (2003)), existe cerca de cem palavras gramaticais, ou seja: preposições, pronomes, artigos, numerias, advérbios, conjunções e interjeições. Em síntese, elas têm a clara função de organizar o texto, seguindo regras mais ou menos fixas. Já as palavras lexicais, mesmo isoladas, despertam em nossa mente uma representação, seja de seres ou de suas qualidades, ação ou modo de ação. Essas palavras formam o maior percentual do léxico de uma língua e apresentam crescimento rápido, principalmente os substantivos, pois é necessário sempre criar novos nomes para representar as coisas do mundo real.

Essas palavras juntas em uma determinada sentença formam significações e, com isso, a capacidade linguística da raça humana. Como o volume da informação manipulada exige uma estrutura cerebral complexa, é necessária uma estrutura física, ou seja, neurológica, que permita manuseá-lo com eficiência, e de forma otimizada. Portanto, a partir dessas definições de complexidade expostas, um texto escrito, então, poderia ser caracterizado como um sistema complexo. De acordo com Caldeira (2005), sistemas complexos são estruturas que, de uma maneira simplificada, podem ser representadas como uma rede complexa, formada por elementos que se interligam a partir de regras de relacionamento, conforme ilustra a Figura 2.1. Portanto, para analisar esses sistemas, utilizamos o modelo de redes complexas

¹Dicionário eletrônico disponível em: <http://educacao.uol.com.br/biblioteca/>. Acesso em 11 jun 2015.

como base para nosso estudo.

2.1.2 Redes Complexas

Cientistas de diversas áreas do conhecimento, nas últimas décadas, aumentaram o interesse pelo estudo de sistemas naturais e sociais que contêm elementos que se relacionam entre si. Utilizadas como modelos que representam sistemas reais, as redes complexas geraram estudos em diversas áreas do conhecimento humano, por exemplo, redes biológicas, tecnológicas, sociais, entre outras (CUNHA, 2013).

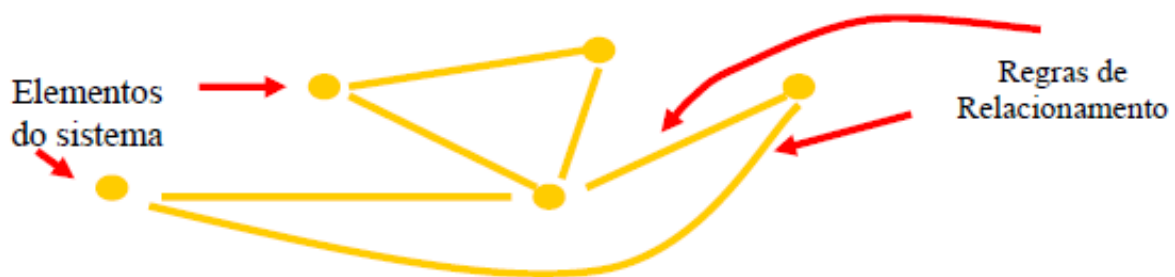


Figura 2.1: Esquema simplificado de uma rede
Fonte: (CALDEIRA, 2005)

O modelo de redes complexas gera estruturas com arquitetura não-linear, aberta, descentralizada, dinâmica, e capaz de auto-organização, caracterizadas fundamentalmente pela sua horizontalidade, isto é, pelo modo de inter-relacionar os elementos sem hierarquia (CALDEIRA, 2005). A matemática pura deu início ao estudo das redes a partir da **teoria dos grafos**, proposto por Euler em 1736.

Teoria dos Grafos

Os primeiros fundamentos da teoria dos grafos surgiram em meados do século XIX, quando Euler se dispôs a responder o enigma das sete pontes. O problema consistia em responder se havia um caminho para cruzar as sete pontes de sua cidade natal, Königsberg, conforme Figura 2.2, passando apenas uma vez por cada uma delas (NEWMAN, 2003).

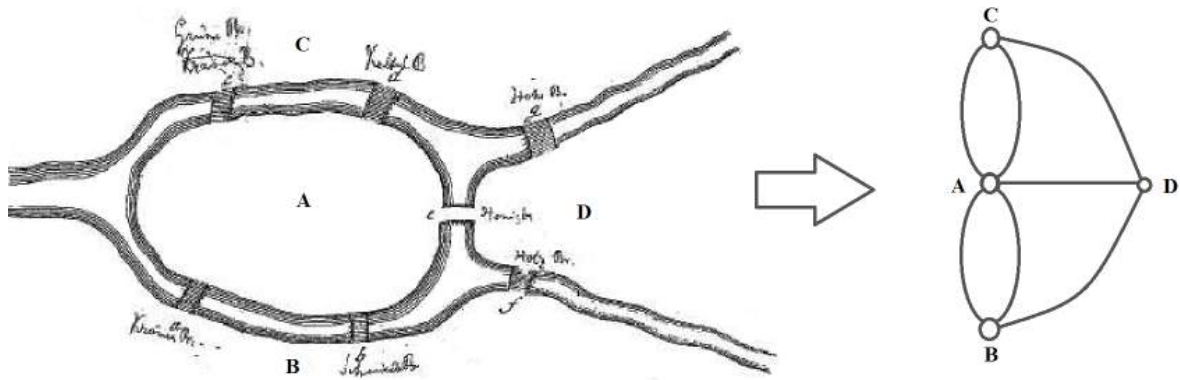


Figura 2.2: Representação semiabstrata das 7 pontes da Prússia (a); Grafo representando as 7 pontes da Prússia (b)

Fonte: (CRILLY, 2007)

Em 1735, Euler apresentou uma solução à Academia Russa, provando que era possível atravessar cada ponte do rio apenas uma vez. Em seu teorema, o passo chave foi tornar a Figura 2.2 (a) mais abstrata, obtendo como resultado (Figura 2.2 (b)), um grafo de linhas e pontos, onde o terreno é representado por pontos e as pontes por linhas.

Euler fez uma observação sobre a caminhada de sucesso. Excluindo o início e o final da caminhada, toda vez que uma ponte é atravessada até uma porção de terra, deve ser possível sair desta através de uma ponte que não foi atravessada anteriormente. Traduzindo esse pensamento em uma figura abstrata, podemos dizer que o encontro de linhas com um ponto deve ocorrer em pares. Excluindo dois pontos que representam o início e término da caminhada, as pontes podem ser atravessadas se, somente se, cada ponto tiver um número par de linhas incidindo nelas. O número de linhas incidindo no ponto é chamado de “grau” do ponto.

O Teorema de Euler afirma que as pontes de uma cidade podem ser atravessadas exatamente uma vez se, excluindo, no máximo, dois, todos os pontos têm o mesmo grau. Logo, um grafo pode ser considerado como um conjunto de pontos, denominados de vértices interligados por linhas, denominadas de arestas (SA; ROCHA, 2010). Segundo Ramos, Neto e Vega (2009), um grafo é um par ordenado (V, A) , onde V é o conjunto de vértices desse grafo e A é o conjunto de arcos que denotam uma relação binária com V ; em que os arcos ou arestas indicam associação entre os vértices.

Existem situações em que uma aresta tem um único vértice como origem e alvo dela.

Esse tipo de conexão é chamado de laço. Existem também arestas "paralelas", ou seja, arestas diferentes com o mesmo par de pontas (essas duas situações não foram consideradas nesta pesquisa).

Um **subgrafo** $S(V_S; A_S)$ é um grafo que pode ser obtido retirando arestas ou retirando vértices de um grafo $G(V; A)$. Dessa forma $V_S \subseteq V$ e $A_S \subseteq A$.

A simplicidade de tais conceitos torna claro que o grafo permite modelar situações concretas muito variadas, como exemplo: as redes de transporte, de comunicação, de citações, de coautoria, entre outras. Mas, apesar dessa simplicidade, o grafo se torna um modelo interessante, pois permite transportar informação suficiente para o esclarecimento ou resolução de uma situação modelada utilizando resultados e algoritmos da teoria dos grafos (SA; ROCHA, 2010).

Cálculo dos Parâmetros de Redes Complexas

Para entender melhor o comportamento de uma rede complexa, utilizando grafos, tanto quantitativamente quanto qualitativamente, utilizamos determinadas métricas. No presente trabalho, as métricas que foram utilizadas e que, segundo Cunha (2013), podemos considerar, são:

- **Grau e Grau Médio** - Grau é a quantidade de arestas em cada vértice, ou seja, informa o número de vezes que as arestas incidem sobre um determinado vértice. E grau médio $\langle K \rangle$ é a média desses valores. A conexão dos vértices em uma rede é devidamente classificada a partir da distribuição das frequências dos graus. Os casos mais considerados na literatura são: Redes Livres de Escala (*scale-free*); Redes de mundo pequeno (*smallworld*) e redes aleatórias. A próxima seção detalha mais esses modelos de redes.
- **Coeficiente de Agrupamento Médio** $\langle C \rangle$ - O coeficiente de agrupamento local de um vértice é calculado como sendo a proporção entre as ligações existentes, entre os seus vizinhos, em relação ao total das ligações possíveis entre eles. O $\langle C \rangle$ é a média dos coeficientes de agrupamento locais para todos os vértices da rede. Em suma, quanto maior for a quantidade de ligações entre os vizinhos de um determinado vértice, maior será a sua aglomeração.

- **Caminho Mínimo Médio $\langle l \rangle$** - Um dos mais importantes índices de grafos a ser considerado em redes é a distância típica entre cada vértice e todos os outros, isto é, o número mínimo de ligações que separam um determinado vértice x de outro y . Portanto, o $\langle l \rangle$ de um grafo é o valor médio para a distância mínima entre quaisquer dois vértices.
- **Diâmetro (DAM)** - O maior caminho mínimo médio é denominado de diâmetro da rede. E representa, em média, a maior distância entre os pontos mais distantes entre si;
- **Densidade** - Primeiramente, é preciso identificar a direção das conexões, para isso, existem dois tipos de grafos, os **direcionados** e os **não direcionados**, conforme ilustra a Figura 2.3. Os primeiros têm suas conexões representadas por flechas que indicam a direção da conexão; já nos últimos, a conexão é mútua e dispensa sua representação por flechas. Em uma rede não dirigida ou não direcionada, quanto mais arestas houver, mais densa ela será. Contudo, a densidade é calculada pelo resultado do total de arestas existentes na rede dividido pelo quantidade total de arestas possíveis.

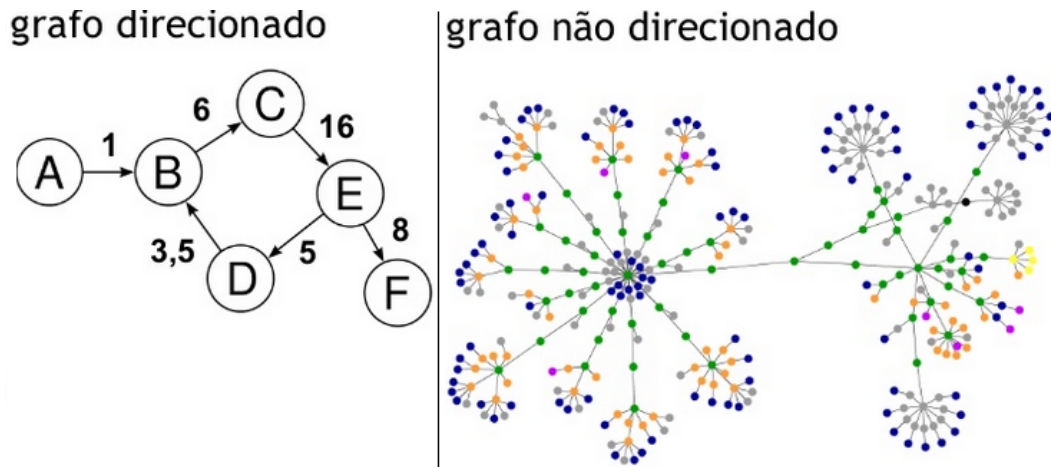


Figura 2.3: Exemplo dos dois tipos de grafos

Existem outros indicadores úteis, porém os supracitados são suficientes para as análises realizadas nessa pesquisa. Esses indicadores são fundamentais para se observar características em redes complexas com muitos nós, mesmo utilizando conceitos e métricas simples. A próxima seção mostra como esses índices de redes ajudam na caracterização de uma rede, quando se percebem padrões na dinâmica de conexão dos nós.

2.1.3 Modelos de Redes

Conforme visto anteriormente, diversas situações do mundo real podem ser convenientemente modeladas através de uma rede complexa que é representada através dos grafos. Essa estrutura é composta por um conjunto de pontos, juntamente com linhas que unem alguns pares de pontos. No entanto, para entender melhor um sistema complexo, é preciso levar-se em conta os relacionamentos entre suas partes. Contudo, ao passar dos anos, foram acrescentadas contribuições importantes aos estudos de Euler, principalmente através do trabalho de dois matemáticos: Erdos e Renyi (1960), a fim de identificar como as redes se formam. Esses estudos levaram à teoria das redes aleatórias (BARABASI, 2003).

Redes Aleatórias

Segundo Barabasi (2003), apud (ERDOS; RENYI, 1960), a melhor forma de explicar o mecanismo de construção das redes seria conectar os vértices aleatoriamente. Portanto, uma rede aleatória é obtida a partir de um conjunto $V(G) = v_1, \dots, v_n$ de vértices e adicionando-se arestas entre eles aleatoriamente. Esse processo de adição aleatório de arestas pode ser dado por uma probabilidade fixa p para cada vértice da rede se conectar com outro, conforme ilustra a Figura 2.4.

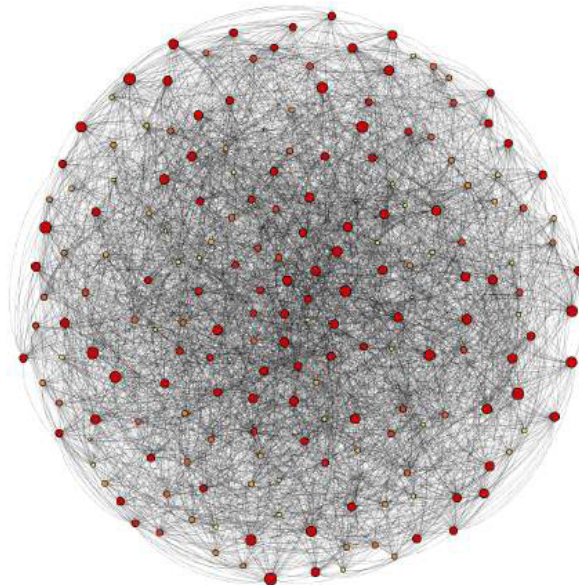


Figura 2.4: Rede aleatória formada a partir do modelo $G(n,p)$, com $n = 200$ e $p = 0,17$
Fonte: (CUNHA, 2013)

Uma das maneiras de identificar se uma rede é aleatória é fazer uma medida de quantos vértices possuem uma, duas, três até k ligações e depois traçar o histograma, que, no caso de uma rede aleatória, segue um modelo de distribuição de *Poisson* (curva normal), conforme ilustra a Figura 2.5 (BARABASI, 2003).

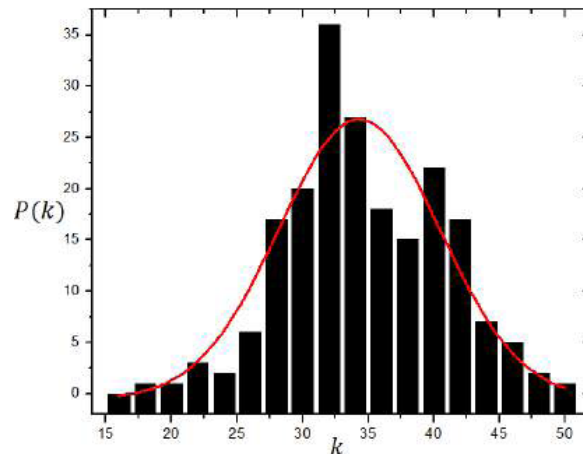


Figura 2.5: Distribuição de Graus de uma rede aleatória formada a partir do modelo $G(n,p)$, com $n = 200$ e $p = 0,17$

Fonte: (CUNHA, 2013)

No entanto, existem evidências encontradas na literatura, destacando a diversidade de sistemas naturais e sociais, cuja redes reais diferem totalmente do modelo de distribuição randômico. Da mesma forma, o CAM quase sempre é muito maior do que o das redes aleatórias. Destarte, outros modelos surgiram para dar conta da diversidade de formas de conexão entre elementos destes sistemas.

