Guilherme Alberto Wachs Lopes
Orientador: Paulo Sérgio Rodrigues
Co-Orientador: Plinio Thomaz Aquino Junior

Um Modelo de Rede Complexa para Análise de Informações Textuais

Dissertação apresentada ao Curso de Mestrado em Inteligência Artificial Aplicada à Automação Industrial do Centro Universitário da FEI, como requisito parcial para a obtenção do grau de Mestre em Engenharia Elétrica.

São Bernardo do Campo, SP
Agosto de 2011

Resumo

Análise de textos é uma tarefa inerentemente humana que envolve processos cognitivos complexos e difíceis de modelar em sistemas computacionais atuais. Esses processos, geralmente paralelos, levam em conta usualmente tanto informações léxicas quanto sintáticas, com o objetivo de situar o texto em um nível hierárquico e semântico adequado. Informações no nível léxico estão mais relacionadas com as regras de uma linguagem para geração de palavras, enquanto o nível sintático está geralmente relacionado ao posicionamento das palavras no texto. O conjunto dessas informações (léxica e sintática) leva à geração das informações semânticas. Diversas áreas de aplicações que envolvem análise automática de textos devem considerar essas informações a fim de atingir uma gama crescente de objetivos, tais como: recuperação de documentos, comparação de textos, geração automática de diálogos, geração de rótulos, indexação de textos, entre outras.

Embora as regras de interpretação de textos sejam conhecidas há bastante tempo, devido a fatores que envolvem principalmente tempo computacional e alta dimensionalidade dos modelos, muitas dessas regras não são levadas em conta em sistemas práticos atuais. Por exemplo, a maioria dos sistemas de recuperação de informações textuais geralmente considera somente a frequência com que as palavras aparecem em um texto, ou o número de links que apontam para uma mesma página de internet, com o objetivo de ordenar documentos por relevância, quando de uma requisição do usuário. Sabe-se, no entanto, que informações léxicas contidas nas *stop-words*, palavras com erros e pontuação, bem como informações sintáticas, como a ordem que as palavras aparecem em um texto, não são geralmente consideradas nesses modelos, motivo que pode levar ao chamado *gap-semântico* entre a requisição do usuário e as informações realmente fornecidas pelo modelo de recuperação. Por outro lado, desde o início da década de 90, estudos em redes complexas vêm ganhando mais e mais atenção dos pesquisadores, sobretudo do ponto de vista da Teoria da Informação Não-Extensiva, devido tratar-se de um problema onde as interações, tanto temporais quanto espaciais das palavras de diversos contextos serem de longo alcance.

Assim, o presente trabalho apresenta um modelo de Redes Complexas que leva em conta não somente as informações de frequência, mas também a ordem das palavras, co-ocorrência das mesmas, *stop-words* e palavras erradas. O preço a pagar para este modelo é a utilização do espaço de armazenamento da ordem de Giga-Bytes, o que o torna inviável para ser tratado em computadores comuns. Modelos dessa grandeza ainda não foram completamente estudados e apresentam comportamentos ainda difíceis de se prever e discutir.

As características das redes complexas estudadas há mais de uma década na literatura

(por exemplo: tipo de rede, coeficiente de clusterização, distribuição de graus, distribuição de pesos, matriz de distâncias, raio, diâmetro, coeficiente espectral, entre outros) permitem o estudo desses modelos para grandes bases de dados. Assim, neste trabalho, propomos o estudo desse modelo aplicado ao contexto da área médica sob o ponto de vista dessas características.

Estudos preliminares mostram que palavras retiradas de um contexto médico, considerando as características léxicas e sintáticas citadas acima, apresentam um comportamento de rede livre de escala. Também apresentamos heurísticas para o cálculo de grandezas físicas muito caras do ponto de vista da ordem computacional ($O(n^3)$), como o coeficiente de clusterização (CC) da rede. Resultados sugerem que é possível o cálculo do CC com erro em torno de 5% para redes densas ou esparsas de até 10.000 palavras.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Em apenas 20 anos de existência, a internet provou ser uma das maiores e mais impressionantes construções humanas de todos os tempos. O fato mais curioso a seu respeito é que ela parece ter uma demanda tão grande quanto totalmente imprevista por quaisquer especialista em tecnologia de épocas passadas.

O impacto do seu surgimento foi tão marcante que forçou o aparecimento quase que simultâneo de diversas áreas de pesquisa e o pleno desenvolvimento de outras, que até então eram apenas especulações teóricas, tais como: recuperação de informação, mineração de dados, e-commerce, hipermídia, esterografia, para citar algumas.

Em meados dos anos 90, as informações tratadas nessas áreas eram quase que exclusivamente textuais. A medida que a tecnologia de hardware foi se desenvolvendo, elementos multimídia como imagem, som e vídeo, foram incorporados às informações tratadas, demandando, na mesma proporção, novos métodos de gerenciamento de dados.

Embora a demanda por pesquisas de novas metodologias e algoritmos para gerenciamento de dados multimídia cresça exponencialmente na mesma velocidade do número de pessoas conectadas mundialmente, dados em formato exclusivamente textuais ainda exigem estudos mais aprofundados, mesmo porque, as aplicações que baseiam-se no conteúdo de textos transcendem a internet e a área de recuperação de informação de documentos escritos.

Como qualquer tipo de dado multimídia, a análise do conteúdo de um texto envolve vários níveis de abstração; os mais conhecidos são: o léxico, o sintático e o semântico [Baeza-Yates and Ribeiro-Neto, 1999]. A habilidade de um sistema de análise de identificar e recuperar informações, em qualquer um desses três níveis, está diretamente relacionada a dois pilares de estudo principais: o modelo de dados e os algoritmos de gerenciamento desses modelos.

O modelo de representação de dados textuais é uma frente de estudo que deve ser muito bem elaborada, uma vez que seu objetivo é representar textos de maneira formal para posterior gerenciamento. Isso significa que o modelo matemático deve ser preciso de tal forma que não perca conteúdo de informações e, ao mesmo tempo, não consuma muitos recursos computacionais. O que se deseja com um modelo computacional é uma transformação de domínio para que o gerenciamento das informações relacionadas, através dos algoritmos, seja o mais eficaz possível.

Há diversas características textuais que podem ser utilizadas para representação de dados dessa natureza, tais como: frequência de palavras, links, palavras-chave, uso de ontologias, uso de thesaurus, relacionamento entre palavras, entre outras [Baeza-Yates and Ribeiro-Neto, 1999]. No entanto, até o início do século XXI, a maioria dos modelos só poderia contemplá-las individualmente, uma vez que a tecnologia de hardware existente na época era insuficiente para tamanha massa de dados. Isso forçava os desenvolvedores a relaxar os modelos de representação e, consequentemente, algumas informações textuais deveriam ser desconsideradas. Um exemplo típico é a eliminação de *stop-words*, tais como preposições, conjunções e artigos, com o objetivo de diminuir o espaço de busca. O preço a pagar era a perda de informação semântica.

A partir dos anos 90, surgiram novas teorias para o tratamento de informações. Inicialmente, a teoria dos grafos aleatórios, introduzida por Erdös and Rényi em 1959 [Erdős and Rényi, 1959], baseava-se em relações ao acaso. Porém, estudos posteriores mostraram que esses modelos apresentavam padrões de crescimento característicos e poderiam ser previsíveis, surgindo então o conceito de Redes Complexas [Newman et al., 2006].

As Redes Complexas são utilizadas em diversos estudos onde deseja-se a compreensão

de como as informações se inter-relacionam e como se desenvolvem. Por exemplo, a rede de interações sociais, de amizades, bem como a rede de relações sexuais podem ser modeladas através de Redes Complexas.

Outra área que experimentou grande desenvolvimento nessa mesma época é a Teoria de Informação. A utilização de entropia como medida de informação, proposta por Claude Shannon em 1948 [Shannon and Weaver, 1948], foi um passo que marcou o início dessa área. Problemas como quantidade de informação que um canal pode transmitir, quantidade de ruído em um sinal, entre outros, puderam ser mensurados. Isso chamou a atenção de muitos pesquisadores que aplicaram a Teoria da Informação em diversos sistemas.

Em meados da década de 80, Constantino Tsallis propôs uma generalização da entropia tradicional de Shannon [Tsallis, 1988] e adicionou um parâmetro de ajuste, abrindo portas para a aplicação da teoria da informação para sistemas que não eram possíveis de serem estudados até então; os chamados sistemas não-extensivos.

O presente projeto propõem uma nova modelagem para análise de informações textuais baseada na co-ocorrência de palavras. A modelagem consta da construção de uma rede complexa ponderada e dirigida, onde os nós são as palavras e as arestas são as co-ocorrências entre elas. Além disso, as arestas são ponderadas e direcionadas, modelando tanto o grau de co-ocorrência quanto a ordem em que aparecem no texto. Esperamos que essa modelagem consiga extrair mais informações textuais que os primeiros modelos conhecidos, uma vez que consideramos diversas características até então com pouca atenção recebida, tais como: ordem das palavras, co-ocorrência de palavras para inferir contexto semântico, *stop-words*, pontuação e acentuação, bem como escrita errada de palavras. Após sua construção, utilizamos heurísticas em conjunto com a teoria da informação para efetuar medidas físicas nessa rede, tais como: coeficiente de clusterização, distribuição de pesos, distribuição de graus, coeficiente espectral, diâmetro, raio, entre outras. De posse dessas medidas, pode-se validar o modelo proposto para aplicação em diversos sistemas práticos que envolvem análise de informação puramente textual.

1.1 Objetivo

Estudar e implementar um modelo que represente o conteúdo de informação de textos que leve em conta: frequência, posição relativa e grau de co-ocorrência de palavras, incluindo *stop-words* e grafias com erros. O modelo proposto baseia-se na teoria das Redes Complexas e Teoria da Informação para análise e proposição de heurísticas para cálculo eficiente de características físicas da rede.

1.2 Principais Contribuições da Dissertação

- Construção e armazenamento de uma rede complexa da ordem de gigabytes.
- Modelagem de contexto baseada na co-ocorrência, palavras com erros, *stop-words* e ordem de ocorrência no texto.
- Elaboração de heurísticas para extração de características físicas de uma rede complexa.
- Utilização da teoria da informação não-extensiva em conjunto com a teoria de redes complexas para análise de características físicas da rede.

Os resultados desse trabalho podem ser utilizados para propor novos modelos de representação de dados textuais, imagens, sons ou até mesmo vídeo. Algumas das aplicações que podem ser beneficiar com o uso dessa modelagem podem ser: sistemas de recuperação de informação, sistemas de correção textual baseada em contexto, sistemas para estudo linguísticos, entre outras.

Capítulo 2

Conceitos Fundamentais

2.1 Definições e Terminologias

Alguns conceitos abordados nesse trabalho são amplamente utilizados na literatura. No entanto, não há uma definição formal para cada um deles, uma vez que podem ser interpretados de diferentes formas de acordo com a aplicação. Nesta seção, apresentamos algumas definições e terminologias de conceitos que serão amplamente utilizados no decorrer do trabalho.

Palavra: Palavra é qualquer cadeia de caracteres que termina com o caracter " "(Espaço) ou caracter de nova linha.[1]

Documento (D_k): Um documento $D_k, k \in \mathbb{N}$ é um vetor de palavras ordenadas na mesma sequência que aparecem no texto.

Documento Relevante: Conceito utilizado na área de Recuperação de Informação. Um Documento Relevante é um documento que satisfaz uma requisição (pesquisa) do usuário.

Base de Documentos: É o conjunto de documentos analisados. $D = D_k, k \in \mathbb{N}$.

[1] A pontuação é considerada como parte integrante da palavra. Assim, a palavra "não" é diferente da palavra "não,".

Matriz de Adjacência (M): A Matriz de Adjacência é um modelo utilizado para descrever as conexões de um grafo. Neste trabalho, utilizamos a matriz de adjacência para descrever uma Rede Complexa de N Palavras. Essa matriz contém N^2 células representadas na notação $M_{i,j}$, onde i é o índice da linha e j é o índice da coluna. Conforme será explicado na Seção 2.3.1, neste trabalho os índices referem-se às palavras do arquivo invertido.

Distância entre palavras ($Dist(D_k, i, j)$): A função de Distância entre palavras em um documento D_k mede a distância de uma palavra j em relação a palavra i. A unidade de medida desta distância é quantidade de palavras entre i e j.

Tamanho do documento ($Size(D_k)$): Essa função recebe um documento como entrada e retorna a quantidade de palavras do mesmo.

Ocorrência de palavra : Diz-se que uma palavra i ocorre em um documento D_k quando i aparece nesse documento.

Co-ocorrência de palavras : A co-ocorrência entre duas palavras i e j é um evento que acontece quando i e j aparecem em um mesmo documento.

Intensidade ou Peso de co-ocorrência ($w(i, j)$): Quando duas palavras i e j co-ocorrem em um documento, o peso $w(i, j)$ é calculado com o objetivo de medir a força (ou probabilidade) dessa co-ocorrência. Matematicamente, a intensidade de peso é definida pela Equação (2.1):

$$w(i,j) = 1 - \frac{\sum_k Dist(D_k, i, j)}{\sum_k Size(D_k)} \tag{2.1}$$

Note que $w_{i,j}$ é um valor real entre 0 e 1, uma vez que é ponderada pela quantidade de palavras do documento ($Size(D_k)$) e k é o índice do documento. Assumimos que, no caso particular quando $Size(D_k) = 0$, $w(i, j)$ é igual a 0.

Clusterização: Neste trabalho, consideramos clusterização como um método de aprendizagem não-supervisionada, que tem por objetivo agrupar elementos semelhantes.

Cluster: Definimos cluster como um conjunto de palavras que co-ocorrem na base de documentos. Um cluster pode possuir sub-clusters e dois clusters podem possuir sobreposição de palavras.

Contexto semântico: Contexto semântico é o conjunto de palavras que possuem alta co-ocorrência entre si. Pode-se concluir que um cluster de palavras pode abranger um contexto.

Aprendizagem Supervisionada: A aprendizagem supervisionada é um tipo de classificação de dados no qual o treinamento do sistema é baseado em um conjunto de dados previamente classificado. Após o aprendizado, novos padrões podem ser classificados [Duda et al., 2000].

Aprendizagem Não-Supervisionada: Na aprendizagem não-supervisionada, a classificação dos dados acontece de forma natural e sem prévio treinamento [Duda et al., 2000].

Aprendizagem por reforço: Na aprendizagem por reforço, são apresentados alguns padrões e o sistema computa uma possível resposta. Porém, o sistema recebe bonificação se a classificação foi feita de maneira correta e uma punição, caso contrário [Duda et al., 2000].

Stop-Words São palavras que ocorrem com muita frequência em uma base de documentos. Artigos, conjunções, preposições e alguns advérbios são fortes candidatos a serem *Stop-Words* [Baeza-Yates and Ribeiro-Neto, 1999].

2.2 Redes Complexas

As Redes Complexas são utilizadas para descrever os mais diversos tipos de sistemas [Newman et al., 2006]. Exemplos de redes que só recentemente foram possíveis de modelar e estudar são: redes sociais (tais como orkut, facebook, comunidades sociais em geral) e redes biológicas (cadeias de DNA, Genoma). Essas redes, até pouco tempo, eram desconhecidas ou difíceis de representá-las matematicamente, ou por falta de um modelo

matemático adequado, ou devido à ausência de hardware com capacidade compatível para sua implementação.

A principal característica das Redes Complexas é que as ligações entre seus elementos baseiam-se em regras[2] previamente estabelecidas (Fig. 2.1) [Newman et al., 2006]. Porém, seu crescimento acontece de uma forma natural e, consequentemente, imprevisível. Assim sendo, definindo-se a regra das ligações, teremos ao final uma rede composta de muitas interações, tornando possível a extração de características físicas importantes para o entendimento do comportamento do sistema, tais como: tamanho, diâmetro, grau de conexão médio, densidade, entre outras.

Figura 2.1: Rede complexa representando todos os produtos comercializáveis [Hidalgo et al., 2007]. Com esse tipo de modelagem é possível tomar decisões estratégicas locais da economia emergente. O tamanho de cada nó representa a grandeza (em dólares) dos produtos e as ligações representam a utilização de um produto para produzir outro.

A modelagem de uma rede complexa pode ser semelhante a de um grafo [Newman et al., 2006]. Há algumas formas de se modelar um grafo [Cormen et al., 2001]. Uma delas é através da

[2]Essas regras são definidas pelo próprio problema proposto. Por exemplo, no caso de rede social, cada nó representa um indivíduo e cada aresta representa amizade entre dois indivíduos.

utilização de lista de adjacência de arestas. Essa lista é representada por uma tabela de 3 colunas, onde cada linha define uma aresta do grafo. A primeira coluna informa qual é o nó de saída da aresta, a segunda informa qual é o nó de destino e a última informa o peso dessa aresta.

