CARACTERIZAÇÃO TOPOLÓGICA DE REDES COMPLEXAS GERADAS NO PROCESSO DE MINERAÇÃO DE DOCUMENTOS

Claudia Abreu Paes

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

_____
Prof. Alexandre Gonçalves Evsukoff, D. Sc.

_____
Prof. Nelson Francisco Favilla Ebecken, D. Sc.

_____
Profª. Beatriz de Souza Leite Pires de Lima, D.Sc.

_____
Prof. Augusto Cesar Noronha Rodrigues Galeão, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
JUNHO DE 2008

PAES, CLAUDIA ABREU

Caracterização Topológica de Redes Complexas geradas no Processo de Mineração de Documentos [Rio de Janeiro] 2008

XI, 50 p. 29,7 cm (COPPE/UFRJ, M.Sc., Engenharia Civil, 2008)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Redes Complexas

2. Mineração de Textos

3. Topologia de Redes Complexas

I. COPPE/UFRJ II. Título (série)

Dedicatória

À Deus por me conceder persistência e amor a vida,
ao meu pai e minha mãe pelo exemplo de vida, cada qual de sua maneira.

Agradecimentos

Agradecer é sempre um movimento de alegria e importância, pois através deste ato estamos nos conscientizando que para toda conquista temos sempre o apoio, a confiança e carinho de todos que nos cercam e nos querem bem.

Nada seria possível sem:

- dedicação e confiança de meu orientador;
- apoio de minha mãe nas responsabilidades com minha filha;
- carinho de minha mãe nas minhas horas de desespero e irritação;
- paciência, compreensão e incentivo de meu marido e minha filha devido às minhas ausências e impaciências;
- colaboração técnica de meus colegas Marina e Marcos, cada qual com seu valor e seu espaço em meu coração;
- incentivo e motivação dadas por meus amigos Mario, Millan e Inês;
- consultas técnicas aos Professores amigos Abel, Júlia e Jorge.
- amor e carinho de todos que me cercam e, que de alguma forma torcem pelo meu sucesso.
- O CNPq que viabilizou financeiramente esta pesquisa.

Portanto, agradeço a todos: MUITO OBRIGADA !!!

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CARACTERIZAÇÃO TOPOLÓGICA DE REDES COMPLEXAS GERADAS NO PROCESSO DE MINERAÇÃO DE DOCUMENTOS

Claudia Abreu Paes

Junho / 2008

Orientador: Alexandre Gonçalves Evsukoff

Programa: Engenharia Civil

Esta dissertação apresenta uma proposta metodológica para caracterização topológica de redes complexas formadas por informações textuais não estruturadas. A partir do processo de pré-processamento de uma coleção de documentos, a rede de documentos é gerada como um grafo no qual cada vértice representa um documento e as arestas são ponderadas pela similaridade entre documentos. A topologia da rede de documentos é analisada por um conjunto de medidas propostas na literatura para caracterização de redes complexas. Foi realizado um estudo de caso com quatro conjuntos de documentos de naturezas variadas quanto ao idioma e volume, para avaliação da metodologia proposta. Os resultados mostram que os conjuntos de documentos apresentam uma topologia característica de redes livres de escala, de forma que uma série de resultados apurados pode ser empregada no processo de extração de conhecimento a partir de coleção de documentos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

TOPOLOGICAL CHARACTERIZATION OF COMPLEX NETWORKS GENERATED IN THE PROCESS OF DOCUMENT MINING

Claudia Abreu Paes

June / 2008

Advisor: Alexandre Gonçalves Evsukoff

Department: Civil Engineering

This dissertation presents a methodological proposal for topological characterization of complex networks formed by non structuralized literal information. From the process of daily pre-processing of a document collection, the document networks is generated as a graph in which each vertex represents a document and the edges are weighed for the similarity between documents. The topology of the document network is analyzed by a set of measure proposals in literature for characterization of complex networks. For evaluation of the methodological proposal, it was carried through a study of case with four sets of documents of varied nature as to the language and volume. The results show that the sets of documents present a topology of scale free networks, so that a series of refined results can be used in the process of extraction of knowledge from document collection.

Índice

1. INTRODUÇÃO 1
2. MINERAÇÃO DE TEXTOS 3
   2.1. Pré-Processamento 4
   2.2. Processamento 8
3. REDES COMPLEXAS 9
   3.1. Elementos da Teoria dos Grafos 9
   3.2. Redes Complexas 10
      3.2.1. Redes Randômicas 11
      3.2.2. Mundo Pequeno 12
      3.2.3. Livre de Escala 13
   3.3. Caracterização Topológica 14
      3.3.1. Grau de Distribuição 14
      3.3.2. Coeficiente de Aglomeração 14
      3.3.3. Centralidade 16
   3.4. Estrutura de Comunidade 16
4. METODOLOGIA PROPOSTA 18
   4.1. Preparação do Ambiente 19
   4.2. Pré-Processamento 20
   4.3. Cálculo da Similaridade 21
   4.4. Apuração das Medidas de Análise 22
   4.5. Visualização 24
   4.6. Análise 25
5. ESTUDOS DE CASOS 27
   5.1. Patentes 27
   5.2. Deputados / RJ 32
   5.3. Resumos de Dissertação de Mestrado-PEC/RJ 37
   5.4. NewsGroup 42
   5.5. Discussão dos resultados 46
6. CONCLUSÃO 48
7. BIBLIOGRAFIA 49

Lista de Figuras

Figura 2.1 – Seqüência de aplicação das técnicas do Pré-processamento 7
Figura 3.1 – Grafo (VxA) 9
Figura 3.2 – Formação de redes Random Graphs [14] 11
Figura 3.3 – Formação redes Mundo Pequeno [14] 12
Figura 3.4 – Freqüência Redes Randômica 14
Figura 3.5 – Freqüência Redes Mundo Pequeno 14
Figura 3.6 – Freqüência Redes Livre de Escala 14
Figura 3.7 – Representação de triângulos 15
Figura 3.8 – Representação de triplas 15
Figura 3.9 – Estrutura de Comunidade [15] 16
Figura 3.10 – Dendrograma – Rede Hierárquica [17] 17
Figura 3.11 – Informações WWW [13] 17
Figura 4.1 – Rede de documentos 18
Figura 4.2 – Etapas da Metodologia proposta 19
Figura 4.3 – Interface Gráfica *Biguá* 21
Figura 4.4 – Resultado medidas de similaridade 22
Figura 4.5 – Matriz de Similaridade 22
Figura 4.6 – Freqüência do grau 24
Figura 4.8 – Representação Matlab 25
Figura 4.9 – Representação Pajek 25
Figura 5.1 – Patentes - Freqüência – Coseno 0.1 29
Figura 5.2 – Patentes - Freqüência – Manhatan 0.1 29
Figura 5.3 - Patentes - Coeficiente aglomeração - Coseno 0.1 29
Figura 5.4 – Patentes - Coeficiente aglomeração - Manhatan 0.1 29
Figura 5.5 – Patentes – Estruturas de Comunidade - Coseno 0.9 30
Figura 5.6 – Patentes - Estruturas de Comunidade - Manhatan 0.9 30
Figura 5.7 – Patentes - Centros - Coseno 0.9 31
Figura 5.8 – Patentes - Centros - Manhatan 0.9 31
Figura 5.9 – Caracterização deputados – Currículo deputados/RJ 32
Figura 5.10 – Deputados - Freqüência - Coseno 0.1 34

Figura 5.11 – Deputados - Freqüência - Manhatan 0.1 34
Figura 5.12 – Deputados Coeficiente aglomeração - Coseno 0.1 34
Figura 5.13 – Deputados Coeficiente aglomeração - Manhatan 0.1 34
Figura 5.14 – Deputados – Estruturas de Comunidade - Coseno 0.1 35
Figura 5.15 – Deputados – Estruturas de Comunidade - Manhatan 0.1 35
Figura 5.16 – Deputados – Identificação dos Centros - Coseno 0.1 36
Figura 5.17 – Deputados – Identificação dos Centros - Manhatan 0.1 36
Figura 5.18 – Dissertação - Freqüência - Coseno 0.1 38
Figura 5.19 – Dissertação - Freqüência - Manhatan 0.1 38
Figura 5.20 – Dissertação - Coeficiente de aglomeração-Coseno 0.1 38
Figura 5.21 – Dissertação - Coeficiente de aglomeração-Manhatan 0.1 38
Figura 5.22 – Dissertação – Estruturas de Comunidade - Coseno 0.3 39
Figura 5.23 – Dissertação – Estruturas de Comunidade - Manhatan 0.3 40
Figura 5.24 – Dissertação – Identificação dos Centros - Coseno 0.3 40
Figura 5.25 – Dissertação – Identificação dos Centros - Manhatan 0.3 41
Figura 5.26 – NewsGroup - Freqüência - Coseno 0.1 43
Figura 5.27 – NewsGroup - Freqüência – Manhatan 0.1 43
Figura 5.28– NewsGroup - Coeficiente aglomeração - Coseno 0.1 44
Figura 5.29 – NewsGroup - Coeficiente aglomeração - Manhatan 0.1 44
Figura 5.30 – NewsGroup – Estruturas de Comunidade - Coseno 0.3 44
Figura 5.31 – NewsGroup - Estruturas de Comunidade - Manhatan 0.3 45
Figura 5.32 – NewsGroup - Centros - Coseno 0.3 45
Figura 5.33 – NewsGroup - Centros - Manhatan 0.3 46

Lista de Tabelas

Tabela 2.1 – BoW 5
Tabela 2.2 – Resultante de aplicação de medida de similaridade 8
Tabela 3.1 – Grau de grafo 10
Tabela 4.1 – Medidas Documento 23
Tabela 4.2 – Medidas Rede 23
Tabela 4.3 – Grau x Coeficiente 24
Tabela 5.0 – Conjuntos de documentos 27
Tabela 5.1 – Distribuição de documentos - Patentes 28
Tabela 5.2 – Transitividade da rede de documentos-Patentes 28
Tabela 5.3 – Distribuição de documentos - Deputado 33
Tabela 5.4 – Transitividade da rede de documentos - Deputado 33
Tabela 5.5 – Distribuição de documentos - Dissertação 37
Tabela 5.6 – Transitividade da rede de documentos – Dissertação 38
Tabela 5.7 – Distribuição de documentos - NewsGroup 42
Tabela 5.8 – Transitividade da rede de documentos- NewsGroup 43
Tabela 6.1 – Apuração conjuntos de documentos – limite 0.1 46
Tabela 6.2 – Transitividade conjuntos de documentos – limite 0.1 47
Tabela 6.3 – Resultados conjuntos de documentos 47

Lista de Símbolos

D = documentos

FT = Freqüência do termo

k = grau de distribuição

m = caminho mínimo entre os vértices

M = número total de termos

N = número total de documentos

f = quantidade de vezes que o termo aparece no documento

p = probabilidade

SC = similaridade Coseno

Sim = similaridade

SM = similaridade Manhatan

T = termos

u = a quantidade de vezes que o termo aparece no documento no índice tf-idf

1. INTRODUÇÃO

Atualmente, a tecnologia e a internet estimulam o crescente volume de produção de documentos, o que faz com que sejam desenvolvidos e aplicados métodos eficientes para organização, armazenamento e recuperação da informação.

A técnica de mineração de textos ("*Text Mining"*), processo definido para extrair conhecimento de informações não estruturadas, é utilizada em vários ambientes de negócio, seja para identificar patentes [21], validar diagnósticos na área médica [20] ou mesmo na localização de textos correlatos à determinada especialização. A utilização da técnica provê agilidade e precisão na identificação, recuperação e utilização dos documentos.

O processo de mineração de textos compreende:

1. Métodos para tratamento da informação, transformando as não estruturadas em estruturadas (numéricas) [5];

2. Métodos de categorização, identificando os conceitos e classificando os documentos em função destes. Desta forma, são definidas classes [1] que separam o conjunto de documentos em categorias de assuntos;

3. Métodos para agrupamento, que estabelecem a união dos documentos similares, facilitando a identificação e recuperação automática da informação.

4. Métodos de recuperação de informação (RI ou "*Information Retrieval*"). Segundo definição [23], a recuperação de informação é uma área que desenvolve modelos para tratamento de grandes coleções de documentos. Tópicos específicos previamente definidos são utilizados para identificar os documentos.

Para valorizar os termos que representam os documentos das coleções e potencializar o valor da ligação dos documentos são utilizadas Métricas[1] de freqüência normalizada e métricas de similaridade, respectivamente.

Após a execução do processo de mineração de textos e aplicação das métricas, o conjunto de documentos forma uma rede complexa, considerando a similaridade como característica da ligação. Uma rede complexa é definida como a representação de um

[1] Métricas são medidas definidas em função das ligações entre os vértices.

conjunto de itens conectados entre si através de ligações qualificadas [4], ou seja, definidas por características.

Estudos [7][8][14][17] demonstram a necessidade de análise das redes complexas a partir de suas características para identificar sua topologia, conhecer o nível de coesão existente entre os documentos, identificar documentos determinantes para manutenção das ligações e extrair conhecimento para tomada de decisão. Foram apresentadas nos trabalhos [1][7][8][10][10][14] as redes complexas formadas a partir das informações disponíveis na internet, que apresentam um elevado número de ocorrências [19] e os caminhos são definidos pela navegação criada nas páginas.

As características das redes complexas são identificadas a partir das métricas: grau de distribuição, ou seja, número de ligações existentes em torno de um vértice; coeficiente de aglomeração ("*clusterizing*"), ou seja, a medida de transitividade, que determina o nível de coesão entre os documentos representados na rede.

Esta dissertação propõe uma metodologia de análise e extração de conhecimentos para redes complexas formadas a partir de uma coleção de documentos. A metodologia é dividida em etapas definidas para realizar todo o processo de análise: tratamento da informação, aplicação das métricas de similaridade, aplicação das métricas de análise para extração das propriedades da rede complexa de documentos, suporte para análise dos resultados e visualização. A metodologia é aplicada em quatro conjuntos de documentos: conjunto Patente formado por descrições de patentes industriais; conjunto Deputados formado por currículo de deputados do Estado do Rio de Janeiro; conjunto Dissertações formado por resumos elaborados para as dissertações de mestrado do PEC/RJ; conjunto NewsGroup formado por artigos de naturezas diversas como: eletrônica, política, religião. Os conjuntos de documentos possuem naturezas variadas quanto ao idioma e volume, para avaliação da metodologia proposta.

O texto está organizado da seguinte maneira. O Capítulo 2 descreve o processo de mineração de textos e as etapas e técnicas aplicadas para tratamento da informação não estruturada. O Capítulo 3 apresenta a revisão bibliográfica de redes complexas, considerando as topologias e medidas para apuração dos resultados. O Capítulo 4 descreve passo-a-passo a metodologia juntamente com a explanação detalhada das etapas a serem realizadas. O Capítulo 5 apresenta a aplicação e resultados apurados da metodologia em quatro estudos de caso. Finalmente, no Capítulo 6, é apresentada a conclusão do trabalho e sugestões para trabalhos futuros.

2. MINERAÇÃO DE TEXTOS

Mineração de textos é uma técnica de processamento de dados e extração de conhecimento para um conjunto de informações não-estruturadas. É necessário utilizar métodos para transformar as informações não estruturadas em informações estruturadas [5], viabilizando a aplicação das técnicas usuais de mineração de textos e recuperação da informação.

O processo de mineração de texto é realizado através de três fases: pré-processamento, processamento e pós-processamento. O pré-processamento tem como objetivo disponibilizar os documentos em uma representação numérica, a partir da análise dos termos constantes dos documentos. A análise dos termos que compõem os documentos segue a seguinte seqüência de procedimentos:

1. Seleção dos termos significativos que definem o conteúdo abordado pelo texto;
2. Apuração da quantidade de vezes em que o termo aparece no texto, transformando as informações textuais em valores quantitativos;
3. Normalização dos valores quantitativos para potencializar os termos significativos;
4. Criação de uma matriz numérica formada pela relação documentos e termos, denominada *BoW* (bolsa de palavras). Por definição, *BoW* é a representação numérica dos termos utilizados nos documentos [5].

O processamento é a etapa em que são utilizados métodos para organizar e disponibilizar os conteúdos de forma ordenada para os processos de recuperação da informação [1]. Os métodos podem ser de classificação de aprendizagem supervisionada ou classificação não supervisionada, como: redes neurais, kNN (k vizinhos mais próximos - "*k Nearest Neighbors"*), Bayesiana e SVM ("*Support Vector Machine"*), árvore de decisão para predizer a categoria dos documentos [2] e agrupamento de documentos com os métodos de aglomeração: "*k-means"*, hierárquico, incremental, entre outros. O pós-processamento objetiva no processo de visualização e análise das informações obtidas.

A seguir, são descritas as fases detalhadamente.

2.1. Pré-Processamento

A etapa Pré-Processamento percorre cada documento do conjunto selecionando os termos significativos, pois estes representam o conteúdo do documento. É preciso eliminar os termos que não representam o conteúdo e são empregados no documento somente para compor a construção gramatical.

São utilizadas técnicas para separar os termos; retirar preposições, artigos, pontuações, verbos e demais palavras de nosso vocabulário que compõem a construção gramatical; reduzir os termos aos seus radicais; contabilizar o número de ocorrências em que o termo é referenciado no documento; armazenar as informações para utilização nas próximas etapas da mineração de texto representadas na Figura 2.1 e descritas a seguir.