Redes de Mundo Pequeno (*Small-World*)

A rede *small-world*, também conhecida como o fenômeno dos "seis graus de separação", segundo Milgram (1967), é um modelo que descreve as redes que possuem caminhos curtos entre seus vértices. Os grafos dessas redes, segundo Watts (1999), devem apresentar as seguintes características:

- **Não direcionado** - conexão mútua, ou seja, sem direção definida;
- **Não ponderado** - A priori o peso das ligações entre os vértices não será levado em conta;

- **Simple** - ligações que unam um vértice a ele mesmo serão desconsideradas;
- **Arestas esparsas** - Conexões distribuídas de maneira dispersa;
- **Conectado** - qualquer vértice pode ser alcançado por qualquer outro vértice, cruzando um caminho composto de um número finito de ligações.

Diante do exposto, Watts (1999) propôs um modelo de construção de redes de mundo pequeno onde, a partir de uma rede regular (rede cujos vértices possuem o mesmo grau), fosse escolhido, aleatoriamente, algum vértice e também de forma aleatória se definisse a qual outro vértice da rede este se conectaria. Se continuamente isso for feito com todos os vértices da rede, surgirá uma rede completamente aleatória, conforme ilustra a Figura 2.6.

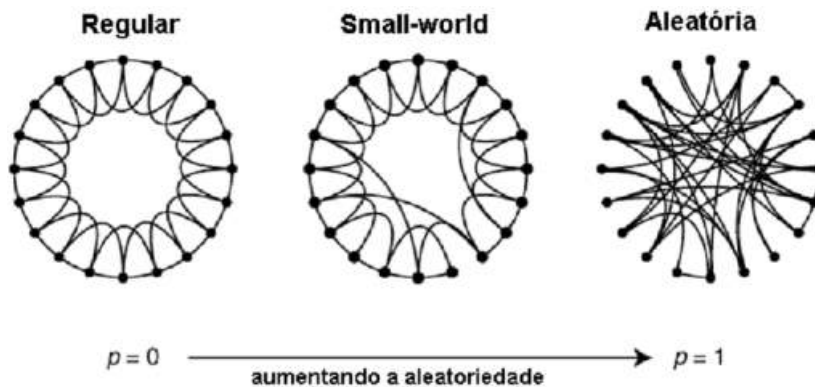


Figura 2.6: Ilustração dinâmica da realização do fenômeno small-world
Fonte: (WATTS, 1999)

Segundo Watts e Strogatz (1998), uma rede de mundo pequeno apresenta um valor de CMM pequeno e um elevado CAM, ou seja, se $\langle C \rangle \gg \langle C \rangle_{rd}$ e se $\langle l \rangle$ é comparável com $\langle l \rangle_{rd}$. Nessa definição, $\langle C \rangle_{rd}$ é o coeficiente de aglomeração médio para uma rede aleatória com mesmo grau médio $\langle K \rangle$ e mesmo número de vértices n . Analogamente, $\langle l \rangle_{rd}$ é o caminho mínimo médio para a rede aleatória correspondente.

Além das redes de mundo pequeno, outro parâmetro de caracterização das redes complexas, no que diz respeito às relações sociais, é a identificação das redes de escala livre.

Redes Livres de Escala (*Scale-Free*)

Identificadas como redes não aleatórias, trata-se da existência de um número relativamente pequeno de vértices que apresentam um grande número de conexões a outros vértices. Os

vértices altamente conectados são chamados de conectores. No entanto, por não apresentarem um valor padrão médio do número de conexões, não exibem uma escala característica. Esse modelo é capaz de descrever inúmeros sistemas naturais, e se caracteriza pela distribuição do número de conexões dos seus elementos, conhecida como distribuição *Power law* (BARABASI; ALBERT, 1999).

A distribuição de graus das redes livres de escala apresentam formato em lei de potência, conforme ilustra a Figura 2.7, que, por sua vez, demonstra a diferença de distribuição com uma rede aleatória.

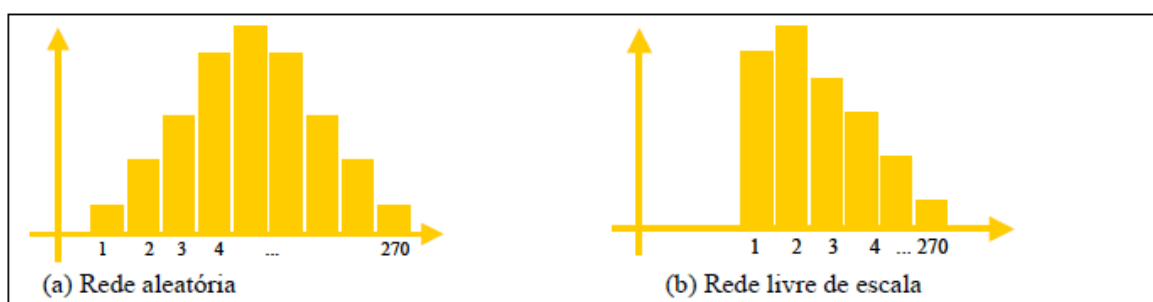


Figura 2.7: Exemplo simplificado de distribuição de graus da diferença entre a rede aleatória e a livre de escala

Fonte: (CALDEIRA, 2005)

Na prática, nas redes de livre escala em processo de crescimento, há sempre uma preferência dos novos vértices se ligarem aos vértices mais antigos que tenham mais ligações. No entanto, mesmo sendo rara a remoção de um vértice com maior grau (Hub), caso ocorra, desestrutura toda a rede fazendo com que vários vértices se desconectem. Nas redes aleatórias, essa possível remoção não desestrutura a rede, já que a maioria dos vértices possuem valores de grau médio muito próximos (BARABASI; ALBERT, 1999).

Além desses modelos abordados anteriormente, existem outros que, atualmente, vêm ganhando destaque na literatura: redes de cliques e as redes semânticas, que podem ou não se ajustar aos três primeiros modelos.

Redes Semânticas

Partindo do princípio de que o elemento principal de uma rede semântica é a palavra, podemos definir uma rede semântica como sendo um conjunto de elementos (palavras ou con-

ceitos) interconectados, que estão relacionados através dos símbolos linguísticos (STERNBERG, 2011).

No geral, são representadas de acordo com a teoria dos grafos: cada vértice da rede representa uma palavra e as arestas representam ligações entre essas palavras, conforme ilustra a Figura 2.8. Segundo Cunha (2013), existem três formas de configurar os relacionamentos entre palavras de um texto em uma rede não dirigida:

- **Linha** - onde os vizinhos de uma palavra são as duas palavras que se encontram imediatamente antes e depois dela;
- **Círculo** - onde a primeira e última palavra de cada sentença são consideradas vizinhas;
- **Cliques** - considera como a vizinhança de uma palavra todas as outras palavras da mesma sentença que ela.

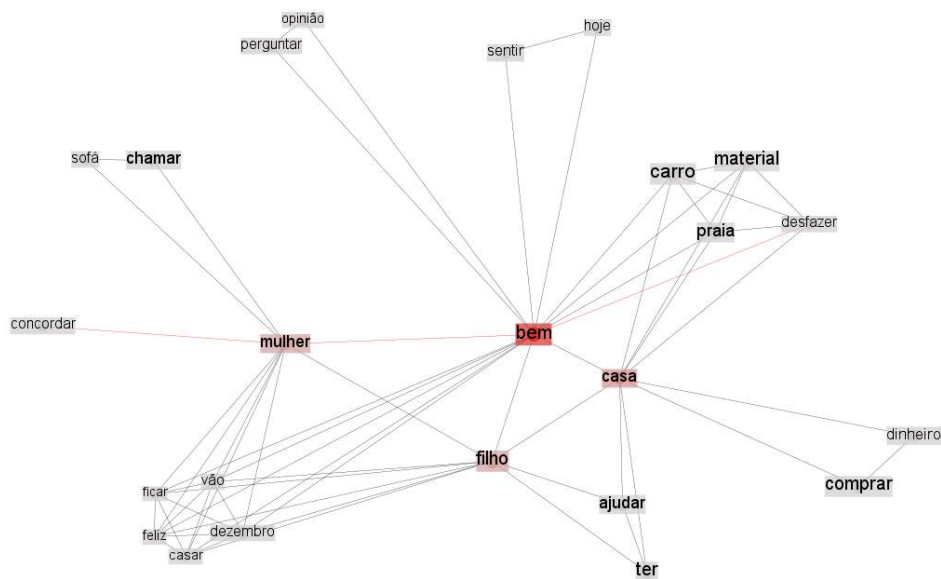


Figura 2.8: Rede semântica formada por cliques
Fonte: (CUNHA, 2013)

Diante do exposto, uma palavra pode apresentar diversos significados a depender de quais palavras ela tem como vizinha, logo, um conjunto de palavras, sintaticamente organizadas, tem significado próprio, compondo uma sentença (CALDEIRA, 2005).

De acordo com o tratamento proposto por Caldeira (2005) (formação de palavras-chave), o processo de junção das cliques, pelo compartilhamento de palavras comuns, é realizado conforme ilustra a Figura 2.9.

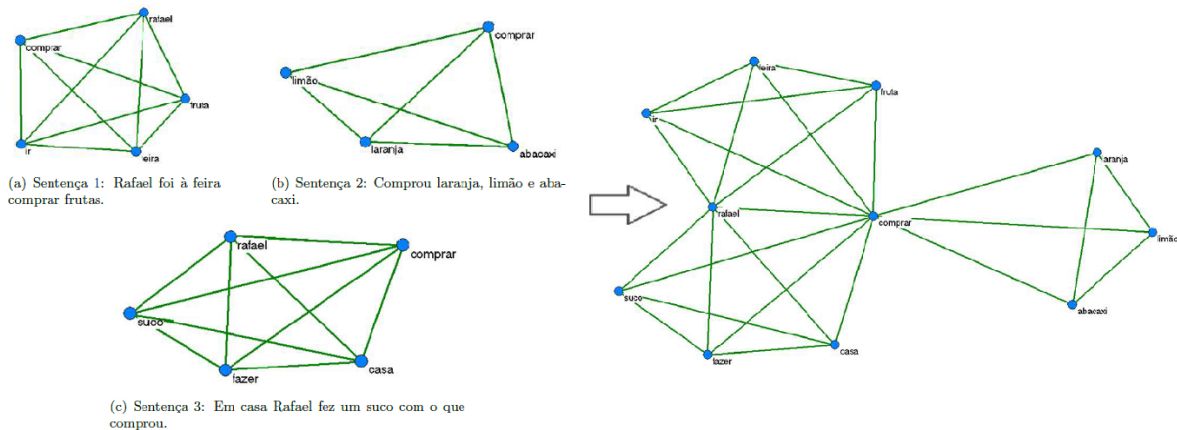


Figura 2.9: Sentenças de um discurso em forma de cliques (a); Após a junção das cliques, formação da rede semântica (b)

Fonte: (CUNHA, 2013)

Tomando como base as sentenças utilizadas por Cunha (2013), para a construção da Figura 2.8, podemos identificar que todas possuem, após o devido tratamento, a mesma palavra em comum "comprar", que, por sua vez, é responsável por conectar as três cliques através do processo de justaposição, formando um sistema único, ou seja, uma rede semântica de cliques.

Dessa forma, o título de um artigo científico transmite ao pesquisador as principais ideias do trabalho, ou seja, sintetiza ao máximo um significado através das palavras que o compõe, na tentativa dos autores de passar, no geral, a(s) principal(ais) contribuição(ões). Todo periódico científico dispõe de um conjunto de títulos que, se representados em forma de rede semântica, pode revelar tendências significativas sobre cada periódico e como suas contribuições estão relacionadas entre si. Portanto, o título serve como atrativo para a leitura do trabalho por outros pesquisadores que buscam agregar novos conhecimentos, bem como usá-los para produzir novas ideias. A análise dos índices das redes complexas é útil para caracterizar, diferenciar os periódicos e dar um caminho para o entendimento de como funciona a difusão do conhecimento das publicações da revista (CUNHA, 2013).

Redes de Cliques

O termo “clique” vem de Luce e Perry (1949), que usaram subgrafos completos em redes sociais para modelar cliques de pessoas, ou seja, grupos de pessoas que conheciam umas às outras. Uma clique, em grafo não direcionado $G = (V, E)$, é um subconjunto do conjunto de vértices $C \subseteq V$, tal que, para cada dois vértices em C , exista uma ponta de ligação entre os dois.

Em suma, clique é uma rede ou sub-rede que possui todos os seus vértices interligados entre si (ERDOS, 1966). Isto é, todos os vizinhos do vértice analisado estão conectados entre si, e este comportamento constitui-se numa ocorrência comum, por exemplo, em redes sociais, representando círculos de amigos, onde cada membro conhece todos os outros (NETO, 2003).

Destarte, os valores dos índices de densidade, coeficiente de aglomeração, caminho mínimo médio e diâmetro para uma clique terão valor igual a 1. São exemplos de aplicações desse tipo de rede: **redes de coautoria**, **redes de atores de filme**, **redes de discursos orais ou escritos** e **redes de títulos** (CUNHA, 2013).

O conjunto de cliques mutuamente conectadas é o elemento básico dessa rede, no entanto, existem duas maneiras de se conectar duas cliques. Segundo Fadigas e Pereira (2013), **justaposição** é o processo no qual duas cliques são ligadas por apenas um vértice comum. Já a **sobreposição** ocorre quando duas cliques são ligadas com dois ou mais vértices comuns, conforme ilustra a Figura 2.10.

Por fim, alinhando às redes complexas ao uso da teoria de redes sociais, pode-se contribuir de forma relevante para mapear a colaboração entre pesquisadores pertencentes a uma mesma comunidade científica, conforme explanaremos na seção a seguir.

2.2 Redes Sociais

2.2.1 Definições e Termos Gerais

Existem diversos tipos de rede espalhados pelo mundo, são exemplos: a *internet* (rede mundial de computadores), redes sociais *online*, conexões entre quaisquer indivíduos, até redes de conexão comercial, redes metabólicas, redes de distribuição tais como: energia elétrica,

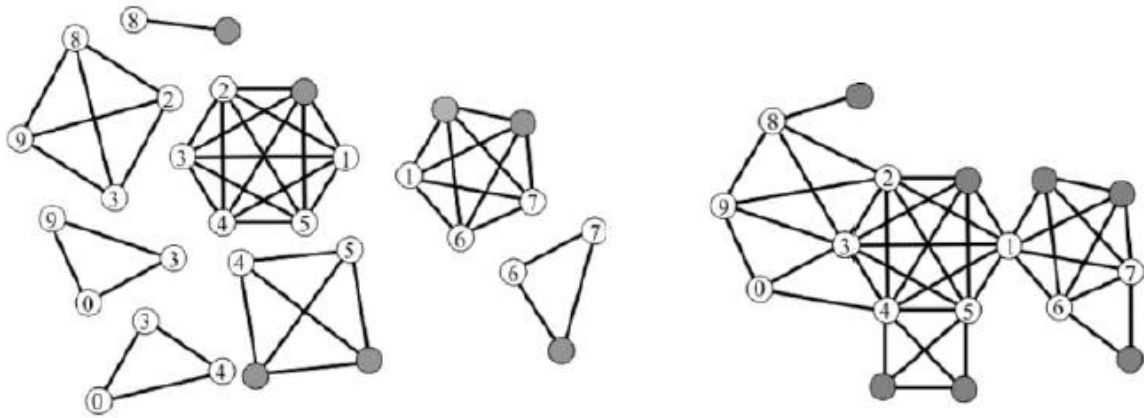


Figura 2.10: Estado inicial de cliques isoladas (a) e uma possível configuração para redes de cliques (b)

Fonte: (FADIGAS; PEREIRA, 2013)

serviços telefônicos, serviços de correio, entre outros (NEWMAN, 2003).