Outro tipo de modelagem, o mais utilizado, cria uma matriz de adjacência (ver Seção 2.1 em *Matriz de Adjacência*). Essa matriz, de dimensão N^2, descreve toda a rede. Cada célula representa uma aresta, podendo assumir os valores 0 ou 1, para redes não ponderadas ou outros valores para redes ponderadas. Além disso, se a rede for não-dirigida essa matriz contém somente $N^2/2$ células, dado que somente a parte superior da diagonal principal pode ser utilizada.

A escolha por qual modelagem deve ser feita está diretamente relacionada à densidade do grafo. Para redes esparsas, a modelagem por lista de adjacência é melhor, uma vez que o consumo de memória é $O(E)$. Para redes densas, a matriz de adjacência é melhor, pois seu consumo é $O(N^2)$.

Na Seção 2.2.1 serão apresentados os três modelos de redes complexas mais estudados na literatura, e na Seção 2.2.2 veremos algumas de suas aplicações.

2.2.1 Modelos de Redes

Nesta seção serão abordados os três tipos de modelos teóricos de Redes Complexas. A primeira delas, Redes Aleatórias, é o modelo mais antigo que possui suas conexões baseadas em uma propriedade conhecida por probabilidade de conexão. O segundo modelo, Redes de Mundo Pequeno, é usado em muitos trabalhos para modelagem de redes sociais e, finalmente, Redes Livres de Escala que estuda a dinâmica do crescimento das redes.

2.2.1.1 Redes Aleatórias

A teoria das Redes Aleatórias foi desenvolvida inicialmente por Solomonoff e Rapoport [Solomonoff and Rapoport, 1951] e posteriormente estudada por [Erdős and Rényi, 1959]. Esse tipo de rede é construída conectando aleatoriamente seus nós em uma proporção

conhecida. Há duas representações matemáticas para este tipo de rede. A primeira delas, $G_{n,p}$, onde n é o número de nós e p é a probabilidade de conexão entre dois quaisquer nós. A segunda representação, $G_{n,m}$, onde n é o número de nós e m é o número de arestas. Essas duas representações matemáticas têm interpretações diferentes, uma vez que em $G_{n,p}$ o número de arestas é alterado proporcionalmente a n e, em $G_{n,m}$, o número de arestas é fixo.

Estudos atuais de Redes Aleatórias encontram algumas características físicas fundamentais para o estudo de seu crescimento, tais como: existência de *giant component*, fase de transição, *small components*, entre outros [Stauffer and Aharony, 1992].

As Redes Aleatórias conseguem modelar algumas redes do mundo real. Porém, alguns fatores podem interferir nos resultados uma vez que, muitas vezes, a distribuição de "graus" dos nós em redes reais tende a ser exponencial, *power-law* ou, até mesmo, distribuição com picos. Isso altera o comportamento da rede e pode trazer resultados imprecisos [Albert et al., 2000, Cohen et al., 2000].

Como será visto mais adiante, as redes aleatórias servem como parâmetro de medição para comparação entre os modelos. A Equação (2.2) fornece uma medida de quanto uma rede tem características de rede aleatória para as medidas de coeficiente de clusterização e diâmetro. Essa equação resulta em 1 para redes aleatórias e um número muito maior que 1 para redes de mundo pequeno [Walsh, 1999].

$$\mu = \frac{C/C_{rg}}{\ell/\ell_{rg}} \tag{2.2}$$

Na Equação (2.2), C e ℓ são respectivamente o coeficiente de clusterização médio e a distância média entre todos os nós da rede estudada, C_{rg} e ℓ_{rg} são respectivamente o coeficiente de clusterização médio e distância média entre todos os nós estimados para uma rede randômica com o mesmo número de nós e arestas (ver também Seção 2.2.3.2 para coeficiente de clusterização).

2.2.1.2 Redes de Mundo Pequeno

Com o objetivo de modelar redes sociais para o estudo da proliferação de doenças, internet, redes metabólicas, entre outras, foi criado um modelo de rede complexa que utiliza uma variedade de técnicas da física estatística [Watts and Strogatz, 1998]. Duas principais características observadas em redes do mundo real levaram à criação das redes de mundo pequeno. A primeira delas é que a média das distâncias entre os nós da rede cresce logaritmicamente de acordo com o número total de nós. Isso significa que a medida que a rede cresce, suas distâncias crescem mais lentamente. Para fazer a medição dessa característica nas redes, deve-se calcular a média aritmética das distâncias entre todos os nós da rede. A segunda característica é que redes de mundo pequeno possuem alto Coeficiente de Clusterização médio (Seção 2.2.3.2). Para isso acontecer, a vizinhança de um nó deve ser altamente conectada entre si.

O cálculo do coeficiente de clusterização [Newman et al., 2006] de um nó é feito a partir da relação entre a quantidade de conexões existentes entre os vizinhos desse nó e a quantidade máxima possível. Considere a rede apresentada pela Fig. 2.2. Podemos calcular o coeficiente de clusterização do nó *"A"* da seguinte forma:

$$CC_A = \frac{\textit{(número de vizinhos conectados entre si)}}{|\,\textit{Vizinhança}(A)| \times (|\,\textit{Vizinhança}(A)| - 1)} \tag{2.3}$$

Onde $\textit{Vizinhança}(A)$ é o conjunto de nós que se conectam a A.

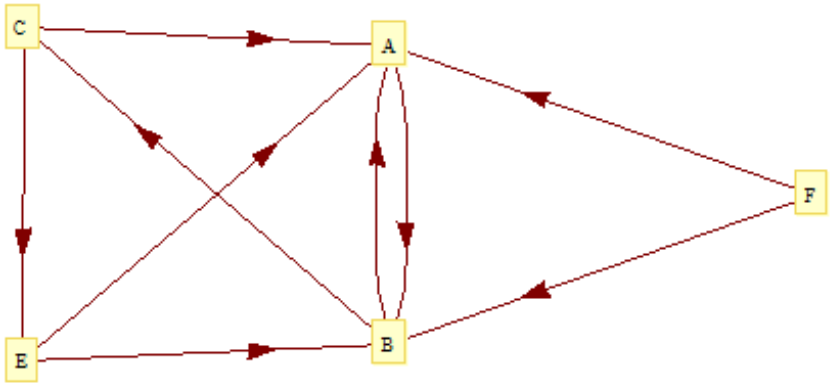


Figura 2.2: Rede com 5 nós

As redes de mundo pequeno assumiram uma importância fundamental para o estudo teórico e prático das redes complexas, uma vez que permitiram o estudo do comportamento de alguns sistemas naturais e a possibilidade de observar o comportamento da distribuição de graus, fundamental para o surgimento das Redes Livres de Escala, abordadas na Seção 2.2.1.3 [Newman et al., 2006].

2.2.1.3 Redes Livres de Escala

As redes livres de escala surgiram a partir da observação da distribuição de graus de alguns modelos estudados nos trabalhos de [Price, 1965, Albert et al., 1999, Faloutsos et al., 1999, Broder et al., 2000][]. Algumas contribuições científicas mostraram que a lei de potência é comum em redes do mundo real, tais como: Web [Albert et al., 1999], atores de filmes [Watts and Strogatz, 1998] e redes de citações científicas [Redner, 1998].

A lei de potência, descrita pela Equação (2.4), contém um parâmetro de ajuste γ e calcula qual a probabilidade de um nó ter grau k.

$$P(k) \sim k^{-\gamma} \tag{2.4}$$

As redes livres de escala também apresentam o efeito de mundo pequeno. Sendo as-

sim, as distâncias entre os nós da rede tendem a crescer logaritmicamente em função da quantidade de nós.

As redes livres de escala são fundamentadas na dinâmica do crescimento de redes naturais; isto é, diferentemente das redes aleatórias, onde o número de nós é fixo e as arestas são inseridas aleatoriamente, as redes livres de escala crescem de acordo com o conceito de *ligação preferencial* a cada inserção de um novo nó. A *ligação preferencial* é uma característica que rege a forma em que novas arestas são inseridas na rede. Mais especificamente, quando um novo novo nó é adicionado, a probabilidade desse nó ligar-se com outro nó de grau elevado é proporcionalmente maior do que ligar-se com um nó de baixo grau [Albert and Barabási, 2002]. A Fig. 2.3 ilustra a evolução de uma rede livre de escala quando novos nós (verde) são inseridos na rede.

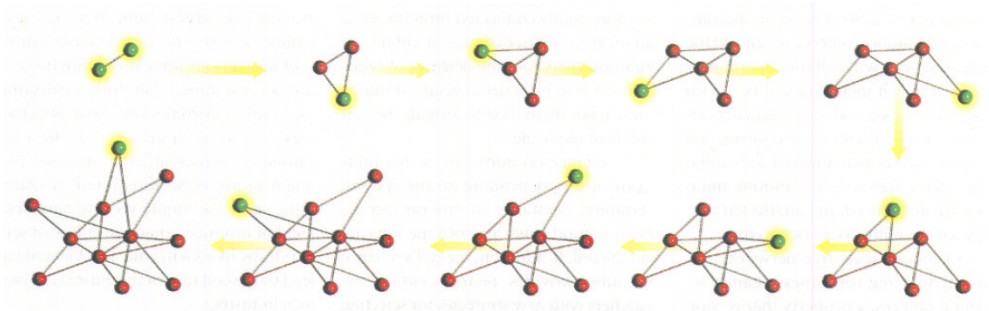


Figura 2.3: Evolução de uma rede livre de escala [Barabasi and Bonabeau, 2003]

2.2.2 Exemplos e Aplicações de Redes Complexas

Todo o ferramental desenvolvido para as Redes Complexas permitiu sua aplicação em diversas áreas. Esta seção abordará algumas delas.

2.2.2.1 Redes Sociais

Redes sociais são construídas para estudos de relacionamentos interpessoais. Nestas redes, os nós representam pessoas e as arestas suas relações. Essas relações podem modelar amizades, doenças, casamentos entre famílias, comunidades de negócio, colaboração

no trabalho, contatos telefônicos, comunicação por e-mail e até mesmo relações sexuais [Liljeros et al., 2001].

2.2.2.2 Redes de Informação

As redes de informação são construídas a partir de bases de conhecimento formal. Nestas redes, os nós representam informações e as arestas a relação entre essas informações.

Em [Redner, 1998], foi estudado um modelo de redes complexas para representar as citações entre artigos acadêmicos. As características dessa rede permitiram entender a dependência entre a distribuição de graus dos artigos, que é descrita por uma lei de potência, e o rank de classificação de acordo com o número de citações.

2.2.2.3 Redes Tecnológicas

Redes tecnológicas são redes complexas utilizadas para modelar a distribuição de facilidade ou recursos, tais como: água, malha elétrica, transporte, linhas aéreas, telefonia (fixa), internet (cabeamento e roteadores), entre outros.

Em [Gov, 2000], foi proposto um modelo de redes complexas para elaboração de heurísticas capazes de aumentar a fidelidade dos roteadores. O mesmo estudo faz uma análise sobre o mapa físico da internet.

2.2.2.4 Redes Biológicas

Alguns sistemas biológicos tais como vascular, nervoso, circulatório, entre outros, são naturalmente identificados como redes complexas. Um exemplo da aplicação dessas redes é o trabalho de [Sporns, 2002], no qual foi analisado o cortex cerebral de primatas através do estudo das características físicas de sua rede.

Outro exemplo da aplicação de redes complexas nessa área é a dependência entre proteínas. Sendo assim, uma rede complexa pode ser modelada utilizando seus vértices para representar as proteínas e as arestas suas dependências para sua produção [Farkas et al., 2002, Guelzim et al., , Shen-Orr et al., 2002].

Cadeia alimentar é outro exemplo onde pode-se aplicar redes complexas para o estudo de diversos ecossistemas.

2.2.3 Características Físicas de Redes Complexas

Há muitas propriedades que podem ser extraídas das redes complexas. Muitas delas são úteis para revelar o tipo de rede que se trabalha: aleatória, mundo pequeno ou livres de escala [Newman et al., 2006]. Aquelas necessárias para este trabalho são citadas a seguir:

2.2.3.1 Grau de Entrada e Saída (*In-Out Degree*)

Dentre muitas características importantes de um nó, pode-se encontrar a quantidade de arestas que chegam ou saem dele: grau de entrada ou grau de saída, respectivamente. Por definição, o somatório do grau de entrada com o de saída resulta no grau de conexão k de um nó. O grau de conexão médio de uma rede complexa é denotado por $<k>$, que é computada através da Equação (2.5), onde E é o número total de arestas e N é o número total de nós da rede.

$$<k> = \frac{E}{N} \tag{2.5}$$

Uma vez que os nós de uma rede complexa podem ter diferentes graus de conexão, utilizando a função de densidade de probabilidade desses graus pode ajudar a entender o tipo específico da rede.

Teorias sólidas desenvolvidas indicam que redes livres de escala seguem a distribuição da lei-de-potência (Equação (2.4)), onde $P(k)$ é a probabilidade do grau k na rede [Newman et al., 2006].

Outra utilidade importante encontrada para o grau de entrada e saída está na checagem de erros na rede. Seja $In(i)$ o grau de entrada de um nó I e $Out(i)$ o grau de saída. A Equação (2.6) mostra o relacionamento entre as duas propriedades físicas.

$$\sum_i In(i) = \sum_i Out(i) \tag{2.6}$$

2.2.3.2 Coeficiente de Clusterização

A média do Coeficiente de Clusterização (CC) é uma importante característica física de redes complexas com implicações em diversas aplicações. Por exemplo, [Watts and Strogatz, 1998] definiu uma rede complexa como mundo pequeno se ela apresenta as duas seguintes propriedades: a) a média das distâncias entre todos os vértices da rede (ℓ) é comparável àquele de redes aleatórias, $\ell/\ell_{rg} \sim 1$, ; e b) o coeficiente de clusterização é muito maior do que o de uma rede aleatória, $CC/CC_{rg} \gg 1$. Ambas propriedades para uma rede de mesma densidade [Newman et al., 2006].

Algumas aplicações importantes como a Internet, World Wide Web, Colaboração Biológica, Co-ocorrência de Palavras, entre outras, apresentam tais características [Newman et al., 2006], que enfatiza a necessidade de computar o coeficiente de clusterização da rede.

O cálculo do coeficiente de clusterização expressa o quanto a vizinhança de um nó i está conectada entre ela mesma. Este valor varia entre 0 e 1, onde 0 é uma vizinhança totalmente desconectada e 1 é uma vizinhança totalmente conectada.

Formalmente, considere i como qualquer nó de uma rede complexa, L_i o conjunto de nós que têm uma conexão com i, e $W = \{w_{u,v} | u, v \in L_i\}$ um conjunto de pesos de arestas que conectam cada nó de L_i a outro nó de L_i. A Equação (2.7) mostra o cálculo do coeficiente de clusterização para uma rede dirigida.

$$CC_i = \frac{|W|}{|L_i| \times (|L_i| - 1)} \tag{2.7}$$

Esta equação relaciona o número $|W|$ de arestas existentes na vizinhança de i com o máximo número de arestas possíveis para a quantidade de L_i nós. A Equação (2.7) é capaz de calcular precisamente o coeficiente de clusterização, porém ela não considera os pesos das arestas, mas apenas a quantidade de elementos. Em casos de redes ponderadas, essa equação é obviamente limitada e não se aplica.

Tendo isso em mente, propomos uma nova equação e adicionamos o peso das arestas, trocando a quantidade $|W|$ na Equação (2.7) das arestas pela soma dos pesos dessas arestas

$\sum W_e$:

$$CC_i = \frac{\sum W_e}{|L_i| \times (|L_i| - 1)} \tag{2.8}$$

Então, para uma rede com N nós, o coeficiente de clusterização médio CC é dado por

$$CC = \frac{1}{N} \sum_i CC_i \tag{2.9}$$

A Equação (2.8) é idêntica à Equação (2.7) com as seguintes modificações: o numerador $|W|$ na Equação (2.7) é o tamanho do conjunto de pesos $w_i \in \{0, 1\}$, e na Equação (2.8) $0 \leq w_i \leq 1$. Isso permite que consideremos valores ponderados entre 0 e 1. Dessa forma, na Equação (2.7), o coeficiente de clusterização é computado apenas para cada nó j vizinho à i conectado por uma aresta $w_{i,j} = 1.0$. Porém, na Equação (2.8), a conexão entre i e um nó vizinho j é ponderada. A consequência dessa modelagem é que o coeficiente de clusterização de i não é somente computado levando em consideração a conexão entre a vizinhança, mas também o quanto i está conectado à ela.

A Fig. 2.4 mostra um exemplo. Para calcular o coeficiente de clusterização do nó $i = 1$, nós devemos primeiramente identificar seus vizinhos, nesse caso $L_1 = \{2, 3, 5, 6\}$. O método usado aqui considera todas as arestas conectando quaisquer dois nós do conjunto L_1, gerando o conjunto de arestas $E = \{2 \rightarrow 3, 3 \rightarrow 5, 6 \rightarrow 2\}$ e o conjunto $W_e = \{0.5, 0.3, 0.4\}$. Então, aplicamos a Equação (2.8), encontrando $CC_1 = 0.1$.