- **Folding:** Processo de transformar as palavras do documento em um mesmo padrão de representação: letras maiúsculas ou letras minúsculas.
- **Tokenização:** Processo utilizado para fragmentar o documento em "*tokens*". O documento é representado por uma única linha sem marcadores, como parágrafos ("*stream"*). O "*token"* se define por um conjunto de caracteres alfanuméricos localizado entre um caracter definido como separador (normalmente é utilizado o caracter espaço " ") e representa a menor unidade do texto processada [9]. O processo de tokenização é fundamental para os procedimentos futuros, pois sem a identificação dos "*tokens*" é difícil retirar informações precisas dos documentos [5].
- **Dicionário:** Processo de pesquisa no conjunto de palavras selecionadas para identificar o contexto abordado pelo documento. O uso do dicionário garante a eficiência do processo de mineração de textos, na medida em que determina o conjunto de termos a serem incluídos na análise [5]. Os dicionários quando possuem seus termos agrupados por conceitos são chamados "*thesaurus"*. A seleção dos termos é feita por um especialista da área abordada.
- **Stopwords:** Processo de elaboração de uma lista de termos cuja construção gramatical não expressa significado relevante ao assunto abordado nos documentos, como: preposição, artigo, pronome, conjunção, advérbio e verbo [16]. Os "*tokens*" constantes da lista de "*stopword*" são retirados do texto.

- **Stemming:** Processo de análise morfológica das palavras reduzindo a representação do termo a seu radical sem influenciar no significado. Os algoritmos de "*stemming*" representam um nível de importância alto no processo de mineração de texto, mas seu uso requer muito cuidado, pois em alguns casos, erros acontecem e distorções são causadas.

- **Contagem:** Processo que contabiliza o número de vezes em que um termo é citado em um documento, refletindo a importância do termo em representação da semântica do documento [24][24]. Neste processo é gerada a informação da quantidade associada a cada termo.

- **Bow (bag of word):** É uma matriz numérica $N \times M$ onde o elemento (i, j), representa um índice relacionado ao termo $T_j \in T$ no documento $D_i \in D$. O conjunto de documentos é representado pelo conjunto $D = \{D_i, i = 1...N\}$ e o conjunto de termos como $T = \{T_j, j = 1...M\}$. A Tabela 2.1 apresenta um exemplo de BoW, em que as linhas são representadas pelos documentos (D) e as colunas pelos termos (T).

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

Tabela 2.1 – BoW

Cada linha é um vetor que representa cada documento na coleção. Ao final do processamento de todos os documentos que formam a coleção, a *BoW* é gerada com a informação resultante da etapa de contagem.

A BoW é uma matriz simétrica e esparsa por não existir a presença de todos os termos em todos os documentos do conjunto.

Weiss, Indurkhya, Zhang e Damerau indicam representar a *BoW* considerando a representação do vetor por uma lista de pares indicando somente as colunas em que $BoW(D_i T_j)$ seja diferente de 0 [5].

Freqüência normalizada: Processo de normalização que tem como objetivo valorizar o termo que expressa o conteúdo do documento. Existem várias métricas de freqüência normalizada, sendo a mais utilizada a "*term frequency–inverse document frequency*" -

TF-IDF, que determina a freqüência relativa dos termos em cada documento comparado pela proporção inversa em relação ao conjunto total de documentos, definindo a importância do termo no conjunto de documentos.

A métrica *TF-IDF* baseia-se em dois fatores: a freqüência dos termos (TF) e o logarítmo do inverso da freqüência do termo (IDF).

A freqüência do termo T_j no documento D_i é:

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

Onde:

f_{ij} = número de ocorrências de T_j em D_i;

f_{ih} = número de ocorrências dos termos em todos os documentos D;

O inverso da freqüência do termo T_j é calculado por:

$$IDF(T_j) = \log\left(\frac{N}{N_j}\right) = \log N - \log N_j \quad (2)$$

$$N_j = \sum_{i=1}^{N} a_{ij} \qquad a_{ij} = \begin{cases} 1, se\ a_{ij} \neq 0 \\ 0, se\ a_{ij} = 0 \end{cases} \quad (3)$$

Onde:

N = número total de documentos na coleção;

N_j = número total de documentos que contém o termo T_j;

A métrica *TF-IDF* (4), armazenada na *BoW,* representa o resultado final da etapa de pré-processamento.

$$BoW(D_i, T_j) = TF(D_i, T_j) * IDF(T_j) \quad (4)$$

- **Filtros:** Processo de filtragem que minimiza os erros e provê uma qualidade maior na seleção dos termos para análise. Os filtros definem regras de redução de termos menos significativos, o que aumenta a qualidade na representação do documento e aperfeiçoa o cálculo de similaridade, pois os termos menos significativos são excluídos da seleção.

No contexto deste trabalho, foram utilizados dois filtros:

a. Uma quantidade mínima de aparição do termo no documento. Após a contagem da ocorrência das palavras no documento, o valor resultado é comparado à informação de quantidade mínima informada como parâmetro, sendo eliminado o termo que não alcançar o mínimo.

b. Aproveitamento de termos segundo a "regra do joelho", onde o termo é considerado representativo na relação entre a lista de freqüência do termo e o número total de documentos do conjunto. O fator de aproveitamento é calculado em função da freqüência (5), onde N é o número total de documentos e *f* é a quantidade de vezes que o termo aparece no documento.

$$FT_j = \sum_{i=1}^{N} f_{ij}^2 - \frac{1}{N}\left[\sum_{i=1}^{N} f_{ij}\right]^2 \qquad (5)$$

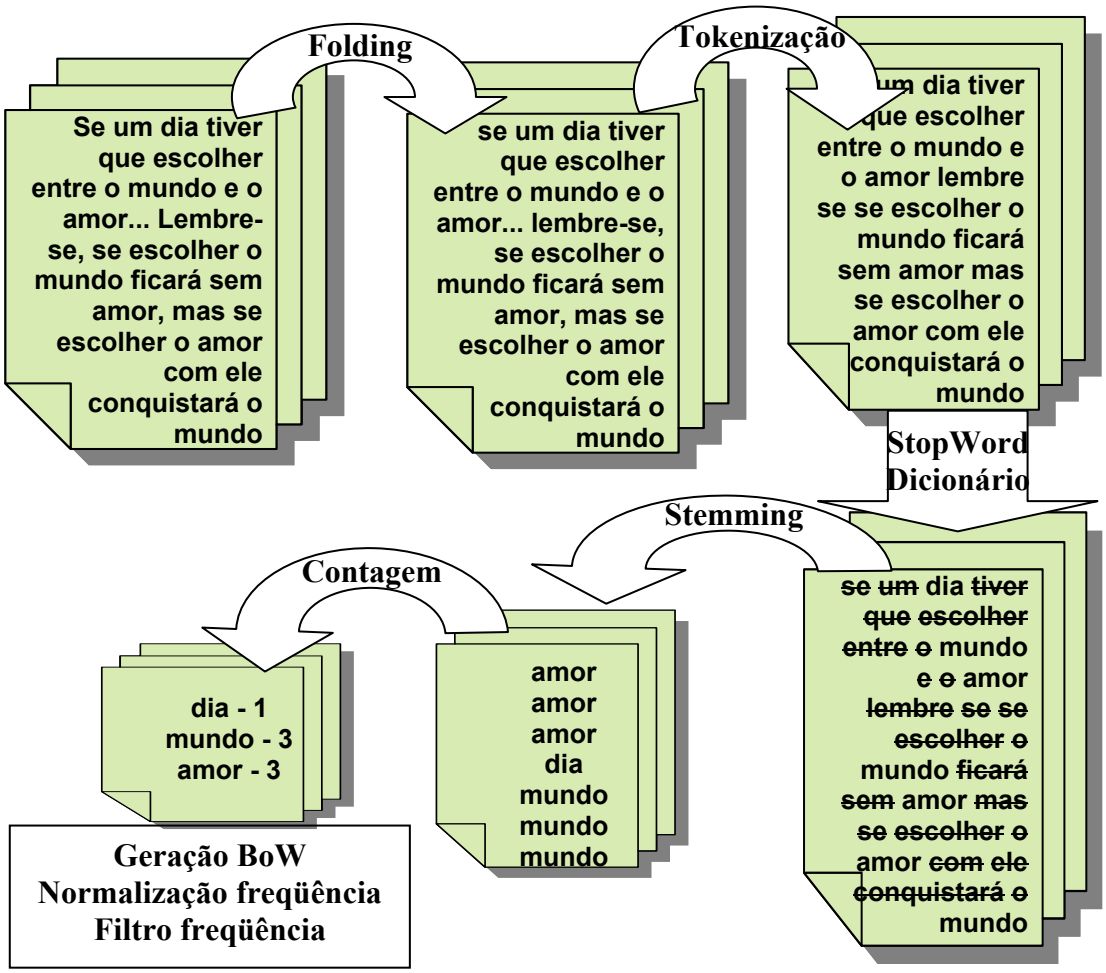

Figura 2.1 – Seqüência de aplicação das técnicas do Pré-processamento

2.2. Processamento

Neste trabalho, o processamento é representado com a aplicação de métricas de similaridade, definindo o peso de semelhança entre os documentos. O processo gera a matriz de similaridade (Tabela 2.2) representando a relação entre documentos. A matriz de similaridade utiliza para armazenamento uma matriz simétrica de dimensões $N \times N$ em que o elemento (i, j), representa um índice de similaridade (Sim_{ij}) relacionado a semelhança entre o Documento D_i e o Documento D_j.

A semelhança entre os termos determina o peso da ligação [11], representado no intervalo [0,1], onde os documentos idênticos possuem o grau 1. O peso possibilita aplicação de cortes na aplicação das técnicas [3].

D\D	***D_j***	***...***	***D_j***
D_i	1		
...	Sim_{ij}	1	
D_i	Sim_{ij}	Sim_{ij}	1

Tabela 2.2 – Resultante de aplicação de medida de similaridade

No escopo do trabalho, foram consideradas as métricas de Coseno e Manhatan para avaliar os resultados.

Considerando os documentos D_i e D_j, normalizado na freqüência *TF-IDF* em vetores u_i e u_j, como:

$$\mathbf{u}_i = \left(u_{D_i}(T_1), u_{D_i}(T_1), \ldots, u_{D_i}(T_M)\right) \qquad (6)$$
$$\mathbf{u}_j = \left(u_{D_j}(T_1), u_{D_j}(T_1), \ldots, u_{D_j}(T_M)\right)$$

A métrica de similaridade Coseno (SC) é baseada na norma da freqüência dos termos entre os documentos e, considera o ângulo entre dois pontos como base:

$$SC(D_i, D_j) = \frac{(u_{\mathbf{i}} \cdot u_{\mathbf{j}}^{\mathbf{T}})}{\|u_{\mathbf{i}}\| \cdot \|u_{\mathbf{j}}\|} \qquad (7)$$

A métrica de similaridade Manhattan (SM) representa a soma da diferença absoluta e, Yu, Amores, Sebe e Tian, relatam que corresponde a situação onde a distribuição dos dados é exponencial [6].

$$SM(D_i, D_j) = \frac{\Sigma |u_i - u_j|}{\Sigma (u_i + u_j)}. \qquad (8)$$

3. REDES COMPLEXAS

3.1. Elementos da Teoria dos Grafos

Um grafo é um conjunto de vértices e um conjunto de arestas, em que cada aresta conecta um par de vértices. Portanto, um grafo é definido por G = {V, A}, onde V é o conjunto de vértices e A o conjunto de arestas definidas entre dois vértices.

Por exemplo, considerando V={1,2,3,4,5} um conjunto finito não vazio e A = {{1,3},{1,2},{1,5},{2,4}}, um grafo pode ser representado conforme Figura 3.1.

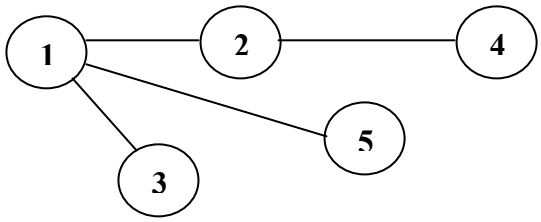

Figura 3.1 – Grafo (VxA)

O grafo pode ser orientado e não orientado. Um grafo orientado determina a direção das arestas entre os vértices e, os não orientados não apresentam direção, pois a aresta é definida nos dois sentidos.

Um caminho é definido pelas arestas que deverão ser percorridas entre um par de vértices. As arestas podem ser ponderadas, introduzindo o conceito de distância, que se refere ao comprimento entre o par de vértices. O significado da distância está diretamente relacionado ao conceito apresentado ao grafo. Para os mapas de um Estado, as cidades representam os vértices e as arestas são as estradas que ligam as cidades umas nas outras. A distância é representada pelos quilômetros a serem percorridos entre duas cidades e o caminho será a soma dos quilômetros entre as cidades escolhidas.

Na teoria de grafos é considerado o problema do caminho mínimo com o objetivo de calcular a menor distância a percorrer entre os vértices. O algoritmo de Dijkstra é mais utilizado para cálculo do caminho mínimo e tem como objetivo eleger um vértice como raiz de busca e calcular o menor caminho para todos os vértices do grafo, considerando as distâncias positivas e não nulas. Estudos demonstram sua utilização em grafos orientados e não orientados.

É possível verificar em um grafo que os vértices podem não estar ligados a todos os vértices, introduzindo o conceito de grau. O grau é o que define o número de arestas conectadas em um vértice.

Considerando o grafo representado na Figura 3.1, o grau é definido como apresentado na tabela 3.1.

Vértice	Grau
1	3
2	2
3	1
4	1
5	1

Tabela 3.1 – Grau de grafo

3.2. Redes Complexas

As redes complexas são caracterizadas por representar um conjunto de elementos em que ligações dependem da característica do estudo que se deseja construir [7].

As redes possuem propriedades próprias, são suportadas pela teoria de grafos e se apresentam de três formas:

- "*directed",* que são ligações estabelecidas em uma única direção, também chamadas de cíclicas, e são representadas por um grafo orientado;
- "*undirected",* que são ligações estabelecidas em duas direções, também chamadas de acíclicas, que são representadas por um grafo não orientado, e
- "*bipartite*", quando existem várias ligações com propriedades diferentes entre pares de vértices.

A análise topológica de redes complexas consiste em mensurar as propriedades estruturais envolvidas na rede, como a conectividade (como e com qual vértice

estabelecem-se as ligações) e, a centralidade [7] (qual vértice possui a melhor conexão ou a maior influência). Cada propriedade é utilizada para caracterização topológica [8].

A topologia das redes complexas é apresentada como "Randômica" ("*Random Graph"*), "Mundo Pequeno" ("*Small-world"*) e "Livre de Escala" ("*Scale-free"*).

3.2.1. Redes Randômicas

Rapoport et. Al. [7] definiram o primeiro modelo para redes randômicas com a denominação de "*Random net*". Em seguida, Erdõs e Rényi redescobriram as redes randômicas e denominaram "Poisson Random Graph" [8].

Albert e, Barabási apresentam a evolução de formação de uma rede randômica (Figura 3.2) [14].

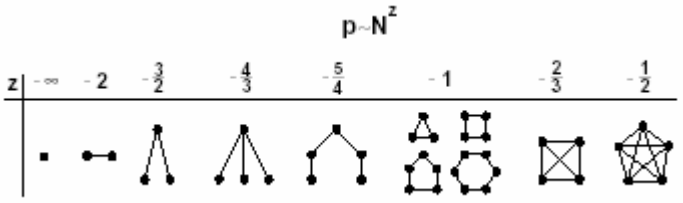

Figura 3.2 – Formação de redes Random Graphs [14]

A rede Randômica é definida pelo número de vértices N e a probabilidade *p* da existência de um link entre dois vértices. O grau esperado de um vértice na rede é [7]:

$$(k) = p(N-1). \tag{9}$$

(k) diverge se *p* é fixada. *p* é escolhida como uma função de N para manter o k fixo:

$$p = \frac{(k)}{(N-1)}. \tag{10}$$

O número de ligação pode ser calculado pela distribuição binomial, tratando-se de número randômico, aleatório.

$$f(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{11}$$

Portanto, foram definidas extensões deste modelo para poder ser aplicado em grau de distribuição arbitrário, denominado "Modelo de Configuração" e "Grafo randômico exponencial" [8].

"Modelo de Configuração" [8] é definido como o conjunto de redes produzidas com peso equivalente, a partir da escolha da seqüência de grau de distribuição com

igual probabilidade. Para definir a propriedade do modelo devem ser conhecidos a *pk* e a quantidade de componentes [8].

"Grafo randômico exponencial", segundo Strauss [8], também chamado de modelo *p**, é uma classe do conjunto de redes com número fixo de vértices, definido analogicamente como conjunto Boltzman de estatística mecânica. O estudo se deu a partir da necessidade de incorporar os efeitos da transitividade nas redes Randômicas. Na teoria de rede Randômica, a probabilidade é definida como uma função do número de documentos definindo o número de nós possíveis por:

$$edges = \frac{N(N-1)}{2} \tag{12}$$

O processo de construção da rede Randômica é:

$$G_0 = p^k (1-p)^{(\frac{N(N-1)}{2})-k} \tag{13}$$

3.2.2. Mundo Pequeno

Proposto por Watts e Strogatz, é o modelo mais popular de redes randômicas [7][14] com tratamentos para alta transitividade, coeficiente de aglomeração (propriedade de redes complexas que define o quanto os documentos estão interligados) e média de distância pequena [18]. Eles partiram do princípio em que redes podem ter um componente geográfico e seus vértices têm posições no espaço. Em muitos casos, é razoável assumir que proximidades geográficas irão impor uma regra para decidir a ligação [8][14], conforme Figura 3.3.