Uma rede, sucintamente, pode ser definida como um conjunto de pontos, que às vezes chamaremos de nós (ciências da computação) ou vértices (matemática), interligados por conexões denominadas arestas (NEWMAN, 2003). Essa morfologia permite identificar relações entre pares de nós, que podem ser, desde páginas web, pessoas, computadores, até mesmo cidades, países ou aeroportos. No caso específico das redes sociais, temos que os pontos podem representar pessoas; e as linhas, uma ligação de amizade; ou ainda os pontos podem ser centros de comunicação; e as linhas, o meio de comunicação entre eles (MIKA, 2007).

Rede social é definida como qualquer conjunto limitado de unidades sociais interligadas (BERKOWITZ, 1982). Essa definição destaca três características importantes das redes sociais:

1. As redes têm limites, isto é, existe algum critério para determinar a associação na rede;
2. Para fazer parte de uma rede social, cada membro deve ter ligações reais ou potenciais com pelo menos outro membro da rede. Essas "conexões" podem ser diretas ou indiretas;
3. Cada componente da rede é um elemento distinto do grupo, ou seja, uma unidade social.

As redes sociais e seus métodos para análise têm atraído consideravelmente a comunidade das ciências sociais e comportamentais com o intuito de analisar os relacionamentos entre entidades sociais, padrões encontrados e implicações desses relacionamentos (WASSERMAN; FAUST, 1994). Para estudar como os comportamentos ou as opiniões dos indivíduos dependem das estruturas nas quais eles se inserem, as unidades de análise deixam de ser os atributos individuais (classe, sexo, idade), passando a ser o conjunto de relações que os indivíduos estabelecem através de suas interações, na rede, uns com os outros (MARTELETO, 1994).

Numa rede social de amizades, por exemplo, é muito frequente que pessoas que não se conhecem tenham amigos em comum, o que torna a distância entre desconhecidos menor (baixo valor de $\langle l \rangle$ ou CMM). Outra característica muito comum é que amigos de uma determinada pessoa da rede se conheçam entre si (Alto $\langle C \rangle$ ou CAM).

2.2.2 Análise de Redes Sociais

A ARS é baseada na descrição formal das redes através de grafos. Através dela, é possível compreender e acompanhar, de forma mais eficaz, a disseminação de informações e a interação entre as pessoas que compõem a rede. Tal acompanhamento pode auxiliar na identificação de conectores de redes isoladas, melhorar a atuação de atores críticos, identificar a fragilidade da rede em relação à comunicação dos membros, dentre outros. Ela pode ser facilmente aplicada a uma vasta gama de unidades sociais como, por exemplo, em prol dos indivíduos, que é o caso das redes de apoio social. Em uma profissão diversa como o trabalho social, este tipo de análise tem aplicações diretas para o estudo da prática clínica, política social, organização comunitária e gestão organizacional (WASSERMAN; FAUST, 1994).

ARS tem como ponto de partida a premissa de que a vida social é criada principalmente por suas relações e os padrões formados por elas. Redes sociais são formalmente definidas como um conjunto de nós que são atados por um ou mais tipos de relações. Um estudioso, ao desejar analisar pesquisadores médicos que estudam doenças do coração, tem que levar em conta que pessoas devem ser consideradas como pesquisadores dessa área, pois isso pode ser comprometedor. Especialmente porque muitos analistas de rede evitam abordagens baseadas em grupos para a compreensão do mundo social. Após identificar os membros da rede, eles devem identificar as relações entre os nós. Estas podem incluir colaborações, amizades, laços

comerciais, links web, citações, fluxo de recursos, troca de apoio social ou qualquer outra ligação possível entre essas unidades especiais (SCOTT; CARRINGTON, 2011).

2.2.3 Componente Gigante

Na teoria de redes, um componente gigante é definido como um componente conectado, de um determinado grafo aleatório, que contém uma fração constante de todos os nós do grafo, conforme ilustra a Figura 2.11.

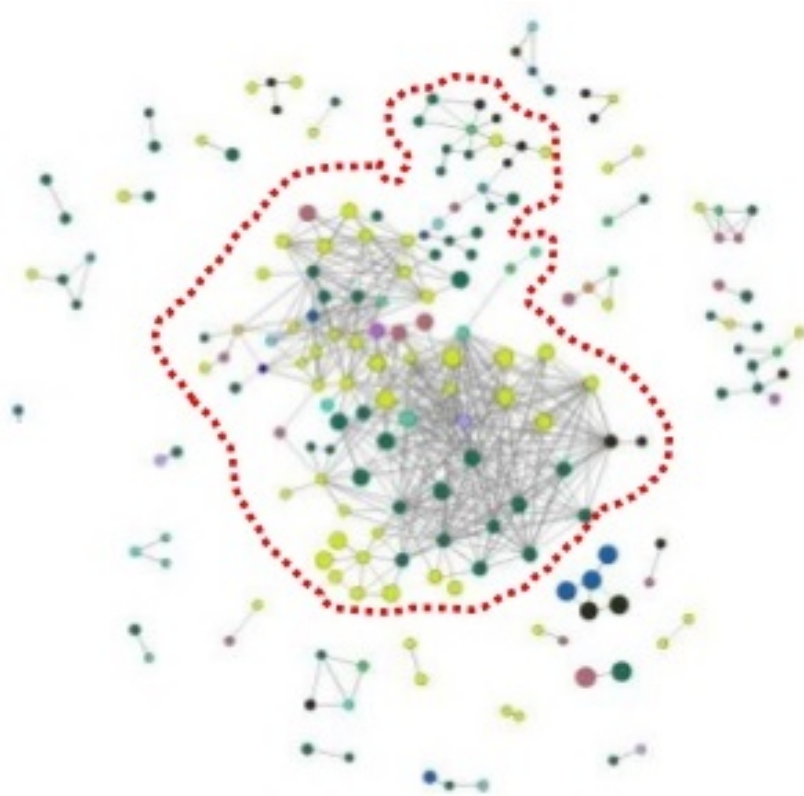


Figura 2.11: Componente Gigante

Na estrutura de um componente gigante, cada possível aresta, ligando pares de um dado conjunto de " n " nós, independentemente de outras arestas, possui uma probabilidade p . Nesse modelo, se $p \leq \frac{1-\epsilon}{n}$ para qualquer constante $\epsilon > 0$ então, teremos alta probabilidade de todos os componentes ligados do grafo terem tamanho $O(\log n)$, e com isso, não haverá nenhum componente gigante. Entretanto, para $p \geq \frac{1+\epsilon}{n}$, teremos uma alta probabilidade de existir um único componente gigante, para todos os outros componentes com tamanho $O(\log n)$. Para $p = \frac{1}{n}$, intermediário entre essas duas possibilidades, o número de vértices no maior

componente do grafo é com alta probabilidade, proporcional ao $n^{2/3}$ (BOLLOBAS, 2001).

2.3 Métricas para Análise das Redes Sociais

Nas seções seguintes serão descritas algumas métricas para estudo das redes que foram interesse dessa pesquisa. Além das métricas citadas na seção 2.1.2, outras foram propostas e podem ser aplicadas em face às características da rede a ser analisada. Uma lista mais extensa de tais métricas pode ser encontrada no trabalho de Newman (2003).

2.3.1 Centralidade de Grau

A centralidade de grau é definida como o número de ligações que incidem sobre o nó. A centralidade de grau de um vértice v_i , para um dado grafo $G = (V, A)$ com $|V|$ vértices e $|A|$ arestas está definido na Figura 2.12.

$$C_D(v_i) = \deg(v_i) = \sum_{j=1}^g x_{ij} = \sum_{j=1}^g x_{ji}$$

Figura 2.12: Equação da Centralidade de Grau
Fonte: (NEWMAN, 2010)

Na equação da Figura 2.12, poderão ser utilizados os valores das linhas i de uma matriz de conectividade, ou das colunas j da mesma matriz, para o cálculo do grau de centralidade de um nó. O valor x_{ij} ou x_{ji} , indica posição de linha/coluna (ij) ou coluna/linha (ji) de uma matriz de conectividade, onde esta é uma matriz binária de uma rede direcionada, em que as linhas representam os nós de partida e as colunas, os nós de chegada. O valor x igual a 1 indica que existe um enlace entre os nós i e j . A soma dos valores de x_{ij} ou x_{ji} indica o valor do grau do nó. O valor g representa o número total de nós ou de linhas/colunas da matriz de adjacências.

Vale ressaltar que, como o foco dessa pesquisa está voltado diretamente com grafos não direcionados, não há necessidade de mostrar os cálculos relativos tanto ao grau de entrada, quanto ao grau de saída.

2.3.2 Centralidade de Intermediação

A centralidade de intermediação é uma métrica utilizada para quantificar o controle de um ser humano sobre a comunicação entre outros numa rede social. Foi observado que os nós da rede que estão num caminho mais curto, escolhidos aleatoriamente entre dois outros nós (também escolhidos aleatoriamente), irão possuir uma elevada intermediação. Portanto, a centralidade de intermediação mede a quantidade de vezes que um nó age como ponte ao longo do caminho mais curto entre dois outros nós. Exemplificando, na Figura 2.13, os vértices A, B, C e D são os que possuem maior grau de centralidade de intermediação (FREEMAN, 1977).

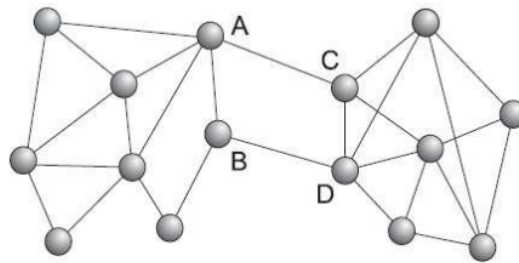


Figura 2.13: Nós com maior grau de intermediação

Para realizar o cálculo da intermediação de um vértice v num grafo $G = (V, E)$ com V vértices, temos que calcular os caminhos mais curtos entre cada par de vértices (s, t) , e determinar a fração de caminhos mais curtos que passam através do vértice em questão (vértice v) e por último somar esta fração de todos os pares de vértices (s, t) (BRANDES, 2001). A Figura 2.14 representa o cálculo da centralidade de intermediação.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Figura 2.14: Equação da Centralidade de Intermediação
Fonte: (BRANDES, 2001)

Onde, σ_{st} é o número total de caminhos curtos desde o nó s ao nó t e $\sigma_{st}(v)$ é o número desses caminhos que passam por v (BRANDES, 2001).

2.3.3 Centralidade de Proximidade

Segundo Freeman (1979), a proximidade está relacionada com o tempo que uma informação leva para ser compartilhada por todos os nós na rede. As medidas de centralidade de pro-

ximidade (*closeness centrality*) e centralidade de intermediação são baseadas na suposição que a informação (ou qualquer conteúdo da ligação) é transmitida somente ao longo dos possíveis caminhos mais curtos, denominados geodésicas. A centralidade de proximidade está relacionada com a distância total de um nó a todos os demais nós do grafo.

Segundo Sadidussi (1966), a centralidade de proximidade é baseada na soma das distâncias de um nó em relação aos demais nós do grafo. Seja G um grafo conexo com n nós e seja v_k um nó de G . A centralidade de proximidade de v_k é dada pelo inverso da soma das distâncias de v_k a todos os demais vértices do grafo, conforme cálculo representado através da Figura 2.15.

$$C_c(v_k) = \frac{1}{\sum_{j=1}^n dist(v_j, v_k)}$$

Figura 2.15: Equação da Centralidade de Proximidade
Fonte: (FREITAS, 2010)

2.3.4 Modularidade

Segundo Newman (2006), a modularidade é uma medida de estrutura de redes ou grafos. Foi designada para medir a força de divisão da rede em módulos (grupos ou comunidades). Redes com alta modularidade têm conexões densas entre os nós dentro de comunidades, mas ligações esparsas entre nós de diferentes comunidades. A modularidade é frequentemente usada em métodos de otimização para a detecção de estrutura da comunidade em redes.

A abordagem frequentemente utilizada na literatura de particionamento de grafos é observar as divisões dos vértices em dois grupos, de modo a minimizar o número de arestas entre os grupos. No entanto, o problema da estrutura de comunidade difere desse particionamento de grafos, no que se refere ao conhecimento antecipado do tamanho das comunidades. O problema é que a simples contagem de arestas não é uma boa maneira de quantificar uma estrutura de comunidades. A boa divisão de uma rede em comunidades não é apenas aquela em que há poucas arestas entre elas, mas sim quando há menos arestas esperadas entre as comunidades (NEWMAN, 2006).

Padrões biológicos e sociais, a *World Wide Web*, redes metabólicas, redes neurais, entre outros, são problemas do mundo real que podem ser matematicamente representados e

topologicamente estudados para revelar características estruturais inesperadas (NEWMAN, 2006). Muitas dessas redes possuem uma certa comunidade que tem importância substancial na construção de um entendimento sobre a dinâmica da rede. Por exemplo, uma comunidade social, intimamente ligada, implicará em uma taxa mais rápida de transmissão de informação ou rumor entre elas do que uma comunidade pouco ligada. Entretanto, se a rede é representada por um número de nós individuais conectados por *links*, significa um certo grau de interação entre os nós. Por isso, pode ser imperativo identificar as comunidades em redes, já que estas podem ter propriedades muito diferentes, tais como o grau do nó, o coeficiente de agrupamento, centralidade de intermediação, entre outros (NEWMAN, 2007). A modularidade é uma dessas medidas que, quando maximizada, leva ao aparecimento das comunidades em uma determinada rede.

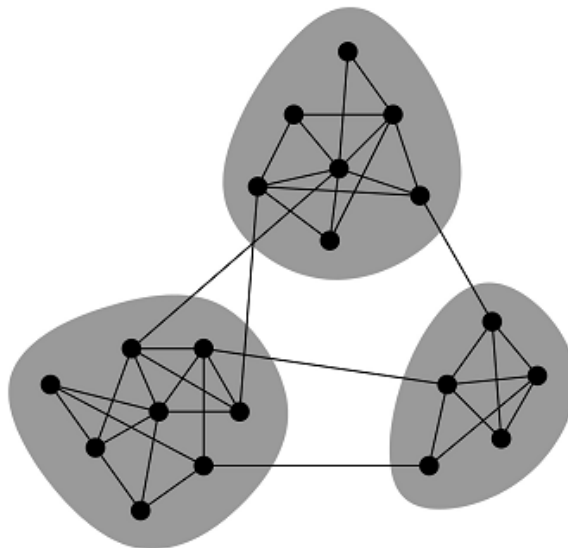


Figura 2.16: Grafo representando 3 comunidades determinadas pela modularidade
Fonte: (NEWMAN, 2006)

A modularidade é uma constante multiplicativa, ou seja, o número de arestas que caem dentro dos grupos, menos o número esperado de uma rede com arestas equivalentes colocadas aleatoriamente. A modularidade pode ser positiva ou negativa. Com valores positivos, indica a possível presença de uma estrutura de comunidade. Assim, podemos procurar essa estrutura, precisamente, olhando para as divisões de uma rede que têm valores positivos, e preferencialmente amplos, de modularidade.

No entanto, mesmo com os resultados obtidos através do uso das métricas supracita-

das, utilizamos outros recursos que pudessem evidenciar a evolução, da menção, das palavras mais relevantes ao longo do tempo em uma determinada área de pesquisa. Para tanto, decidiu-se aplicar um método para se estimar o valor esperado de um determinada variável. Portanto, abordaremos na seção a seguir, de forma conceitual, sobre o método utilizado nesta pesquisa.

2.4 Análise de Tendências

Para se obter uma previsão, existe uma série de métodos disponíveis, mas pode-se subdividi-los em dois grandes grupos: os métodos qualitativos e os quantitativos. No presente trabalho, focou-se diretamente em gerar uma análise temporal baseada nos dados estatísticos encontrados através das métricas de ARS, portanto, o método de previsão utilizado foi o quantitativo. Destarte, para realizar o cálculo de previsão sobre uma determinada demanda, uma das formas é utilizar um modelo matemático conhecido como regressão linear (MILORE; ANGELINI, 1995).