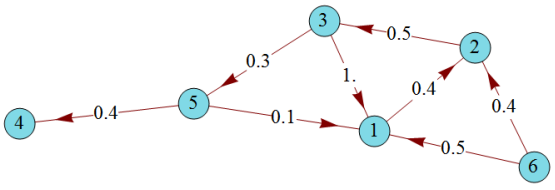


Figura 2.4: Coeficiente de Clusterização

Note que, para o coeficiente de um nó ser igual a 1, é necessário que todas a arestas possíveis da vizinhança tenham $w_{i,j} = 1$. Nesse caso, a Equação (2.8) se reduz à Equação (2.7), portanto, ela é uma generalização (veja Fig. 2.5).

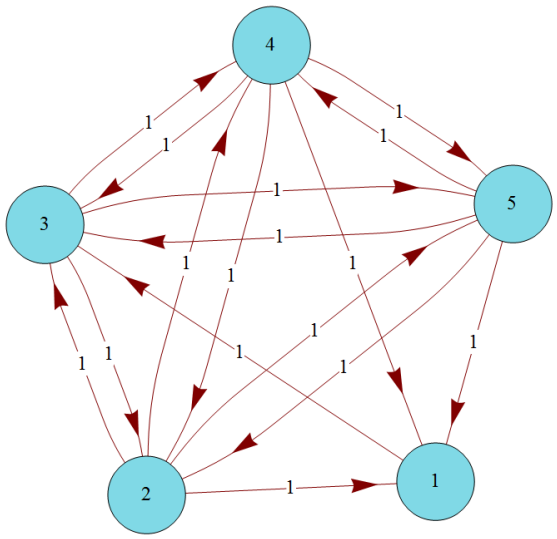


Figura 2.5: Coeficiente de Clusterização Máximo

Note que a Equação (2.8) também pode ser utilizada para redes não dirigidas. Para isso, teremos a metade do total de arestas possíveis entre os vizinhos. Sendo assim, o denominador da equação será $|L_i| \times (|L_i| - 1)/2$ e a Equação (2.8) pode ser reescrita como:

$$CC_i = \frac{2 \times \sum W_e}{|L_i| \times (|L_i| - 1)} \tag{2.10}$$

2.2.3.3 Densidade de Conexão Média

A Densidade de Conexão Média (K_{den}) de uma Rede Complexa é uma medida física que varia entre 0 e 1 de acordo com a densidade das conexões da rede. Essa medida varia de forma abrangente dependendo da topologia da rede [Sporns, 2002], tornando-a uma importante medida física para sua análise. A Equação (2.12) apresenta o cálculo do K_{den}:

$$K_{den} = \frac{|E|}{n^2 - n} \tag{2.11}$$

onde $|E|$ é o conjunto de células da matriz de adjacência com pesos diferentes de zero. Isso significa que a Equação (2.12) relaciona a quantidade de arestas existentes com a quantidade máxima de arestas possíveis.

2.2.3.4 Índice de Semelhança de Conexão

O Índice de Semelhança de Conexão $m_{i,j}$ (para $i \neq j$) de uma Rede Complexa mede o quanto o conjunto de conexões de um nó i é semelhante ao conjunto de conexões de um nó j [Hilgetag et al., 2000, Sporns, 2002]. O cálculo de $m_{i,j}$ é feito somando-se a quantidade de nós semelhantes de i e j (que se conectam aos mesmos nós) e dividindo pela quantidade total de de nós conectados a i e j. Suponha A o conjunto de nós conectados a i e B o conjunto de nós conectados a j. O índice de semelhança de conexão $m_{i,j}$ é dado pela equação

$$m_{i,j} = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{2.12}$$

2.2.3.5 Conexões Recíprocas

Conexões Recíprocas são pares de arestas que conectam dois nós em ambos os sentidos. Formalmente, uma conexão recíproca existe se $M_{i,j} > 0$ e $M_{j,i} > 0$ para $i \neq j$ [Sporns, 2002]. A divisão da quantidade de arestas recíprocas pela quantidade de arestas da rede é uma medida conhecida como "Fração de Conexões Recíprocas" e é simbolizada por ρ.

2.2.3.6 Probabilidade de Ciclos

Ciclos são caminhos que conectam um nó j a ele mesmo com vértices e arestas distintas. Basicamente, esta medida informa a probabilidade de um caminho ser cíclico [Sporns, 2002].

2.2.3.7 Matriz de Distâncias, Excentricidade, Raio, Diâmetro

A Matriz de Distâncias armazena o tamanho do menor caminho, d_{ij}, entre um nó i e um nó j [Harary, 1969]. Se nenhum caminho existe entre i e j então $d_{i,j} = \infty$.

No modelo proposto por esse trabalho, consideramos o peso de co-ocorrência de acordo com a Equação (2.1). Quanto maior esse peso, maior a relação entre duas palavras e consequentemente menor a distância entre elas. Dessa forma, considere que os pesos da matriz de adjacência são interpretados pelo algoritmo de distância como:

$$d_{i,j} = 1 - w_{i,j} \tag{2.13}$$

A Excentricidade de um nó i é a distância finita máxima para todos os outros nós da rede. Dessa forma, pode-se obter a excentricidade de um nó i a partir da linha da matriz de distâncias: $ecc(i) = \max_{j=1}^{N}\{d_{i,j}\}$ para $d_{i,j} \neq \infty$.

O Raio de uma Rede Complexa é a excentricidade mínima da rede: $\min_{i=1}^{N}\{ecc(i)\}$. O Diâmetro de uma Rede Complexa é a excentricidade máxima da rede: $\max_{i=1}^{N}\{ecc(i)\}$.

2.2.3.8 Matriz de Alcance

A Matriz de Alcance informa se existe pelo menos um caminho que conecta um nó i a um nó j. Se o caminho existe, denotado por $r_{i,j}$, ele recebe o valor 1 (caso contrário, recebe 0) [Sporns, 2002].

2.3 Modelos de Armazenamento

2.3.1 Arquivos Invertidos

A indexação de palavras com base em um conjunto de documentos é essencial para algumas aplicações. Por exemplo, um sistema de recuperação de textos pode requerer documentos onde uma determinada palavra ocorre. Arquivos invertidos é uma técnica de indexação para agilizar esse processo. Mais especificamente, um arquivo invertido é composto por uma lista de palavras (ou termos léxicos) onde cada palavra p_i possui uma lista de ponteiros que indica em quais documentos p_i ocorre [Witten et al., 1999]. A Tabela 2.1 mostra um exemplo.

índice	termo	documentos
1	Redes	$< D_1, D_5, D_6, D_7 >$
2	Complexas	$< D_1, D_5, D_7 >$
3	Nó	$< D_1, D_3, D_4 >$
4	Aresta	$< D_2, D_4 >$

Tabela 2.1: Exemplo de um arquivo invertido

A Tabela 2.1 informa que a palavra (termo) *Nó* (índice 3) ocorre nos documentos D_1, D_3 e D_4.

Podemos ainda procurar documentos através do conjunto de palavras. Por exemplo, para encontrar um documento específico D_k onde ocorreram as palavras *Nó* e *Aresta* (índice 4) basta fazer intersecção dos dois conjuntos apontados por nó e aresta: $\{D_1, D_3, D_4\} \bigcap \{D_2, D_4\} = \{D_4\}$. Portanto, somente o documento D_4 possui conjuntamente as palavras *Nó* e *Aresta*.

Os arquivos invertidos podem ser considerados como uma tabela onde cada tupla (ou linha) contém o índice da palavra, a palavra e a lista de documentos onde cada palavra ocorreu [Witten et al., 1999]. Uma prática comum é substituir cada termo, não pela palavra em si, mas pela frequência com que cada termo ocorre na base de dados.

No entanto, algumas limitações são encontradas quando se fala em recuperação textual através de Arquivos Invertidos. A mais relevante é que as palavras do conjunto de busca devem ser exatamente aquelas que estão nos documentos que devem ser recuperados. Isso significa que não há uma modelagem contextual das palavras (ver Seção 2.1 para conceito de *contexto*). Também deve-se destacar que a ordem das palavras não interfere no resultado final. Sua modelagem envolve a construção de uma tabela que contém vetores para cada documento, gerando um consumo e gerenciamento de memória excessivos. Outro ponto negativo pode ser observado nas palavras chamadas *stop-words* (ver seção 2.1 para conceito de *stop-words*). Estas palavras aparecem na maioria dos documentos e, em muitos casos, não alteram o resultado da pesquisa se forem eliminadas do documento [Baeza-Yates and Ribeiro-Neto, 1999]. Sendo assim, em muitos trabalhos como em [Ferrer et al., 2001, Sebastiani, 2002], estas palavras devem ser desconsideradas quando se

constrói o arquivo invertido.

Neste projeto, utilizamos os arquivos invertidos e tiramos proveito do índice da palavra como sendo o índice do nó em nossa rede complexa. Dessa forma, como nossa rede é descrita por uma matriz de adjacência M, o número da linha e da coluna correspondem aos índices dessas palavras. A Seção 2.2 explicará com mais detalhes sobre a utilização desses índices na rede considerada.

Na análise contextual, o uso ou não de *stop-words*, bem como, a indexação de documentos e palavras através da frequência de ocorrência nos textos, são termos inerentemente ligados à semântica dos documentos. Na Seção 2.4 tratamos mais detalhadamente esse assunto.

2.3.2 Árvore Patricia (Pratical Algorithm To Retrieve Information Coded In Alphanumeric)

Árvores de pesquisa binária são muito vantajosas quando se fala em grandes bases de dados. O principal recurso dessas árvores está relacionado com o tempo computacional envolvido para a tarefa de busca ($O(logn)$). A Árvore Patricia é uma árvore binária que utiliza nós de decisão para definir um caminho para que se encontre o dado desejado [Ziviani, 1996]. A construção da Árvore Patricia é feita com base nos bits diferentes entre uma chave de busca e outra buscada. Por exemplo, dado duas chaves $k_1 = 01100$ e $k_2 = 00011$, o primeiro bit que se diferencia é o segundo. Dessa forma, teremos uma árvore como representado pela Fig. 2.6.

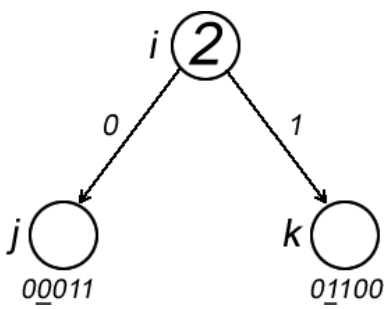


Figura 2.6: Árvore Patricia

Assim, ao fazer uma busca pelo índice 01100 a partir da raiz da árvore, escolhe-se o caminho de acordo com o valor do bit indicado pelo nó de decisão. Neste caso, o bit de decisão do nó i é 2, conforme ilustrado. Sendo assim, deve-se ir à direita uma vez que o bit 2 do elemento a ser buscado é 1. O que se obtém no final é uma Árvore de Pesquisa Binária (Trie). Utilizamos esta árvore para fazer a busca de uma palavra para obter seu índice correspondente. De posse desses índices, podemos operar nas linhas e colunas da matriz de adjacência da rede complexa (ver Seção 2.2).

Conforme mencionado anteriormente, o tempo computacional para uma busca pela árvore Patricia é $O(\log n)$ na média. Porém, há alguns fatores negativos que devem ser levados em conta quando for utilizada. A principal desvantagem é que a estrutura de dados envolvida e sua própria construção exigem um consumo maior da memória principal. Outro fator negativo é que ela não é uma árvore balanceada. Isso significa que nem sempre o tempo computacional para a busca será $O(\log n)$ e, em seu pior caso, será $O(n)$.

2.4 Análise de Informações Textuais

Atualmente, as máquinas de buscas têm ganhado grande importância com o crescimento da rede mundial de computadores. Responsáveis por encontrar as mais diversas informações espalhadas na Internet, as máquinas de busca devem lidar com um conjunto de dados cada vez maior e a uma demanda excessiva, gerando um enorme desafio para os cientistas da área de "Recuperação de Informação", que tentam criar novos métodos para encontrar resultados que satisfaçam os usuários.

A área de recuperação de informação de textos considera o conceito de "Documentos Relevantes". Em análise de textos, um "Documento Relevante" pode ser uma página html ou um arquivo de texto sem formato (geralmente não proprietário), que é a unidade básica de uma máquina de busca. Tradicionalmente, na área de recuperação de informação, um documento é mais relevante quanto mais ele satisfaz a requisição do usuário, que geralmente baseia-se em palavras-chave [Baeza-Yates and Ribeiro-Neto, 1999].

A eficiência de uma máquina de busca é geralmente medida de acordo com a quantidade

e qualidade dos documentos relevantes recuperados [Eakins, 2002]. Embora muitos trabalhos tenham apresentado resultados satisfatórios, ainda existe uma demanda na diminuição da taxa de erros do número de documentos retornados. Essa diminuição está estreitamente relacionada com a capacidade do modelo computacional de busca de reconhecer o conteúdo semântico de um documento.

De uma maneira geral, as máquinas de busca tentam modelar a semântica por meio de diferentes métodos: frequência de palavras, relevance feedback, palavras-chave e ontologia são alguns bem conhecidos [Baeza-Yates and Ribeiro-Neto, 1999]. Historicamente, o desafio da área de recuperação de informação é modelar a semântica de tal forma que a taxa de erros no número de documentos retornados seja a menor possível.

É interessante notar que a análise semântica de um documento também influencia quando os elementos de busca são imagens. Essa área é conhecida como "Recuperação de Imagens com Base no Conteúdo" (RIBC). Por exemplo, o trabalho de J. P. Eakins [Eakins, 2002] discute sobre a necessidade de alguns usuários relacionadas com sistemas de recuperação de imagens. O autor comenta que há um desencontro das informações de entrada, fornecida pelos usuários, e o que os sistemas de recuperação de imagens realmente esperam como entrada. É argumentado que a principal diferença sobre essas informações está relacionada com a semântica. Por exemplo, muitas vezes o usuário gostaria de procurar uma imagem informando o nome de um objeto, fenômeno ou evento. Porém, a maioria dos mecanismos de busca de imagens consegue trabalhar somente em um nível de informação mais baixo, tais como: cores, formas e texturas. Isso significa que os usuários entram com informações muitas vezes pobres e incompletas para as quais o sistema espera, gerando o chamado *gap semântico*. Assim, Eakins definiu três níveis de busca. O primeiro nível, mais baixo, engloba algumas características primitivas da imagem, tais como: cor, forma e textura. Por sua vez, o segundo nível, por meio de inferências, engloba objetos e regiões isoladamente. Finalmente, o terceiro nível, mais alto, é o mais difícil de ser alcançado, requerendo algumas aplicações da área de IA. Esse nível representa solicitações com atributos abstratos, tais como: significado dos objetos ou descrições de cenas.

À luz dessas mesmas idéias, na area de análise textual podemos fazer uma analogia

sobre os 3 níveis de semântica envolvidos em um documento de texto:

1. Baixo: Compreende os caracteres (padrão e especiais), pontuações e palavras (Análise léxica).

2. Médio: Compreende as orações do documento, compostas geralmente por: Sujeito, Verbo, Predicado e outros termos integrantes da oração (Análise sintática).

3. Alto: Compreende a análise de todo o documento. Neste nível, existe a compreensão do documento de uma forma mais abstrata. Por exemplo, documentos textuais com caracteres e palavras totalmente diferentes podem ter o mesmo significado semântico, uma vez que podemos usar palavras diferentes para comunicar a mesma ideia.

Sabe-se, no entanto, que muitas das informações semânticas de um documento (texto, imagem, som, etc...) estão intimamente relacionadas ao contexto que o documento está inserido. Entretanto, o conceito de contexto, assim como o de semântica, é altamente intuitivo e dependente de subjetividade. Não conhecemos na literatura, até o momento, trabalhos que abordem uma modelagem contextual abrangente. De maneira intuitiva e informal, contexto pode ser tudo aquilo que está relacionado de alguma forma com as idéias contidas em um documento, seja ele texto, imagem, música, ou qualquer outra forma de comunicação humana. Os modelos tradicionais baseiam-se principalmente na frequência de palavras como principal conteúdo de informação que descreve um documento. Porém, como apontam indiretamente Eakins e outros autores, a não-modelagem contextual é um dos principais motivos que está por de trás do chamado *gap-semântico*. Além disso, outras informações, tais como ordem com que as palavras aparecem no texto, pontuação, stop-words e palavras com grafia errada também podem influenciar no contexto, e no entanto, não receberam grande atenção dos pesquisadores até o momento.

Existem pelo menos duas razões principais pelas quais o contexto é geralmente ignorado: a falta de um modelo matemático adequado; e poder computacional para seu processamento, uma vez que envolve muitas variáveis e consequentemente possui alta demanda computacional.