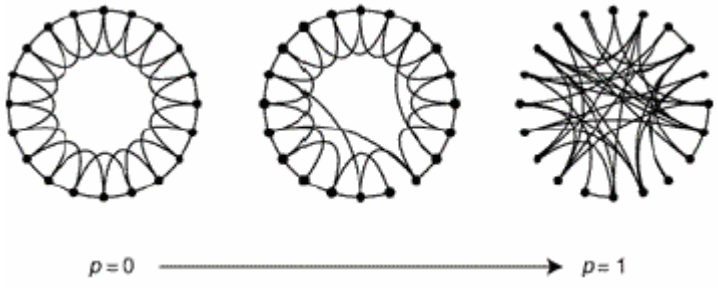

Figura 3.3 – Formação redes Mundo Pequeno [14]

A *p*=0 liga os vértices em pares, denominada rede regular. Na medida em que *p* aumenta o número de ligações cresce apresentando-se como rede Mundo Pequeno e, ao chegar na *p*=1 representa uma rede Randômica. Uma rede Mundo Pequeno é um

modelo comum entre muitos domínios e apresenta um número grande de ligação entre três vértices, isto é, se vértice i está conectado a j e x, existe uma grande possibilidade de j estar conectado a x;

3.2.3. Livre de Escala

Os modelos apresentados anteriormente relacionam propriedades relativas à distribuição regular de conectividade, seja com ligações determinadas ou aleatórias. O modelo Livre de Escala apresenta uma distribuição irregular das ligações. O conjunto de ligações a um mesmo vértice indica o aparecimento de vértices centrais ("*hubs"*), ou seja, vértices que concentram ligações com muitos outros vértices. Em redes "*directed", os centros* representam os vértices que possuem muitos vértices filhos [12] e, em redes "*undirected",* os centros são representados por vértices que concentram ligações com outros vértices.

A expansão da rede se dá pela adição de vértice a vértice já existente. A escolha do vértice na ligação segue a regra de junção linear preferencial, onde os vértices centrais têm a maior probabilidade de receber novos vértices, segundo o paradigma "o rico fica cada vez mais rico", chamado, por Derek de Solla Price, de vantagem acumulativa [14].

O grau de distribuição segue uma regra exponencial, onde P(k) é a probabilidade de um dado nó ter grau igual a k:

$$P(k) \sim k^{-y} \qquad (14)$$

onde:

k = grau de distribuição;

y = constante.

Portanto, considera-se característica desta rede: a existência de vértices centrais; distribuição das ligações não uniforme; a expansão da rede pela adição de vértice aos vértices já existentes e a existência de poucos documentos com grau de distribuição alto e muitos documentos com grau de distribuição baixo.

3.3. Caracterização Topológica

3.3.1. Grau de Distribuição

O grau de distribuição de uma rede representa o número de ligações existentes de um vértice com outros vértices, também referenciado por Centralidade do Grau. O grau do vértice *i*, denominado k_i, é o número de ligações, isto é, a cardinalidade do conjunto de vértices, também chamada, na física, de conectividade.

Albert e Barabási apresentam a freqüência em função do grau de distribuição para redes randômicas (Figura 3.4), redes Mundo Pequeno (Figura 3.5) e redes Livre de Escala (Figura 3.6), para definição da topologia da rede [14].

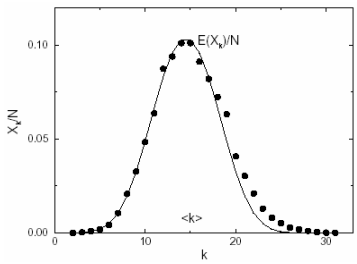

Figura 3.4 – Freqüência Redes Randômica

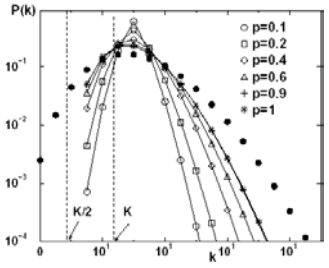

Figura 3.5 – Freqüência Redes Mundo Pequeno

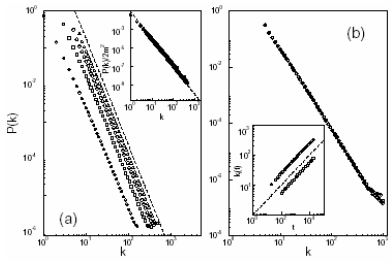

Figura 3.6 – Freqüência Redes Livre de Escala

A representação da freqüência em redes Randômica (Figura 3.4) configura a existência de muitos vértices com o grau de distribuição médio. Para redes Mundo Pequeno (Figura 3.5), o número de vértices sem ligação aumenta, mas ainda possui um grande número de vértices com grau na média. Para as redes Livre de Escala (Figura 3.6), o gráfico representa uma maior quantidade de vértices em grau de distribuição baixo, que diminui na medida em que o grau de distribuição aumenta.

3.3.2. Coeficiente de Aglomeração

O coeficiente de aglomeração é utilizado para definir a transitividade e indicar a densidade da rede, ou seja, formar grupos de vértices coesos [7][8] e identificar vértices centrais. Valor alto de coeficiente de aglomeração indica que seus vizinhos estão conectados entre si. O cálculo do coeficiente de aglomeração é obtido através da relação entre o número de triângulos e o número de triplas que o vértice participa. [8] (18), em outras palavras, a soma da relação entre o número de triângulos-conjunto de três vértices interligados entre si e número de triplas de cada documento [8] (19) .

Coeficiente de aglomeração do documento:

$$C_i = \frac{NT}{NL} \tag{18}$$

Coeficiente de aglomeração da Rede:

$$C_r = \frac{\sum NT}{\sum NL} \tag{19}$$

onde:

- NT = número total de triângulos para um documento;
- NL = Número de triplas = $\frac{k!}{2}$

Os triângulos são identificados pela ligação fechada entre três documentos (Figura 3.7) e as triplas como o total de pares de ligações definidos para cada documento (Figura 3.8), estabelecendo uma relação fatorial em relação ao grau de distribuição *k* [8]. Na Figura 3.7, é totalizado um triângulo para o vértice A, C, E e nenhum para os demais vértices. Na Figura 3.8 o vértice A contém três triplas, os vértices B, C e E possuem uma tripla e o vértice D não possui tripla.

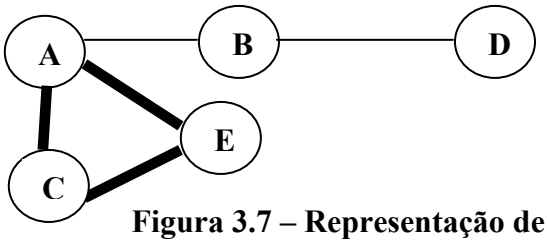

Figura 3.7 – Representação de triângulos

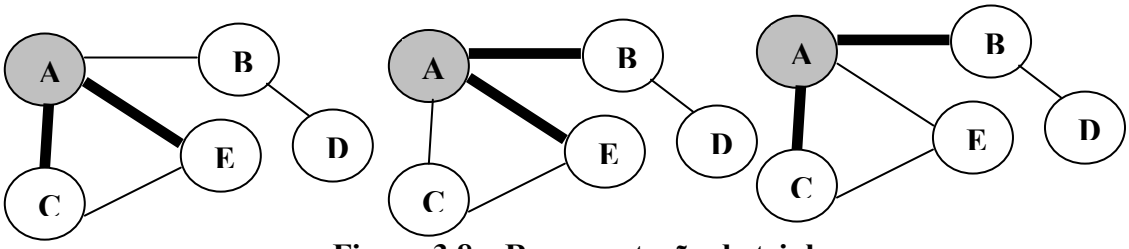

Figura 3.8 – Representação de triplas

3.3.3. Centralidade

A centralidade refere-se ao ponto de localização de cada vértice e sua importância em relação a seus vizinhos. Ela identifica o documento central e pode ser vista como uma medida de elasticidade da rede [8].

O documento central possui grau de distribuição alto, número de vizinhos também elevados e menores caminhos em suas ligações. O número de vizinhos representa o grau de distribuição dos documentos relacionados ao par da ligação em análise. Esses documentos são candidatos a gerar uma redefinição na rede quando de sua retirada.

A centralidade indica a importância do vértice na rede, ou seja, o que possui uma melhor conexão e maior influência [8]. A remoção de um vértice com alta centralidade provoca a queda de desempenho da rede [7].

3.4. Estrutura de Comunidade

Uma "Estrutura de Comunidade" em uma rede é um conjunto de vértices interligados fortemente [7][8], mas com a existência de poucas ligações entre os grupos. As estruturas de comunidade são analisadas para identificar grupos coesos interligados, ou seja, buscam verificar se a rede complexa de documentos é formada por grupos de mesma característica, interligados por documentos afins.

Newman [8] sugere agrupamento hierárquico como método para extração de comunidades (Figura 3.9) e em [15] apresentam a medida de modularidade para quantificar a organização entre as ligações e avaliar a existência de estruturas de comunidade.

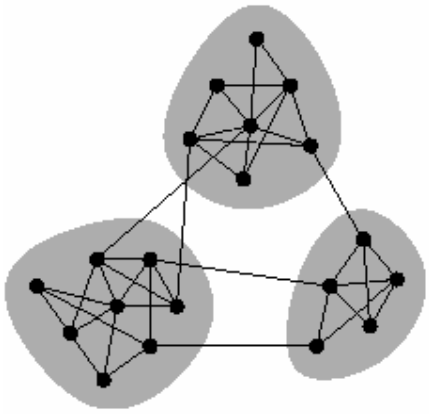

Figura 3.9 – Estrutura de Comunidade [15]

Assim como as redes hierárquicas, as estruturas de comunidade podem ser representadas por uma árvore ou dendrograma (Figura 3.10) [8], representando os grupos de documentos aninhados.

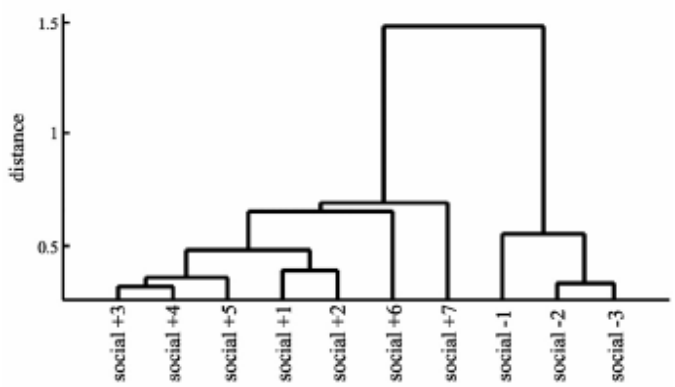

Figura 3.10 – Dendrograma – Rede Hierárquica [17]

Pastor-Satorras, Rubi e Diaz-Guilera descrevem a identificação de redes hierárquicas em função do coeficiente de aglomeração *C* e o grau de distribuição *k* (24).

$$C(k) \sim k^{-1} \tag{24}$$

Ele identifica uma topologia Livre de Escala e uma organização hierárquica para redes formadas a partir de páginas www conectadas através das chamadas internas. Em alguns casos apresenta grupos coesos, mas com baixa aglomeração e em outros casos, o *k* independente de *C* (Figura 3.11).

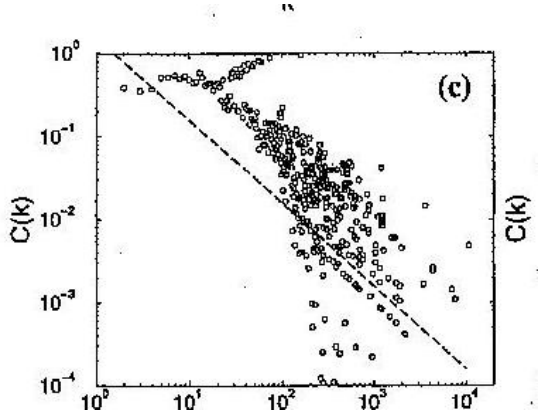

Figura 3.11 – Informações WWW [13]

4. METODOLOGIA PROPOSTA

A metodologia proposta neste trabalho define as etapas necessárias para desenvolver a análise de redes complexas para coleção de documentos em geral, sejam livros, artigos, teses, dissertações, revistas ou qualquer documento em formato texto. A literatura apresenta experiências para informações disponíveis na internet.

A análise de redes complexas de coleção de documentos na metodologia proposta extrai propriedades como topologia e comportamento da rede formada pelo conjunto de documentos com objetivo de entender as ligações existentes entre esses documentos.

A rede complexa para coleção de documentos (Figura 4.1) é classificada como "*undirected*" e é representada pela estrutura de grafo não ordenado, definido por G (D,S), onde os documentos D representam os vértices e S as arestas formadas a partir da similaridade calculada em função da freqüência normalizada dos termos nos documentos. Portanto, a representação é uma matriz simétrica ao longo da diagonal principal, indicando que a informação Sim_{ij} é igual à informação Sim_{ji}.

O valor apurado para cada ligação representa o peso da ligação, tornando possível a formação de redes em limites diferenciados e a navegação entre vértices baseada na definição de caminhos mínimos. Somente são considerados pesos com valores diferentes de zero.

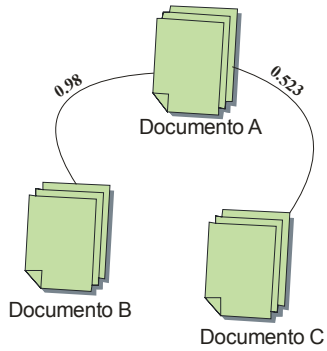

Figura 4.1 – Rede de documentos

Definida a rede de documentos, o próximo passo é extrair suas propriedades aplicando as métricas de análise. As métricas de análise utilizadas estão relacionadas basicamente pelo grau de distribuição e coeficiente de aglomeração. A freqüência do

grau de distribuição sugere a topologia da rede e o coeficiente de aglomeração o nível de coesão dos documentos.

Complementarmente, são gerados grafos para identificação dos documentos centrais e estruturas de comunidade e gráficos para análise da topologia.

A metodologia é apresentada em seis etapas, conforme Figura 4.2. As cores utilizadas na representação das etapas indicam a natureza das atividades realizadas. As etapas identificadas pelo número I requerem envolvimento direto de usuário, as de número II são atividades desenvolvidas no escopo deste trabalho e as de número III utiliza software comercializado.

Cada etapa tem suas atividades definidas e é realizada seqüencialmente, mas a retroação entre as etapas é considerada sempre que for necessário melhorar os resultados finais.

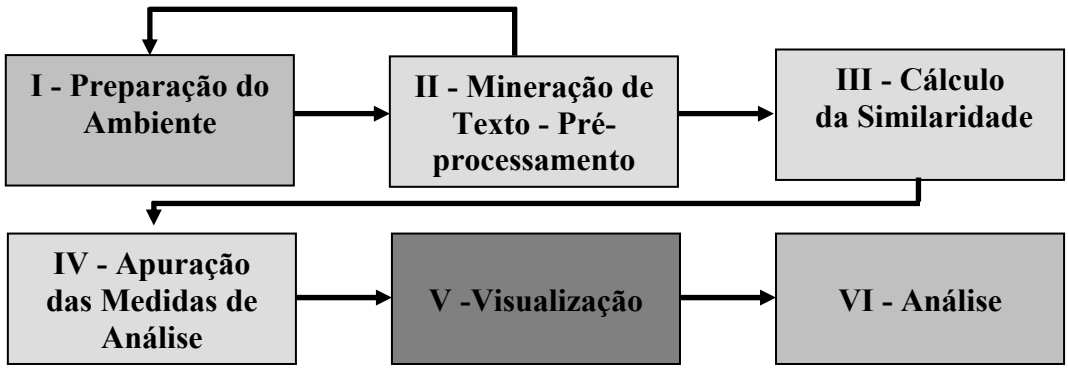

Figura 4.2 – Etapas da Metodologia proposta

4.1. Preparação do Ambiente

A etapa de preparação do ambiente para execução das etapas de análise de redes complexas de documentos define o conjunto de documentos, seleciona termos e cria o banco de dados.

No desenvolvimento deste estudo, são utilizados quatro conjuntos de documentos. Os conjuntos de documentos não possuem padronização de formatos e língua. O critério utilizado para seleção é a validação dos resultados com trabalhos já desenvolvidos.

A seleção de termos é a etapa utilizada para definir os termos que representem o conjunto de documentos, compondo um dicionário. Neste trabalho não foi elaborado dicionário previamente.

As informações são armazenadas em banco de dados relacional, devido ao alto volume de dados utilizado e gerado no processo de mineração de texto.

A preparação do ambiente é uma etapa importante no processo, pois representa a estrutura que dará suporte a toda execução.

4.2. Pré-Processamento

No pré-processamento do processo de mineração de textos são realizadas duas atividades: tratamento dos documentos e análise do conjunto de documentos.

A atividade de tratamento carrega os documentos a partir de um diretório e converte-os para o formato txt, gerando o documento livre de pontuação, tipos e tamanhos de letras, representando-o por uma única linha.

Cada termo empregado no documento em análise é separado em "*tokens'* e entregues às técnicas de retirada de "*StopWords*" e "*Stemming*", resultando na geração da lista de termos efetivamente utilizado pelo documento.

O processo contabiliza o número de vezes em que o termo aparece no documento, aplica os filtros descritos na seção 2.1. e gera a *BoW*. O processo se repete para todos os documentos do conjunto. Ao final do processamento de todos os documentos, é calculada a normalização da freqüência relativa aos documentos, utilizando a formulação "*Tf-Idf*", descrita no item 2.1. O resultado é armazenado no banco de dados para todo termo do documento.

A execução das atividades de pré-processamento é realizada pelo software Biguá, desenvolvido por um grupo de alunos do Programa de Engenharia Civil (PEC) – Coppe/RJ, para suprir os processos de mineração de texto. Essa ferramenta suporta conversão, tratamento, geração da tabela de freqüência, exportação de dados e integração com outros módulos como agrupamento e análise de redes complexas. A ferramenta está desenvolvida na linguagem de programação Java, utilizando o software de banco de dados PostGreSQL 8.1. para armazenamento dos dados.

O Biguá utiliza uma interface gráfica para informação de localização dos documentos, idioma, arquivos de tratamento e aplicação das técnicas de "*folding*", "*stopword*" e "*stemming*", conforme Figura 4.3.