2.4.1 Regressão Linear Simples

O método dos mínimos quadrados é uma maneira de se obter a melhor reta que possa ser ajustada aos dados experimentais. Basicamente, é um procedimento que busca o mínimo de uma função de duas variáveis construída a partir da distância entre os pontos experimentais e os pontos de uma reta. A regressão linear é um caso simples em que se utiliza o método dos mínimos quadrados para investigar e modelar o relacionamento existente entre as diversas variáveis de um processo. Portanto, esse método consiste em uma variável chamada dependente, por está relacionada a uma ou mais variáveis independentes numa equação linear (WERKEMA; AGUIAR, 1996). Para se obter o cálculo da equação da reta basta aplicar a seguinte equação:

$$Y = a + bX$$

Figura 2.17: Equação da regressão linear.

Na equação 2.17, “Y” refere-se a variável dependente e “X” a variável independente. O “a” representa interseção da linha no eixo Y (coeficiente linear) e “b” a inclinação da linha

Capítulo 3

Trabalhos relacionados

O objetivo deste capítulo é apresentar os trabalhos relacionados que tiveram participação, direta ou indireta, ao tema da pesquisa. Temos o interesse principal de analisar de que modo as redes semânticas estão sendo utilizadas, com objetivo de fornecer meios para caracterização de áreas de pesquisa. É importante mencionar que, os trabalhos relacionados citados nas próximas seções, foram objeto de uma revisão sistêmica da literatura.

Iniciamos este capítulo apresentando um dos trabalhos de grande relevância, na literatura acadêmica, na área de análise topológica de textos escritos através das métricas utilizadas em redes complexas (Seção 3.1). Em seguida apresentaremos os principais trabalhos relacionados com a análise de redes semânticas através de títulos de artigos científicos (Seção 3.2). Finalmente, na Seção 3.3, abordamos alguns pontos críticos relacionados com estudos desta natureza, bem como, os diferenciais alcançados.

3.1 Análise de redes complexas

3.1.1 *The network of concepts in written texts*

O artigo de Caldeira et al. (2006) estudou o relacionamento entre textos escritos individuais, usando-os como ponto de partida para a construção de redes significativas. Projetando ambos os conceitos presentes no texto, bem como a forma como eles são relacionadas entre si em uma rede, dá a oportunidade de usar as ferramentas e conceitos desenvolvidos em seu âmbito, e caracterizar, de forma quantitativa, como os conceitos em um texto escrito aparecem, quão

ordenado e conectado eles são, e quão próximos estão, uns dos outros, dentro do texto. Uma vez que as palavras significativas estão relacionadas com conceitos, em mente, do autor, resulta em como as redes de texto podem descobrir padrões, em processos de comunicação e de linguagem, que ocorrem na mente.

Os resultados obtidos indicam que as redes analisadas têm conectividade altamente esparsa, com pequenos valores para as métricas de densidade e do caminho mínimo médio, mas com altos valores para métrica referente ao coeficiente de agrupamento médio, que constituem evidências de um cenário de rede *small-world*.

3.2 Análise de redes semânticas baseada em títulos de artigos científicos

3.2.1 Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática

O artigo de Fadigas et al. (2009), apresentou um processo de tratamento manual das palavras e também com ferramentas computacionais, em suma, aborda sobre a divulgação científica em educação matemática a partir de um diagnóstico quantitativo e qualitativo fundamentado em redes sociais e complexas. Após os devidos tratamentos, foi construída a rede semântica, analisando, segundo os conceitos de ARS, métricas de centralidade, a fim de identificar a importância, a frequência, das palavras nos títulos extraídos de artigos científicos.

A pesquisa mostrou que é possível agrupar as redes semânticas de palavras usadas nos títulos dos periódicos em basicamente dois grupos distintos, se a análise for fundamentada em índices de redes complexas. Se a análise for baseada em índices de redes sociais, outros dois grupos de revistas são observados, porém de composições diferentes dos dois primeiros. Com isso, ele contribui ao oferecer suporte para definição de estratégias que captem mais leitores, ajudando no processo de difusão do conhecimento em campos específicos.

Os resultados dessa análise mostram que o núcleo de palavras importantes dos títulos para a interseção das duas centralidades, compõe-se de “matemática”, “ensino”, “professor”

e “educação matemática”, acompanhando o padrão de outras redes.

Uma importante contribuição deste trabalho foi a de oferecer suporte para a definição de estratégias que captem mais leitores, ajudando no processo de difusão do conhecimento em campos específicos.

3.2.2 Redes de títulos de artigos científicos variáveis no tempo

Cunha et al. (2013), apresentaram uma abordagem para investigar a existência de padrões de relacionamento em uma comunidade científica e suas tendências ao longo do tempo, através de redes formadas por cliques a partir dos títulos de artigos científicos publicados no periódico *Nature*. Para análise das tendências ao longo do tempo, foi aplicado um método baseado na teoria encontrada em *Time-Varying Graph* (TVG), que segundo Casteigts et al. (2011), é um grafo estático $G = (V; E)$ acrescido de outros parâmetros que representam funções ou conjuntos temporais: ς (i.e. função de latência), Υ (i.e. função de presença) e Γ (i.e. tempo de vida), formando assim uma quintupla $G = (V, E, \varsigma, \Upsilon, \Gamma)$. A análise foi feita, semana por semana, utilizando indicadores atemporais.

Neste estudo, os resultados comparam o padrão de conexão de vértices de cada rede do TVG em épocas diferentes. As redes apresentam o fenômeno *small-world*.

3.2.3 Semantic networks based on titles of scientific papers

Pereira et al. (2011) apresentam a estrutura topológica de redes semânticas com base em títulos de artigos publicados em revistas científicas. Entretanto, a análise de uma rede de palavras é uma tentativa de compreender as associações entre um texto científico e seu título. Assim, o foco principal é a interpretação da linguagem humana, ou seja, como podemos usar essa linguagem para fazer uma conexão entre a ideia central de um artigo científico, geralmente expressa em seu título, e a descrição completa da proposta. Com isso, o principal objetivo da análise dos periódicos científicos, apresentados neste artigo, é mostrar que a ARS e a teoria de redes complexas são ferramentas práticas para o estudo de redes semânticas. Essas ferramentas podem contribuir para a difusão de conhecimento através de uma melhor compreensão de uma determinada revista e seu grau multidisciplinar.

Dentro deste contexto, é importante ressaltar que a análise da rede semântica apresenta

algumas reflexões sobre como o uso de modelos de redes sociais complexas podem contribuir para a difusão do conhecimento.

3.2.4 *A network approach based on cliques*

Fadigas e Pereira (2013), destacam que o uso da teoria de redes sociais contribui de forma relevante para mapear a colaboração entre pesquisadores pertencentes a uma mesma comunidade científica. Como resultado, eles investigaram propriedades (processos de justaposição e/ou sobreposição de cliques) e deram um novo sentido aos índices para redes exclusivamente formadas por cliques, por exemplo, redes de títulos, redes de coautoria, redes de atores de filmes. Contudo, na formação de redes de títulos, através de cliques isoladas, eles mostram o quão um índice de rede clássico varia em relação ao seu respectivo na configuração inicial. Dentro deste contexto, pode-se medir e interpretar de forma mais adequada as redes estudadas.

3.3 Considerações

A partir da revisão literária abordada anteriormente, todo trabalho que se propõe a desenvolver algum tipo de pesquisa relacionada às redes semânticas, deverá ter ciência das seguintes premissas:

1. Independente da biblioteca digital onde os metadados dos artigos científicos são extraídos, a rede de títulos deverá ser separada por idioma (em caso de haver mais de um), ou traduzir todos os títulos em um único idioma, a fim de alcançarmos resultados mais legíveis;
2. Em relação a criação das redes semânticas, as palavras-chave em sua maioria não mantêm uma padronização, fazendo que se torne difícil um processo automatizado para sua geração, com isso, faz-se necessário, a realização de um processo manual de padronização;
3. Atualmente, as ferramentas disponíveis para obtenção, análise dos dados e criação da rede ainda não estão devidamente integradas. Portanto, faz-se necessário, a utilização

isolada de algumas ferramentas (abordadas no Capítulo 4) para obtermos informações íntegras sobre uma determinada área de conhecimento.

De acordo com as premissas elencadas anteriormente, identificamos alguns pontos críticos em estudos desta natureza. Desta forma, a partir da análise dos principais estudos relacionados diretamente ao nosso tema e, visando avançar o estado da arte nesse tipo de pesquisa, nosso trabalho irá apresentar os seguintes diferenciais:

1. Nosso trabalho se propõe a integrar, um novo artefato a ferramenta desenvolvida por Júnior (2015), que possibilite realizar uma análise temporal, baseado nos títulos dos artigos, a fim de caracterizar uma determinada área de pesquisa;
2. Nossa metodologia permitirá que, a partir de um termo de busca, tenha-se como saída dados quantitativos e qualitativos necessários para análise temporal da rede semântica.
3. Trabalharemos com conceitos da teoria dos grafos e das redes complexas para inferirmos sobre a relevância de palavras-chave em um determinado contexto, proximidade, intermediação, CMM, CAM e peso na ligação entre essas palavras (densidade);
4. Nosso estudo permitirá a caracterização de áreas de pesquisa através da identificação de comunidades existentes dentro da rede semântica, destacando as palavras-chaves mais relevantes em cada uma delas.

Capítulo 4

Procedimentos Metodológicos

O modelo do processo de desenvolvimento metodológico, nesta pesquisa, pode ser caracterizado como iterativo e incremental, ou seja, durante o desenvolvimento dos estudos de caso, explanados no próximo Capítulo, a metodologia foi aprimorada de forma evolutiva.

Portanto, conforme mencionado no Capítulo 1, o presente trabalho tem interesse em obter respostas para a seguinte questão geral de pesquisa: “De que forma seria possível caracterizar meios científicos através da análise de palavras-chave, extraídas dos títulos, em artigos científicos?”. Com objetivo de responder a citada questão de pesquisa, foi proposta uma metodologia, utilizada no último estudo de caso (Seção 5.3), para análise de redes semânticas, criadas a partir de títulos de artigos científicos. Tal metodologia é baseada em redes sociais e está centrada nas seguintes atividades: (i) formalização do processo para criação da rede semântica e (ii) provimento de suporte automatizado para criação da rede semântica. Nas seções a seguir descreveremos em detalhes sobre cada uma dessas atividades.

4.1 Processo para criação e análise da rede semântica

No Programa de Pós-Graduação em Informática (PPGI) criou-se um grupo de pesquisa voltado a área de ARS, onde foi desenvolvido uma nova ferramenta, chamada *Paper Crawler*, que, segundo Júnior (2015), serve para realizar coleta dos metadados de uma página *HTML*, por exemplo, título e ano de publicação de um artigo científico, em uma determinada biblioteca digital acadêmica (*IEEEExplore*). Contudo, tomamos como referência inicial o processo metodológico desenvolvido por Júnior (2015), conforme ilustra a Figura 4.1.

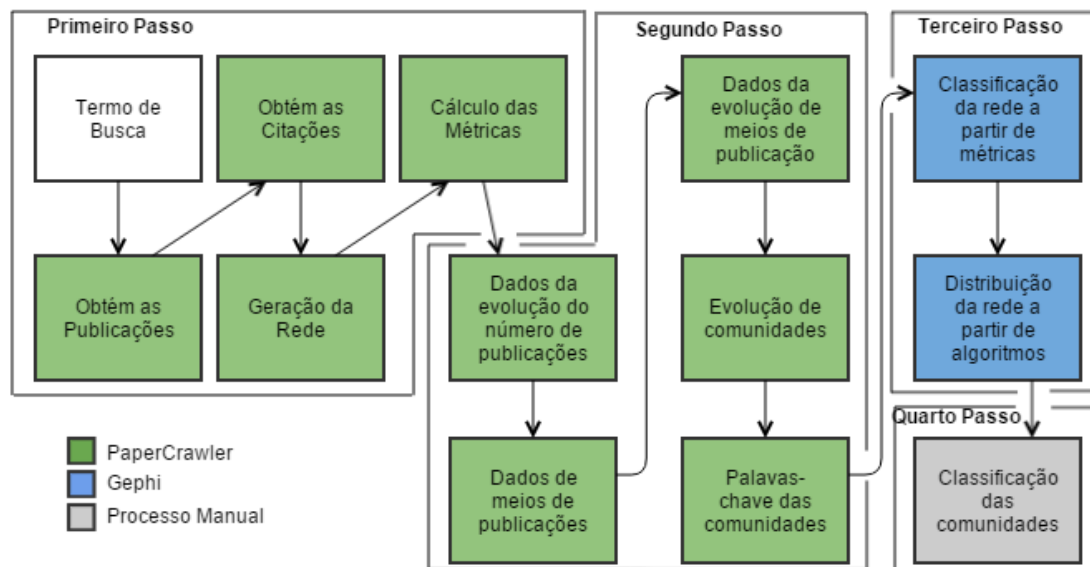


Figura 4.1: Processo Metodológico Referencial
Fonte: (JÚNIOR, 2015)

Essa metodologia referencial, é dividida em quatro passos, que por sua vez, são formados através de várias ações realizadas desde o envio do termo de busca, pela ferramenta, até o processo final de classificação das comunidades:

Passo 1: Modelagem da rede - Obtém as publicações e suas citações a partir de um termo de busca, com isso pode-se gerar a rede de publicações e realizar cálculo das métricas de ARS, como grau, grau de entrada, grau de saída, centralidade de intermediação, modularidade e *Page Rank*.

Passo 2: Análise da rede - Analisa os dados obtidos referentes a área de conhecimento como os meios de publicações, evolução das publicações ao longo dos anos, evolução do número de publicações para os meios de publicações ao longo dos anos, evolução das comunidades ao longo dos anos e por último obtém as *keywords* de cada comunidade.

Passo 3: Classificação e distribuição da rede - Obtém uma visualização amigável da rede para então analisar a área de conhecimento como um todo.

Passo 4: Definição e classificação de comunidades - A partir das *keywords* identificadas em cada comunidade, analisa de forma manual qual seria a melhor classificação para as cinco

maiores e mais relevantes comunidades.

Um dos objetivos principais deste trabalho de pesquisa é a formalização de um processo metodológico para identificação e análise temporal de artigos científicos, através de palavras-chave extraídas dos títulos, bem como para o mapeamento das comunidades relevantes de uma área de conhecimento. Para tanto, é necessário definirmos um novo processo metodológico para construção da rede semântica, que, por sua vez, pode ser interpretada como um sistema de representação do conhecimento variável no tempo, baseado em grafos. A criação deste processo envolveu uma série de etapas detalhadas, conforme mostra a Figura 4.2.