Dentre esses dois motivos, a escolha de um modelo matemático é sem dúvida a mais crítica, uma vez que o processo cognitivo de tratamento das informações contextuais ainda não é perfeitamente compreendido pelos pesquisadores, tanto na área da linguística quanto neurociência e cognição. No entanto, sabe-se que a inferência de idéias baseadas em informações advindas do contexto depende de diversos fatores, tais como: co-ocorrência e posição espacial entre primitivas de baixo nível, informações *a priori* a respeito do objetivo cognitivo e memórias de curto e longo tempo.

Todos esses fatores receberam pouca atenção dos pesquisadores da área de recuperação de informação, levando ao *gap-semântico* e, consequentemente, baixa precisão e revocação de documentos relevantes, como resultados de busca.

Na presente dissertação de mestrado, propomos um modelo de informações contextuais que leva em conta dois dos principais fatores citados acima: a co-ocorrência entre as primitivas de baixo nível, bem como o posicionamento espacial entre elas. Esse modelo também contempla informações tradicionais como frequência de palavras no texto. O objetivo principal é a análise e estudo do modelo proposto, que se baseia em Redes Complexas e Teoria da Informação. Assim, independente do debate científico que ainda existe sobre o tema, é importante deixar claro o que, formalmente, significa contexto na presente dissertação.

Sendo assim, definimos neste trabalho o conceito de contexto como sendo um conjunto de palavras que co-ocorrem frequentemente em diversos documentos da base textual. Por exemplo, é intuitivo imaginar que duas palavras, como *carro* e *veículo*, co-ocorram muitas vezes nos mesmos documentos. Sendo assim, essas duas palavras pertencem ao mesmo contexto. É importante notar que a palavra *veículo* também pode estar inserida em contextos diferentes. Por exemplo, com a palavra *meio* (quando se trata de um veículo de comunicação). Mais formalmente, para se definir o contexto, é necessário introduzir os conceitos de vocabulário, documento e cluster (estes dois últimos já foram definidos na Seção 2.1). Um vocabulário de n palavras pode ser definido como: $V = \{p_1, p_2, p_3, \ldots, p_n\} \bigcup \beta$, onde β é uma palavra especial que representa o espaço em branco. Então, um Documento será definido de forma recursiva. Considere γ como uma palavra vazia (tamanho 0):

Definição de Documento:

i) Base: $\gamma \in V^*$

ii) Passo recursivo: Se $P \subseteq V^*$ e $p_j \in V$, então $P \cup p_j \cup \beta \in V^*$

iii) Fecho: $P \in V^*$ apenas se ele pode ser obtido a partir de γ por uma aplicação recursiva finita de ii).

Definição formal de Cluster:

Seja $C_k = \{C_1, C_2, C_3, \ldots, C_n\}$ um cluster composto de sub-clusters em que, no caso especial para $|C_k| = 1$, $C_k = \{p_i\}$ (contém uma única palavra de V). Dizemos que duas palavras p_i e p_j pertencem ao mesmo contexto se e somente se $\exists k | p_i, p_j \in C_k$. Suponha os seguintes documentos: $D_1 = \{p_1, p_3, p_7, p_2\}$, $D_2 = \{p_3, p_2, p_7, p_1, p_8\}$ e $D_3 = \{p_2, p_3, p_4, p_8\}$. Supondo que a base de documentos seja formada somente por D_1, D_2 e D_3, sabemos que as palavras p_2 e p_3 estão em um mesmo cluster, uma vez que co-ocorrem em todos os documentos da base. Porém, as palavras p_2 e p_8 fazem parte de um outro cluster, uma vez que co-ocorrem em 2 documentos. Nota-se então que uma palavra pode pertencer a mais de um cluster.

Neste trabalho consideramos que um documento pode ser modelado em 3 níveis hierárquicos de abstração: baixo (1), médio (2) e alto (3). Por exemplo, dois documentos de um mesmo contexto (nível semântico 3) podem apresentar o mesmo conteúdo semântico e, consequentemente, a mesma referência para uma mesma busca. Porém, a alternância das orações (nível semântico 2) pode torná-los como sendo de contextos diferentes (ver definição de *contexto semântico* na Seção 2.1); da mesma forma que duas orações de mesmo conteúdo semântico (nível hierárquico 2) podem pertencer a contextos diferentes devido a alternância de palavras (nível hierárquico 1).

É importante notar que os principais modelos propostos até hoje (Seção 2.4.1), que tratam da análise de textos, inclusive as máquinas de busca, são baseados na clusterização dos documentos no nível hierárquico 1, assim como para recuperação de imagens baseada em contexto. Para os níveis semânticos 2 e 3, há uma demanda por novos modelos de clusterização. Dessa forma, os resultados das máquinas de busca podem ser mais eficientes

a medida que diminuímos o *gap semântico*, ou a medida que subimos no nível hierárquico.

A clusterização de palavras é uma tarefa fundamental quando se deseja trabalhar no nível hierárquico 3. Na Seção 2.6.1 serão descritos alguns dos principais métodos de clusterização abordados neste trabalho.

2.4.1 Modelos de Recuperação de Informações

Esta seção reúne alguns modelos de recuperação de informações (R.I.) clássicos. Porém, para estudá-los, é necessário primeiramente definir o que é um modelo de R.I.

Um modelo de R.I. é formado por 4 entidades: *Document*, *Query*, *Framework* e *Ranking*. A primeira delas, *Document*, é uma representação de um documento da base. A segunda entidade, *Query*, representa as necessidades de informações do usuário. O *Framework* modela e relaciona as entidades *Document* e *Query*. Finalmente, *Ranking* é uma função que define a ordem dos documentos de acordo com uma *Query* do usuário.

2.4.1.1 Modelo Booleano

Utilizado nos primeiros sistemas bibliográficos comerciais, o Modelo Booleano de R.I. baseia-se nos conceitos de Teoria dos Conjuntos e Álgebra Booleana. A principal vantagem desse modelo está em sua simplicidade e formalismo [Baeza-Yates and Ribeiro-Neto, 1999].

O Modelo Booleano indexa um documento através de "termos de indexação". Esses termos são palavras que abrangem o assunto do documento e devem ser estrategicamente escolhidos de tal forma a não serem palavras que apareçam em todos os documentos e ao mesmo tempo que apareçam somente em poucos documentos. Esse assunto será discutido com mais detalhes na Seção 2.4.1.2. O processo de busca é iniciado com a solicitação do usuário através de "termos" (ou palavras) de pesquisa.

De uma forma geral, a técnica de recuperação Booleana classifica um documento como sendo Relevante ou Não-Relevante (veja Seção 2.1), sendo essa sua principal desvantagem, uma vez que os documentos não podem ser ordenados por grau de relevância. Dessa forma, o sistema verifica se cada termo contido na *Query* existe em cada documento da

base. Se todos os termos existirem no documento, o documento é considerado relevante e é retornado ao usuário, caso contrário, o documento é considerado não-relevante e não é retornado.

É importante notar que, como um sistema de R.I. booleano é preciso nas informações contidas na *Query*, muitas vezes é difícil traduzir a necessidade do usuário corretamente em uma *Query*. Na tentativa de melhorar esse modelo, foi proposto o Modelo Vetorial (ver Seção 2.4.1.2) que acrescenta ponderação nos "termos de indexação".

2.4.1.2 Modelo Vetorial

O modelo vetorial é o sucessor do modelo booleano. Nele, cada "termo de indexação" é relacionado a um documento através de um peso entre 0 e 1. Esses pesos são utilizados para calcular o *grau de similaridade* entre cada documento da base e a *Query* (solicitação) do usuário.

No modelo vetorial, um documento é representado através de um vetor $\overrightarrow{d_k} = \{w_{1,k}, w_{2,k}, \ldots, w_{t,k}\}$, onde t é o número total de "termos de indexação" e a *Query* é modelada como um vetor $\overrightarrow{q} = \{w_{1,q}, w_{2,q}, \ldots, w_{t,q}\}$. Tanto em $\overrightarrow{d_k}$ quanto em $\overrightarrow{q}$, os termos $w_{i,k}$ e $w_{i,q}$ representam a frequencia com que cada palavra ocorreu respectivamente no documento k e *Query* q. Portanto, os Documentos e a *Query* estão em um espaço *t-dimensional*. Isso significa que para se comparar os documentos com a *Query*, basta medir a diferença entre esses vetores. Uma das maneiras mais utilizada é medir o ângulo entre esses dois vetores por meio da sua correlação (Equação (2.14)).

$$sim(d_k, q) = \frac{\sum_{i=1}^{t} w_{i,k} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,k}^2 \times \sum_{i=1}^{t} w_{i,q}^2}} \tag{2.14}$$

2.4.1.3 Modelo Probabilístico (Redes Bayesianas)

Nos trabalhos de [Baeza-Yates and Ribeiro-Neto, 1999, Ribeiro and Muntz, 1996, Ribeiro et al., 2000] foram apresentados modelos probabilísticos com Redes Bayesianas baseados em frequência de palavras para recuperação de textos na internet. Esses trabalhos mostram o quanto as

Redes Bayesianas são flexíveis para modelagem de problemas na área de recuperação de informação.

Basicamente, as Redes Bayesianas (ou Redes de crença) são um grafo sem ciclos e direcionado, no qual os nós modelam os eventos de um problema e as arestas representam a probabilidade condicional de transição entre os eventos.

Adaptando o problema de Recuperação de Informação ao modelo de Redes Bayesianas, os trabalhos de [Baeza-Yates and Ribeiro-Neto, 1999, Ribeiro and Muntz, 1996, Ribeiro et al., 2000] consideraram cada palavra como um evento e as relações entre elas as suas respectivas probabilidades condicionais entre si.

O processo de recuperação inicialmente consiste na construção de uma base de dados com suas respectivas probabilidades de ocorrência (eventos *a priori*). Esse processo é supervisionado e tem por objetivo relacionar as probabilidades de transição entre as palavras. Após a construção da Rede Bayesiana, dado um conjunto de palavras Q, é calculada a probabilidade conjunta de todas as palavras ocorrerem em cada documento. Então, posteriormente, os documentos são ordenados de acordo com suas respectivas probabilidades e retornados ao usuário por grau de relevância.

Apesar de ser uma abordagem interessante para o problema, é muito importante saber escolher quais características extrair do texto para modelar uma Rede Bayesiana. Nos trabalhos de Ribeiro-Neto [Ribeiro and Muntz, 1996, Ribeiro et al., 2000], uma das características escolhidas foi a frequência de palavras. Sabe-se que o modelo de frequência de palavras pode contribuir para o *gap semântico*, conforme explicado na Seção 2.4. Portanto, a busca por novas características e modelagens com uso de Redes Bayesianas pode trazer melhorias significativas na área de R.I.

2.4.2 Informações Semânticas Latentes

Os modelos de recuperação de informação introduzidos na Seção 2.4.1 abordam o desafio da indexação de documentos. A escolha por "termos" de indexação para representação de documentos pode gerar perdas de informações, uma vez que os termos escolhidos podem

não descrever o contexto de todo o documento [Baeza-Yates and Ribeiro-Neto, 1999].

Na tentativa de obter um modelo que perdesse menos informações do texto e conseguisse identificar o contexto geral de um documento, foi proposto o modelo de indexação de documentos baseado na análise de informações semânticas latentes [Baeza-Yates and Ribeiro-Neto, 1999, G. W. Furnas and Lochbaum, 1988]. Nesse modelo, uma matriz M é gerada a partir da relação documentos-termos da base, decompondo-se M em 3 componentes de tal forma que $M = KSD^t$, onde K é a matriz de autovetores gerada a partir da matriz de correlação termo-a-termo, S é a matriz diagonal de valores singulares e D é a matriz de autovetores obtida a partir da matriz documento-documento (M^tM). Após essa decomposição, são selecionados os s maiores valores de S. Assim, defini-se um espaço de menor dimensão que o da matriz M. Dessa forma, espera-se que a recuperação textual em menor dimensionalidade possa ser superior a do espaço de termos original [Baeza-Yates and Ribeiro-Neto, 1999].

A solução por redução de dimensionalidade na área de imagens já obteve bons resultados [Ben, 2005] . Porém, um problema relacionado ao modelo de representação do documento é gerado quando se deseja obter autovalores de uma matriz documentos-termos. Essa matriz representa os documentos de acordo com a frequência que os termos ocorrem nos documentos. Contudo, se permutarmos as colunas dessa matriz (termos), os autovetores também se modificam. Sendo assim, há perda de informações (*gap semântico*) na modelagem de um documento que pode interferir no sistema de recuperação.

2.5 Entropia

Esta seção foi subdividida em três partes. A primeira delas, Seção 2.5.1, introduz o conceito de entropia em sistemas tradicionais, as primeiras formas e suas respectivas aplicações em ordem cronológica. A Seção 2.5.2 abrange o tema de entropia não-extensiva e mostra os trabalhos na área. Finalmente, a Seção 2.5.3 aborda a medida de distância entre duas distribuições considerando os dois tipos de entropia.

2.5.1 Entropia Tradicional

O conceito de Entropia nasceu na área da termodinâmica clássica criada por Rudolph Clausius no estudo do engenho a vapor, desenvolvido inicialmente por Carnot. Nessa época, concluiu-se que uma parte da energia gerada pela máquina a vapor era desperdiçada em forma de calor residual. Sendo assim, foi proposta uma equação que relacionava calor com a perda de energia:

$$\Delta S = \frac{\Delta Q}{T} \tag{2.15}$$

Note que esta medida é relativa. Desta forma, somente era possível medir a entropia na mudança dos estados do sistema.

Após a aplicação na área da termodinâmica, Ludwing Boltzmann, em 1872, trouxe o conceito de entropia para a Mecânica Estatística. Utilizando estudos de James Maxwell e probabilidades das propriedades dos átomos, Boltzmann aplicou o conceito de entropia para formular uma lei probabilística conhecida como segunda lei da termodinâmica. De uma forma abrangente, essa lei diz que, quando uma parte de um sistema fechado interage com outra parte do sistema, a energia divide-se até que o sistema entre em um equilíbrio térmico. A Equação (2.16) é conhecida como a segunda lei da termodinâmica.

$$S = k \ln \omega \tag{2.16}$$

Na Equação (2.16), k é chamada de constante de Boltzmann e ω é o número de microestados do sistema. A interpretação da equação criada por Boltzmann remete a sua capacidade de medir a quantidade de desordem do sistema. Dessa forma, se há um elevado número de microestados, a desordem é alta e, consequentemente, a entropia também.

Impulsionado pelo significado da medida de entropia, J. Gibbs criou uma equação mais abrangente para representá-la:

$$S = -k \sum_{i=1}^{\omega} p_i \ln p_i \tag{2.17}$$

Nessa equação, J. Gibbs modela o sistema mecânico através da probabilidade p_i de cada

estado i do sistema ocorrer. Dessa maneira, se todos os estados tiverem a mesma probabilidade $p_i = 1/\omega$, o sistema reduz-se à Equação (2.16), proposta inicialmente por Boltzmann.

Tome como um exemplo um sistema que contém 2 estados como, por exemplo, o lançar de uma moeda. Nesse sistema, se a moeda não for "viciada", temos as probabilidades $p_1 = 0.5$ e $p_2 = 0.5$. Nesse caso, o sistema se comporta de forma totalmente aleatória e não temos certeza de qual estado a moeda pode cair. Dessa forma, o sistema é imprevisível e a quantidade de informação é máxima. Porém, caso a moeda sempre caia com o mesmo lado em todas as jogadas, temos um sistema previsível e a quantidade de informação é baixa. A Fig. 2.7 ilustra um gráfico do resultado da entropia em função da probabilidade p_i.

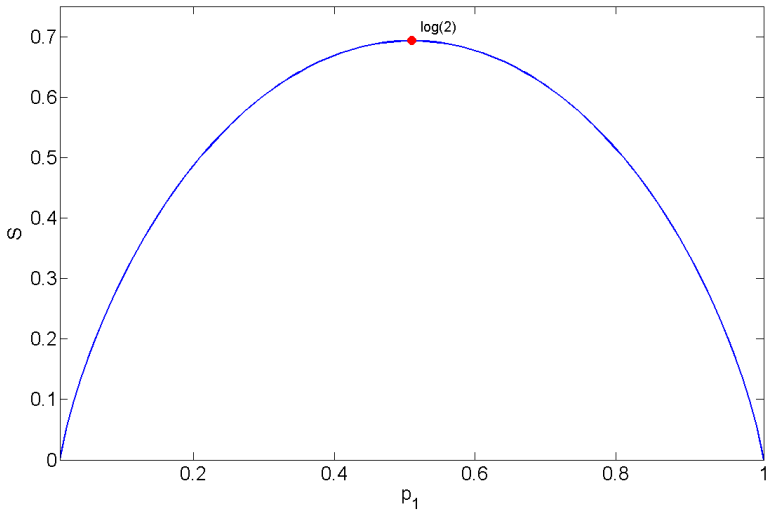


Figura 2.7: Entropia máxima para um sistema de 2 estados

Note que a entropia máxima $S = log(w)$ é alcançada quando as probabilidades dos estados são iguais. Pode-se concluir então que a entropia está relacionada com a quantidade de desordem do sistema.