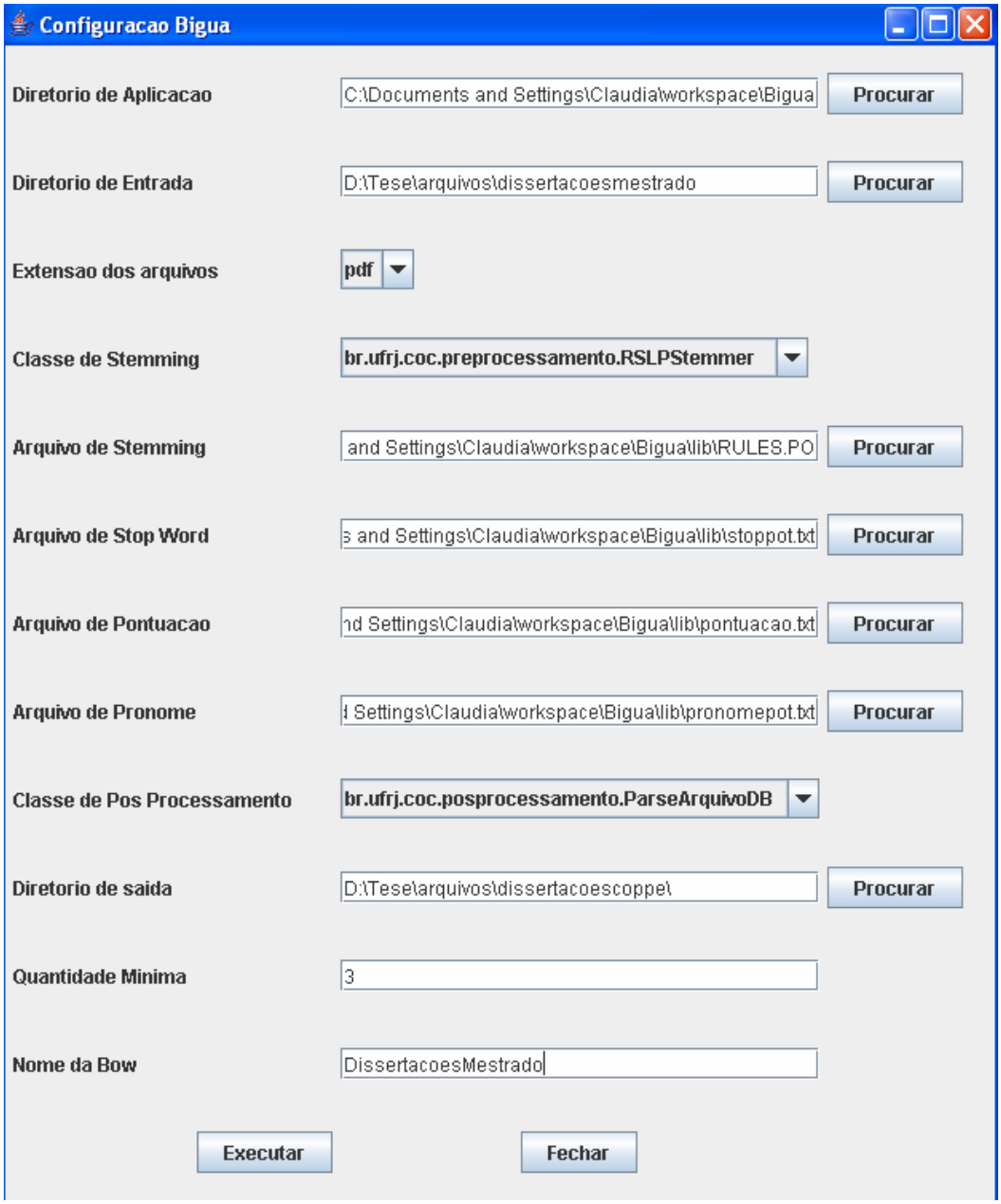

Figura 4.3 – Interface Gráfica *Biguá*

4.3. Cálculo da Similaridade

O cálculo da similaridade entre os documentos é desenvolvido no Biguá, onde são aplicadas as medidas de Coseno e Manhatan. O cálculo baseia-se nas informações armazenadas ao final da normalização "*Tf-Idf*" e resulta em um valor no intervalo {0,1}, determinando o peso da ligação entre os documentos. Os resultados são armazenados

no banco de dados, como exemplificado na Figura 4.4, onde cada linha da tabela de similaridade representa os dois documentos relacionados e o valor da similaridade. Não são armazenados os documentos que não possuem qualquer similaridade, ou seja, que tenham similaridade igual a zero.

Documento	Documento	Similaridade
PI03133249.txt	PI9912882 9.txt	0.991093
PI03133249.txt	PI9905270 9.txt	0.991016
PI03133249.txt	PI9917568 1.txt	0.989278
PI03133249.txt	PI9910996 4.txt	0.984914
PI03133249.txt	PI9910524 1.txt	0.979856
PI03133249.txt	PI9915774 8.txt	0.97388
PI03133249.txt	PI9916388 8.txt	0.973149

Figura 4.4 – Resultado medidas de similaridade

A matriz de similaridade é uma matriz simétrica, portanto, apenas a matriz triangular inferior (ou superior) é armazenada (Figura 4.5).

1	2	3	4	5	6	7	8	9	10	11	12	13	14	1	16	17	18	19	20	21	22	23	24
1.00																							
0.82	1.00																						
0.81	0.87	1.00																					
0.79	0.76	0.79	1.00																				
0.85	0.86	0.83	0.81	1.00																			
0.80	0.82	0.81	0.79	0.87	1.00																		
0.75	0.81	0.75	0.78	0.78	0.75	1.00																	
0.77	0.78	0.76	0.73	0.76	0.71	0.74	1.00																
0.70	0.76	0.70	0.78	0.76	0.74	0.75	0.71	1.00															
0.72	0.76	0.69	0.75	0.76	0.70	0.79	0.74	0.78	1.00														
0.73	0.75	0.73	0.72	0.81	0.75	0.76	0.70	0.78	0.79	1.00													
0.76	0.79	0.78	0.72	0.76	0.76	0.72	0.75	0.76	0.79	0.79	1.00												
0.72	0.75	0.68	0.70	0.75	0.76	0.79	0.67	0.72	0.73	0.78	0.78	1.00											
0.76	0.77	0.75	0.74	0.74	0.72	0.72	0.71	0.72	0.72	0.70	0.74	0.72	1.00										
0.68	0.72	0.70	0.74	0.70	0.67	0.79	0.71	0.72	0.77	0.75	0.70	0.73	0.74	1.00									
0.81	0.81	0.81	0.78	0.83	0.83	0.81	0.73	0.73	0.75	0.76	0.79	0.84	0.77	0.78	1.00								
0.82	0.83	0.81	0.75	0.78	0.75	0.76	0.76	0.75	0.71	0.68	0.78	0.69	0.78	0.69	0.81	1.00							
0.78	0.81	0.75	0.71	0.73	0.71	0.76	0.73	0.67	0.70	0.69	0.76	0.73	0.75	0.73	0.75	0.76	1.00						
0.79	0.82	0.75	0.71	0.72	0.71	0.75	0.73	0.67	0.68	0.68	0.73	0.72	0.77	0.71	0.75	0.76	0.98	1.00					
0.70	0.73	0.71	0.70	0.70	0.65	0.78	0.67	0.73	0.76	0.72	0.72	0.80	0.75	0.85	0.77	0.68	0.77	0.76	1.00				
0.70	0.73	0.70	0.70	0.73	0.68	0.70	0.65	0.76	0.70	0.73	0.72	0.78	0.71	0.70	0.76	0.73	0.74	0.72	0.76	1.00			
0.76	0.80	0.76	0.68	0.72	0.71	0.70	0.74	0.68	0.70	0.69	0.75	0.77	0.75	0.71	0.76	0.79	0.76	0.76	0.72	0.76	1.00		
0.73	0.75	0.72	0.73	0.68	0.68	0.70	0.71	0.65	0.64	0.57	0.67	0.65	0.74	0.65	0.72	0.70	0.76	0.79	0.68	0.70	0.75	1.00	
0.71	0.73	0.72	0.71	0.67	0.68	0.74	0.67	0.66	0.70	0.65	0.71	0.73	0.70	0.75	0.79	0.76	0.73	0.72	0.72	0.74	0.81	0.70	1.00

Figura 4.5 – Matriz de Similaridade

No escopo deste trabalho, o peso da ligação está sendo utilizado para avaliar o comportamento da rede. O limite do peso da ligação considerado para apresentação dos resultados é 0.1.

4.4. Apuração das Medidas de Análise

A apuração das medidas de análise consiste no processo de extração das informações de grau de distribuição e coeficiente de aglomeração dos documentos e da rede como um todo, a partir das ligações de similaridades dos documentos criadas pela

fase de pré-processamento. O processo foi desenvolvido no escopo deste trabalho, utilizando a linguagem Java como linguagem de programação, e é compatível com o software Biguá. A apuração é realizada em três partes.

Na primeira parte, o procedimento percorre todos os documentos calculando para cada um o grau de distribuição, o número de triângulos e o número de triplas (Tabela 4.1).

Documento	Grau	Triângulos	Triplas
PI0313324 9.txt	43	249	903
PI9905430 2.txt	15	21	105
PI9905466 3.txt	45	334	990
PI9906278 0.txt	32	132	496
PI9906295 0.txt	13	9	78
PI9907479 6.txt	9	16	36
PI9907868 6.txt	19	44	171
PI9908450 3.txt	7	21	21
PI9909578 5.txt	9	24	36
PI9910161 0.txt	6	11	15
PI9910238 2.txt	57	585	1596
PI9910238 2.txt	57	585	1596
PI9910238 2.txt	57	585	1596

Tabela 4.1 – Medidas Documento

Na segunda parte o procedimento calcula as medidas da rede, gerando informações de freqüência, coeficiente de aglomeração e totais de documentos distribuídos (Tabela 4.2).

Limite	Total documentos	Total documentos ligados	Total de ligações	Total sem ligação	Coeficiente aglomeração	Grau Máximo
0.1	803	755	21090	48	0.449303711	127
0.3	803	576	3918	227	0.819141253	46
0.5	803	267	1808	536	0.937395126	38
0.7	803	95	1478	708	0.945787041	37
0.9	803	56	748	747	0.936725096	24

Tabela 4.2 – Medidas Rede

Na terceira parte o procedimento consiste em disponibilizar as informações para visualização. As informações são extraídas do banco de dados em formato texto para representação em gráfico (Figura 4.6 e Tabela 4.3) e geração dos grafos com o software Pajek.

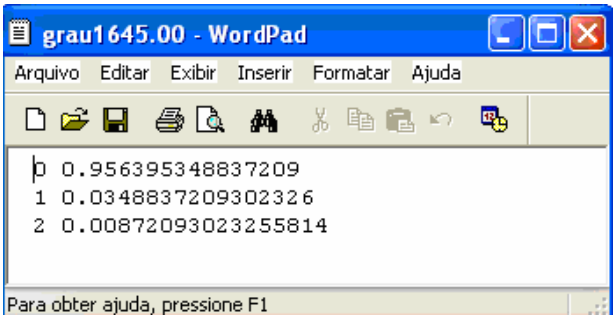

Figura 4.6 – Freqüência do grau

Grau	Coeficiente aglomeração
17	0.470588235
18	0.222222222
20	0.421052632
26	0.098461538
27	0.145299145
39	0.951417004
40	0.947435897
41	0.924390244

Tabela 4.3 – Grau x Coeficiente

A geração do arquivo com os dados a serem utilizados no software Pajek é dividido em duas partes: uma com a lista dos documentos e outra com as ligações e o valor referente a medida de similaridade. As informações para utilização no software Pajek são extraídas na totalidade independente de limite, pois este disponibiliza recursos para manipulação desta natureza com as informações.

4.5. Visualização

As redes complexas de conjunto de documentos são visualizadas a partir dos softwares Matlab 7.0 na geração dos gráficos de freqüência do grau de distribuição e em função do coeficiente de aglomeração e grau de distribuição (Figura 4.8) e no software Pajek 1.19 na geração do grafo, tornando possível a visualização das conexões estabelecida entre os documentos, centralidade e estrutura de comunidades (Figura 4.9). A visualização deste software é dinâmica, interativa e colorida, o que torna possível experimentar situações referentes ao comportamento da rede. A facilidade apresentada nos movimentos interativos nem sempre expressa a mesma apresentação clara no desenvolvimento textual.

A visualização oferece suporte para análise dos comportamentos. Em grande volume de informação o efeito visual pode não ser eficiente na extração de conhecimento justificando a aplicação das medidas e plotagem numérica para análise mais apurada.

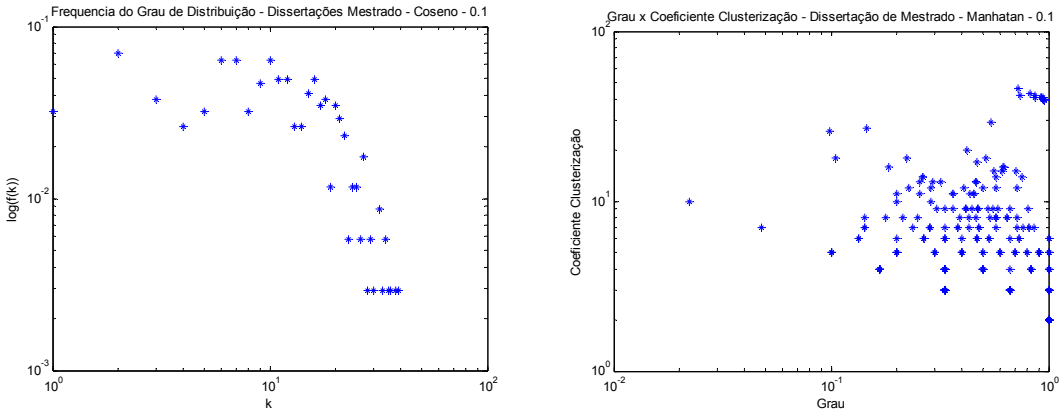

Figura 4.8 – Representação Matlab

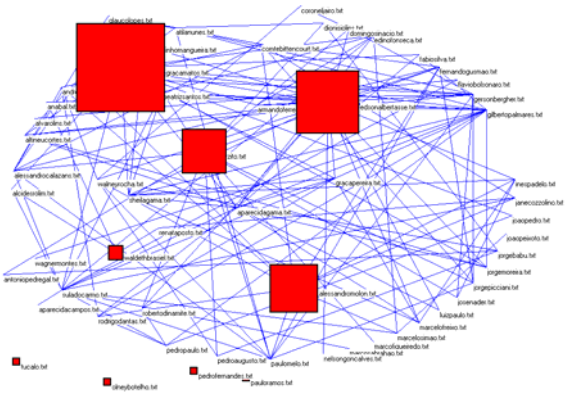

Figura 4.9 – Representação Pajek

4.6. Análise

A análise dos resultados tem como objetivo extrair conhecimento das redes para:

- ✓ Definir a topologia da rede;
- ✓ Identificar os documentos centrais;
- ✓ Identificar os documentos com maior vulnerabilidade da conectividade na rede; e
- ✓ Identificar a estrutura de comunidade.

Cada etapa tem sua importância no desenvolvimento da metodologia e é avaliada em função das características dos documentos analisados para obter melhores resultados.

A retroação entre as etapas da metodologia é realizada, seja para utilização de medidas de similaridades diferentes, inclusão de palavras na lista de "*StopWords"* ou termos no dicionário em busca de resultados melhores.

5. ESTUDOS DE CASOS

O estudo de caso é desenvolvido sobre quatro conjuntos de documentos: currículo de deputados / RJ, Patentes, Newsgroup e resumos de dissertações de mestrado do PEC.

Os conjuntos de documentos foram tratados de forma individualizada em função das naturezas diferentes de formatação e representação que possuem. O conjunto de deputados e patentes está em português e no formato texto; o conjunto de dados resumos de dissertações de mestrado do PEC está em português e no formato pdf e o conjunto de dados "*NewsGroup*" está em inglês e no formato texto (Tabela 5.0).

Conjunto de documentos	Total documentos	Natureza	Idioma
Newsgroup	942	txt	Inglês
Patentes	769	txt	Português
Deputados	59	txt	Português
Dissertações	344	pdf	Português

Tabela 5.0 – Conjuntos de documentos

As seções deste capítulo descrevem a caracterização e os resultados obtidos com a aplicação da metodologia proposta em cada conjunto de documentos.

5.1. Patentes

O conjunto de dados de patentes refere-se a descrições de patentes industriais, utilizado em [22], que consiste em aplicar técnicas de agrupamento para geração de grupos de documentos.

a. Pré-Processamento

As informações foram extraídas em formato html e convertidos para txt. Foram geradas 4.088 palavras para 769 documentos e aplicadas as medidas de similaridade Coseno e Manhatan.

b. Apuração das Medidas

Os resultados obtidos na apuração das medidas, variando os limites dos pesos e similaridades, podem ser conferidos na Tabela 5.1.

Medida de similaridade	Limite	Total de documentos conectados		Total de Documentos desconectados		Total de ligações
Coseno	0.1	755	98%	14	2%	20696
Manhatan	0.1	745	97%	24	3%	21184
Coseno	0.3	575	75%	194	25%	3900
Manhatan	0.3	398	52%	371	48%	2504
Coseno	0.5	267	35%	502	65%	1806
Manhatan	0.5	145	19%	624	81%	1560
Coseno	0.7	95	12%	674	88%	1478
Manhatan	0.7	67	9%	702	91%	1400
Coseno	0.9	56	7%	713	93%	748
Manhatan	0.9	47	6%	722	94%	688

Tabela 5.1 – Distribuição de documentos - Patentes

O uso da similaridade Coseno resultou em um maior número de documentos conectados, mas com um número menor de ligações em comparação com o Manhatan no limite de 0.1. Nos demais limites, a medida de similaridade Coseno superou o nível de conexão.