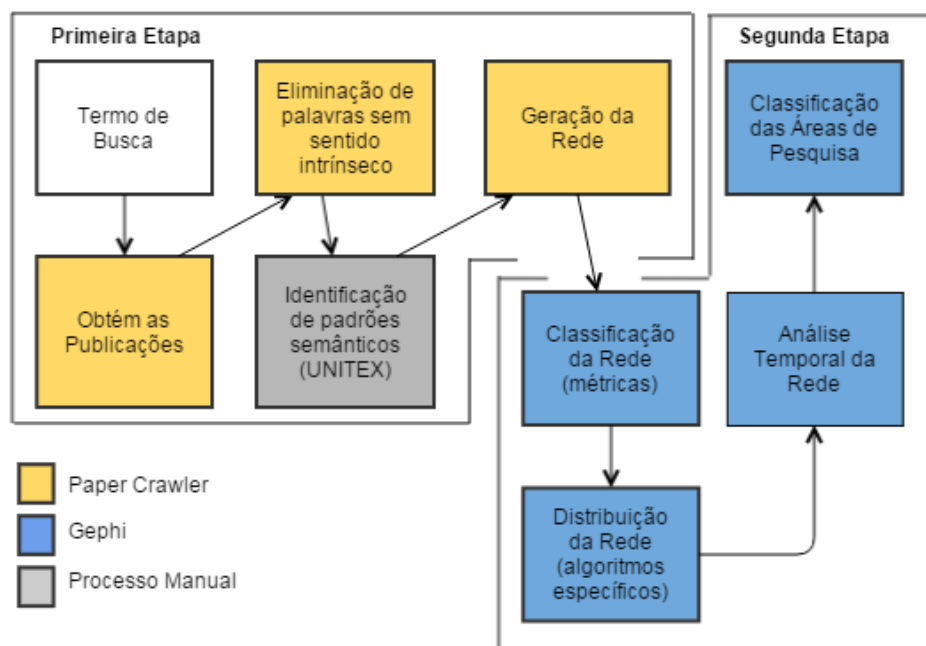


Figura 4.2: Processo Metodológico Proposto

A seguir descreveremos os significados das ações apresentadas na Figura acima (fluxo de caixas):

- **Consulta** - Informa um termo de busca qualquer (*query*) a ferramenta *Paper Crawler*;
- **Publicações** - A ferramenta retorna todos os títulos das publicações, bem como seus respectivos anos, encontrados no *IEEE Xplore*;
- **Refinando o texto** - A ferramenta elimina as palavras sem significados intrínsecos (artigos, pronomes pessoais e possessivos, adjetivos possessivos, demonstrativos, interrogativos, advérbios, caracteres especiais, acentuação etc.) e ao final é gerado um arquivo texto;

- **Formação de Padrão Semântico** - Manualmente, abrimos o arquivo texto, e alteramos algumas palavras que apesar de isoladas terem sentidos diferentes, quando juntas possuem sentido único, formando uma única palavra (forma padronizada), que por sua vez, são identificadas através do software livre utilizado nesta pesquisa, chamado *UNITEX*;
- **Geração da Rede Semântica** - Finalizado os devidos ajustes no arquivo texto, realiza-se o carregamento deste na aplicação (*Paper Crawler*), por fim, gera um novo arquivo, contendo a estrutura de um grafo não direcionado, com extensão *Graph Modeling Language* (GML), que por sua vez, é reconhecido pelo programa *Gephi*;
- **Classificação da Rede** – No *Gephi*, classificamos os nós e as arestas do grafo, por exemplo, quanto a sua cor (modularidade) e/ou tamanho (grau);
- **Distribuição da Rede** – Utilizamos algoritmos de distribuição existentes no *Gephi* (*Yifan hu proportional* e *Noverlap*) a fim de obter uma melhor visualização da rede;
- **Análise Temporal** – Através da criação de vários arquivos com extensão GML separados por ano, cria-se um projeto no *Gephi* unindo todas as redes semânticas criadas, formando uma única rede variável no tempo, e com isso, pode-se analisar a evolução das palavras, bem como das comunidades de uma determinada área de pesquisa;
- **Classificação das Comunidades** – Por fim, através das comunidades identificadas pelo *Gephi*, podemos analisar as palavras-chave que estão diretamente conectadas a uma outra de maior relevância em cada área, ou seja, classificamos um conjunto de palavras-chave, ligadas diretamente a uma outra palavra-chave de maior relevância. Contudo, o pesquisador terá a possibilidade de nomear comunidades relevantes dando um contexto numa determinada área de pesquisa.

Em suma, dois procedimentos principais são necessários: o tratamento manual e o tratamento com o uso de programas computacionais. A primeira parte, caracterizada como fase inicial da formação de dados, realiza-se uma "limpeza", por exemplo, remoção de *stopwords*, a fim de detectar termos que não farão parte da rede semântica. Baseado em Fadigas et al. (2009) seguimos as seguintes regras de normalização:

- Foram removidos dos títulos, para uma formação clara do grafo, preposições, pronomes, advérbios, artigos, acentos e caracteres especiais. As conexões entre os vértices ocorrerão de acordo com a predominância das ligações existentes entre palavras-chave nos títulos;
- Palavras repetidas no mesmo título foram excluídas, restando apenas uma ocorrência (processo automatizado);
- Conversão de duas palavras-chave consecutivas, com sentido único, em uma única palavra-chave, comumente conhecidas como *n-gramas*. Por exemplo, a expressão *real time* deve ser convertida para *realtime* (processo manual).

4.2 Criação do suporte automatizado

O processo de formatação do texto para tratar questões como: eliminação das ambiguidades, palavras compostas, caracteres especiais, entre outros, comumente conhecidas como *stopwords*, foi mapeado através do programa *UNITEX*, disponibilizado (*open source*) pela Rede Relex Brasil, que faz parte de um programa francês do *Laboratoire d'Automatique Documentaire et Linguistique (LADL)* (CALDEIRA et al., 2006). Contudo, através da identificação das palavras no *UNITEX*, foram geradas expressões regulares para substituição e/ou eliminação de determinadas palavras-chave no texto, objetivando assim uma padronização. Uma vez finalizada esta última etapa, de acordo com as regras propostas por Fadigas et al. (2009) e Pereira et al. (2011), através da ferramenta *Paper Crawler*, passamos então à construção propriamente dita das redes semânticas, realizada com base em duas operações bastante simples:

- i Cada título de artigo dá origem a uma "clique" onde todas as palavras, representadas por "nós", estão conectadas entre si;
- ii Existe uma aresta ligando "nós" de cliques distintas, se estes "nós" representam a mesma palavra em forma normalizada.

As conexões entre os nós ocorrerão de acordo com a predominância das ligações existentes entre palavras-chave nos títulos. O objetivo dessa atividade é a definição da implementa-

ção do suporte automatizado (i.e. *Paper Crawler*) para criação da rede semântica, descritas a seguir.

Funcionamento da *Paper Crawler* na criação de uma Rede Semântica

Foi adicionado a essa ferramenta um novo módulo, que por sua vez, está diretamente focado na análise de agrupamentos semânticos (palavras-chave) extraídos dos títulos de artigos científicos. Basicamente podemos descrever o funcionamento da aplicação com base na arquitetura descrita na Figura 4.3.

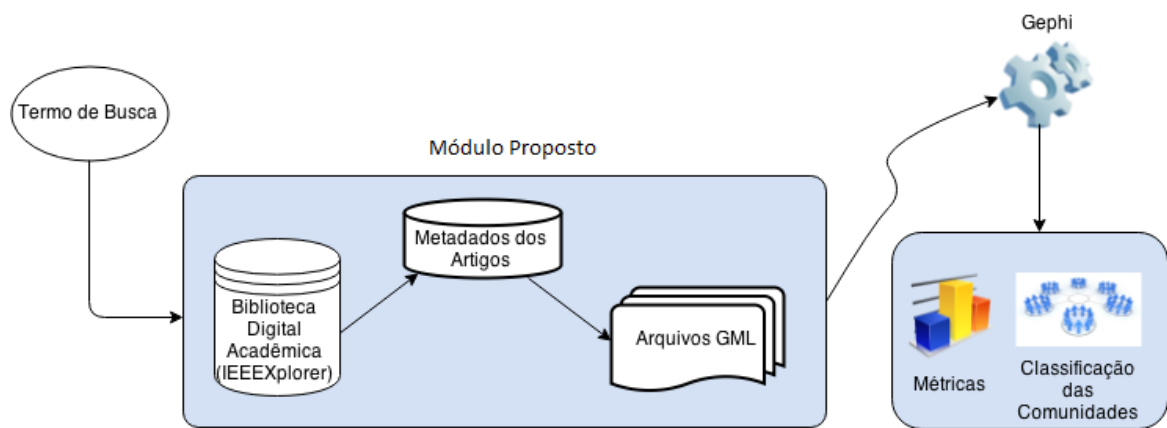


Figura 4.3: Módulo de Criação da Rede de Termos

Inicialmente, segundo Júnior (2015), o usuário escolhe um termo a ser pesquisado, a ferramenta *Paper Crawler* realiza requisições *HTTP* para a biblioteca digital acadêmica, por exemplo, *IEEEExplore*¹. O resultado destas requisições são páginas *HTML* com as listagens dos resultados, nelas a ferramenta realiza leituras para extrair os metadados dos artigos retornados, como, por exemplo, título, ano da publicação, entre outros. Contudo, em posse dos títulos e seus respectivos anos, manualmente, cria-se um novo arquivo a fim de realizar as devidas formatações do texto, conforme dito anteriormente. Destarte, após esse último procedimento, inicia-se a geração da rede semântica, onde através do novo módulo, é gerado um grafo separado por ano, isto é, um arquivo de extensão *GML*, para cada ano encontrado, com a estrutura da rede nele. Por fim, através do *Gephi*, cria-se um projeto agregando todos os arquivos gerados pelo módulo e, com isso, pode-se realizar um conjunto de análises

¹<http://ieeexplore.ieee.org>

quantitativas com o intuito de identificar a evolução das palavras-chave ao longo dos anos, o grau de relevância de cada palavra para cada comunidade, bem como caracterizar uma determinada área de pesquisa.

4.3 Ferramentas e tecnologias

Nas seções a seguir iremos expor as principais tecnologias utilizadas na criação do novo módulo integrado a ferramenta *Paper Crawler*, com foco na criação de redes semânticas variáveis ao longo do tempo. Adicionalmente, iremos expor as principais razões da escolha de tais tecnologias.

4.3.1 Java

Em 1991, a *Sun Microsystems* iniciou um projeto chamado *Green*. O resultado foi uma linguagem de programação baseada em C/C++, denominada de Java. A *Sun* expôs esta linguagem oficialmente em 1995, chamando a atenção do mundo dos negócios por causa da capacidade de desenvolver sistemas em plataforma *web*. Atualmente, Java é uma das linguagens mais utilizadas para o desenvolvimento de aplicativos de grande porte, para o aprimoramento de servidores *Web*, para o desenvolvimento de aplicações móveis e etc (DEITEL, 2005).

Contudo, como a primeira versão da ferramenta *Paper Crawler* foi desenvolvida toda em Java, por Júnior (2015), este módulo foi implementado na mesma linguagem.

4.3.2 Gephi

Gephi é uma ferramenta para pessoas que desejam explorar e entender grafos. Fazendo uma alusão com o *photoshop*, o usuário pode interagir com a representação gráfica, manipular as estruturas, formas e cores para revelar propriedades escondidas. É projetada de maneira modular dividido em diferentes módulos. Todas as características são envoltas em módulos separados, por exemplo, tem como funcionalidades para apoio a ARS: gerar a visualização em tempo real, *layout*, métricas, análise de redes dinâmicas, criação de cartografias, clusterização, grafos hierárquicos, entre outros.

O objetivo é dar suporte aos analistas de dados fazerem hipóteses, intuitivamente, descobrir padrões, isolar singularidades na estrutura ou falhas. É uma ferramenta complementar para estatísticas tradicionais, como o pensamento visual com interfaces interativas, é agora reconhecido para facilitar o raciocínio (BASTIAN; HEYMANN; JACOMY, 2009).

Capítulo 5

Estudos de caso

Neste capítulo iremos expor os principais resultados obtidos no contexto desta pesquisa. Sendo assim, apresentaremos os resultados associados a três estudos de caso: Resultados do estudo da análise de redes de palavras baseada em títulos de *tickets* extraídos de um sistema de atendimento (Seção 5.1), resultados da análise das edições do evento intitulado *Webmedia* (Seção 5.2) e os resultados do estudo na área de Sistemas Embarcados (Seção 5.3).

Na seção 5.1, a base de dados foi fornecida pela própria instituição, referente a um período de dois anos, contando, ao final, com um montante de 14.406 títulos dos chamados. Para tratamento do texto, por exemplo, verificação das *stopwords*, *n-gramas*, entre outros, utilizou-se o UNITEX. Quando necessário, as substituições eram realizadas no próprio texto, através do uso de expressão regulares, por fim, para formação da rede utilizou-se o mesmo algoritmo criado, que posteriormente seria adicionado a ferramenta *Paper Crawler*.

Na seção 5.2, os títulos, tanto em inglês quanto em português, foram coletados manualmente através de uma página disponibilizada pelo próprio evento. Para tratamento do texto foi empregada a mesma metodologia utilizada na seção 5.1.

Finalmente, na seção 5.3 foi utilizada exatamente a metodologia proposta no Capítulo 4. É importante destacar que todo o processo de análise, cálculo das métricas e visualização das redes foram realizados com suporte de uma ferramenta de terceiros, isto é, do *Gephi*.

5.1 Análise de redes de palavras baseada em títulos extraídos de um sistema de atendimento

Nessa seção, apresentaremos os resultados obtidos através da análise da rede semântica criada a partir das palavras-chave extraídas de títulos de um sistema de atendimento (*Help Desk*) na área de Tecnologia da Informação (TI). Com objetivo de melhor organizar a apresentação dos resultados associados a esse estudo de caso, dividimos o conteúdo nas seguintes seções: Na seção 5.1.1 será apresentada uma análise quantitativa em cima dos dados obtidos por cada setor solicitante do atendimento, na seção 5.1.2 serão expostos os resultados obtidos referentes a rede semântica, e com isso, ao peso existente nas ligações entre as palavras-chave (densidade das arestas). Em ambas as seções os resultados estão divididos em dois casos relacionados aos métodos de captura de referências descritos no Capítulo 3. Finalmente, na seção 5.1.3 descrevemos algumas conclusões alcançadas neste estudo.

5.1.1 Análise quantitativa

Nos títulos de cada chamado são fornecidas duas informações, o setor que originou a solicitação, e o assunto relativo ao problema que motivou a origem do chamado. Os vértices ou nós, representam cada palavra-chave, e o peso de cada aresta a quantidade de ligações entre dois vértices, ou seja, representa quantas vezes duas palavras-chave apareceram juntas nos títulos dos chamados. Porém, vale ressaltar que a simples quantificação não estabelece o mais importante na ARS, mas a relação estabelecida entre as palavras (métricas de centralidade).

A Tabela 5.1 mostra todos os quantitativos relativos ao número de solicitações por setor ou unidades administrativas, e seus respectivos índices de centralidade.

Consequentemente, de posse desses resultados, a direção do setor de T.I. poderá facilmente identificar que, por exemplo, o CCEN (Centro de Ciências Exatas e da Natureza) foi o centro que mais gerou ordem de serviço, ou seja, problemas referentes à área de T.I., em um período de dois anos. Podemos concluir também que, esse setor, através das métricas de centralidade, além de possuir o maior número de referências (grau), a maioria dos problemas passam por ele como "ponte" pelo caminho mais "curto" entre outras palavras-chave (intermediação).

Tabela 5.1: Quantitativos por setor e seus respectivos índices de centralidade

Setor	Índices						
	Nº Nós	Nº Arestas	Nº Títulos	Nº Palavras	Grau	Proximidade	Intermediação
CCEN	902	1895	1489	5708	351	2,239	463,176
CCHLA	891	1867	1273	4963	327	2,262	410,055
CCS	782	1739	1289	5176	278	2,325	298,971
CE	598	1147	726	2742	224	2,398	199,061
CI	273	459	236	940	105	2,535	42,450
CT	629	1359	902	3579	224	2,373	186,894

5.1.2 Análise da Rede Semântica dos Títulos das Solicitações

De acordo com uma pequena amostra da rede semântica criada a partir dos títulos das solicitações, conforme ilustra a Figura 5.1, seguindo o critério de centralidade de grau, as palavras **email** e **acessar** apresentam uma maior importância nas ligações entre os problemas apontados. Destarte, conseguimos evidenciar possíveis "gargálos" encontrados na área de T.I., bem como seus relacionamentos (peso das arestas) com outras palavras-chave. Contudo, pode-se concluir que as palavras **email**, **senha** e **acessar** são aquelas mais importantes, de acordo com os três tipos de métricas de centralidade abordadas nesta pesquisa. A interpretação da ARS, aplicada às redes de títulos, mostra que, essas três palavras citadas são aquelas que estão mais próximas de suas vizinhas na rede, bem como mais próximas de todas as outras (FADIGAS et al., 2009).