Dado o significado relevante de sua medida, a entropia chamou atenção de diversos cientistas, abrindo possibilidades de novas aplicações em diversas áreas. No final da década de 40, a entropia teve sua primeira aplicação na área da Teoria da Informação, proposta por Claude Shannon [Shannon and Weaver, 1948]. A ideia de Shannon era medir a quantidade

de informação transmitida em uma mensagem (Equação (2.18)). De forma mais específica, Shannon considerou um microestado (da termodinâmica) como sendo a probabilidade de um possível acontecimento. Se a probabilidade de uma mensagem ocorrer for pequena, então o sistema contém muita informação (problema da moeda não viciada). Porém, se uma mensagem ocorre muito frequentemente, o sistema terá pouca informação (problema da moeda viciada). Na Equação (2.18), k é a constante de Boltzmann e p_i é a probabilidade da mensagem i ocorrer, e n o número de estados possíveis.

$$H = -k \sum_{i=1}^{n} p_i \ln p_i \tag{2.18}$$

Uma propriedade importante da entropia de Shannon é conhecida por "Aditividade". Essa propriedade considera que, para dois sistemas totalmente independentes A e B, a entropia do sistema composto é dada por

$$S(A \oplus B) = S(A) + S(B) \tag{2.19}$$

onde $S(A)$ e $S(B)$ são as entropias dos sistemas A e B.

2.5.2 Entropia não-extensiva

Conforme abordado na Seção 2.5.1, sabe-se que a entropia proposta por Boltzmann-Gibbs [Boltzmann, 1864] é capaz de explicar diversos sistemas físicos clássicos no campo da termodinâmica. No entanto, para alguns sistemas com características específicas como memória de longo alcance, interações de longo alcance e comportamento fractal nas fronteiras, o formalismo de Boltzmann é apenas uma aproximação. Discussões mais detalhadas sobre essas ideias podem ser encontradas na literatura como em [Tsallis, 2001, Tsallis, 1999, Boghosian, 1995].

Em meados da década de 80, C. Tsallis propôs um novo formalismo, que ficou conhecido como entropia de Tsallis ou estatística de Tsallis [Tsallis, 2001, Tsallis, 1999], definido pela

seguinte equação:

$$S_q = k\frac{1-\sum_{i=1}^{n} p_i^q}{q-1} \tag{2.20}$$

onde k é a constante de Boltzmann, n é o número de estados do sistema físico considerado, p_i é a probabilidade do estado i e q é o parâmetro entrópico ajustável ou parâmetro de não-extensividade.

Da mesma forma como foi abordado na Seção 2.5.1, a Fig. 2.8 ilustra a entropia não-extensiva com diversos valores de q para o sistema de lançamento de moeda.

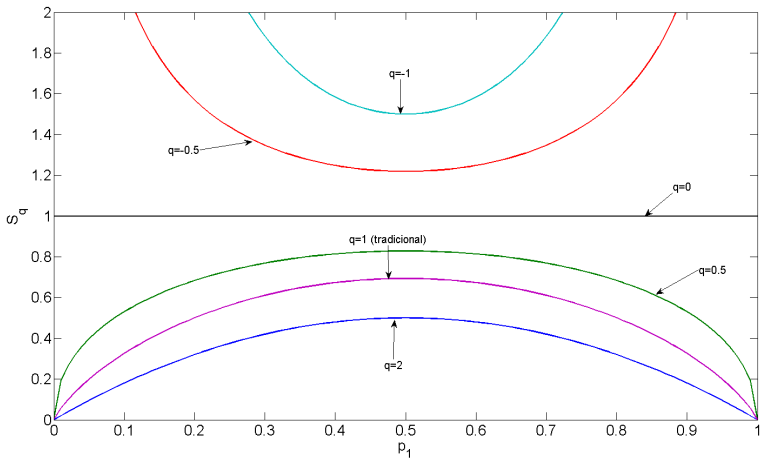


Figura 2.8: Distribuições de entropias para diferentes valores de q em um sistema de 2 estados

A nova entropia proposta não vai de encontro a entropia tradicional de Boltzmann-Gibbs, uma vez que trata-se de uma generalização, podendo ser aplicada a uma gama maior de sistemas físicos que até então não poderiam ser explicados de maneira precisa pelo formalismo tradicional. No trabalho de [de Pinho Tavares, 2003]pode-se encontrar uma prova matemática formal mostrando que a Equação (2.20) é reduzida à Equação (2.18) no limite de $q = 1.0$. Um conjunto de aplicações da entropia de Tsallis pode ser encontrado no endereço http://tsallis.cat.cbpf.br/biblio.htm, onde é constantemente atualizado.

Imediatamente após a descoberta de Tsallis, várias teorias correlatas surgiram ao

mesmo tempo que diversas aplicações foram propostas como exemplos do novo formalismo. No campo da Teoria da Informação, foi imediato generalizar as equações existentes propostas por Claude Shannon [Shannon and Weaver, 1948], gerando grande interesse na área.

Em [Santos, 1997], o teorema de Shannon foi generalizado para sistemas não extensivos. Dessa forma, uma equação foi proposta para satisfazer a mesma propriedade de aditividade encontrada para sistemas extensivos, conhecida como equação de pseudo-aditividade:

$$S_q(A \oplus B) = S_q(A) + S_q(B) + (1-q) \cdot S_q(A) \cdot S_q(B) \tag{2.21}$$

onde $S_q(A)$ e $S_q(B)$ são as entropias de Tsallis dos sistemas A e B, e q é o parâmetro entrópico do sistema composto. No trabalho de [Rodrigues and Giraldi, 2009], foi proposto um novo método de segmentação baseado em entropia não-extensiva e cálculo automático do parâmetro q. Os resultados foram comparados com uma base de imagens segmentadas manualmente. Os autores concluíram que as segmentações efetuadas com o cálculo automático de q se aproximaram daquelas efetuadas manualmente.

2.5.3 Entropia Relativa

Definida em 1951 por Kullback e Leibler para sistemas tradicionais, a Entropia Relativa é uma medida de distância estatística entre duas distribuições probabilísticas. Alguns trabalhos científicos referem-se à Entropia Relativa também como "Distância de Kullback-Leibler", "Divergência I" e "Ganho de Informação de Kullback-Leibler". A Entropia Relativa é definida como sendo:

$$D_{KL}(p, p') = \sum_{i=1}^{k} p_i \cdot \log \frac{p_i}{p'_i} \tag{2.22}$$

onde p e p' são as distribuições e k o número de estados do sistema físico considerado. É importante destacar que, para aplicar a Equação (2.22), o alfabeto das distribuições deve ser o mesmo.

A Entropia Relativa isoladamente não deve ser considerada como uma medida de distância métrica, uma vez que não atende à propriedade da desigualdade triangular:

$$D_{KL}(p, p') \neq D_{KL}(p', p) \tag{2.23}$$

Dessa forma, em [Jeffreys, 1939] foi proposta uma versão simétrica para entropia relativa, a "Divergência J":

$$D(p, p') = D_{KL}(p, p') + D_{KL}(p', p) \tag{2.24}$$

Em 1998, [Borland et al., 1998] propôs a generalização da entropia relativa para sistemas não extensivos, adicionando o parâmetro entrópico q à comparação estatística entre duas distribuições, na forma da seguinte equação:

$$D_{KL_q}(p, p') = \sum_{i=1}^{k} \frac{p_i^q}{1-q} \cdot (p_i^{1-q} - p_i'^{1-q}) \tag{2.25}$$

Na Seção 3.5 apresentaremos o uso da entropia de Tsallis no modelo proposto por esta dissertação.

2.6 Clusterização de Dados

Clusterização é um conceito abrangente que tem sido utilizado em diversas áreas de aplicação. De maneira informal, podemos definir clusterização com sendo a tarefa de identificar grupos dentro de um conjunto de dados. Especificamente, clusterização pode ser destacada como uma forma não-supervisionada de aprendizado, uma vez que o problema remete à identificação de estruturas das classes com o objetivo de explorar o conjunto de dados de entrada [Jain, 2010, Duda et al., 2000] sem qualquer treinamento prévio.

A definição de um cluster é subjetiva, uma vez que pode representar entidades diferentes para cada problema. Assim, de acordo com o tipo de entidade, podem haver diferentes

tipos de ligações entre os dados que estão sendo clusterizados, gerando diferentes tipos de topologias.

Dessa forma, um cluster depende do conhecimento do domínio de um problema. Por exemplo, em [Rodrigues and Giraldi, 2010] foi criado um modelo onde o cluster representa uma região de interesse da imagem (primeiro ou segundo plano). Sabe-se, no entanto que, dependendo das características consideradas de uma imagem, ela pode ser subdividida em regiões e sub-regiões (clusters) que dependem fortemente tanto dessas características, quanto do método de clusterização usado.

É importante observar que dois clusters não são necessariamente um conjunto de dados disjuntos, uma vez que podem compartilhar alguns elementos, o que é conhecido na literatura científica como *overlapping* [Nicosia et al., 2010]. Na Seção 2.6.1 são apresentados os principais algoritmos de aprendizado não-supervisionados, que serão abordados neste trabalho.

2.6.1 Métodos de Clusterização

Podemos dividir os métodos de clusterização em 2 tipos [Jain, 2010]: Particional e Hierárquica. O método de clusterização particional encontra os clusters de forma simultânea. Isso significa que o conjunto de entradas é decomposto diretamente no conjunto de clusters. O tipo hierárquico de clusterização considera que um cluster pode conter sub-clusters, resultando em uma classificação mais específica ou genérica dos dados, dependendo do nível desejado.

Os algoritmos de clusterização hierárquicos são normalmente implementados utilizando recursão. Eles iniciam considerando que cada cluster tem apenas 1 elemento e, recursivamente, mesclam os clusters mais semelhantes (método aglomerativo), ou, consideram que todos os elementos pertencem a um único cluster, dividindo-os em clusters menores (modo divisívo).

Na Seção 2.4 definimos formalmente o que significa um cluster nessa dissertação. Dentre algumas características de cluster, podem-se destacar o seu tamanho e sua densidade. Mais

especificamente, quanto maior o número de palavras e quanto mais palavras conectadas entre si, mais denso é o cluster. Formalmente, menor é o valor do critério de Fischer: a razão entre a distância intra-cluster (elementos do mesmo cluster) pela distância inter-clusters (elementos de clusters diferentes) [Duda et al., 2000]. No presente trabalho, consideramos por enquanto, a distância entre duas palavras de um mesmo cluster como sendo o grau de co-ocorrência num conjunto de documentos. Outra característica de cluster que consideramos nesse trabalho é o chamado overlapping, que significa que uma palavra pode pertencer a diversos clusters (contextos).

Uma parte fundamental de qualquer método de clusterização é a medida de similaridade, que compara tanto os elementos intra-clusters (dentro de um cluster) quanto inter-clusters (entre clusters). Se representarmos os elementos com f características no domínio espacial de dimensão f, essa medida de similaridade é referida como distância. Existem algumas medidas de distância que já foram exploradas pelas mais variadas aplicações. Abaixo são listadas algumas conhecidas.

Nas descrições a seguir, utilizamos a notação $D(i, j)$ para representar a distância entre dois elementos i e j, f para a quantidade de características (dimensão), x_{il} (ou x_{jl}) para o valor do elemento i(ou j) na dimensão l:

Distância Manhattan: Caso especial da distância Minkowski para $q = 1$. Representa a distância do menor caminho entre i e j onde cada segmento é paralelo a uma coordenada do sistema.

$$D(i, j) = \sum_{l=1}^{f} |x_{il} - x_{jl}| \tag{2.26}$$

Distância Euclidiana: Caso especial da distância Minkowski para $q = 2$. Representa a distância do menor caminho entre i e j em linha reta.

$$D(i, j) = (\sum_{l=1}^{f} |x_{il} - x_{jl}|^2)^{\frac{1}{2}} \tag{2.27}$$

Distância Minkowski: Métrica utilizada no espaço euclidiano. Representa a generalização das distâncias de Manhattan e Euclidiana:

$$D(i,j) = (\sum_{l=1}^{f} |x_{il} - x_{jl}|^q)^{\frac{1}{q}} \tag{2.28}$$

Distância Vetorial: A distância vetorial considera o módulo, direção e sentido de dois vetores para medir a similaridade entre eles. Através do cosseno, um vetor é projetado sobre outro gerando uma proporção entre 0 e 1.

$$cos(i,j) = \frac{\sum_{l=1}^{f} x_{il}x_{jl}}{\sqrt{\sum_{l=1}^{f} x_{il}^2 \sum_{l=1}^{f} x_{jl}^2}} \tag{2.29}$$

Divergência J: Parte do conceito da distância de Kullback-Leibler, porém com a característica de simetria. Também conhecida por Entropia Relativa, é uma distância utilizada para comparar distribuições de probabilidade.

$$D(i,j) = \sum_{l=1}^{f} x_{il} \log \frac{x_{il}}{x_{jl}} + \sum_{l=1}^{f} x_{jl} \log \frac{x_{jl}}{x_{il}} \tag{2.30}$$

2.6.1.1 K-Means

K-Means é um dos algoritmos de clusterização mais conhecidos e utilizados atualmente. Criado em 1956 [Steinhaus, 1956, Lloyd, 2003, Macqueen, 1967], é de fácil implementação, eficiente e simples. O K-Means é classificado como um algoritmo particional por trabalhar sincronizadamente os clusters em cada iteração.

Seja $X = \{x_1, x_2, x_3, \ldots, x_n\}$ um conjunto de dados e $C = \{c_1, c_2, c_3, \ldots, c_k\}$ para $k << n$, um conjunto de clusters, onde cada elemento x_i pode pertencer a somente um elemento c_i. O K-Means é um método iterativo que agrupa os n elementos de X pelos k clusters de C.

Para a execução do algoritmo K-Means, deve-se informar a quantidade de clusters (k)

que deseja-se obter do conjunto de dados. Com essa informação, k centros de clusters, mais conhecidos como centróides (μ_i), são escolhidos aleatoriamente para dar início às iterações. Assim, pode-se medir a qualidade de um cluster c_j avaliando a seguinte equação:

$$J(c_j) = \sum_{x_i \in c_j} ||x_i - \mu_j||^2 \tag{2.31}$$

A Equação (2.31) soma as distâncias quadráticas de cada elemento do cluster c_k em relação ao centro do cluster μ_k. O objetivo do algoritmo é minimizar a equação para todos os clusters de tal forma que a Equação (2.32) seja mínima:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} ||x_i - \mu_k||^2 \tag{2.32}$$

Um trabalho envolvendo clusterização de documentos [Li et al., 2008] explorou o K-Means e o comparou com outros métodos. Nele, foi constatado que a clusterização textual envolve dados de alta dimensionalidade, e mostrando também que esse fato prejudica a qualidade dos resultados obtidos pelo K-Means, além de exigir tempo computacional elevado. Outro trabalho [Rodrigues and Giraldi, 2010] analisou e investigou como a distribuição dos dados pelos clusters pode afetar os tamanhos dos clusters encontrados pelos K-Means.

Uma vez que a topologia da distribuição de clusters é um fator fundamental na clusterização de textos, pode-se especular que o resultado do K-Means é altamente afetado por essa topologia.

2.6.1.2 Fuzzy K-Means

Como visto na Seção 2.6.1.1, o algoritmo K-Means classifica cada elemento x_i da base em apenas uma classe C_j. Porém, se um elemento encontra características pertencentes a duas classes, ele poderá ser classificado de forma incompleta. O Fuzzy K-Means é recomendado justamente para os casos onde um elemento tem uma probabilidade de ser classificado ao mesmo tempo em grupos distintos. O resultado final do Fuzzy K-Means é uma matriz

de n linhas por k colunas, onde n é o número de elementos e k é o número de grupos. Essa matriz contém valores entre 0 e 1 que representam a pertinência de um determinado elemento em relação aos k grupos, com a restrição de que a soma da linha da matriz deverá sempre ser igual a 1.

Assim como o K-Means, O Fuzzy K-Means minimiza uma função global. Porém, a função de minimização contém um parâmetro b que pode ser ajustado para controlar o quanto um cluster pode se "sobrepor" a outro (Equação (2.33)).

$$J_{fuz} = \sum_{i=1}^{k} \sum_{j=1}^{n} \hat{P}(C_i|\mathbf{x}_j)^b ||\mathbf{x}_j - \mu_i||^2 \tag{2.33}$$

Na Equação (2.33), a função $\hat{P}(C_i|\mathbf{x}_j, \hat{\theta})$ está normalizada entre 0 e 1 e representa o grau de pertinência do elemento x_j para o cluster C_i e μ_i é o centro do cluster i. Para encontrar o centro de cada cluster, o algoritmo, a cada iteração, utiliza a relação entre probabilidades:

$$\mu_j = \frac{\sum_{j=1}^{n} [\hat{P}(C_i|x_j)]^b x_j}{\sum_{j=1}^{n} [\hat{P}(w_i|x_j)]^b} \tag{2.34}$$

Sendo que:

$$\hat{P}(C_i|x_j) = \frac{(1/d_{ij})^{1/(b-1)}}{\sum_{r=1}^{c} (1/d_{rj})^{1/(b-1)}} \tag{2.35}$$

e:

$$d_{ij} = ||x_j - \mu_i||^2 \tag{2.36}$$

A versão nebulosa (fuzzy) do K-Means permite modelar a sobreposição entre os clusters. Entretanto, o modelo não considera hierarquia de clusters e, da mesma forma que o K-Means, o número de clusters deve ser informado previamente.