A apuração das medidas no conjunto de patentes está apresentando o movimento de transitividade (Tabela 5.2), onde é observado que a classificação apresentada pela medida de similaridade Manhatan com 0.7 de limite demonstra, efetivamente, que os poucos documentos conectados formam uma rede coesa, com muitos triângulos formados, se comparado ao limite de 0.5.

Pode-se observar, ainda, que o fato de ter um maior número de ligações não gera obrigatoriamente uma rede coesa, como demonstra a medida de similaridade Manhatan para o limite 0.1 (Tabela 5.2).

Medida de similaridade	Limite	Grau de distribuição	Coeficiente de aglomeração	Grau Máximo
Coseno	0.7	1.921976593	0.945787041	37
Manhatan	0.9	0.894668401	0.945568401	23
Manhatan	0.5	2.028608583	0.945247056	37
Coseno	0.5	2.348504551	0.937544953	38
Coseno	0.9	0.972691808	0.936725096	24
Manhatan	0.7	1.820546164	0.929120337	37
Manhatan	0.3	3.256176853	0.884735732	38
Coseno	0.3	5.071521456	0.821276496	46
Manhatan	0.1	27.54746424	0.489865437	144
Coseno	0.1	26.91287386	0.447655498	127

Tabela 5.2 – Transitividade da rede de documentos-Patentes

Nesta base de dados, o menor coeficiente de aglomeração está relacionado a similaridade Coseno (Tabela 5.2) com o limite 0.1, indicando a menor transitividade, maior número de ligações e grau máximo, mas o limite 0.7, mesmo com o total documentos interligados menor, apresenta uma transitividade alta.

c. Freqüência do Grau

A rede complexa formada a partir do conjunto de documentos de Patentes apresenta características de topologia do modelo Livre de Escala, onde o número de graus entre os vértices é variável e decresce na medida em que o grau aumenta. (Figura 5.1 e Figura 5.2). As medidas de similaridades aplicadas causaram pouca alteração nos resultados para esses conjuntos.

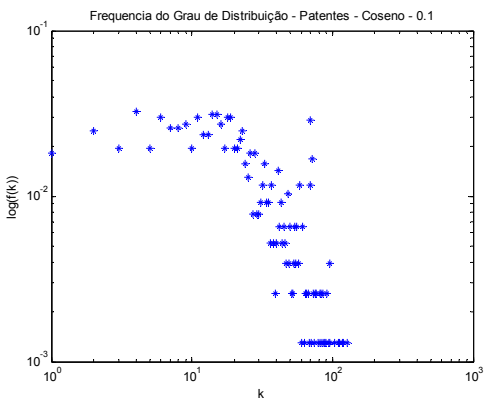

Figura 5.1
Patentes - Freqüência – Coseno 0.1

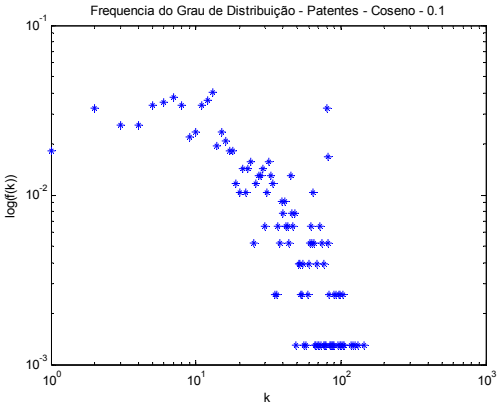

Figura 5.2
Patentes - Freqüência – Manhatan 0.1

d. Coeficiente de aglomeração x Grau de Distribuição

As Figuras 5.3 e 5.4 apresentam a rede complexa do conjunto de documentos "Patentes" como uma rede coesa. Na similaridade Coseno (Figura 5.3) a rede apresenta uma maior coesão em relação a similaridade Manhatan (Figura 5.4). A aparência assemelha ao experimento de informações WWW descrita na seção 3.4 indicando característica hierárquica.

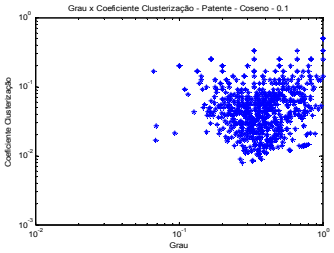

Figura 5.3 - Patentes
Coeficiente aglomeração - Coseno 0.1

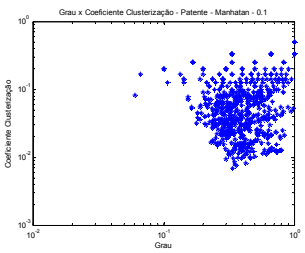

Figura 5.4 – Patentes
Coeficiente aglomeração - Manhatan 0.1

e. Estruturas de Comunidade

Na rede complexa do conjunto de documentos Patentes pode ser visualizada a presença de estruturas de comunidade na similaridade Coseno (Figura 5.5). Dois grupos de documentos estão interligados por um documento identificado por PI0405535, em destaque na Figura 5.5, definindo uma estrutura de comunidade. A retirada deste documento desconecta a rede.

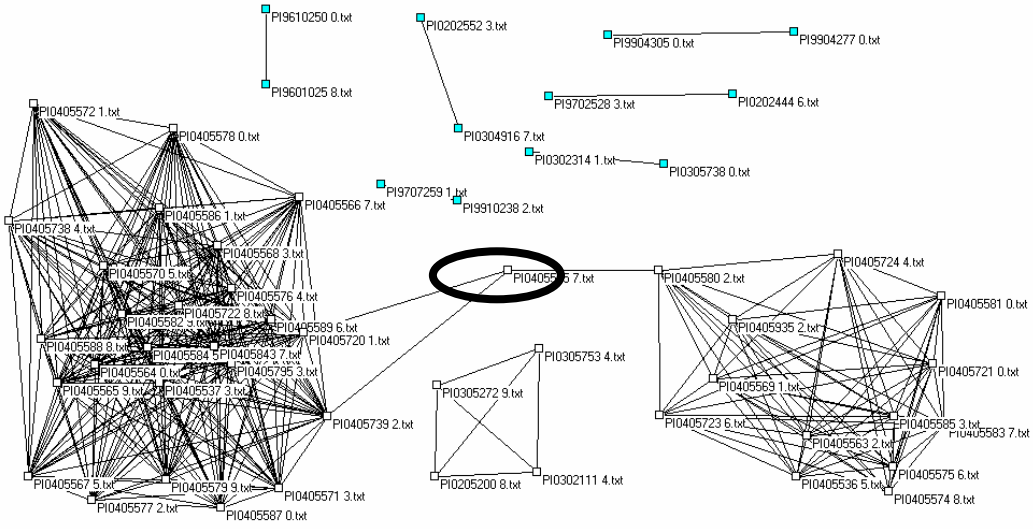

Figura 5.5 – Patentes – Estruturas de Comunidade - Coseno 0.9

A mesma rede representada na similaridade Manhatan (Figura 5.6) os grupos coesos existem, mas desconectados, o que desconfigura a existência da estrutura de comunidade.

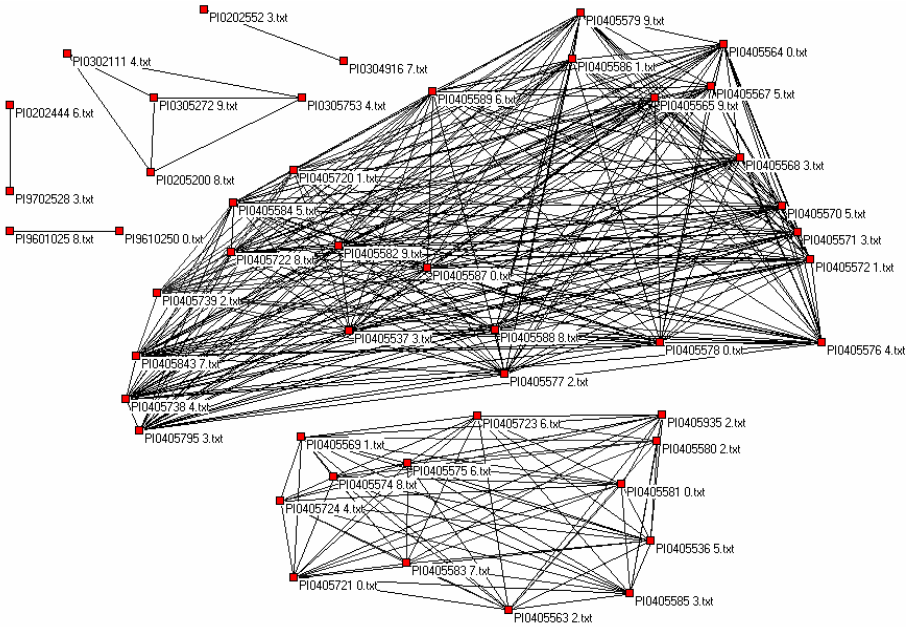

Figura 5.6 – Patentes - Estruturas de Comunidade - Manhatan 0.9

f. Identificação dos Centros

Os centros são identificados nos grafos pela figura geométrica quadrada (Figuras 5.7 e 5.8). O tamanho dos quadrados aumenta conforme a representação do documento no conjunto, ou seja, em função da quantidade de ligações existentes entre o documento central e seus vizinhos.

Os documentos centrais garantem a definição do caminho e navegação entre os documentos.

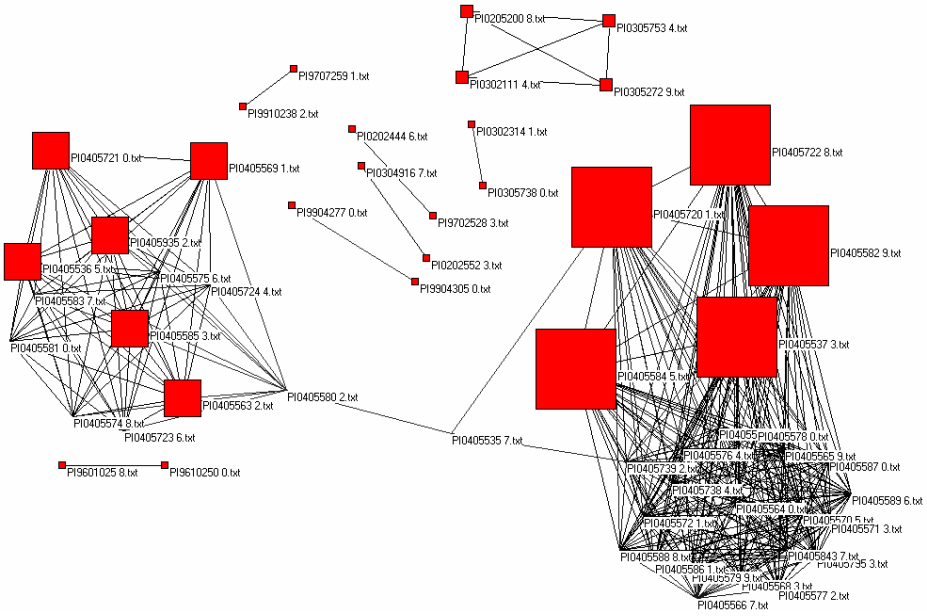

Figura 5.7 – Patentes - Centros - Coseno 0.9

Os documentos centrais identificados nas redes geradas para similaridade Coseno com limite 0.9 não são alterados na similaridade Manhatan com limite 0.9.

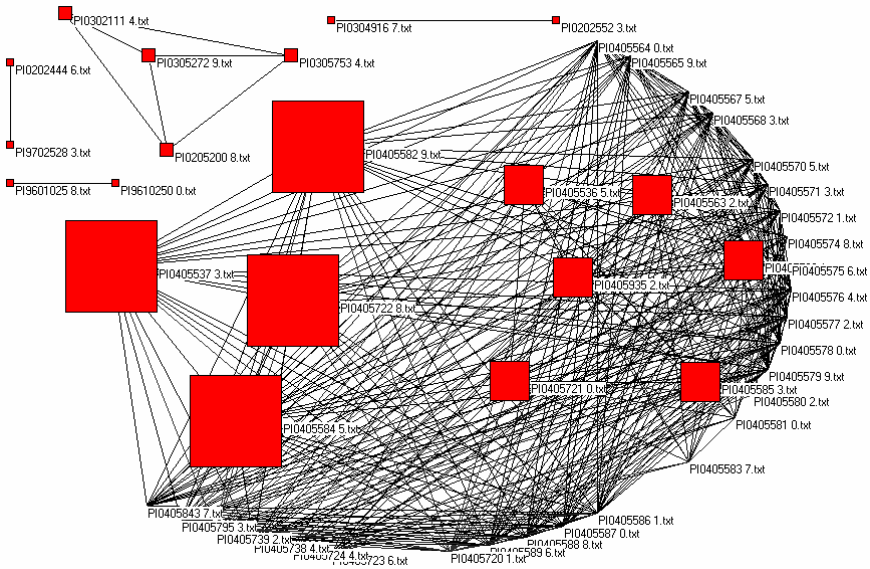

Figura 5.8 – Patentes - Centros - Manhatan 0.9

5.2. Deputados / RJ

As informações de cada deputado descrevem basicamente sua formação acadêmica, política e partido que pertence (Figura 5.9). As informações não são descritas em um único formato, mas seguem a seguinte indicação:

Nome completo:
Nome parlamentar:
Data de nascimento:
Local de nascimento:
Partido:
<Descrição de sua vida profissional e política>
Endereço/telefone/e-mail para contato:

Figura 5.9 – Caracterização deputados – Currículo deputados/RJ

As informações foram obtidas através da internet no endereço www.camara.gov.br onde estão disponíveis informações gerais de todos os deputados do Estado do Rio de Janeiro. Foram selecionados 59 deputados, pois os demais não possuíam informações registradas.

Com esse grupo de informação é possível conhecer os deputados de mesmo partido, mesma formação acadêmica e/ou mesma formação política.

a. Pré-Processamento

As informações foram extraídas em formato html e convertidos para txt. No pré-processamento do processo de mineração de texto foram incluídas na lista de Stop-Word termos do tipo: deputado, estado, ano, e-mail e outros, por serem considerados termos sem valor qualitativo. Foram geradas 2.396 palavras para 59 documentos, não havendo incidência forte a nível quantitativo, provavelmente em função do pequeno número de elementos analisados.

b. Apuração das Medidas

Medida de similaridade	Limite	Total de documentos conectados		Total de Documentos desconectados		Total de ligações
Coseno	0.1	37	63%	22	37%	76
Coseno	0.3	4	7%	55	93%	4
Coseno	0.5	0	0%	59	100%	0
Coseno	0.7	0	0%	59	100%	0
Coseno	0.9	0	0%	59	100%	0
Manhatan	0.1	55	93%	4	7%	296
Manhatan	0.3	0	0%	59	100%	0
Manhatan	0.5	0	0%	59	100%	0
Manhatan	0.7	0	0%	59	100%	0
Manhatan	0.9	0	0%	59	100%	0

Tabela 5.3 – Distribuição de documentos - Deputado

O conjunto de documentos apresenta baixa similaridade em todos os limites, sinalizando a medida de similaridade Manhatan de limite 0.1 com maior número de conexões. A medida de similaridade coseno com limite 0.1 apresenta um comportamento inferior a medida de similaridade Manhatan de mesmo limite, mas em compensação supera no limite 0.3.

Quanto a transitividade o conjunto apresenta baixa transitividade, a maior coesão está na medida Manhatan com limite 0.1 (Tabela 5.4).

Medida de similaridade	Limite	Total de ligações	Grau de distribuição	Coeficiente aglomeração	Grau Máximo
Manhatan	0.1	296	5.0169	0.305893	16
Coseno	0.1	76	1.2881	0.16	6
Coseno	0.3	4	0.0678	0	1
Manhatan	0.3	0	0	0	0
Coseno	0.5	0	0	0	0
Manhatan	0.5	0	0	0	0
Coseno	0.7	0	0	0	0
Manhatan	0.7	0	0	0	0
Coseno	0.9	0	0	0	0
Manhatan	0.9	0	0	0	0

Tabela 5.4 – Transitividade da rede de documentos - Deputado

c. Freqüência do Grau de Distribuição

Este conjunto de documentos possui pouca similaridade. Na medida de similaridade Coseno os limites 0.5, 0.7, 0.9 não têm representação e, na medida de similaridade Manhatan os limites 0.3, 0.5, 0.7, 0.9 não têm representação. Mesmo assim pode ser observado que o gráfico do conjunto de documentos na similaridade Manhatan (Figura 5.11) decresce em função do aumento do grau apresentando uma característica da topologia Livre de Escala,

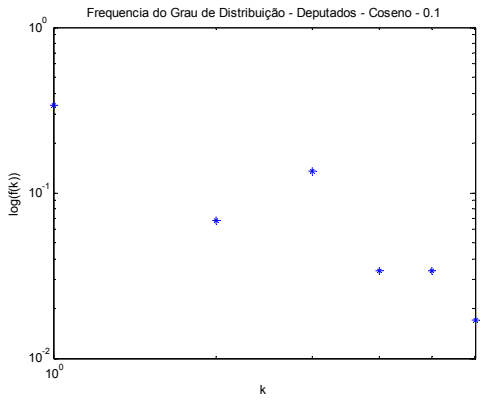

Figura 5.10 – Deputados Freqüência - Coseno 0.1

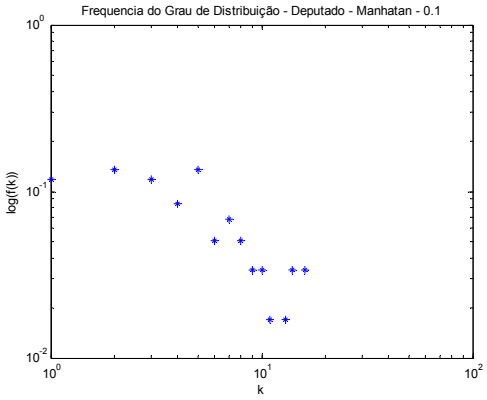

Figura 5.11 – Deputados Freqüência - Manhatan 0.1

d. Coeficiente de aglomeração x Grau de Distribuição

Esta medida define grupos não coesos no conjunto de documentos Deputados para a similaridade Coseno (Figura 5.12). Na similaridade Manhatan (Figura 5.13) o conjunto de documentos apresenta pouca coesão, mas com características da rede hierárquica. Somente foi definida representação para limite 0.1. Os demais limites e medidas de similaridade ficaram zerados.