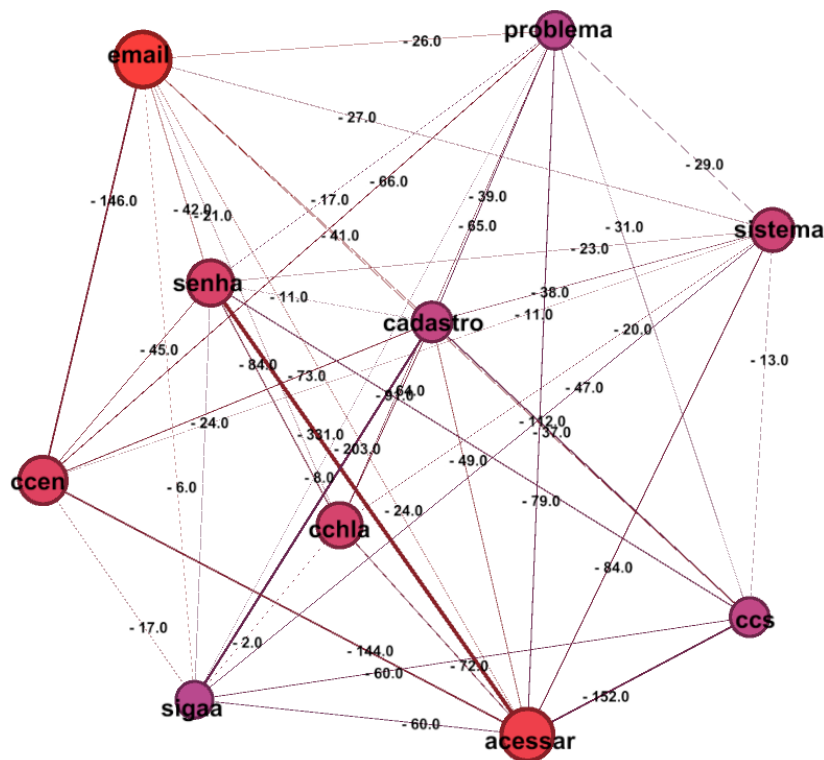


Figura 5.1: Sub-rede formada contendo os pesos das arestas

Outra análise interessante é que boa parte dos problemas relacionados com o CCEN estão diretamente ligados com a palavra-chave **email**. Portanto, pode-se apontar evidências de como a alta gestão poderá atuar de forma a dirimir possíveis problemas relacionados ao contexto mencionado naquele determinado setor.

A Tabela 5.2 mostra as ligações, não direcionadas, entre palavras dos chamados e o peso entre elas. O peso denota a quantidade de vezes que essa ligação existiu, podendo ser chamado também de densidade da aresta. Podemos evidenciar, a partir desse estudo, que a ligação origem-destino, por exemplo, entre as palavras-chave **email** e **institucional**, apresenta o maior número de ocorrência, destacando assim que, boa parte dos problemas referentes à TI estão diretamente ligados ao contexto **email institucional** e, com isso, ao concentrar esforços sobre este, diminuir especificamente cada ponto de estrangulamento que ocorre na área de TI.

Tabela 5.2: Relação dos pesos em relação às arestas que conectam os vértices

Origem/Destino	Origem/Destino	Peso
Email	Institucional	426
PortalCapes	Acessar	423
Acessar	Senha	331

5.1.3 Considerações

Com este estudo, podemos afirmar que o uso da metodologia de ARS juntamente com as características de exploração visual disponibilizadas na ferramenta *Gephi*, permitiram e facilitaram a combinação de elementos para entendimento e avaliação do modo de comunicação, troca de informações e problemas encontrados na comunidade estudada. O padrão da comunicação dos vértices, ou palavras-chave, responsáveis pela dinamização da rede, foi devidamente avaliado e identificado. Estes indivíduos atuam como mediadores, colaborando com a identificação de problemas.

A visão fornecida, neste estudo, complementou as informações geradas através da base de dados do sistema de *Help Desk* de uma determinada instituição de ensino referente à área de TI. Introduziu uma visão dos problemas direcionados por setor, bem como das palavras-chave mais mencionadas nos títulos dos chamados. A pesquisa mostrou que, o núcleo de problemas mais recorrentes nos títulos dos chamados, de acordo com o grau de centralidade, compõe-se das palavras-chave **email**, **acessar** e **senha**, na ordem de maior incidência para menor respectivamente, bem como, os contextos, **email institucional** e **portalcapas acessar**, apresentam um maior número de ligações entre palavras-chave (peso das arestas).

Contudo, através deste mapeamento, podemos destacar que o uso de redes semânticas baseadas em títulos, no geral, auxilia no entendimento da integração das palavras-chave. Desta forma, pode-se apontar evidências de como a alta gestão poderá atuar no processo de tomada de decisão sobre possíveis soluções, no que diz respeito aos problemas que ocorrem com maior frequência em um determinado local.

5.2 WebMedia

O Simpósio Brasileiro de Sistemas Multimídia e Web, *WebMedia*, chegou em 2014 à sua vigésima edição consolidado como o principal evento nacional em sua área de atuação. Com

o objetivo de caracterizar a comunidade científica atuante na área, analisamos o corpus de 453 artigos científicos publicados nas dezoito edições anteriores do evento realizadas entre os anos de 1995 e 2012.

Este estudo se desenvolveu focado na caracterização das áreas de pesquisa atuantes no evento, isto se deu com base em uma rede montada com palavras significativas presentes nos títulos dos artigos publicados no evento, formando uma rede de títulos.

Foram utilizadas técnicas de ARS para permitir a avaliação não só de componentes isolados como também, e principalmente, dos relacionamentos entre estes componentes, permitindo-nos apontar contextos mais relevantes/atuentes nas edições anteriores do evento de acordo com uma série de diferentes métricas de análise.

Conforme mencionado anteriormente, neste experimento exploramos o conteúdo dos títulos dos artigos publicados no *WebMedia* obtidas a partir dos anais das dezoito edições anteriores. As listas de títulos foram salvas em arquivos texto separadas por ano e utilizadas para gerar uma rede semântica para cada ano do evento.

Estas redes foram modeladas como grafos não direcionados onde os vértices representam as palavras-chave e as arestas indicam as relações entre elas. Contudo, na seção seguinte será apresentado os resultados obtidos com a análise das redes semânticas de títulos.

5.2.1 Análise das Redes Semânticas de Títulos

Foram criadas duas redes semânticas a partir das palavras nos títulos dos artigos publicados no *WebMedia*, uma para os títulos em português e outra para os títulos em inglês. As tabelas 5.3 e 5.4 apresentam dados referentes a evolução dessas duas redes ao longo do tempo.

Tabela 5.3: Estatísticas referentes à rede semântica de títulos em português

Métrica	1995-1997	1995-2000	1995-2003	1995-2006	1995-2009	1995-2012
N	113	269	389	527	708	875
K	7,664	10,022	11,069	12,679	14,218	14,768
C	0,839	0,835	0,825	0,811	0,803	0,809
L	2,701	2,592	2,785	2,683	2,666	2,703
D	0,068	0,037	0,029	0,024	0,02	0,017

Os dados apresentados em tais tabelas são relacionados às seguintes métricas referentes a topologia das redes de títulos utilizadas neste estudo:

Tabela 5.4: Estatísticas referentes à rede semântica de títulos em inglês

Métrica	1995-1997	1995-2000	1995-2003	1995-2006	1995-2009	1995-2012
N	48	91	148	313	389	485
K	8,167	7,758	7,689	9,502	10,242	10,742
C	0,917	0,893	0,882	0,852	0,833	0,825
L	2,035	2,466	3,077	2,882	2,917	2,985
D	0,174	0,086	0,052	0,030	0,026	0,022

- **Número de vértices (N);**
- **Grau Médio (K);**
- **Coefficiente de Agrupamento Médio (C);**
- **Caminho Mínimo Médio (L);**
- **Densidade (D).**

A partir dos dados apresentados nestas duas tabelas podemos concluir que desde o seu início as palavras utilizadas na elaboração dos títulos dos artigos publicados no *WebMedia*, tanto em português quanto em inglês, formam uma rede mundo-pequeno (*small-world*). Essa conclusão se deve aos altos valores de C, $0,803 \leq C \leq 0,839$ para títulos em português e $0,825 \leq C \leq 0,917$ para títulos em inglês, alinhados aos pequenos valores observados para L, $2,592 \leq L \leq 2,703$ para títulos em português e $2,035 \leq L \leq 3,077$ para títulos em inglês, isto quando comparados com os valores esperados de C de uma rede aleatória com as mesmas quantidade de vértices e valores semelhantes para K e L.

O estudo da topologia da rede semântica de títulos nos dá dicas em relação a disciplina coberta pelos artigos publicados na conferência. Cada área tem seu conjunto de jargões e termos técnicos e muitas das propriedades destas redes são baseadas na diversidade do vocabulário utilizado para elaborar os títulos dos artigos. A densidade (D) e o caminho mínimo médio (L) são pontos de partida iniciais para analisar tais propriedades.

A densidade reflete a quantidade de ligações entre as palavras encontradas nos títulos e indica a tendência de palavras aparecerem juntas em um grande conjunto de palavras. Os dados nas tabelas 5.3 e 5.4 mostram a evolução da densidade das redes de títulos do *WebMedia* até chegar aos valores de 0,017 e 0,022 para as redes em português e inglês respectivamente durante todo o período analisado. A título de comparação, as redes de títulos dos artigos

publicados na *Nature* e *Science* apresentam uma densidade inferior a 0,002, destacando o caráter extremamente multidisciplinar destas publicações. Já a densidade da rede de títulos de artigos publicadas na Revista Brasileira de Informática na Educação (RBIE) se aproxima de 0,035, caracterizando um meio de publicação voltado a um tema mais específico (PEREIRA et al., 2011).

O fenômeno de mundo-pequeno observado nas redes semânticas de títulos indica que as palavras estão muito próximos em termos de distância em um grafo. Ou seja, valores pequenos para L implicam em distâncias muito pequenas para palavras pertencentes aos títulos dos artigos científicos publicados no *WebMedia*. Apesar da baixa densidade, a distância mínima média entre as palavras nos títulos dos artigos publicados no *WebMedia*, 2,7 para os títulos em português e 2,9 para inglês, indicam que o evento tende a seguir uma linha multidisciplinar uma vez que esta distância se aproxima bastante das observadas nas redes de títulos de revistas como *Nature* e *Science*, ambas na casa dos 2,9 (PEREIRA et al., 2011).

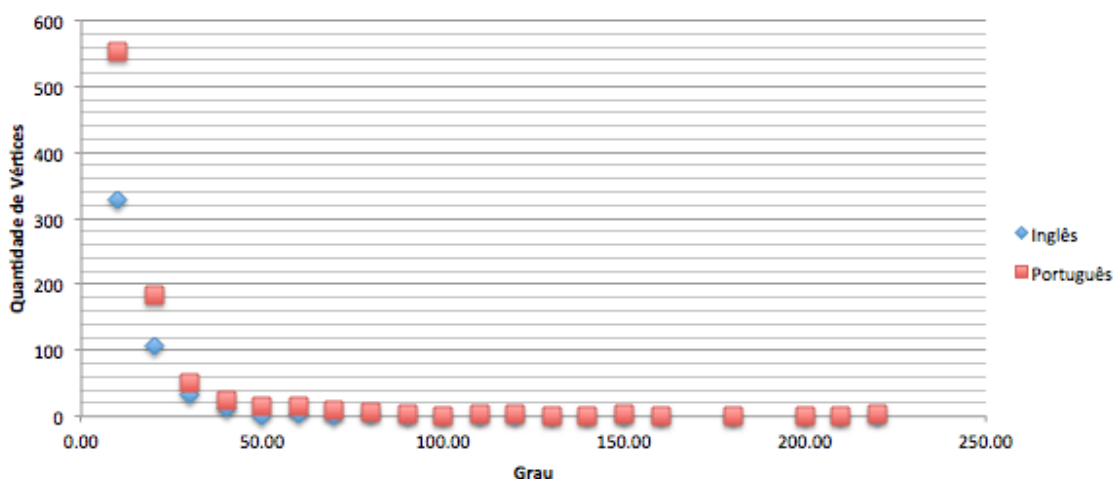


Figura 5.2: Distribuição do grau para os vértices das redes semânticas

As distribuições de grau das redes semânticas tanto em português quanto em inglês, exibidas na Figura 5.2, apresentam um comportamento semelhantes: ambas seguem uma distribuição *power-law*. Esta distribuição aparenta ser uma tendência em redes livres de escala como pode ser observado nos resultados apresentados em (PEREIRA et al., 2011). Este efeito se explica pela existência, tanto na rede para os títulos escritos em português quanto em inglês, de um pequeno número de vértices com grau elevado e de um grande número de vértices com grau reduzido.

Tabela 5.5: Estatísticas referentes à distribuição do grau das redes semânticas.

Português	Inglês
Web(219)	<i>Web</i> (116)
Multimidia(219)	<i>Application</i> (73)
Aplicacao(119)	<i>Multimedia</i> (73)
Ambiente(154)	<i>Environment</i> (57)
Servico(142)	<i>Approach</i> (54)
Digital(140)	<i>System</i> (49)
Suporte(115)	<i>Hypermedia</i> (39)
Video(114)	<i>Document</i> (36)
TV(113)	<i>Video</i> (36)
Sistema(112)	<i>Support</i> (33)

A Tabela 5.5 lista os dez vértices de maior grau, representando palavras em sua forma normalizada, presentes nas redes de títulos em português e inglês com seus respectivos graus. É possível observar que há uma certa uniformidade entre as palavras mais utilizadas nos títulos dos artigos independente do idioma e este conjunto tende a evidenciar uma natureza mais aplicada nos resultados de pesquisa reportados no *WebMedia* ao apresentar grande destaque para termos como aplicação, ambiente, serviço, suporte, sistema, *environment*, *approach*, *system* e *support*.

5.2.2 Considerações

Este estudo processou os metadados dos 453 artigos publicados nas 18 edições do *WebMedia* realizadas entre os anos de 1995 e 2012 com o intuito de analisar a evolução da estrutura da comunidade científica responsável pelos artigos publicados no evento.

Caracterizamos grupos diferentes de periódicos a partir da densidade, coeficiente médio de aglomeração e do caminho mínimo e comparamos os resultados com periódicos escritos em português e inglês. O uso de técnicas de ARS juntamente com exploração visual manual, permitiram e facilitaram a combinação de elementos para entendimento e avaliação das estruturas das redes semânticas dos títulos dos artigos científicos. A análise dessas redes semânticas, através da aplicação das métricas de redes complexas, por exemplo, a densidade, evidenciou que o *WebMedia* apresenta características de evento multidisciplinar com foco em pesquisa aplicada.

Um fator que chama atenção é a baixa inserção e impacto internacional do evento. Ape-

nas 28% dos artigos publicados nestas 18 edições anteriores foram escritos em inglês.

Os resultados mostram que as redes têm topologias semelhantes, independente da língua, ou seja, basicamente "livre de escala" e "mundo pequeno". Assim, os resultados dos principais índices de redes complexas utilizados nesse trabalho e os modelos de rede provocam reflexões importantes na contribuição para a difusão do conhecimento.

5.3 Sistemas Embarcados

Esta seção é dedicada à apresentação dos resultados obtidos pela ferramenta *Paper Crawler* através do termo de busca, “*Embedded Systems*”, na base de publicações *IEEE Xplore*, realizado no final do ano de 2014.

Na seção 5.3.1 explanamos os resultados obtidos referentes a análise da rede semântica utilizando o processo metodológico descrito no Capítulo 4, na seção 5.3.2 descrevemos os resultados obtidos durante a classificação de comunidades da rede, na seção 5.3.3 realizou-se uma análise temporal em cima das palavras-chave mais referenciadas nos títulos dos artigos científicos obtidos através da ferramenta *Paper Crawler*. E na seção 5.3.4 descrevemos algumas considerações alcançadas neste estudo.

5.3.1 Análise da Rede Semântica de Títulos

A partir dos resultados obtidos nos últimos trabalhos, resolvemos consolidar esta pesquisa aplicando nossa ferramenta em outro contexto como estudo de caso. Portanto, nosso passo inicial foi obter as publicações, para que, uma vez a rede gerada, pudéssemos realizar a análise da área de sistemas embarcados. A ferramenta *Paper Crawler*, através do termo *Embedded Systems*, realizou a busca no repositório acadêmico *IEEE Xplorer*, que nos retornou mais de dez mil títulos de artigos.

No entanto, para alcançarmos um melhor resultado em nossa análise, foram retirados dos títulos encontrados, todas as palavras-chave *embedded* e *system*, por fazerem parte do termo de busca. Destarte, através dos títulos e seus respectivos anos de publicação encontrados pela ferramenta, montamos a rede semântica variável ao longo do tempo.