2.6.1.3 Markov Clustering (MCL)

Uma cadeia de Markov é um processo estocástico em que o estado futuro de um sistema depende do seu estado atual [Schaeffer, 2007]. Para descrever as probabilidades de um estado ir para outro é utilizada uma matriz chamada "Matriz de Transição". Podemos visualizar a "Matriz de Transição" como um grafo ponderado e direcional, onde cada nó representa um estado e cada aresta representa uma transição. Dessa forma, podemos utilizar conceitos envolvidos na "Cadeia de Markov" para a clusterização de grafos. Na literatura, encontramos alguns trabalhos que denominam essa técnica de clusterização como "Markov clustering algorithm" (MCL) [van Dongen, 2000].

O trabalho de [van Dongen, 2000] utilizou cadeias de Markov para a clusterização de grafos. Aplicando diversas operações na matriz de transição, foi possível "filtrar" as relações mais importantes dessa matriz, restando somente as transições mais prováveis e, consequentemente, identificando os clusters.

Para clusterização de grafos, o algoritmo MCL utiliza conceitos de caminhos aleatórios através de uma rede de interação por meio de dois operadores: Expansão e Inflação. Inicialmente, o algoritmo insere pesos máximos nas células vazias da matriz de adjacência para que haja a formação de ciclos no grafo e todos os nós consigam "enxergar" todos os outros. Colocando pesos máximos, faz-se uma conexão fraca entre os nós da rede. Esta matriz é então transformada em uma matriz "Markoviana" estocástica, que representa a probabilidade de transição entre quaisquer pares de nós.

A operação de Expansão eleva a matriz "Markoviana" a um expoente n para que seja calculada a probabilidade de um caminho de tamanho n entre quaisquer dois nós. Em um cluster, a probabilidade de ter caminhos longos entre seus elementos é maior do que em clusters diferentes. Essa característica leva à elaboração da segunda operação: Inflação. A operação de Inflação seleciona os maiores valores da matriz "Markoviana" e reescala cada coluna para que a matriz continue estocástica (soma da linha igual a 1). Isso permite que os caminhos longos tenham uma probabilidade alta de continuar na matriz. As duas operações são alternadas até que os clusters sejam identificados e não existam mais caminhos entre

clusters.

A clusterização Markoviana não considera a sobreposição de clusters. Desta forma, um elemento poderá apenas pertencer a um único cluster. Da mesma forma que o K-Means e o Fuzzy K-Means, a modelagem não permite a representação de hierarquia nos elementos.

2.6.1.4 Clusterização Espectral

A clusterização espectral é largamente utilizada na area de visão computacional e já teve resultados positivos na área de redes complexas [Albert and Barabási, 2002]. O coeficiente espectral de uma rede complexa é o conjunto de autovalores da matrix de adjacência e está intimamente ligado às características topológicas da rede, de acordo com os trabalhos de Wigner [Wigner, 1955, Wigner, 1957, Wigner, 1958]. Isso significa que as características espectrais de uma rede complexa também podem ser utilizadas para análise de sua estrutura. O trabalho de Barabási aborda a importância da análise espectral para a identificação do tipo de rede que se estuda. O autor afirma que quanto mais próximo o gráfico espectral da rede se assemelha a um semi-circulo, mais randômica são suas ligações [Albert and Barabási, 2002].

A maioria dos métodos de clusterização baseia-se no espaço euclidiano para a formação dos clusters. Diferentemente dos tradicionais, os métodos de clusterização espectral são mais flexíveis, pois possibilitam a captura de uma gama maior de geometrias para clusterização [Yan et al., 2009].

Alguns trabalhos mostram que o segundo menor autovetor é suficiente para minimizar a função objetivo da formulação Ncut (utilizado para clusterização hierárquica). Porém, o tempo computacional para calcular os autovetores de uma matriz n dimensional é $O(n^3)$, o que torna inviável a utilização desse método para redes com milhares de nós [Yan et al., 2009]. Entretanto, novas heurísticas devem ser propostas para minimizar o tempo computacional sem afetar significantemente os resultados.

2.6.1.5 Clusterização por Algoritmos Genéticos

De acordo com [Newman, M. E. J. and Girvan, M., 2004], encontrar uma solução ótima para a clusterização de uma rede complexa pode ser considerado como um problema NP-difícil. Por esse motivo, muitas heurística acabam aparecendo com o objetivo de simplificar o problema e diminuir o tempo computacional envolvido na operação.

Observando o problema de clusterização mais a fundo, podemos interpretá-lo como um problema de otimização. Isso significa que podemos modelar, através de equações, quais são as características esperadas de um cluster. Em sua essência, a característica fundamental de um cluster é o alto número de conexões entre os nós internos e baixo número de conexões entre os nós internos e externos a ele.

Dado a interpretação do problema de clusterização como um problema de otimização, podemos utilizar algoritmos alternativos baseados em otimização. Nesse caso, Algoritmos Evolutivos, como é o caso de Algoritmos Genéticos, podem ser uma forma viável de resolver o problema.

Levando isso em consideração, o trabalho de [Nicosia et al., 2010] propôs um método de clusterização baseado em algoritmos genéticos. O trabalho vai além das formas tradicionais de clusterização, uma vez que trata também o conceito de *overlapping*, ou sobreposição de clusters (conforme abordado na Seção 2.6.1.2), dando graus de pertinência de um nó para cada cluster.

O trabalho de [Nicosia et al., 2010] utiliza a equação proposta por [Newman, M. E. J. and Girvan, M., 2 com o objetivo de medir a qualidade de um cluster. Tal medida é chamada de modularidade (veja Seção 2.6.2.2).

A solução para o problema de sobreposição de clusters foi obtida considerando que cada nó i da rede complexa tem um grau de pertinência $\alpha_{i,c}$ para cada cluster c, de tal forma que

$$\sum_{c=1}^{|C|} \alpha_{i,c} = 1 \tag{2.37}$$

Adaptando o problema para a solução por algoritmos genéticos, [Nicosia et al., 2010]

considerou que cada cromossomo representa uma matriz $M = (\alpha_{i,c})$ onde $i = 1, \cdots, |V|$ e $c = 1, \cdots, |C|$. Cada elemento $\alpha_{i,c}$ representa o quanto um nó i pertence ao cluster c. Para satisfazer a Equação (2.37), a cada iteração (época), todos os vetores são normalizados.

O computacional do algoritmo é $O(|C| * N^2)$, o que é aceitável para grandes redes complexas. Isso significa que o método pode ser aplicado à rede proposta nessa monografia. Além disso, a modelagem de sobreposição de clusters abordada no trabalho de [Nicosia et al., 2010] é fundamental para nosso trabalho, uma vez que um nó de nossa rede pode pertencer a vários clusters.

2.6.2 Medidas de Clusterização

As medidas de clusterização servem para quantificar a qualidade dos clusters. É sabido que, quando há muitas conexões dentro dos clusters (clusters compactos) e poucas entre os clusters (grande distância entre clusters), a qualidade da clusterização é melhor. Essa é a ideia do critério de Fisher. Nesta seção abordaremos algumas dessas medidas para validar a clusterização.

Segundo [Legány et al., 2006], existem três técnicas para se avaliar o resultado de um algoritmo de clusterização . A primeira, chamada de *"Critério Externo"* mede a qualidade baseada em algumas informações do usuário. A segunda, chamada de *"Critério Interno"*, se baseia em métricas do conjunto de dados e da clusterização propriamente dita. Finalmente, a terceira técnica, chamada de *"Critério Relativo"*, compara o resultado de diversos algoritmos de clusterização para o mesmo conjunto de dados. Os métodos Interno e Externo demandam maior tempo computacional em relação ao Relativo.

Em [Legány et al., 2006], os autores introduzem algumas notações, tais como: número de clusters (n_c), número de dimensões (d), distância entre dois elementos ($d(x, y)$), valor esperado na dimensão j ($\overline{X_j}$), conjunto de elementos do i-nésimo cluster (c_i), ponto central do i-nésimo cluster (v_i) e número de elementos no i-nésimo cluster ($|c_i|$). Estas notações pode ser utilizadas para validar a qualidade de clusterização. Porém, tais medidas são normalmente úteis para clusterização sem sobreposição de clusters.

2.6.2.1 Índice S_Dbw

O índice S_Dbw é uma medida da qualidade de clusterização baseada no quanto os clusters estão compactos e o quanto estão separados. Para se obter este resultado, esta medida considera a variância dos elementos da base de dados (Equação (2.38)) e a variância dos elementos dentro dos clusters (Equação (2.39)).

$$\sigma_x^p = \frac{1}{n}\sum_{k=1}^{n}\left(x_k^p - \overline{x^p}\right)^2 \qquad \sigma(x) = \begin{bmatrix} \sigma_x^1 \\ \vdots \\ \sigma_x^d \end{bmatrix} \tag{2.38}$$

$$\sigma_{v_i}^p = \frac{1}{|c_i|}\sum_{k=1}^{n}\left(x_k^p - \overline{v_i^p}\right)^2 \qquad \sigma(v_i) = \begin{bmatrix} \sigma_{v_i}^1 \\ \vdots \\ \sigma_{v_i}^d \end{bmatrix} \tag{2.39}$$

Onde n é o número de elementos, x_k^p é o elemento k na dimensão p, v_i^p é o ponto central do cluster i na dimensão p e $|c_i|$ é o número de elementos no cluster i.

Essas variâncias são utilizadas como base para estimar a qualidade da clusterização. Relacionando as variâncias de cada cluster com a variância de toda a base, obtemos a variação relativa intra-clusters, de acordo com a Equação (2.40).

$$Var_{intra} = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{||\sigma(v_i)||}{||\sigma(x)||} \tag{2.40}$$

Onde $||X|| = \sqrt{X^T X}$. A variância inter-cluster é calculada a partir de um ponto médio euclidiano entre dois clusters (u_{ij}). A Equação (2.41) descreve o cálculo.

$$Var_{inter} = \frac{1}{n_c(n_c-1)}\sum_{i=1}^{n_c}\left(\sum_{\substack{j=1,\\ i\neq j}}^{n_c}\frac{densidade(u_{ij})}{\max\{densidade(v_i), densidade(v_j)\}}\right) \tag{2.41}$$

Onde n_c é o número de clusters e a função *densidade* conta o número de pontos dentro de uma hiper-esfera cujo raio é a média dos desvios padrão de todos os clusters (Equação (2.42)).

$$raio = \frac{1}{n_c}\sqrt{\sum_{i=1}^{n_c} ||\sigma(v_i)||} \tag{2.42}$$

O valor final do índice S_Dbw é obtido a partir da Equação (2.43).

$$S\_Dbw = Var_{intra} + Var_{inter} \tag{2.43}$$

Interpretando as Equações (2.40) e (2.41), podemos assumir que quanto menores seus valores, mais elementos estão contidos nos clusters e menos elementos estão posicionados entre os clusters. Desta maneira, pode-se concluir que quanto menor o valor de S_Dbw, melhor a qualidade dos clusters.

2.6.2.2 Modularidade

Em [Newman, M. E. J. and Girvan, M., 2004] a Equação (2.44) foi proposta como medida de qualidade da clusterização e também foi apresentado e discutido várias aplicações onde essa medida foi usada com sucesso.

A modularidade baseia-se na comparação de Redes Aleatórias com o modelo observado. Sendo assim, um conjunto de nós é um cluster se a quantidade de arestas entre eles é maior do que aquela esperada se a rede fosse totalmente aleatória. A qualidade de um cluster s para um conjunto de nós $V \in s$ pode ser obtida através da Equação (2.44).

$$Qs = \sum_{i,j \in V} [\frac{A_{ij}}{2m} - P_{ij}] \tag{2.44}$$

Onde P_{ij} é a probabilidade esperada de conexão entre os nós i e j para uma rede aleatória de mesmas características (quantidade de nós (N) e densidade de arestas (K_{den})), m é a soma dos pesos da matriz de adjacência e A_{ij} é o peso de uma aresta que sai do nó i em direção ao nó j.

Nesta dissertação, usaremos a Equação (2.44) para medida de qualidade de clusterização de palavras do modelo proposto.

Capítulo 3

Proposta

3.1 Bases de Dados

Com o objetivo de estudar características físicas, algumas informações textuais são retiradas de uma base de artigos médicos chamada *ohsumed*[Hersh et al., 1994] (veja Fig. 3.1). Como destacado nesse exemplo, esta base está organizada com os seguintes rótulos: resumo (.W), título (.T), autores (.A) e palavras-chave (.M). Apesar de ser uma base com diversas informações, nesse projeto apenas o resumo é de nosso interesse. O resumo é visto como um documento na etapa de construção da rede, uma vez que o resumo (.W) trata-se de uma descrição de informações médicas.

3.2 Construção da Rede

Na rede proposta, uma palavra da base é modelada como um nó. Além disso, definimos documento como sendo o texto colocado entre o rótulo $.W$ e o rótulo imediatamente posterior. Isso resulta em uma rede de 221175 palavras obtida a partir de 54710 documentos.

Ao escanear a base de dados, os nós são conectados de acordo com suas co-ocorrências em um mesmo documento (veja a Fig. 3.2 para um exemplo). Essas conexões (arestas) têm um peso $w_{i,j}$ como atributo computado de acordo com a Equação (3.1), onde $Dist(D, i, j)$

```
.I 3246
.U
87077345
.S
Br G Wachs 8704; 73(12):1012-4
.M
Example
.T
A simple title.
.P
JOURNAL ARTICLE.
.W
This is a small text document.
.A
Author 1; Author 2; Author 3.
```

Figura 3.1: Exemplo de um documento TREC (Ohsumed)

fornece a distância entre as palavras i e j dentro de um documento D_k. Para normalizar o valor do peso $w_{i,j}$ no intervalo $[0:1]$, dividimos o somatório da Equação (3.1) pelo tamanho do documento, dado pelo número máximo de palavras que aparecem, $Size(D)$. Note que, quanto menor a razão $Dist/Size$, maior o peso, indicando que a palavra i está fortemente conectada à palavra j.

$$w_{i,j} = 1 - \frac{\sum_{D} Dist(D, i, j)}{\sum_{D} Size(D)} \tag{3.1}$$

A rede complexa proposta é descrita como uma Matriz de Adjacência (AD), onde as linhas e colunas representam palavras distintas que foram previamente indexadas em um arquivo invertido de tamanho N, na sequência em que aparecem na base. Considerando que AD é uma matriz de adjacência de tamanho $N \times N$, $AD_{(i,j)}$, com $i \neq j$, é a co-ocorrência das palavras de indices i e j. Uma vez que esta é uma rede dirigida e ponderada, assumimos que i é um índice da palavra de uma aresta de entrada e j o índice da palavra de uma aresta de saída. Quanto maior $w_{i,j}$, maior é a frequência com que as palavras i e j co-ocorrerem na base.

Como a modelagem proposta gera alta probabilidade de conexão, ela produz uma matriz

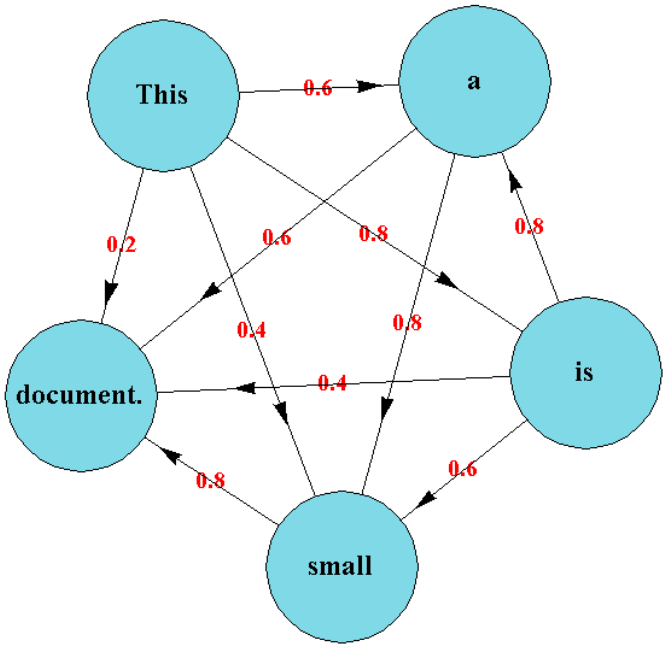


Figura 3.2: Rede de palavras obtido a partir do documento da Fig. (3.1)

densa da ordem de $O(N^2)$. Além disso, é esperado que aconteça uma clusterização natural das palavras que estejam no mesmo contexto a medida que novas conexões vão aparecendo durante a varredura da base (construção da rede). Além disso, ela preserva a ordem em que as palavras co-ocorrem, requisito fundamental quando se deseja trabalhar com aplicações que envolvem o estudo linguístico de palavras [Ferrer et al., 2001].