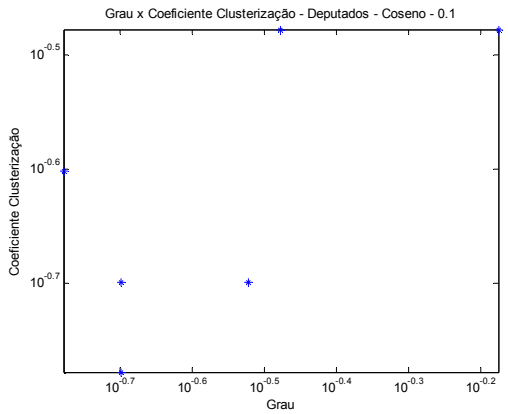

Figura 5.12 – Deputados Coeficiente aglomeração - Coseno 0.1

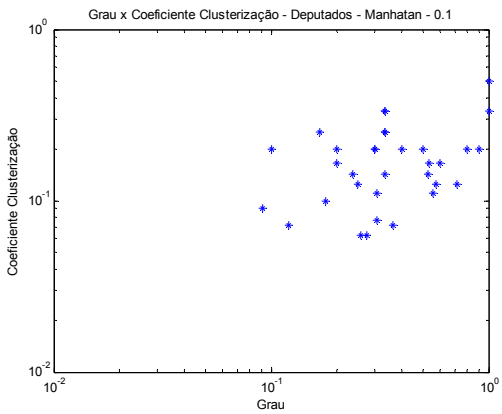

Figura 5.13 – Deputados Coeficiente aglomeração - Manhatan 0.1

e. Estruturas de Comunidade

A rede complexa do conjunto de documentos Deputado não possui estrutura de comunidade. Em similaridade Coseno limite 0.1 (Figura 5.14) não se observa grupos coesos e, em similaridade Manhatan limite 0.1 (Figura 5.15) observa-se um único grupo interligado, mas com poucas conexões entre os deputados.

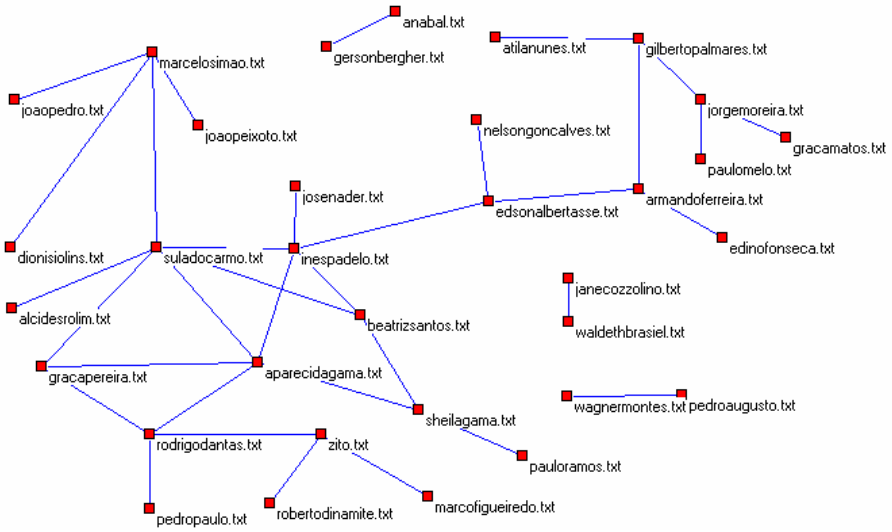

Figura 5.14 – Deputados – Estruturas de Comunidade - Coseno 0.1

A visualização Manhatan (Figura 5.15) sugere o aparecimento de vários triângulos confirmando o índice de coeficiente de aglomeração, que é maior do que o Coseno no mesmo limite, mas não é identificada estrutura de comunidade. Os documentos estão apenas ligados, não formando grupos.

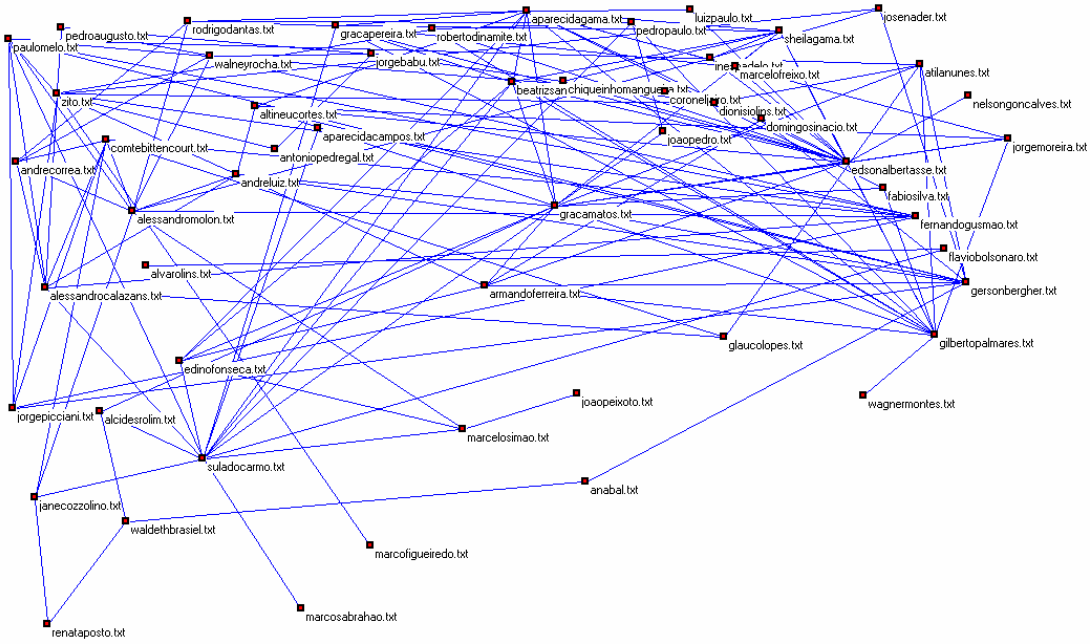

Figura 5.15 – Deputados – Estruturas de Comunidade - Manhatan 0.1

f. Identificação dos Centros

Os documentos centrais alteram em função do limite considerado. A similaridade Manhatan (Figura 5.17) no mesmo limite 0.1 da similaridade Coseno (Figura 5.16) gera um número superior de conexões, mas com menos documentos centrais.

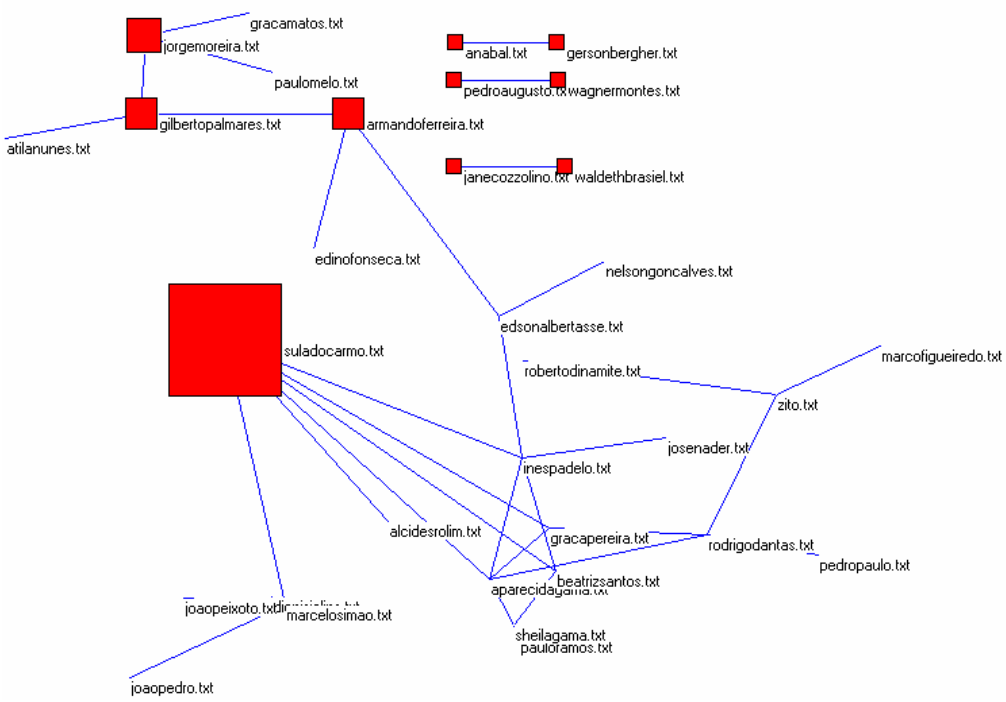

Figura 5.16 – Deputados – Identificação dos Centros - Coseno 0.1

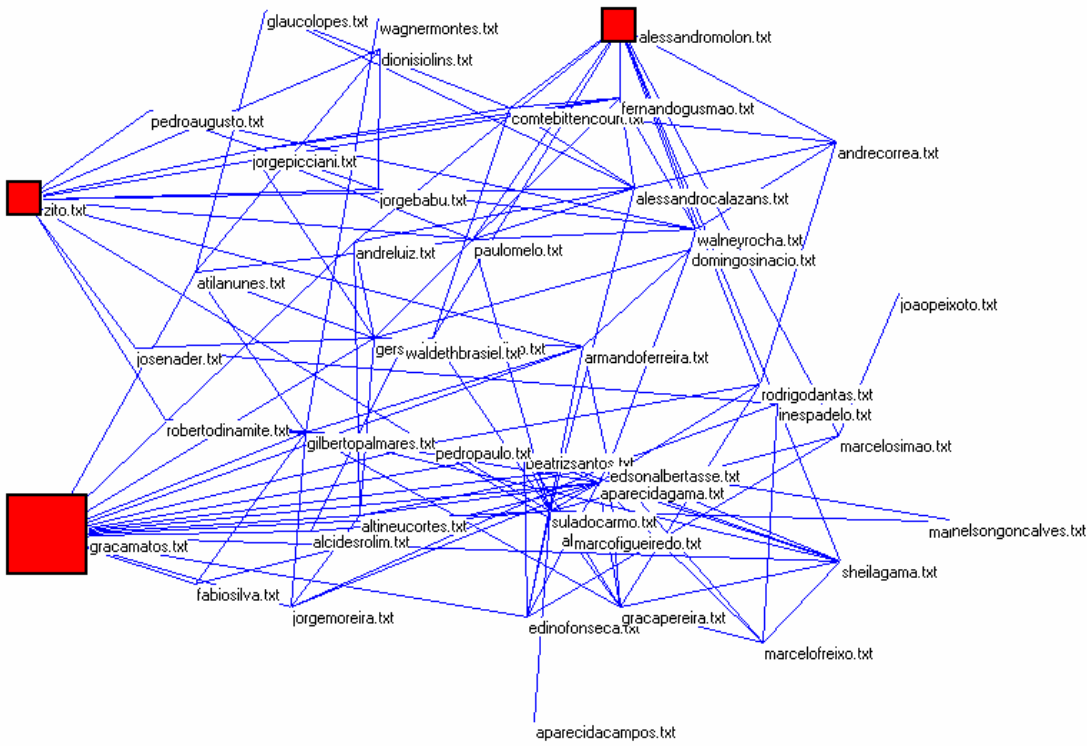

Figura 5.17 – Deputados – Identificação dos Centros - Manhatan 0.1

5.3. Resumos de Dissertação de Mestrado-PEC/RJ

O conjunto de documentos resumos de dissertação de mestrado-Pec/RJ, denominado "Dissertação" nesta seção, refere-se aos resumos elaborados para as dissertações de mestrado do PEC/RJ. Os documentos são formados por textos dissertativos de assuntos diversos ligados a engenharia. Foram obtidos através do acervo do PEC/RJ.

a. Pré-Processamento

As informações foram extraídas em formato PDF e convertidos para txt. No pré-processamento do processo de mineração de texto foram utilizados os processos de Stop-Word e stemming. Foram carregados 344 documentos e 2.236 palavras.

b. Apuração das medidas

Medida de similaridade	Limite	Total documentos conectados		Total documentos desconectados		Total de ligações
Coseno	0.1	336	98%	8	2%	3464
Manhatan	0.1	335	97%	9	3%	4100
Coseno	0.3	137	40%	207	60%	302
Manhatan	0.3	90	26%	254	74%	148
Coseno	0.5	50	15%	294	85%	64
Manhatan	0.5	15	4%	329	96%	18
Coseno	0.7	15	4%	329	96%	18
Manhatan	0.7	15	4%	329	96%	18
Coseno	0.9	15	4%	329	96%	18
Manhatan	0.9	13	4%	331	96%	16

Tabela 5.5 – Distribuição de documentos - Dissertação

O conjunto de documentos apresenta pouca similaridade, mas é possível identificar ainda documentos similares em limites 0.3 e até 0.9 (Tabela 5.5).

As medidas de similaridade e limites com Coeficiente de aglomeração igual a 0 determinam que os documentos conectados não formam triângulos e sim pares. A informação de transitividade (Tabela 5.6) para a medida Manhatan é maior com limite 0.3, não correspondendo a maior quantidade de ligação. A medida de similaridade Coseno no limite 0.5 apresenta um nível de coesão superior a limites inferiores, o que significa que as redes formadas na similaridade Coseno e limite 0.5 possuem maior densidade em comparação a rede formada na similaridade Coseno e limite 0.1. O grau de distribuição analisado isoladamente não determina a densidade e coesão da rede.

Medida de similaridade	Limite	Total de ligações	Grau de distribuição	Coeficiente aglomeração	Grau Máximo
Coseno	0.3	302	0.877906977	0.49122807	9
Manhatan	0.3	148	0.430232558	0.45	4
Coseno	0.5	64	0.186046512	0.4	3
Coseno	0.1	3464	10.06976744	0.361186434	36
Manhatan	0.1	4100	11.91860465	0.309471234	39
Manhatan	0.5	18	0.052325581	0	2
Coseno	0.7	18	0.052325581	0	2
Manhatan	0.7	18	0.052325581	0	2
Coseno	0.9	18	0.052325581	0	2
Manhatan	0.9	16	0.046511628	0	2

Tabela 5.6 – Transitividade da rede de documentos – Dissertação

c. Freqüência do Grau de Distribuição

Esta medida define a topologia de rede complexa do tipo livre escala nas duas medidas de similaridade Coseno (Figura 5.18) e Manhatan (Figura 5.19).

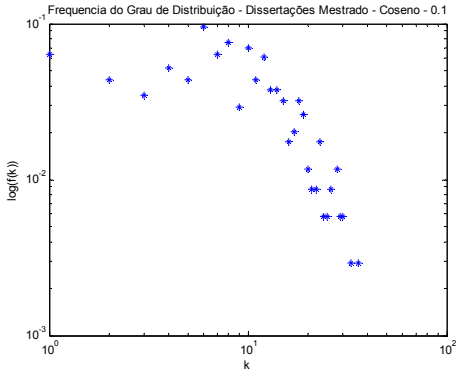

Figura 5.18 – Dissertação Freqüência - Coseno 0.1

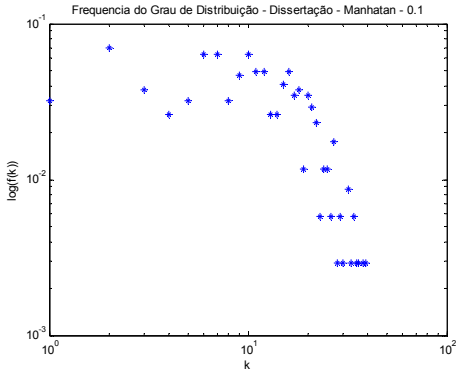

Figura 5.19 – Dissertação Freqüência - Manhatan 0.1

d. Coeficiente de aglomeração x Grau de Distribuição

O conjunto de documentos Dissertação apresenta grupos pouco coesos.

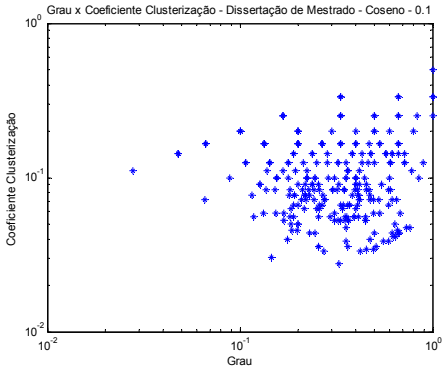

Figura 5.20 – Dissertação Coeficiente de aglomeração-Coseno 0.1

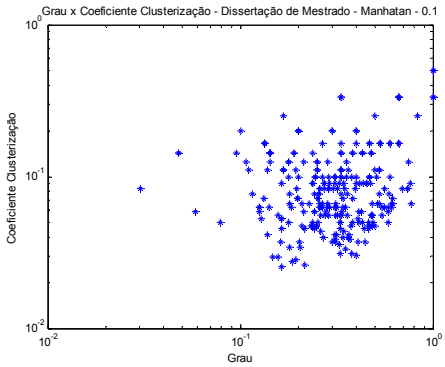

Figura 5.21 – Dissertação Coeficiente de aglomeração-Manhatan 0.1

e. Estruturas de Comunidade

A medida de similaridade indicada como a rede mais coesa, torna as ligações visualmente misturadas (Figura 5.23), mas o software faz verificar que os documentos estão efetivamente interligados. A similaridade Coseno no limite 0.3 apresenta coesão (Figura 5.22).