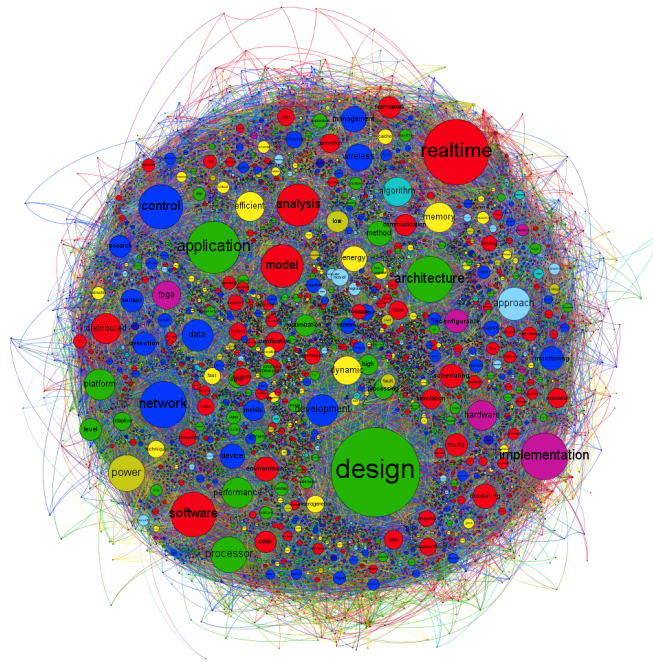


Figura 5.3: Rede Semântica

Para obter a visualização da rede, inicialmente, particionamos utilizando a métrica de modularidade, deixando assim, visível as comunidades encontradas, em seguida era necessário destacar os nós mais importantes classificando-os pelo tamanho, através da métrica centralidade de grau, e finalmente, distribuimos a rede para que se tornasse visualmente mais fácil de analisar, para isso utilizamos dois algoritmos de distribuição o *yifan hu* proporcional (HU, 2005) e *noverlap*. Portanto, conforme ilustrado na Figura 5.3, a cor representa as comunidades e o tamanho das circunferências o grau.

Tabela 5.6: Estatísticas referentes ao grau dos vértices da rede semântica dos títulos

Palavra-chave	Grau
<i>Design</i>	2661
<i>Realtime</i>	1950
<i>Application</i>	1548
<i>Implementation</i>	1391
<i>Architecture</i>	1380
<i>Network</i>	1374
<i>Software</i>	1315
<i>Control</i>	1312
<i>Analysis</i>	1254
<i>Model</i>	1253

Tabela 5.7: Estatísticas referentes à rede semântica de títulos

Métricas	1981-1985	1981-1990	1981-1995	1981-2000	1981-2005	1981-2010	1981-2014
Nº Títulos	11	88	325	1187	3694	9217	11298
Nº Palavras	88	699	2564	9993	32631	85660	106675
Nº Nós	43	286	738	1898	4309	8081	9429
Nº Arestas	118	1038	3481	13376	43379	109346	135213
Grau Médio	5,488	7,259	9,434	14,095	20,134	27,062	28,68
CAM	0,899	0,874	0,82	0,787	0,77	0,767	0,77
CMM	2,864	3,191	3,052	2,875	2,79	2,735	2,726
Densidade	0,025	0,105	0,013	0,07	0,005	0,003	0,003

A Tabela 5.6 lista os dez vértices de maior grau, representando palavras-chave em sua forma normalizada, presentes na rede de títulos dos artigos científicos publicados na área de sistemas embarcados com seus respectivos graus. É possível observar que há uma certa uniformidade entre as palavras mais utilizadas nos títulos dos artigos e, este conjunto tende a evidenciar uma natureza mais aplicada nos resultados de pesquisa reportados no *IEEE Xplorer*, para referida área, ao apresentar grande destaque, por exemplo, para termos como *design*, *realtime*, *application*, *implementation* e *architecture*.

A Tabela 5.7 apresenta os dados referentes à evolução da rede com progressão de cinco em cinco anos, bem como os valores referentes as métricas aplicadas a topologia da rede de títulos. A partir dos dados apresentados nesta tabela podemos concluir que, de acordo com Pereira et al. (2011), as palavras utilizadas na elaboração dos títulos dos artigos científicos relacionados com a área de sistemas embarcados, formam uma rede mundo-pequeno (*small-world*), pois ao longo do tempo o coeficiente de agrupamento médio (C) mostra um grande valor ($0,767 \leq C \leq 0,899$) e pequenos valores para o caminho mínimo médio (L) ($2,726 \leq L \leq 3,194$), quando comparado com o C de uma rede aleatória com a mesma quantidade de vértices e valores semelhantes para K e L.

A densidade reflete a quantidade de ligações entre as palavras encontradas nos títulos e indica a tendência de palavras que aparecerem juntas em um grande conjunto de palavras. Os dados mostrados na Tabela 5.7, com base em Pereira et al. (2011), demonstram que dos anos oitenta até metade da década de noventa, pegando como referência a métrica D, os meios de publicação eram voltados a um tema mais específico, enquanto que ao final do século vinte

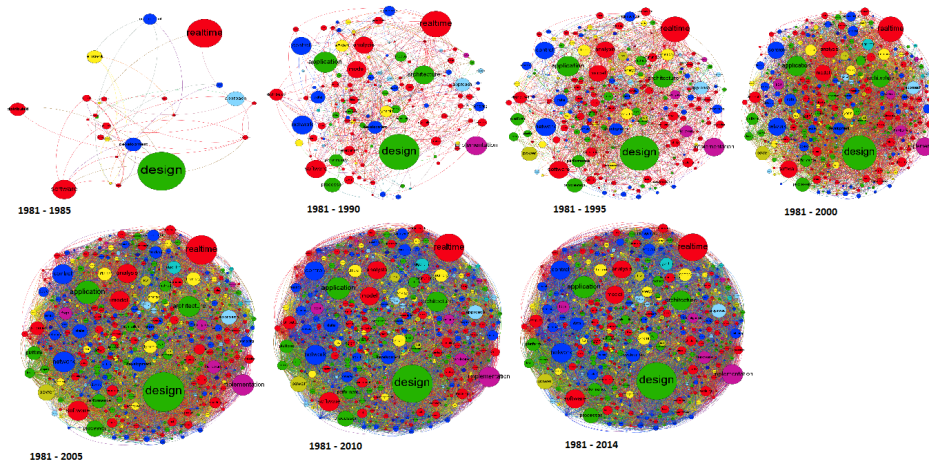


Figura 5.4: Evolução da rede semântica dos títulos na área de sistemas embarcados

as publicações foram se tornando de caráter extremamente multidisciplinar, o que converge com a interpretação do último índice L de 2,7.

5.3.2 Classificação das Comunidades

Na Figura 5.4, podemos observar a evolução das interações entre as palavras-chave dos títulos de cinco em cinco anos, em média, desde 1981 até 2014. Inicialmente, entre 1981 e 1985, a rede possuía 43 vértices (palavras-chave), 118 arestas e com apenas 5 comunidades. Considerando todo o período avaliado, de 1981 a 2014, a rede evoluiu para 9429 vértices, 135213 arestas e com 27 comunidades, um crescimento de, aproximadamente, 661,45% vértices por ano. Contudo, através da evolução da rede, pode-se observar uma sub-rede conectando uma grande parte dos nós da rede original, logo, há uma formação de **componente gigante**, ou seja, sub-redes conectando uma grande parte dos vértices da rede, o que indica que a comunidade de palavras-chave dos títulos na área de sistemas embarcados é bastante coesa. Observa-se também que, na comunidade “A” (cor verde), houve uma evolução de destaque, desde a década de noventa até 2014, relacionando essa comunidade diretamente com as palavras-chave *application* e *architecture*. Portanto, através deste mapeamento, podemos classificar este grupo com as palavras-chave de maior relevância.

Foram identificadas 27 comunidades, das quais consideramos como mais relevantes neste estudo as que possuíam o maior número de vértices e arestas, sendo as cinco maiores descritas na Tabela 5.8.

Tabela 5.8: Cinco maiores comunidades

Comunidade	Nº Vértices	Nº Arestas
A	2122	18454
B	2068	14014
C	1895	10859
D	965	4654
E	823	3062

Contudo, realizamos a classificação das comunidades conforme descrito nas Tabelas 5.9, 5.10, 5.11, 5.12 e 5.13.

Tabela 5.9: Palavras-chave mais relevantes para comunidade A

Palavra-chave	Número de Referências
Design	2661
Application	1548
Architecture	1380
Processor	1062
Platform	911

Tabela 5.10: Palavras-chave mais relevantes para comunidade B

Palavra-chave	Número de Referências
Realtime	1950
Software	1315
Analysis	1254
Model	1253
Distributed	878

Tabela 5.11: Palavras-chave mais relevantes para comunidade C

Palavra-chave	Número de Referências
Network	1374
Control	1312
Development	926
Data	919
Wireless	789

Tabela 5.12: Palavras-chave mais relevantes para comunidade D

Palavra-chave	Número de Referências
Implementation	1391
Hardware	872
Fpga	818
Reconfigurable	687
Interface	336

Tabela 5.13: Palavras-chave mais relevantes para comunidade E

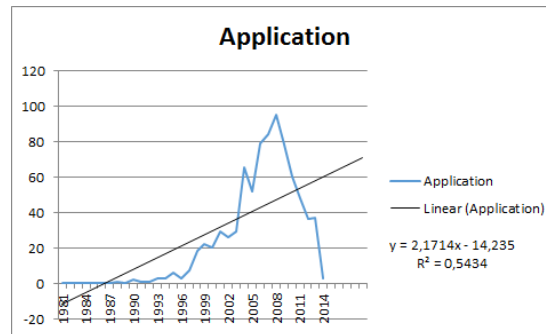
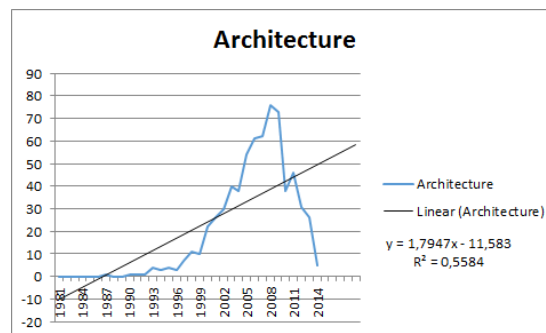
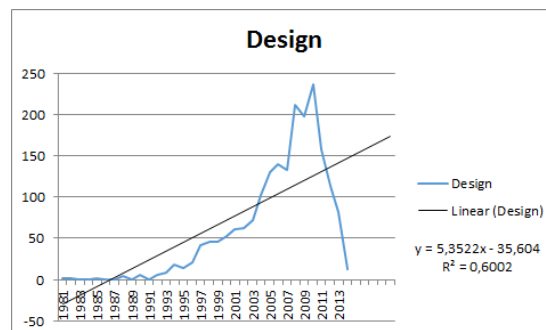
Palavra-chave	Número de Referências
Memory	886
Efficient	859
Dynamic	836
Energy	740
Heterogeneous	511

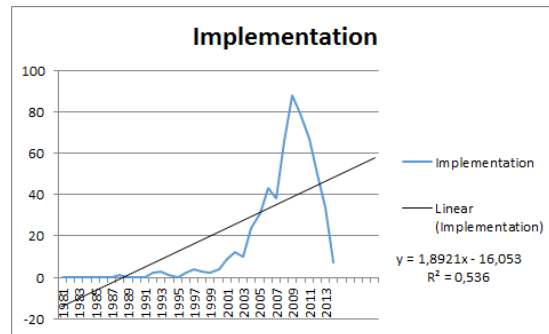
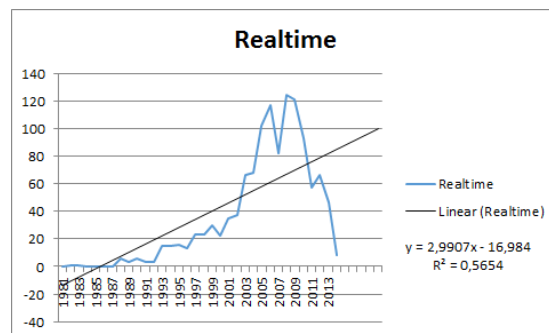
Cada comunidade foi classificada conforme as palavras-chave de maior relevância. Com isso, o pesquisador, poderá mapear em qual grupo determinado contexto se encaixa dentro da área pesquisada, que neste caso foi sistemas embarcados.

5.3.3 Análise temporal das palavras-chave

A partir dos dados retornados pela ferramenta realizamos a análise da evolução das palavras-chave dos títulos dos artigos publicados na área de sistemas embarcados ao longo dos anos.

Para demonstrar o processo evolutivo de algumas palavras ao longo do tempo, decidiu-se selecionar as cinco mais relevantes, ou seja, as cinco palavras-chave com maior grau. Contudo, para realizarmos uma análise temporal resolvemos aplicar um método para se estimar a condicional (valor esperado) de uma determinada variável y , conhecido como regressão linear, determinada por uma reta ($Y = aX + b$). Para nosso estudo de caso, o eixo Y é representado pela quantidade de menções de uma determinada palavra e o eixo X representa o ano.

Figura 5.5: Tendência do termo *Application*Figura 5.6: Tendência do termo *Architecture*Figura 5.7: Tendência do termo *Design*

Figura 5.8: Tendência do termo *Implementation*Figura 5.9: Tendência do termo *Realtime*

Para geração das retas de tendência de cada palavra mais relevante, segundo a métrica de centralidade de grau utilizada em ARS, gerou-se a equação com previsão linear para os próximos cinco anos. Isto é, através da linha de tendência, as cinco palavras-chave com maior grau de relevância estão, para a área de sistemas embarcados, a partir de 2010, em período de baixa menção. No entanto, de acordo com as Figuras 5.5, 5.6, 5.7, 5.8, 5.9, é possível observar que, quanto maior foi a relevância do termo, maior foi o coeficiente angular da reta, ou seja, mesmo com a queda de referência de todos os cinco termos, no período de 2010 a 2014, a previsão é que contextos relacionados a essas palavras-chave devam continuar sendo utilizados.

Tabela 5.14: Estatísticas referentes aos índices do coeficiente angular

Palavra-chave	Coeficiente Angular
Design	5,95
Realtime	2,99
Application	2,17
Implementation	1,8
Architecture	1,79

A Tabela 5.14 lista os coeficientes angulares, obtido através do método de regressão linear, que denota tendência de crescimento das palavras-chave de maior relevância, em sua forma normalizada, presentes na rede de títulos dos artigos científicos publicados na área de sistemas embarcados.

Diante do exposto, podemos concluir que, para os próximos cinco anos, a tendência é que, por exemplo, palavras-chave como *design* e *realtime* sejam utilizadas com maior frequência nas pesquisas realizadas na área de sistemas embarcados, ou seja, o contexto **design de sistemas em tempo real** possui uma tendência com crescimento significativo para pesquisa.

5.3.4 Considerações

Este estudo processou os metadados dos 11303 artigos publicados em diversos meios de publicação realizados entre os anos de 1981 e 2014 com o intuito de caracterizar os termos mais relevantes, durante este período e as comunidades que foram mais relevantes, a fim de caracterizar a área de pesquisa abordada, no caso, sistemas embarcados.

É importante ressaltar que em todas as fases deste estudo a metodologia como também a ferramenta foram de crucial importância para o seu desenvolvimento.

A análise das redes semânticas de títulos evidenciou que a área de sistemas embarcados na base de publicações do *IEEE Xplore* apresenta características de evento multidisciplinar com foco em pesquisa aplicada.

O método utilizado para classificar as comunidades através das palavras-chave pertencentes a cada publicação destas comunidades facilitou o processo. A classificação das comunidades relevantes da rede nos mostrou que, por exemplo, na comunidade "A", a palavra-chave de maior relevância está diretamente relacionada aos termos *application*, *architecture*, *processor* e *platform*, ou seja, com isso podemos mostrar ao pesquisador quais as subáreas de pesquisa mais alinhadas ao contexto *design* dentro da área de sistemas embarcados.

5.4 Publicações

Até o presente momento nosso trabalho resultou em três publicações. A primeira publicação apresentou genericamente a proposta inicial de trabalho, revisão bibliográfica sobre

redes semânticas, as metodologias existentes para análise e os resultados preliminares obtidos através dos dados referentes a um sistema de atendimentos na área de TI. O artigo foi intitulado “Análise de redes de palavras baseada em títulos extraídos de um sistema de atendimento”, apresentado durante o III *Brazilian Workshop on Social Network Analysis and Mining* (III BRASNAM) componente do Congresso da Sociedade Brasileira de Computação (CSBC) realizado no mês de julho do ano de 2014 na cidade de Brasília-DF.