Mesmo que o resultado seja uma matriz com muitas palavras que co-ocorram com baixa frequência, gerando pesos que podem ser arredondados para zero, permitindo matrizes esparsas e, consequentemente, menos tempo de processamento, em alguns estudos considerar estes pesos baixos pode ser vital quando há uma necessidade de navegabilidade entre os clusters. Portanto, a modelagem proposta preserva todas as arestas com peso, mesmo que seja muito próximo a zero. Consequentemente, esta rede complexa tem como principais características um grande número de nós e grande número de arestas. Em contrapartida, o tempo computacional utilizado pela maioria dos algoritmos para a extração de características físicas consome tempo $O(N^3)$.

Outra característica importante sobre esta modelagem, porém relacionada à base de da-

dos *ohsumed*, é o número de palavras. Em muitas aplicações de análise textual [Urbain et al., 2009, Li et al., 2008], para se reduzir o número de palavras do banco de dados, as palavras chamadas de "stop-words" são removidas. *Stop-words* são normalmente preposições, advérbios, conjunções, números e símbolos especiais que aparecem com alta frequência em todos os textos da base. Porém, em muitas aplicações que envolvem estudos linguísticos, é necessário considerar estas palavras, conforme visto na Seção 2.4. Grafias erradas também podem ser consideradas.

Uma estimativa em nossa base de dados indica que o número de palavras, considerando "stop-words", símbolos e grafias erradas, alcança até 250.000 nós, resultando em $450GB$ de dados para armazenar. Claramente, esta matriz deve ser fracionada para ser carregada na memória principal de um computador comum. Nesse sentido, propomos a divisão da matriz de adjacência em matrizes menores, aqui referenciadas como ADm.

3.3 Extração de Características Físicas

De acordo com a Seção 3.2, a construção da rede requer algoritmos alternativos para lidar com o carregamento parcial da matriz de adjacência de uma grande base de dados. Nesta seção são apresentados alguns algoritmos utilizados para obtenção de algumas características físicas (veja Seção 2.2.3) extraídas da rede complexa proposta nesse trabalho.

3.3.1 Grau de Entrada e Saída

Conforme visto na Seção 2.2.3.1, o grau de entrada de um nó i esta relacionado à quantidade de arestas que partem de outros nós e se conectam a um determinado nó i. De maneira análoga, o grau de saída de um nó i está relacionado com a quantidade de arestas que saem de i e se conectam a outros nós.

Para a extração dos graus de todos os nós da rede, é necessário percorrer todas as células da matriz de adjacência AD proposta nesse trabalho. Utilizamos dois algoritmos para se obter o grau de saída de um nó i. O primeiro deles conta a quantidade de arestas

(de entrada ou saída) que possuem valores acima do valor 0. O segundo soma os pesos de todas as arestas (de entrada ou saída).

Sendo assim, as Equações (3.2, 3.3, 3.4, 3.5) calculam respectivamente as seguintes características de um nó i: grau de saída, grau de saída ponderado, grau de entrada e grau de entrada ponderado.

$$D_{Out}(i) = \sum_{j} 1 \ \forall \ AD_{i,j} > 0 \tag{3.2}$$

$$Dw_{Out}(i) = \sum_{j} AD_{i,j} \tag{3.3}$$

$$D_{In}(i) = \sum_{i} 1 \ \forall \ AD_{i,j} > 0 \tag{3.4}$$

$$Dw_{In}(i) = \sum_{i} AD_{i,j} \tag{3.5}$$

Os resultados preliminares gerados a partir dessas equações são apresentados e discutidos na Seção 4.2.

3.3.2 Coeficiente de Clusterização

Nesta seção, apresentamos duas técnicas propostas para computar o Coeficiente de Clusterização, apresentado na Seção 2.2.3.2, em tempo computacional e uso de memória reduzidos. No Capítulo 4, serão apresentados alguns resultados obtidos a partir desses métodos.

3.3.2.1 Simplificação Baseada em Imagens

De acordo com a Equação (2.9), o coeficiente de clusterização de um nó i é uma medida física que depende principalmente de como sua vizinhança está conectada. Porém, em termos geométricos, nada previne que um vizinho esteja longe de i. Esta é uma característica encontrada em matrizes de adjacência densas e esparsas. Mas é fato que tanto arestas de peso baixo quanto as de peso alto podem afetar o valor do coeficiente de clusterização. Portanto, qualquer método que minimize o número de multiplicações e/ou a soma do coeficiente de clusterização deve levar em conta pesos de qualquer magnitude, tornando a

tarefa de redução de complexidade computacional uma tarefa difícil.

Neste trabalho, representamos a matriz de adjacência como uma imagem digital em escalas de cinza, onde os pesos são mapeados no intervalo dos valores da luminância de acordo com [Noel and Jajodia, 2005, Sporns, 2002].

Embora a vizinhança de um ponto (x, y) em uma matriz de adjacência ponderada muitas vezes não signifique exatamente a vizinhança de uma aresta na rede (uma vez que a distribuição espacial dos valores em AD não seja exatamente as vizinhanças reais entre palavras), aplicando transformações afins de redução geralmente preserva o relacionamento entre os pesos, resultando em uma redução do tempo computacional para o cálculo do coeficiente de clusterização para uma rede complexa com uma quantidade de nós N muito elevada. Isso sugere que transformações afins, como *redução de matriz*, podem representar a matriz original mantendo a aparência da vizinhança da rede complexa original. Como descrito acima, podemos mapear os pesos da matriz de adjacência em imagens digitais e estudar alguns métodos para redução de imagem com o objetivo de computar o coeficiente de clusterização em um tempo computacional aceitável.

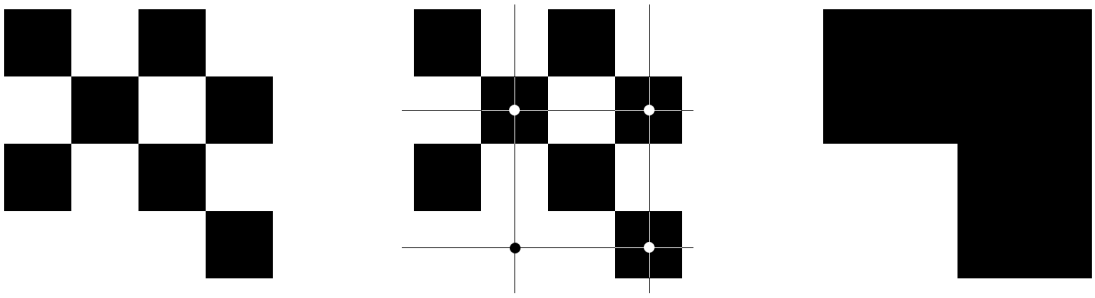
Neste método, a matriz ADm é convertida em uma imagem de tons de cinza, gerando uma matriz $AD_{img}(i, j) \in [0, L]$, onde $L = 255$ é comumente utilizado como valor máximo, e $AD_{img}(i, j)$, como já foi dito, é o peso de uma aresta que sai de um nó i em direção ao nó j como na matriz ADm. Em seguida, reduzimos a imagem AD_{img} utilizando uma das seguintes técnicas de interpolação: *Nearest-Neighbors*, *Bilinear* ou *Bicubic*. Porém, a aplicação de algumas dessas transformações pode resultar em perda de informações em torno da vizinhança de um pixel; consequentemente, isso pode alterar a rede original. Os efeitos dessas mudanças no cálculo do Coeficiente de Clusterização serão analisados, mensurados e discutidos no Capítulo 4.

A interpolação bilinear é uma extensão natural da interpolação linear, porém para funções de duas variáveis em uma malha regular [Gonzalez and Woods, 2002]. Portanto, esta é uma maneira direta de aplicar este algoritmo em processamento digital de imagens. Apesar de sua simplicidade, este é um dos métodos mais utilizados e mais conhecidos para interpolação bidimensional. A ideia básica é executar esta interpolação primeiramente

em uma direção, e então aplicar novamente em outra direção. Apesar de cada passo ser realizado de maneira linear sobre os valores amostrados, a interpolação não é linear como um todo, mas quadrática no local da amostra. A interpolação bilinear considera a vizinhança 2×2 mais próxima de valores de pixel conhecidos em torno de um pixel desconhecido. Em seguida, calcula a média ponderada dos 4 pixels conectados para chegar ao seu valor interpolado final. Isso resulta em uma imagem suavizada.

A redução bicubic vai um passo além da bilinear considerando os 4×4 vizinhos mais próximos de pixels conhecidos para um total de 16 pixels. Uma vez que estes pixels estão em distâncias variadas a partir do pixel desconhecido, para pixels mais próximos são dados pesos maiores no cálculo. A bicubic produz imagens mais nítidas do que o método bilinear, e é, talvez, a combinação ideal entre tempo de processamento e qualidade de saída. Por esta razão, é considerado padrão em muitos programas de edição de imagens, *drivers* de impressoras e interpolação em câmeras.

A interpolação *Nearest-Neighbors* é uma das mais fáceis de se implementar [Parker et al., 1983]. O valor da amostra de uma imagem reduzida é escolhida de acordo com o pixel mais próximo à ela. Quando uma imagem é reduzida, alguns pixels da imagem original que estão fora da grade regular são desconsiderados. A Fig. 3.3 mostra um exemplo de imagem com a grade e a imagem reduzida (Ampliada para visualização).



(a) Imagem de exemplo

(b) Fig. 3.3(a) com a grade de redução (pontilhada). Pixels próximos a esta grade serão escolhidos para compor a imagem reduzida

(c) Imagem (ampliada 2x) da Fig. 3.3(a) reduzida 50% com interpolação *Nearest Neighborhood*

Figura 3.3: Redução *Nearest-Neighbors*

Como dito anteriormente, o algoritmo NN escolhe o pixel da imagem original que tem suas coordenadas proporcionalmente próximas da coordenada do pixel na imagem de saída. Isto significa que o tom de cinza que está concentrado em um região é mais provável de ser escolhido para representar aquela região, sem mudança de seu valor original. Esta é a característica mais importante desse algoritmo que permitiu o sucesso do cálculo de CC em nosso trabalho. Iremos chamar esta propriedade como *Propriedade fundamental do* NN. Uma discussão mais detalhada sobre esta observação será feita no Capítulo 4.

Os três algoritmos de interpolação apresentados foram escolhidos por serem largamente utilizados na literatura. Qualquer um deles pode ser utilizado para reduzir uma imagem de matriz de adjacência AD_{img}. Portanto, é interessante observar o valor do Coeficiente de Clusterização médio (CC) sob esta redução. A ideia geral de redução de imagem para computar o coeficiente de clusterização é a seguinte: seja AD_{img} uma imagem de matriz de adjacência com N linhas e N colunas (matriz quadrada). Aplicando o algoritmo de redução sobre AD_{img} gera uma nova matriz ADR_{img} com $N \cdot (1 - \eta)$ linhas e $N \cdot (1 - \eta)$ colunas, onde $\eta \leq 1.0$ é o fator de redução utilizado no algoritmo. O impacto da redução

de AD_{img} sob o valor do CC será experimentalmente discutido no Capítulo 4. O Algoritmo 1, apresentado abaixo, apresenta os passos necessários para a aplicação do método:

Entrada: AD_{img}, η

Saída: Coeficiente de Clusterização

Inicialização;

$ADR_{img} \leftarrow ReduzImagem(\eta, Method)$;

$CCSum \leftarrow 0$;

para *cada nó* i *em* ADR_{img} **faça**
| $CCSum \leftarrow CCSum + CC(i)$;

fim

retorna $\frac{CCSum}{Size(ADR_{img})}$

Algorithm 1: Cálculo do CC com base em simplificação de imagens

3.3.2.2 Simplificação Estatística

Analisando a Rede Complexa de N nós em termos de probabilidade, um nó i tem a probabilidade $1/N$ de ser aleatoriamente escolhido. Considerando que a Equação (2.9) é uma média sobre todos os coeficientes de clusterização, podemos computar o coeficiente de clusterização global aproximado a partir de alguns nós tomados aleatoriamente.

Este método escolhe $Nr << N$ nós aleatoriamente e calcula seus respectivos coeficientes de clusterização, de acordo com a Equação (2.8), considerando a rede de tamanho original (AD). Após esta etapa, utilizando a Equação (2.9), calcula-se a média do coeficiente de clusterização dos Nr nós. Note que esta é uma aproximação estatística do coeficiente de clusterização real.

$$CC = \frac{1}{Nr} \sum_{i=0}^{Nr} CC_i \tag{3.6}$$

Como a rede proposta por esta trabalho está dividida em matrizes de adjacência menores ADm (de acordo com a Seção 3.2), é necessário desenvolver algoritmos capazes de computar parcialmente as informações do coeficiente de clusterização para cada matriz ADm carregada na memória principal.

Uma estrutura de dados contendo informações sobre quais nós foram escolhidos, quais são seus vizinhos e qual é o peso acumulativo de sua vizinhança durante o escaneamento das matrizes foi desenvolvido baseado na própria equação do coeficiente de clusterização (Equação (2.8)). Primeiramente, um gerador de números aleatório sorteia $Nr << N$ nós entre os indices 1 e N. Estes nós são armazenados em um vetor $V1$ para pesquisa futura. Após isso, o algoritmo escaneia cada matriz ADm para encontrar os vizinhos de todos os Nr nós. A vizinhança de cada nó é armazenada em um vetor $V2_i$ para cada nó $i \in V1$. Em seguida, o algoritmo escaneia novamente as matrizes ADm e acumula, em um vetor V_3, a relação (pesos) entre cada possível conexão (permutação) em $V2_i$ para cada nó $i \in V1$. Nesse ponto, $V3$ contém o numerador da Equação (2.8) para cada nó. O denominador é computado utilizando o tamanho do vetor $V2_i$: $|V2_i| * (|V2_i| - 1)$. Então, aplicamos a Equação (2.8) e armazenamos os resultados em um vetor $V4$, de acordo com a Equação (3.6). Finalmente, o coeficiente de clusterização é obtido calculando a média de $V4$.

Ao contrário da simplificação baseada em redução de imagem, que reduz a dimensão da matriz original e calcula a média dos coeficientes de clusterização após a redução, o método proposto escolhe um conjunto aleatório de nós e, para cada nó, calcula seu Coeficiente de Clusterização correspondente (CC_i) na matriz original. O coeficiente de clusterização médio é a média desses CC_is escolhidos aleatoriamente. O Algoritmo 2 Mostra os passos necessários para computar o CC aleatório.

Entrada: ADm, Nr

Saída: Coeficiente de Clusterização

Inicialização;

Escolha Nr nós entre 1 e N;

$V1 \leftarrow RandomInt(1, N, Nr)$;

para *cada nó* $i \in V1$ **faça**

  Obs.: O algoritmo EncontraVizinhos carrega cada matriz ADm ao invés de toda matriz AD para caber na memória;

  $V2_i \leftarrow EncontraVizinhos(AD, i)$;

  $V3_i \leftarrow 0$;

fim

para *cada nó* $k \in V1$ **faça**

  para *cada nó* $i \in V2_k$ **faça**

    para *cada nó* $j \in V2_k$ **faça**

      se $ArestaExiste(AD, i, j)$ **então**

        $V3_k \leftarrow V3_k + AD(i, j)$;

      fim

    fim

  fim

  $V4_k \leftarrow \frac{V3_k}{size(V3_k)*(size(V3_k)-1)}$;

fim

retorna $Mean(V4)$

Algorithm 2: Cálculo do CC com base em simplificação estatística

3.3.3 Densidade de Conexão Média

Para o cálculo da Densidade de Conexão Média (K_{dem}), é necessário um conjunto de nós e arestas para descobrir a densidade através da Equação (2.12). Isso possibilita a verificação de quanto agrupado está o conjunto de nós e, consequentemente o quanto um cluster está agrupado. Se o valor do K_{dem} for alto, isso significa que há muitas arestas e que há um

grande relacionamento entre os nós do grupo, tornando-o válido como um cluster. Porém, se o K_{dem} for baixo, então os nós não possuem relacionamento suficiente para que esse grupo seja considerado como um cluster.

Utilizando este mesmo conceito para a modelagem desse trabalho, podemos medir o quanto um conjunto de palavras tem característica de um cluster a partir do valor de seu K_{dem}.

3.3.4 Índice de Semelhança de Conexão

O Índice de Semelhança de Conexão é um importante valor obtido a partir da comparação entre dois nós da rede complexa (veja Seção 2.2.3.4). Em [Schaeffer, 2007], a autora comenta sobre clusterização baseada em similaridade de vértices. Um dos métodos abordados por ela consiste em comparar a vizinhança dos vértices e agrupá-los de acordo com a semelhança dessas vizinhanças. Isso gera um algoritmo de ordem $O(N^3)$.