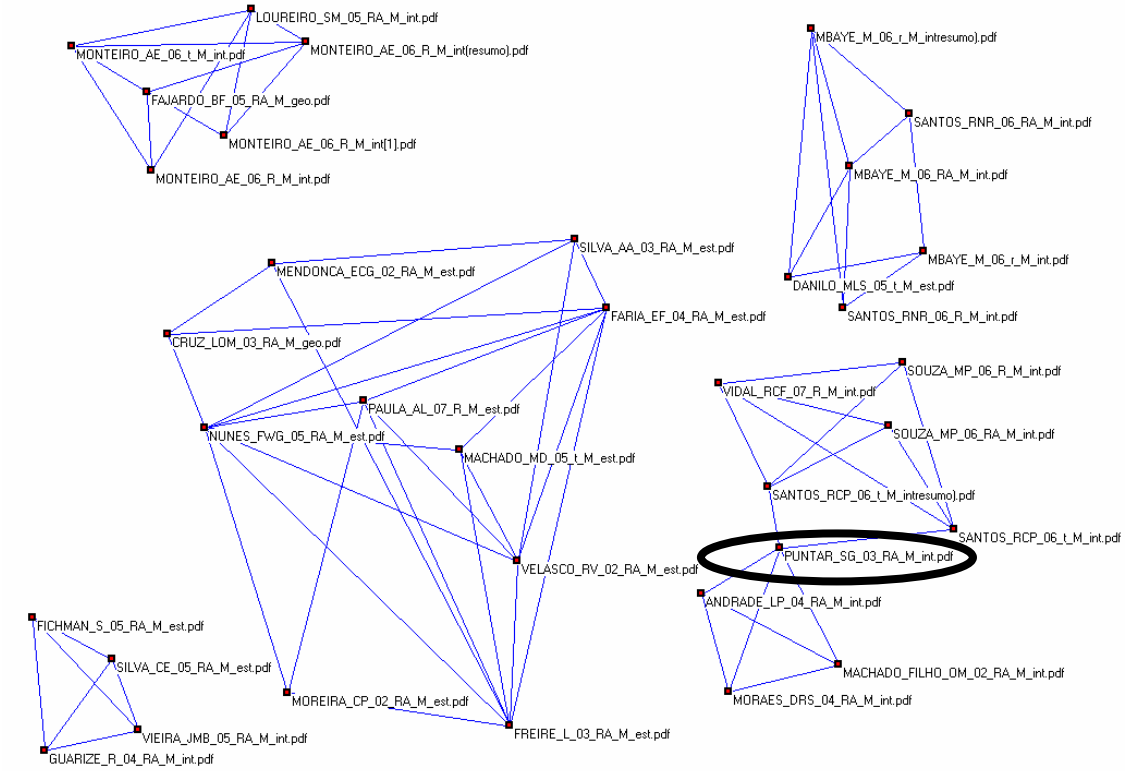

Figura 5.22 – Dissertação – Estruturas de Comunidade - Coseno 0.3

A rede de documentos na similaridade Coseno 0.3 (Figura 5.22) apresenta uma estrutura de comunidade interligada pelo documento "PUNTAR_SG_03_RA_M_intpdf" em destaque na figura.

A rede de documentos na similaridade Manhatan 0.3 (Figura 5.23) não apresenta estrutura de comunidade e possui menos documentos interligados.

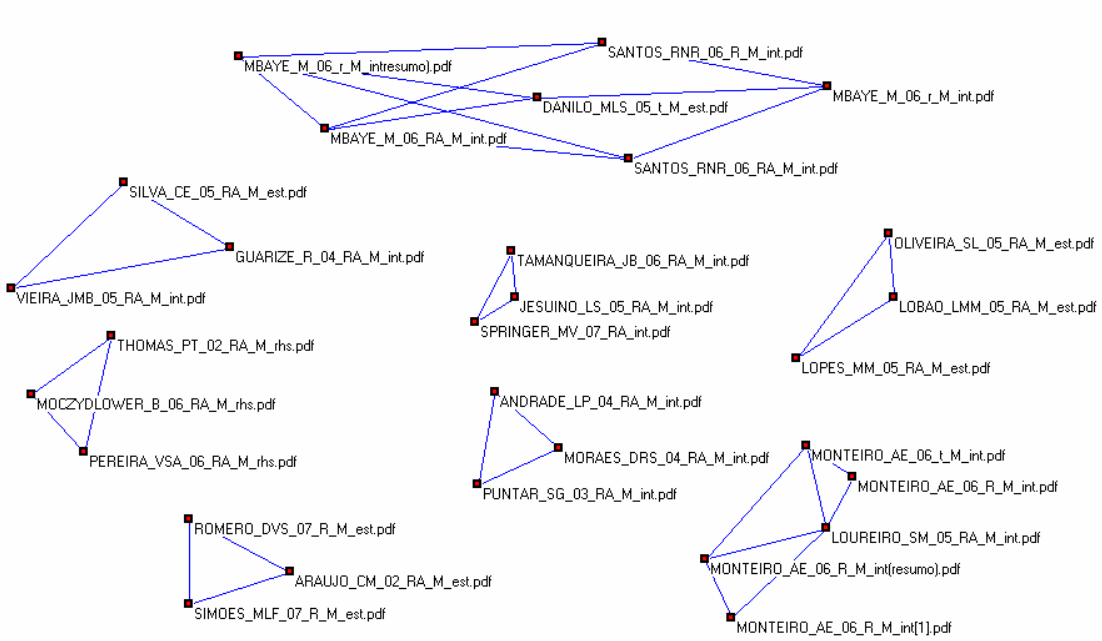

Figura 5.23 – Dissertação – Estruturas de Comunidade - Manhatan 0.3

f. Identificação dos Centros

O limite de conexão analisado foi 0.3, pois 0.5, 0.7 e 0.9 possui um baixo grau de distribuição. Com poucos laços.

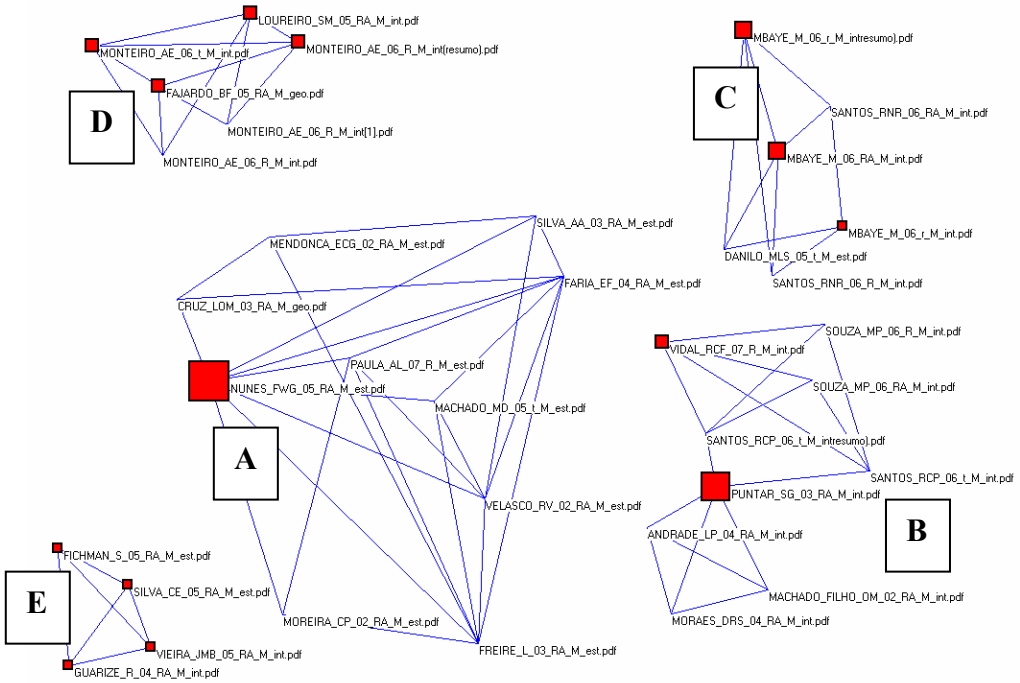

Figura 5.24 – Dissertação – Identificação dos Centros - Coseno 0.3

O conjunto de documentos “Dissertações” apresenta cinco grupos. Analisando a natureza de cada grupo é possível identificar que no grupo A temos a maioria de

documentos da área de estrutura e somente um documento da área de geotecnia. No grupo B os documentos se referem a área interdisciplinar. No grupo C os documentos são da área interdisciplinar com apenas um da área de estrutura. No grupo D os documentos são da área interdisciplinar com apenas um da área de geotecnia e, por fim o grupo E que possui documentos da área interdisciplinar e geotecnia. Com essa amostra pode-se dizer que a área de estruturas e a área interdisciplinar possuem similaridade entre suas dissertações de mestrado e a área de geotecnia usa assuntos relacionados às áreas estrutura e interdisciplinar.

Na similaridade Manhatan-0.1 (Figura 5.25) a rede é modificada e apresenta menos documentos conectados, mas as dissertações da área interdisciplinar e estrutura se apresentam com similaridade e conexão.

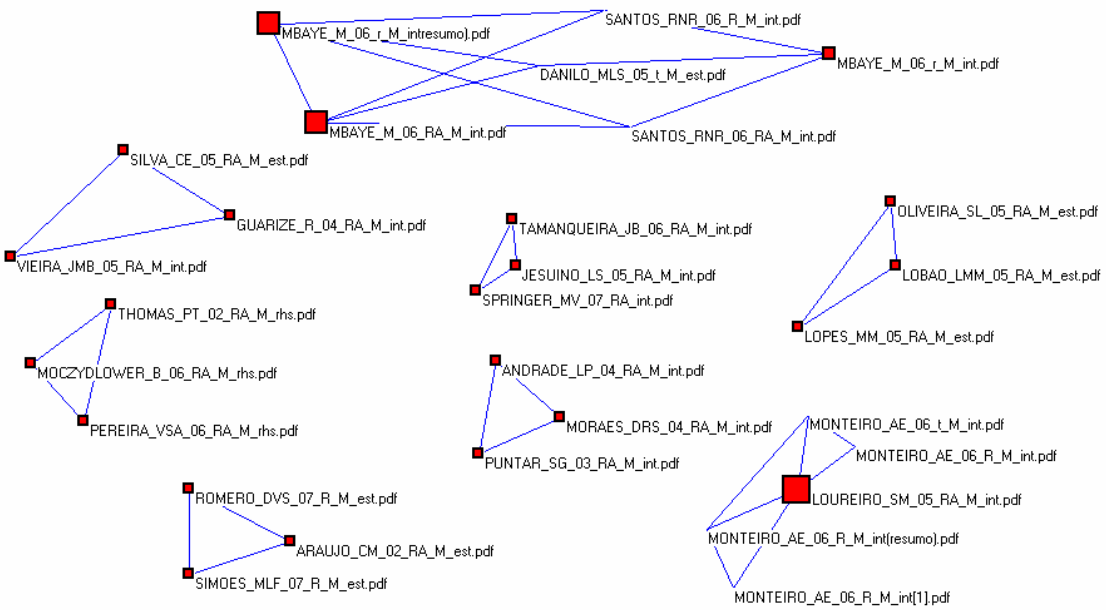

Figura 5.25 – Dissertação – Identificação dos Centros - Manhatan 0.3

5.4. NewsGroup

O conjunto de documentos NewsGroup contempla artigos de naturezas diversas como: eletrônica, política, religião, ..., foi montado por Tom Mitchell na escola de ciência da computação da Universidade de Carnegie Mellon. Está disponível no endereço http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html, para fins educacionais.

a. Pré-Processamento

As informações foram extraídas em formato txt e geradas 4.622 palavras para 972 documentos e aplicadas as medidas de similaridade Coseno e Manhatan.

b. Apuração das Medidas

Os resultados obtidos na apuração das medidas variando os limites dos pesos e similaridades podem ser conferidas na Tabela 5.7.

Medida de similaridade	Limite	Total documentos conectados		Total documentos desconectados		Total de ligações
Manhatan	0.1	943	97%	29	3%	19206
Coseno	0.1	940	97%	32	3%	16044
Coseno	0.3	604	62%	368	38%	2772
Manhatan	0.3	433	45%	539	55%	1984
Coseno	0.5	284	29%	688	71%	756
Manhatan	0.5	156	16%	816	84%	434
Coseno	0.7	89	9%	883	91%	166
Manhatan	0.7	47	5%	925	95%	74
Coseno	0.9	19	2%	953	98%	30
Manhatan	0.9	15	2%	957	98%	22

Tabela 5.7 – Distribuição de documentos - NewsGroup

O uso da similaridade Manhatan resultou em um maior número de documentos conectados e também em no total de ligações. Nos demais limites a medida de similaridade Coseno superou o nível de conexão e total de ligações.

A apuração das medidas no conjunto NewsGroup está apresentando o movimento de transitividade alto (Tabela 5.8), onde é observado que o ranking apresentado pela medida de similaridade Coseno com 0.9 de limite demonstra efetivamente que os poucos documentos conectados formam uma rede coesa. Pode-se observar ainda que o fato de ter um maior número de ligações não gera obrigatoriamente uma rede coesa, como demonstra a medida de similaridade Manhatan

para o limite 0.1, que possui o maior grau, mais não apresenta um índice de transitividade alto.

Nesta base de dados o menor coeficiente de aglomeração está relacionado a similaridade Manhatan (Tabela 5.8) com o limite 0.7, indicando uma rede com muitas presenças de pares de documentos.

Medida de similaridade	Limite	Grau de distribuição	Coeficiente de aglomeração	Grau Máximo
Coseno	0.9	0.030864	1	3
Manhatan	0.9	0.022634	0.75	3
Manhatan	0.3	2.041152	0.728287	39
Coseno	0.3	2.851852	0.624581	50
Coseno	0.1	16.50617	0.608315	97
Manhatan	0.1	19.75926	0.596311	100
Manhatan	0.5	0.446502	0.583046	20
Coseno	0.7	0.170782	0.559701	11
Coseno	0.5	0.777778	0.550546	26
Manhatan	0.7	0.076132	0.509434	5

Tabela 5.8 – Transitividade da rede de documentos- NewsGroup

c. Freqüência do Grau

Analisando o gráfico (Figura 5.26) para o conjunto NewsGroup é possível verificar uma topologia de rede Livre de Escala, pois a medida em que o grau aumenta a freqüência diminui.

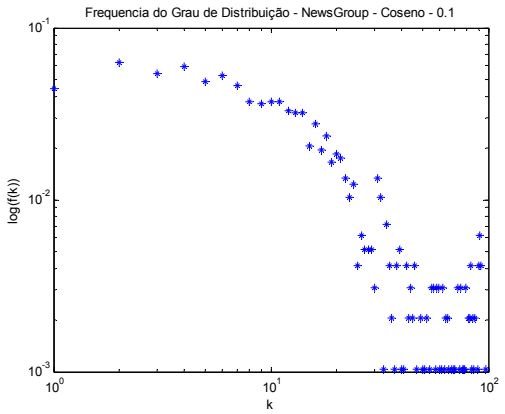

Figura 5.26 – NewsGroup Freqüência - Coseno 0.1

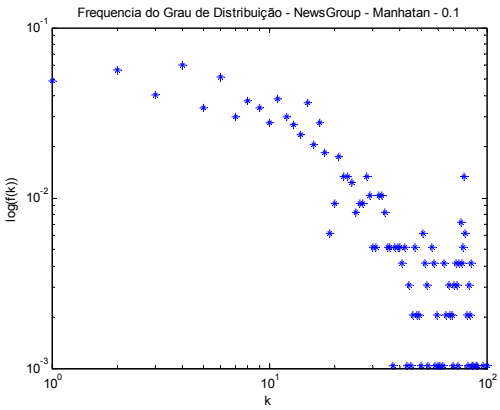

Figura 5.27 – NewsGroup Freqüência – Manhatan 0.1

d. Coeficiente de aglomeração x Grau de Distribuição

A Figura 5.28 e 5.29 apresentam a rede complexa do conjunto NewsGroup como uma rede que possui formação de grupos coesos.

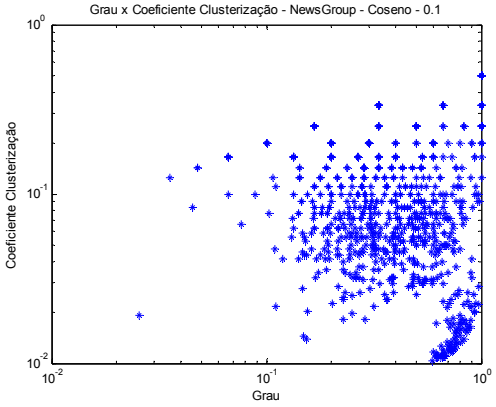

Figura 5.28– NewsGroup Coeficiente aglomeração - Coseno 0.1

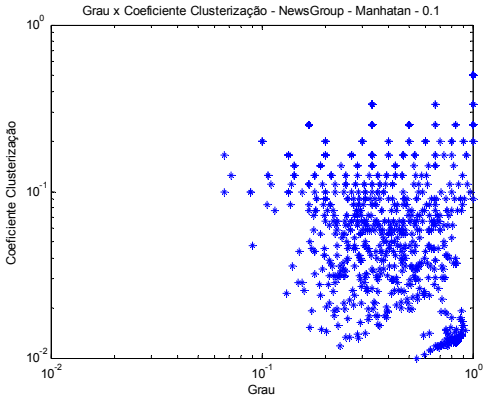

Figura 5.29 – NewsGroup Coeficiente aglomeração - Manhatan 0.1

e. Estruturas de Comunidade

Na rede complexa do conjunto NewsGroup limite de 0.3 da similaridade Coseno (Figura 5.30), identificamos a presença duas estruturas de comunidade. Os documentos 10-44 em destaque, representam a união do grupo menor com os grupos maiores. Para melhor visualização foram retirados os pares da visualização.