Na segunda publicação, relacionada ao nosso projeto de pesquisa, conseguimos aprofundar os resultados obtidos no trabalho publicado descrito anteriormente, partindo para análise em títulos de artigos científicos, a fim de conseguir levantar novas hipóteses para o pesquisador a cerca de caracterização de determinados meios científicos. O artigo foi denominado “WebMedia XX Quem Somos e o Que Temos Feito nas Últimas Duas Décadas” e foi aceito e apresentado durante as atividades do Simpósio de Multimídia e Web (WebMedia 2014) realizado no mês de novembro do ano de 2014 João Pessoa-PB.

Na terceira publicação, resolvemos aplicar os conhecimentos adquiridos na última publicação, voltados a área de sistemas embarcados. Foi desenvolvida uma ferramenta para extração, caracterização e montagem de redes utilizando a base do *IEEE Xplore*. Ao final, mais de dez mil artigos foram extraídos e analisados, e a área de pesquisa foi devidamente caracterizada. O artigo foi intitulado “Análise de redes de termos em Sistemas Embarcados através de títulos extraídos de trabalhos científicos”, apresentado durante o IV *Brazilian Workshop on Social Network Analysis and Mining* (IV BRASNAM) componente do Congresso da Sociedade Brasileira de Computação (CSBC) realizado no mês de julho do ano de 2015 na cidade de João Pessoa-PB.

Finalmente, encerrando o ciclo de artefatos gerados com este tema de estudo, pretendemos compilar os resultados obtidos da área de Sistemas Embarcados em um artigo completo que deverá ser submetido para alguma revista ligada a área.

Capítulo 6

Considerações Finais

Neste capítulo serão discutidas as principais conclusões obtidas a partir dos estudos de caso apresentados no Capítulo 5. Desta forma o presente capítulo encontra-se dividido em três seções: Na seção "Ameaças à validade" serão apontados os principais fatores que dificultaram o desenvolvimento da presente pesquisa; em "Discussão" ampliaremos o debate acerca das conclusões obtidas e expostas; por fim, em "Trabalhos Futuros" estão descritas algumas sugestões de pesquisas.

6.1 Ameaças à validade

Embora a presente pesquisa tenha obtido relevantes resultados, estes podem ter sido afetados por diferentes tipos de fatores. Os fatores estão relacionados a (i) validade externa: Formação do texto, escolha do repositório de artigos (ii) validade interna: Escolha de métricas. Cada um dos fatores, bem como as ações para mitigar seus impactos sob os resultados da presente pesquisa, são descritos abaixo:

1. **Formatação do texto** - Em todos os estudos de caso abordados nas seções 5.1, 5.2 e 5.3 obtivemos uma quantidade considerável de títulos para serem analisadas, porém, em todos os casos, encontramos dificuldades no processo de reajuste manual das palavras-chave desses títulos, fazendo com que tivéssemos que, em alguns momentos, refazer nossa análise geral. Para atingir a padronização das palavras-chave, além da remoção das palavras sem sentido intrínseco, a maior dificuldade foi em agregar palavras consecutivas, num mesmo título, com intuito de obter um sentido único. Contudo, para

sanar este problema, aperfeiçoamos ao máximo nossa ferramenta tornando a análise íntegra e, conseqüentemente, bem mais interessante.

2. **Escolha dos repositórios de artigos** - Especificamente, para o estudo de caso abordado na seção 5.3, a escolha do repositório *IEEE Xplorer* se deu por que este era o mais simples (com relação a coleta dos dados) de realizar o processo de paginação e assim obter o resultado, mas apesar disso essa decisão fez com que limitássemos nossa base de dados. Para sanar essa limitação, deixamos a ferramenta modularizada, caso houvesse a necessidade de incluir novos repositórios. Outro problema é que os repositórios mudam com alguma frequência a estrutura de seus sites, invalidando os analisadores de *html* utilizados.
3. **Escolha de métricas** - Existem inúmeras métricas (i.e grau de entrada, centralidade de intermediação, centralidade de proximidade, modularidade, entre outras) quando se trata de ARS. Neste estudo procuramos utilizar métricas conhecidas e de simples entendimento que nos desse um resultado passível de validação, mesmo assim o foco do estudo não é a métrica em si, mas provar que, uma vez sendo devidamente aplicada a rede semântica, é possível caracterizar uma determinada área de pesquisa através de uma metodologia bem definida.

6.2 Discussão

Com base nos dados apurados, pode-se afirmar que a partir da construção e análise de uma rede composta de palavras-chave, extraídas de títulos, existe a possibilidade de fazer uso de métricas de ARS para obter informações importantes a respeito da produção científica de determinada área de pesquisa. Este estudo se mostrou satisfatório, uma vez que conseguiu identificar contextos relevantes e classificar comunidades de uma área de conhecimento, tudo isso utilizando uma metodologia específica para análise de redes semânticas, possibilitando assim dar suporte a pesquisadores que estejam iniciando alguma pesquisa e precisem de algum aporte.

A partir da utilização da ferramenta *Paper Crawler*, alinhada ao uso do *Gephi*, conseguimos todo suporte necessário para análise dos dados obtidos, suas funcionalidades em

conjunto com a metodologia definida permitiram a capacidade de automatizar quase todo o processo analisado nos estudos de caso descritos no Capítulo 5. Apesar da existência de outras métricas para clusterização de rede, a utilização da métrica modularidade nos foi satisfatória, sua capacidade de identificar grupos na rede fez com que pudéssemos visualizar a interação entre as comunidades encontradas.

Nossa questão de pesquisa foi “De que forma seria possível caracterizar meios científicos através da análise de palavras-chave em artigos científicos?”, através do decorrer deste estudo tentamos respondê-la de acordo com a metodologia aplicada no Capítulo 4, e através dos estudo de caso abordados no Capítulo 5.

Inicialmente, na identificação dos problemas recorrentes relacionados a área de TI (seção 5.1), utilizamos as seguintes métricas de centralidade:

- **Grau** - Permitiu identificar a frequência que uma determinada palavra-chave foi mencionada, a fim de apontarmos os problemas de maior incidência;
- **Proximidade** - Permitiu evidenciar o quanto uma palavra-chave é mais próxima de todas as outras dentro da rede semântica, ou seja, apontando os problemas mais centrais;
- **Intermediação** - Quantificou o número de vezes que uma palavra-chave agiu como "ponte" ao longo do caminho mais curto entre dois outros problemas.

Contudo, podemos afirmar que, a partir dos resultados explanados nos estudos de caso, através das propriedades encontradas nas redes semânticas, foi possível analisar a evolução das áreas de pesquisa ao longo do tempo, como observado na seção 5.2, onde identificamos uma mudança de caráter voltado a um tema específico para multidisciplinar, referente a comunidade científica *WebMedia*.

Por fim, no estudo de caso da seção 5.3, foi possível observar que a utilização das palavras-chave, com maior grau de relevância, contidas nos títulos dos artigos científicos relacionados a área de sistemas embarcados retirados da base do *IEEE Xplorer*, podem ser utilizadas para nomear as comunidades encontradas na rede semântica, e com isso, o pesquisador poderá caracterizar subáreas de pesquisa dentro da área de conhecimento analisada, o que responde proficuamente a questão desta pesquisa.

6.3 Trabalhos Futuros

Como trabalhos futuros, acreditamos que resolver os problemas referente a formatação adequada do texto, por exemplo, adicionando a ferramenta utilizada nesta pesquisa, conceitos sobre inteligência artificial, mais especificamente, um novo módulo contendo agentes inteligentes, seria de fundamental importância para alcançarmos uma melhor análise sobre as redes semânticas, com um maior volume de dados possíveis.

É importante destacar também que, buscar novas métricas que possibilitem a caracterização de áreas de pesquisa em redes semânticas, ou até mesmo comparar o desempenho destas com as métricas utilizadas nesta pesquisa, seria outra abordagem bastante interessante. E por fim, modificar a ferramenta *Paper Crawler* para que fique apropriada para um ambiente *WEB*.

Bibliografia

BARABASI, A. L. *How everything is connected to everything else and what it means for Business, Science, and everyday life*. [S.l.]: Cambridge: Plume, 2003.

BARABASI, A. L.; ALBERT, R. Emergence of scaling in random networks. *Science*, v. 286, p. 509–512, 1999.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. Disponível em: <<http://www.aiai.org/ocs/index.php/ICWSM/09/paper/view/154>>.

BERKOWITZ, S. D. *An introduction to structural analysis: The network approach to social research*. [S.l.]: Butterworth-Heinemann, 1982.

BOLLOBAS, B. The evolution of random graphs—the giant component. In: *Random Graphs*. Second. Cambridge University Press, 2001. p. 130–159. ISBN 9780511814068. Cambridge Books Online. Disponível em: <<http://dx.doi.org/10.1017/CBO9780511814068.008>>.

BRANDES, U. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, v. 25, p. 163–177, 2001.

CALDEIRA, S. M. G. *Caracterização da rede de signos linguísticos: Um modelo baseado no aparelho psíquico de Freud*. Dissertação (Mestrado) — Centro de Pós-Graduação e Pesquisa Visconde de Cairu, Fundação Visconde de Cairu., 2005.

CALDEIRA, S. M. G. et al. The network of concepts in written texts. *The European Physical Journal B*, v. 4, n. 49, p. 523–529, April 2006.

CASTEIGTS, A. et al. Time-varying graphs and dynamic networks. In *Frey, H., Li, X., and Ruhup, S., editors, ADHOC-NOW*, v. 6811, p. 346–359, Aug 2011.

CRILLY, T. *50 Mathematical Ideas You Really Need to Know*. Book Sales, 2007. ISBN 9781847241474. Disponível em: <<http://books.google.com.br/books?id=XfYSKAAACAAJ>>.

CUNHA, M. V. *Redes semânticas baseadas em títulos de artigos científicos*. Dissertação (Mestrado) — PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL E TECNOLOGIA INDUSTRIAL, 2013.

CUNHA, M. V. et al. Redes de títulos de artigos científicos variáveis no tempo. In: *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. [S.l.: s.n.], 2013.

DEITEL, H. M. *Java: como programar*. Pearson Prentice Hall, 2005. ISBN 9788576050193. Disponível em: <<http://books.google.com.br/books?id=U5AyAgAACAAJ>>.

ERDOS, P. On cliques in graphs. *Science*, v. 4, p. 233–234, 1966.

ERDOS, P.; RENYI, A. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci.*, v. 5, p. 17–61, 1960.

FADIGAS, I. et al. Análise de redes semânticas baseada em títulos de artigos de periódicos científicos: o caso dos periódicos de divulgação em educação matemática. *Educação Matemática Pesquisa, EMP*, v. 11, n. 1, p. 167–193, Sept 2009.

FADIGAS, I.; PEREIRA, H. A network approach based on cliques. *Physica A: Statistical Mechanics and its Applications*, v. 392, n. 10, p. 2576–2587, 2013.

FREEMAN, L. C. A set of measures of centrality based on betweenness. *Sociometry*, v. 40, p. 35–41, 1977.

FREEMAN, L. C. Centrality in social networks: Conceptual clarification. *Social Networks*, v. 1, p. 215–239, 1979.

FREITAS, L. Q. *Medidas de Centralidade em Grafos*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Produção., 2010.

HU, Y. F. Efficient and high quality force-directed graph drawing. *The Mathematica Journal*, v. 10, p. 37–71, 2005.

JÚNIOR, M. L. M. *Metodologia para Análise de Relevância de Publicações através de Rede de Citações*. Dissertação (Mestrado) — Universidade Federal da Paraíba - Programa de Pós-graduação em Informática, 2015.

LUCE, R. D.; PERRY, A. D. A method of matrix analysis of group structure. *Pychometrika*, out. 1949.

LUNA, S. V. *Planejamento de pesquisa – Uma introdução*. [S.l.]: EDUC, São Paulo, 1999.

MARTELETO, R. M. Cultura da modernidade: discursos e práticas informacionais. *Revista da Escola de Biblioteconomia da UFMG*, v. 23, n. 2, p. 115–137, 1994.

MARTINS, N. S. *Introdução à estilística: a expressividade na língua portuguesa*. [S.l.]: Biblioteca universitária de língua e lingüística; v. 8, 2003.

MIKA, P. *Social Networks and the Semantic Web*. [S.l.]: New York: Springer, 2007. ISBN 9780387710006.

MILGRAM, S. The small world problem. *Psychology today, New York*, v. 2, n. 1, p. 60–67, 1967.

MILORE, G.; ANGELINI, F. *Estatística Aplicada*. [S.l.]: São Paulo: Atlas, 1995.

- MIRANDA, D.; PEREIRA, M. O periódico científico como veículo de comunicação: uma revisão de literatura. *The New Palgrave Encyclopedia of Economics*, v. 25, p. 375–382, 1996.
- NETO, P. O. B. *Grafos: teoria, modelos, algoritmos*. [S.l.]: Edgard Blücher, 2003.
- NEWMAN, M. E. J. The structure and function of complex networks. *SIAM review*, JSTOR, v. 45, n. 2, p. 167–256, 2003. ISSN 0036-1445.
- NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. In: *Proceedings of the National Academy of Sciences*. [S.l.: s.n.], 2004. p. 5200–5205.
- NEWMAN, M. E. J. Mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2007.
- NEWMAN, M. E. J. *Networks: An Introduction*. New York: Oxford University Press, 2010.
- NEWMAN, M. J. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, v. 103, n. 23, p. 8577–8582, jun. 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=retrieve&db=pubmed&list_uids=16723398&dopt=AbstractPlus>.
- PEREIRA, H. et al. Semantic networks based on titles of scientific papers. *Physica A: Statistical Mechanics and its Applications*, v. 6, n. 390, p. 1192–1197, Mar 2011.
- PROCÓPIO, P.; LAENDER, A. H.; MORO, M. M. Análise da rede de coautoria do simpósio brasileiro de bancos de dados. *Anais do XXVI Simpósio Brasileiro de Bancos de Dados*, p. 131–138, 2011.
- RAMOS, M. V. M.; NETO, J. J.; VEGA, I. S. *Linguagens Formais: Teoria, Modelagem e Implementação*. 1. ed. Porto Alegre: Bookman, 2009. 656p p. ISBN 978-85-7780-453-5.
- SA, C. C.; ROCHA, J. *Treze Viagens Pelo Mundo da Matemática*. U. PORTO, 2010. ISBN 9789898265340. Disponível em: <<https://books.google.com.br/books?id=25kri1SrOAUc>>.
- SADIDUSSI, G. The centrality index of a graph. *Psychometrika*, v. 31, p. 581–603, 1966.
- SCOTT, J.; CARRINGTON, P. J. *The SAGE Handbook of Social Network Analysis*. SAGE Publications, 2011. ISBN 9781446250112. Disponível em: <<https://books.google.com.br/books?id=mWlsKkIuFNgc>>.
- STERNBERG, R. *Psicologia cognitiva*. [S.l.]: Porto Alegre, RJ: Artes Medicas Sul, 2011.
- WASSERMAN, S.; FAUST, K. *Social network analysis: Methods and applications*. Cambridge university press, 1994. v. 8. Disponível em: <http://scholar.google.com/scholar.bib?q=info:gET6m8icitMJ:scholar.google.com/&output=citation&hl=en&as_sdt=0,5&as_vis=1&ct=citation&cd=0>.
- WATTS, D. J. *Small Worlds: The Dynamics of Networks between Order and Randomness*. [S.l.]: USA : Princenton University Press, 1999.
- WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. *Nature*, v. 393, n. 6684, p. 409–10, 1998.

WERKEMA, M. C. C.; AGUIAR, S. *Análise de regressão: como entender o relacionamento entre as variáveis de um processo*. [S.l.]: BELO HORIZONTE: Fundação Christiano Ottoni, 1996.

ZIPF, G. K. *The Principle of Least Effort: an introduction to human ecology*. [S.l.]: New York: Hafner Publishing Company, 1972.