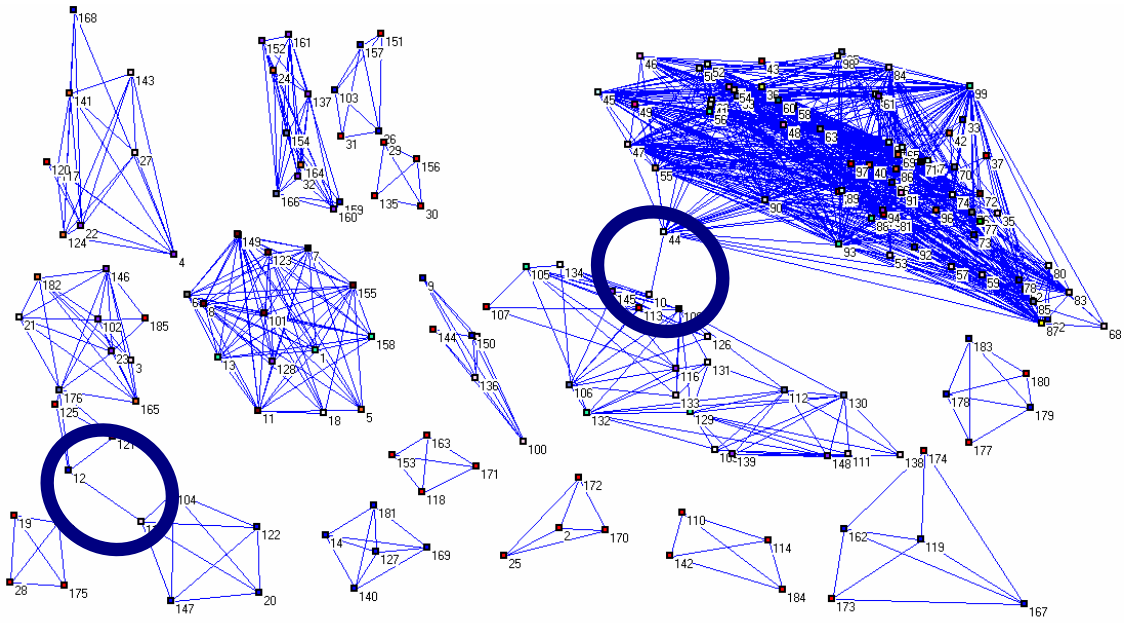

Figura 5.30 – NewsGroup – Estruturas de Comunidade - Coseno 0.3

Na similaridade Manhatan (Figura 5.31) a rede não possui estruturas de comunidade.

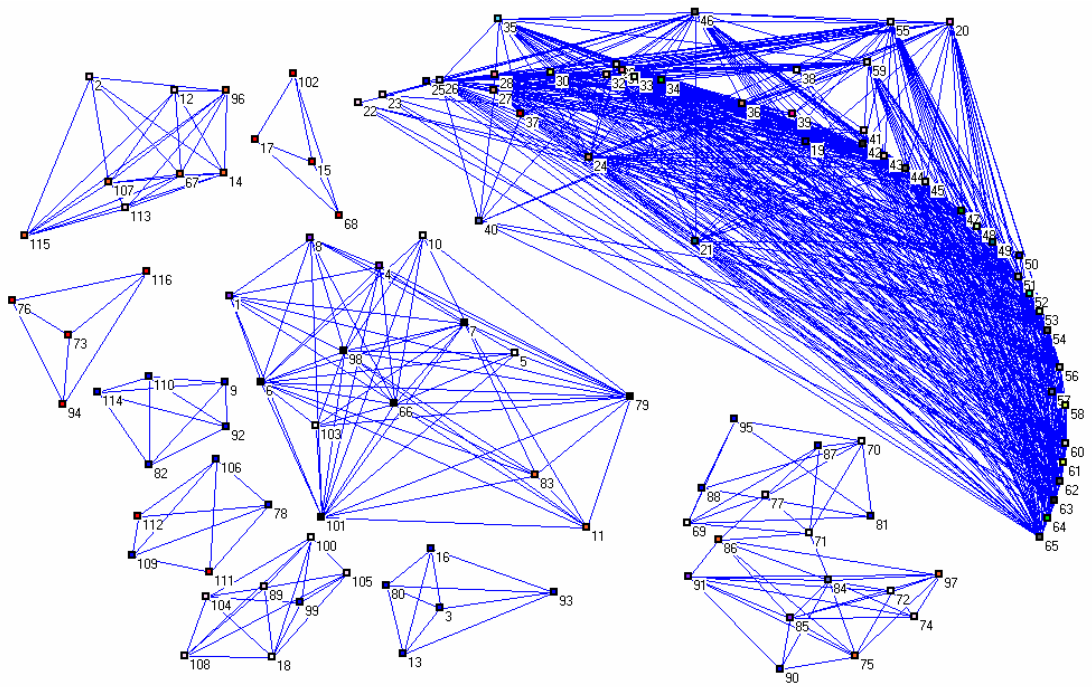

Figura 5.31 – NewsGroup - Estruturas de Comunidade - Manhatan 0.3

f. Identificação dos Centros

Os centros são identificados nos grafos pela figura geométrica quadrada (Figuras 5.32 e 5.33). O tamanho dos quadrados aumenta conforme o nível de centralidade do documento. As redes possuem centros diferentes.

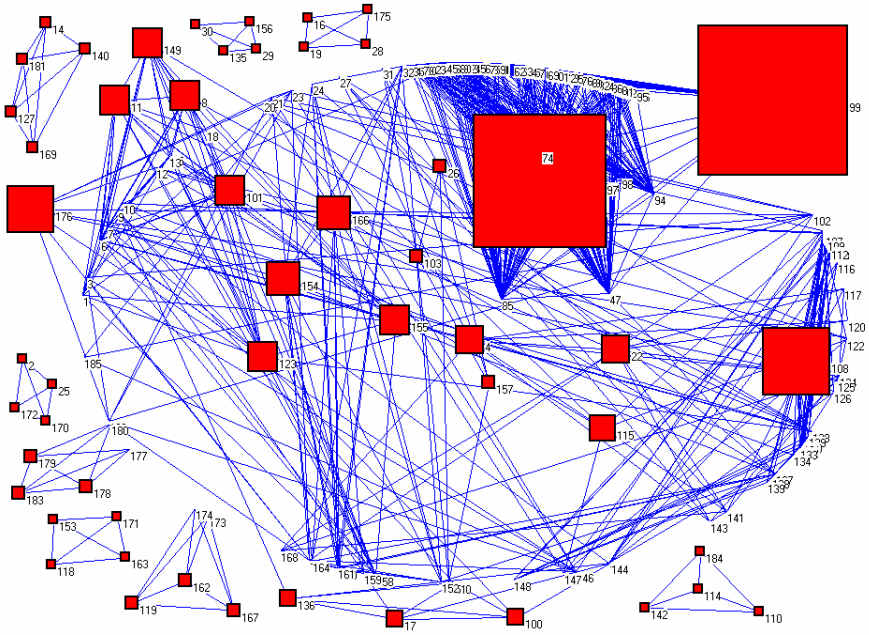

Figura 5.32 – NewsGroup - Centros - Coseno 0.3

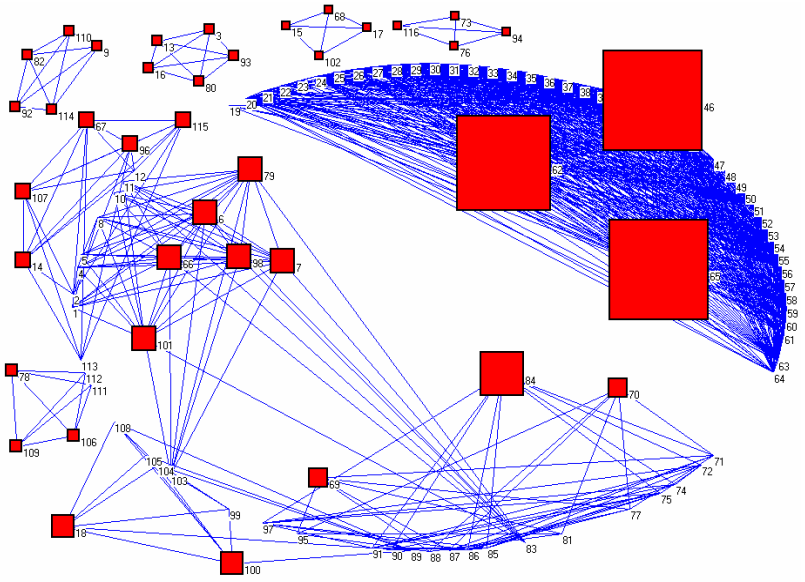

Figura 5.33 – NewsGroup - Centros - Manhatan 0.3

5.5. Discussão dos resultados

O trabalho utilizou quatro conjuntos de documentos de natureza e volumes diferentes, como pode ser verificado na Tabela 6.1. Com os resultados é possível concluir que: o número de ligações não determina a coesão da rede; o tratamento eficiente da informação é fundamental para transformação da informação não estruturada em estruturada; a transitividade deve ser verificada em função do coeficiente de aglomeração e coeficiente do grau de distribuição; as medidas de similaridade apresentaram poucas diferenças, predominando a similaridade Coseno com um índice de aproveitamento maior para quantidade de ligações nos conjuntos de maior volume (Tabela 6.1) e maior transitividade (Tabela 6.2).

Conjunto de documento	Similaridade	Total documentos	Total documentos conectados		Total documentos desconectados		Total de ligações
Newsgroup	Manhatan	972	943	97%	29	3%	19206
Newsgroup	Coseno	972	940	97%	32	3%	16044
Patentes	Coseno	769	755	98%	14	2%	20696
Patentes	Coseno	769	745	97%	24	3%	21184
Dissertações	Coseno	344	336	98%	8	2%	3464
Dissertações	Manhatan	344	335	97%	9	3%	4100
Deputados	Manhatan	59	55	93%	4	7%	296
Deputados	Coseno	59	37	63%	22	37%	76

Tabela 6.1 – Apuração conjuntos de documentos – limite 0.1

Conjunto de documento	Similaridade	Grau de distribuição	Coeficiente de aglomeração	Grau Máximo
Newsgroup	Coseno	16.5062	0.6083	97
Newsgroup	Manhatan	19.7593	0.5963	100
Patentes	Coseno	27.5475	0.4899	144
Patentes	Coseno	26.9129	0.4477	127
Dissertações	Coseno	10.0698	0.3612	36
Dissertações	Manhatan	11.9186	0.3095	39
Deputados	Manhatan	5.0169	0.3059	16
Deputados	Coseno	1.2881	0.1600	6

Tabela 6.2 – Transitividade conjuntos de documentos – limite 0.1

A metodologia proposta define uma sistemática para análise de todo conjunto de documento de qualquer natureza e tamanho.

Após execução da metodologia e análise dos resultados é verificado que todos os conjuntos apresentam topologia do modelo Livre de Escala; o conjunto de documentos "Patentes" apresenta-se mais coeso; "Deputados" tem muitas conexões, mas não tem transitividade; "Dissertação" embora tenha muitos documentos interligados formando grupos não apresenta índice de transitividade e, "NewsGroup" é um conjunto que tem muitas ligações e transitividade (Tabela 6.3).

	Modelo	**Transitividade**	**Estrutura de comunidade**
Newsgroup	Livre de Escala	Sim	Sim
Patentes	Livre de Escala	Sim	Sim
Deputados	Livre de Escala	Não	Não
Dissertações	Livre de Escala	Não	Não

Tabela 6.3 – Resultados conjuntos de documentos

6. CONCLUSÃO

Nesta dissertação foi apresentada uma metodologia de análise de redes complexas formadas a partir de conjunto de documentos com o objetivo de identificar a topologia, a centralidade, a transitividade e as estruturas de comunidade. A metodologia utiliza as técnicas de mineração de textos para tratamento das informações, normalização "*Tf-Idf*", cálculo de similaridades Coseno e Manhatan, métricas de análise de redes complexas e softwares de visualização. A análise é baseada nas medidas grau de distribuição e coeficiente de aglomeração a partir de gráficos de freqüência e coeficiente de aglomeração e representações geradas pelo software Pajek para visualização da centralidade e existência de estruturas de comunidade.

O tratamento da informação, etapa inicial de todo processo apresentado neste trabalho, é determinante na extração das informações e na garantia do sucesso da aplicação nas demais etapas. A medida de similaridade Coseno apresenta uma maior coesão entre os documentos em comparação a medida Manhatan.

Os resultados obtidos confirmam a utilização de medidas de rede complexa para análise e conhecimento do conjunto de documentos.

Como trabalho futuro pode-se sugerir: Desenvolvimento em processamento paralelo do cálculo da distância geodésica utilizando o algoritmo de *Dijkstra*. A medida distância geodésica é importante para determinar a vulnerabilidade e a eficiência global da rede; Desenvolvimento de procedimento que identifique os documentos que estabelecem a ligação entre os grupos.

Desta forma é interessante desenvolver processos para identificar automaticamente nas redes complexas características determinantes ao tratamento e manipulação do conjunto de documentos que a compõe.

7. BIBLIOGRAFIA

[1] FREEMAN, Richard T., YIN Hujun. “Web Content Management by Self-Organization”. IEEE transactions on neural networks, vol. 16, no. 5, September 2005.

[2] KOPRINSKA Irena, POON Josiah, CLARK James, CHAN Jason. “Learning to classify e-mail”. Information Sciences 177 (2007) 2167–2187.

[3] XIAOFENG He, HONGYUAN Zh, DINGB Chris H.Q., SIMONB Horst D.. “Web document clustering using hyperlink structures”. Computational Statistics & Data Analysis 41. 2002.

[4] DIAS, Leila Christina; SILVEIRA, Rogério Leandro Lima da. (orgs). Redes, sociedades e territórios. Santa Cruz do Sul: EDUNISC. [Capítulos 1 e 2, p.11-73]. [PGG 303.4833 R314]. 2005.

[5] WEISS, SHOLOM M.; INDURKHYA , Nitin; ZHANG , Tong; DAMERAU, Fred J.. “Text Mining – Preditive Methods for Analyzing Unstructered Information”. [Capítulos 1 e 2, p.1-84]. ISBN 0-387-95433-3. Springer Science+Business Media, Inc. 2005

[6] YU Jie, AMORES Jaume, SEBE Nicu, TIAN Qi.. “A new study on distance metrics as similarity measurement”.2006.

[7] COSTA, L. da F., RODRIGUES F. A., TRAVIESO, G. e VILLAS BOAS, P.R.., “Characterization of Complex Networks: A survey of measurements”. Instituto de Física de São Carlos, Universidade de São Paulo. June 30, 2005.

[8] NEWMAN, M. E. J.. “The Structure and Function of complex networks”. University of Michigan. Department of Physics, University of Michigan, Ann Arbor, MI 48109, U.S.A. and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, U.S.A.

[9] KONCHADY, Manu, “Text Mining Application Programming / Manu Konchady”. ISBN: 1-58450-460-9. Charles River Media, Boston, Massachusetts. 2006.

[10] NEWMAN, M. E. J. and WATTS, D. J.. “Renormalization group analysis of the small-world network model”. Phys. Lett. A 263, 341-346. 1999.

[11] AVRAMENKO Yuri, KRASLAWSKI Andrzej, “Similarity concept for case-based design in process engineering”, Computers and Chemical Engineering. Science Direct. 2005.

[12] BALDI Pierre, FRASCONI Paolo, SMYTH Padhraic, “Modeling the Internet and the Web”, ISBN 0-470-84906-1. Wiley. 2003.

[13] PASTOR-SATORRAS R., RUBI M., DIAZ-GUILERA A., “Statistical Mechanics of Complex Networks”. Springer. ISBN 3-540-40372-8. March 2003.

[14] ALBERT Réka, BARABÁSI Albert-László, “Statistical Mechanics of Complex Networks”, arXiv:cond-mat/0106096v1. Jun 2001.

[15] NEWMAN M. E. J.. “Modularity and community structure in networks”. Department of Physics and Center for the Study of Complex Systems, Randall Laboratory, University of Michigan, Ann Arbor, MI 48109–1040. 2006.

[16] ANTIQUEIRA L., NUNES M.G.V., OLIVEIRA O.N. Jr., COSTA L. da F., “Strong correlations between text quality and complex networks features”, Science Direct. Physica A 373 (2007) 811–820.

[17] SAKATA Shuzo, YAMAMORI Tetsuo, “Topological relationships between brain and social networks”. Science Direct. Neural Networks 20 (2007) 12–21.

[18] GUO Qiang, ZHOU Tao, LIU Jian-Guo, BAI Wen-Jie, WANG Bing-Hong, ZHAO Ming, “Growing scale-free small-world networks with tunable assortative coefficient”. Science Direct. Physica A 371 (2006) 814–822.

[19] Dragos Arotaritei, Sushmita Mitra. “Web mining: a survey in the fuzzy framework. Fuzzy Sets and Systems”. 148, pp. 5-19. 2004.

[20] Christian Gieger;Hartwig Deneke and Juliane Fluck. “The future of text mining in genome-based clinical research. Biosilico”. Vol. I, No.3. July. 2003.

[21] Byungun Yoon, Yongtae Park. “A text-mining-based patent network: Analytical tool for high-tecnology trend”. The journal of high technology management research, 15, 37-40. 2004.

[22] Graziella Martins Caputo, “Sistema computacional para o processamento textual de patentes industriais”. Tese de mestrado. Abril/2006.

[23] Santos Lopes, Maria Célia. Mineração de Dados textuais utilizando técnicas de clustering para o idioma português. Tese de Doutorado, UFRJ. Setembro, 2004.

[24] Michael W. Berry, Zlatko Drmac, Elizabeth R. Jessup, “Matrices, Vector Spaces, and Information Retrieval”; Siam Review, Vol.41, No. 2, pp. 335-362, 1999.