Trazendo esta mesma ideia para a área de Redes Complexas, podemos comparar dois nós através de suas conexões e propor uma clusterização baseada nesta medida. Porém, sabe-se que a Rede Complexa gerada neste trabalho apresenta dimensão muito elevada, o que inviabilizaria o tempo computacional envolvido na clusterização.

Propomos, então, um método que reduz a dimensionalidade da Matriz de Adjacência utilizando descritores estatísticos. Diversos descritores podem ser utilizados, tais como: média, mediana, desvio-padrão, moda, variância, percentis, entre outros. Afim de se obter uma maior precisão sobre a comparação das distribuições, calculamos os descritores em algumas faixas da distribuição original. Por exemplo, considere v_i um vetor com os pesos de conexão de um nó i. Para representar essa distribuição, desmembramos o vetor em p partes. Para cada uma dessas partes calculamos todos os descritores. Esse processo é repetido para todos os nós da rede, gerando-se uma matriz de descritores, onde cada linha representa as características da distribuição de vizinhança de um nó da rede complexa e cada coluna um descritor.

Inicialmente, o método de clusterização cria a matriz de descritores, separando a base

em dois grupos com maior diferença diferença possível entre si. Da mesma forma que os métodos de clusterização hierárquica, cada um dos os dois grupos são divididos em mais dois grupos, porém os descritores são aplicados em cada grupo separadamente até que o número de níveis seja alcançado ou reste somente 1 elemento em cada grupo.

3.3.5 Conexões Recíprocas

Conforme descrito na Seção 3.2, a construção da Rede Complexa proposta nesse trabalho leva em conta a ordem em que as palavras aparecem em um documento. Isso significa que, uma palavra A pode se conectar a uma palavra B, porém sem que a palavra B se conecte à A. Uma palavra que apresenta esta característica claramente apresenta uma regra sintática para sua colocação em uma oração. Por exemplo, não podemos colocar um artigo após um substantivo para referenciá-lo. Dessa forma, o artigo da frase "A casa é grande" não pode ser colocado em outra posição, como na frase "Casa a é grande". Essa rigidez sintática na linguagem pode ser obtida através da análise do número de Conexões Recíprocas (Seção 2.2.3.5).

Analisando o efeito da sintática sobre a semântica, podemos ainda especular que, quanto mais a sintaxe apresenta regras bem definidas, menos ambiguidade será encontrado em frases. Isso possibilitará uma descrição melhor da mensagem que se deseja passar, tornando o entendimento da frase mais claro. Por exemplo, a frase "O aluno estava conversando com o professor parado" não se sabe quem estava parado. Isso porque a ordem em que a palavra "parado" aparece não muda o sentido da frase, tornando-a ambígua.

3.3.6 Probabilidade de Ciclos

De acordo com a Seção 2.2.3.6, a Probabilidade de Ciclos mede o quanto um caminho de uma Rede Complexa pode ser um ciclo ou não. Para uma Rede de Palavras, como a proposta por este trabalho, isso significa a capacidade de uma palavra (nós) retornar a ela mesma através de outros nós. Esse caminho percorrido pode passar por diversos clusters tornando-a uma palavra com alto grau de navegabilidade. Dessa forma, podemos calcular

as palavras que estão presentes em diversos clusters, sendo as principais portas de entrada para os mais diversos clusters.

3.3.7 Matriz de Distâncias, Excentricidade, Raio e Diâmetro

A Seção 2.2.3.7 apresentou os conceitos envolvendo Matriz de Distâncias, Excentricidade, Raio e Diâmetro. Para a rede de palavras proposta neste trabalho, isso representa o quanto uma palavra está distante semanticamente da outra. Porém, a obtenção da Matriz de Distâncias é uma tarefa que apresenta alto custo computacional e, para uma rede, como a proposta por este trabalho, algumas heurísticas devem ser desenvolvidas.

Partindo de uma Rede Complexa Clusterizada, ou seja, com informações sobre quais nós pertencem a cada uma das comunidades, podemos calcular uma distância aproximada apenas medindo a distância entre os nós centrais das comunidades envolvidas. Por exemplo, suponha as comunidades $C_i = \{p_1, p_3, p_4\}$ e $C_j = \{p_2, p_5\}$ e os respectivos centros como p_3 e p_2. Para encontrar as distâncias entre quaisquer dois nós entre as comunidades C_i e C_j, efetua-se o cálculo da distância entre os nós centrais p_3 e p_2. Dessa forma, espera-se encontrar valores próximos da distância real, uma vez que os nós internos aos clusters devem estar fisicamente próximos entre si (propriedade discutida na Seção 2.6.1).

Entretanto, conforme visto na Seção 2.6.1, há diversos métodos para se clusterizar uma rede e alguns desses métodos utilizam o conceito de hierarquia de clusters. Nossa proposta é estimar o nível hierárquico que minimize o tempo computacional para computar uma matriz de distâncias com uma margem de erro aceitável.

3.4 Exemplo de Construção da Rede Complexa

Nesse trabalho modelamos uma palavra como um nó da rede, como proposto em [Ferrer et al., 2001]. Porém, através de varreduras nas bases textuais, os nós são conectados na medida em que co-ocorrem em um mesmo documento da base (Fig. 3.4). Essas conexões (arestas)possuem como atributo um peso $w_{i,j}$ que é calculado de acordo com a Equação (3.1)

, onde $Dist(D_k, i, j)$ é a distância entre as palavras i e j no documento D_k e $Size(D)$ é o tamanho, em palavras, do documento. Note que, quanto menor a relação $Dist/Size$, maior será o peso, indicando que uma palavra i está fortemente conectada com uma palavra j (De acordo com a Seção 3.2).

A Fig. 3.4 mostra uma rede composta por apenas 8 palavras. Nessa figura, pode-se identificar dois clusters: $A = \{$"João","e","honesto"$\}$ e $B = \{$"este", "um","documento"$\}$. Após a construção da rede, espera-se encontrar grupos de palavras que se relacionam entre si. Denominamos nesse projeto que contexto é o conjunto de palavras que se co-relacionam (conexões altas entre si), possuindo maior probabilidade de aparecerem no mesmo documento. Tomemos como exemplo os seguintes documentos:

Doc 1: Este é um simples documento.

Doc 2: João é simples e honesto.

Para montar um rede a partir desses dois documentos, aplica-se a Equação (3.1) para cada co-ocorrência de palavras nos documentos. Por exemplo: a ocorrência da palavra "Este" com "é" pode ser quantificada como $w = 1 - 1/5$, dado que o documento tem um tamanho de 5 palavras e a distância entre elas é de 1 palavra. No final da análise dos documentos, teremos um rede como a Fig. 3.4. É importante notar que a co-ocorrência das palavras "é" e "simples" apareceu nos dois documentos e, por esse motivo, foi calculada uma média aritmética dos seus valores (0.6 e 0.8), resultando em um reforço na conexão para 0.7.

Como já foi dito, a rede é descrita como uma Matriz de Adjacência, cujas linhas e colunas representam palavras distintas que foram previamente indexadas em um arquivo invertido de tamanho n, na ordem em que aparecem na base. Dessa forma, considerando AD uma matrix de adjacência de dimensão $N \times N$, $AD_{i,j}$, com $i \neq j$, é uma co-ocorrência entre palavras de índices i para j. Por se tratar de uma rede complexa direcional e ponderada, assumimos que i é o índice da palavra em que uma aresta sai e, j, o índice da palavra em que a aresta chega. Para considerar a ordem em que as palavras co-ocorrem

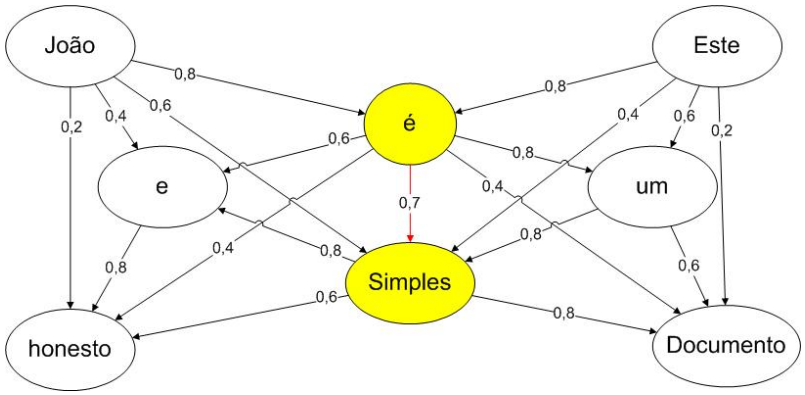


Figura 3.4: Rede de palavras demonstrando a modelagem proposta. Os nós representam as palavras e as arestas representam a frequência de co-ocorrência entre elas em um mesmo documento.

no documento, consideramos a rede como dirigida, ou seja, se a palavra i aparece primeiro que a palavra j, o sentido da aresta será da palavra i para palavra j.

De uma maneira geral, o que se tem ao final da construção da rede são *clusters* de palavras que possuem algum relacionamento entre si. Cada *cluster* pode, então, representar vários documentos que, mesmo não possuindo rigorosamente as mesmas palavras, podem possuir um contexto semelhante. Com a rede construída é possível se valer de todo o ferramental teórico estudado e atualmente disponível, como a estatística de Tsallis, para sua manipulação.

3.5 Aplicação da Estatística de Tsallis na Clusterização

Esta seção abordará a utilização da Teoria da Informação não-extensiva para análise da rede complexa proposta por este trabalho. A análise da rede é composta por três etapas. A primeira delas é a validação do modelo para representação de dados textuais (Seção 3.5.1). A segunda etapa compreende na aplicação da entropia de Tsallis para encontrar o parâmetro de não-extensividade dos dados (Seção 3.5.2). Finalmente, utilizando os conceitos abordados na Seção 2.5.3, a Divergência J é utilizada para identificação dos centros de clusters (Seção 3.5.3).

3.5.1 Medida de Qualidade de Clusterização

De acordo com a Seção 3.2, a rede complexa proposta neste trabalho é baseada na co-ocorrência de palavras. Porém, deve-se mensurar o quanto este modelo consegue representar os dados. Com este objetivo, elaboramos uma medida baseada na própria base de documentos que gerou a rede.

Nesta avaliação, assumimos que a qualidade da clusterização é melhor quanto mais as palavras de um documento se concentrarem em um único cluster da rede. Isso significa que se todas as palavras de um documento estiverem em um único cluster, a qualidade da clusterização é maior. Porém, se as palavras estiverem dispersas em varios clusters, a qualidade diminui. Essa medida é feita considerando todos os documentos da base e, posteriormente, normalizadas entre 0 e 1. A Equação (3.7) é utilizada para esse cálculo.

$$Q = \frac{\sum_{d \in D} \frac{S(distrClusters(d))}{log(||C||)}}{||D||} \tag{3.7}$$

onde $distrClusters(d)$ é a distribuição de probabilidade das palavras do documento d estarem em cada um dos clusters, $S(x)$ é a entropia de Shannon (veja Seção 2.5.1) de uma distribuição x, $||C||$ é o número de clusters e $||D||$ é o número de documentos. Note que a entropia calculada é dividida pela entropia máxima, resultando em um número entre 0 e 1. A normalização também é feita após a somatória da avaliação de todos os documentos.

O que se espera no final da avaliação para todos os documentos é que a o valor de Q seja próximo à zero. Isso significa que a distribuição das palavras nos clusters deve estar concentrada em algum cluster para que a entropia seja mínima.

3.5.2 Medida de Não-Extensividade dos Dados

Com o objetivo de encontrar o melhor valor do parâmetro não-extensível q para cada cluster da rede e, consequentemente, podendo ser útil para classificar novas palavras em um cluster, é feita a comparação da distribuição de pesos de cada um dos nós de um mesmo cluster da rede.

O método consta em encontrar um valor de q, dado pela Equação (2.20), que minimize a diferença de entropia entre elementos de um mesmo cluster. A Equação (3.8) mostra de maneira formal como deve-se encontrar o valor de q.

$$q_k = argmin \sum_{i \in C_k} \sum_{\substack{j \in C_k, \\ i \neq j}} |S_{q_k}(pesos(i)) - S_{q_k}(pesos(j))| \tag{3.8}$$

onde q_k é o valor de q que minimiza a diferença das entropias entre todos os elementos de um mesmo cluster C_k, S_{q_k} é a entropia não-extensiva de Tsallis (veja Seção 2.5.2) e $pesos(i)$ é a distribuição de pesos de todas as conexões de um nó i com os outros nós do mesmo cluster.

De posse dos valores de q para cada cluster, é possível utilizar a comparação de entropias das distribuições de pesos para classificar um nova palavra na rede complexa. Isso pode permitir maior eficiência na busca de um cluster ideal para essa palavra.

3.5.3 Identificação dos Centros de Clusters

É fundamental encontrar elementos de um cluster que consigam representá-lo de forma abrangente. Isso permite eficiência na navegação, comparação e representação dos dados. Na maioria dos trabalhos que envolvem representação de clusters, um elemento central é criado a partir da média dos elementos do mesmo cluster. Neste projeto, propomos um método semelhante para encontrar os elementos centrais de cada cluster, porém utilizamos a divergência J (veja Seção 2.5.3) para localizar o nó mais próximo da cada um desses elementos centrais.

Inicialmente, um histograma h_k é criado a partir da média de todos os histogramas de pesos entre os elementos presentes em um mesmo cluster k. Consideramos que esse histograma representa o centro do cluster k. Após essa operação, comparamos através da Divergência J qual é o elemento do cluster k que apresenta distribuição de pesos mais próxima à h_k.

Capítulo 4

Resultados Preliminares

Neste capítulo serão apresentados e discutidos os resultados preliminares obtidos a partir dos métodos propostos no Capítulo 3. Todos os experimentos foram realizados em um computador convencional com memória principal de 4 GB. Dessa forma, todos os algoritmos para a extração dessas características tiveram que ser adaptados para carregar a matriz AD em pequenas matrizes, AD_m, conforme visto na Seção 3.3.

Outras características físicas propostas na Seção 3.3 serão implementadas futuramente para uma análise mais minuciosa da rede. Isso permitirá uma compreensão melhor, tornando possível a escolha ou desenvolvimento de algoritmos específicos como, por exemplo, de clusterização.

4.1 Distribuição de Pesos

O histograma da distribuição de pesos de uma rede complexa permite uma análise abrangente sobre como os nós se conectam entre si. Além disso, o histograma ilustra sua densidade, o que pode ser viável para decisões de modelagem. Por exemplo, a Fig. 4.1 representa o histograma da rede complexa proposta neste trabalho.

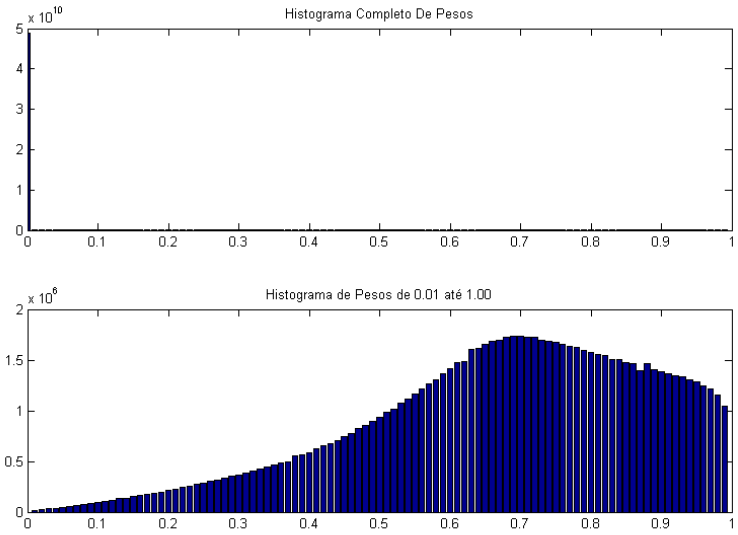


Figura 4.1: Histograma de pesos da rede complexa para a base *oshumed*. Topo: incluindo os pesos "zeros". Baixo: pesos diferentes de zero.

O primeiro histograma da Fig. 4.1 compreende todas as faixas de pesos possíveis para uma aresta (entre 0 e 1). Nota-se que há uma grande quantidade de arestas que possuem valor 0 e por este motivo fica claro que a rede é esparsa. Com o objetivo de ilustrar os outros pesos possíveis da rede, eliminamos do histograma a primeira faixa de valores (de 0 à 0.01) e geramos o segundo histograma. É possível perceber que o segundo valor mais frequente está na faixa de 0.7. Conclui-se então que, quando dois nós se conectam, esta conexão é normalmente forte o que nos indica que pode haver possível formação de clusters.

4.2 Distribuição de Graus

Conforme explicado na Seção 2.2.1.3, o modelo de rede livre de escala possui a característica da lei de potência em sua distribuição de graus. Por esse motivo, é importante observar o comportamento dessa distribuição para entender se o sistema se comporta de acordo com as definições de redes livres de escala. A Fig. 4.2 ilustra o histograma de graus para a rede proposta nesse trabalho.

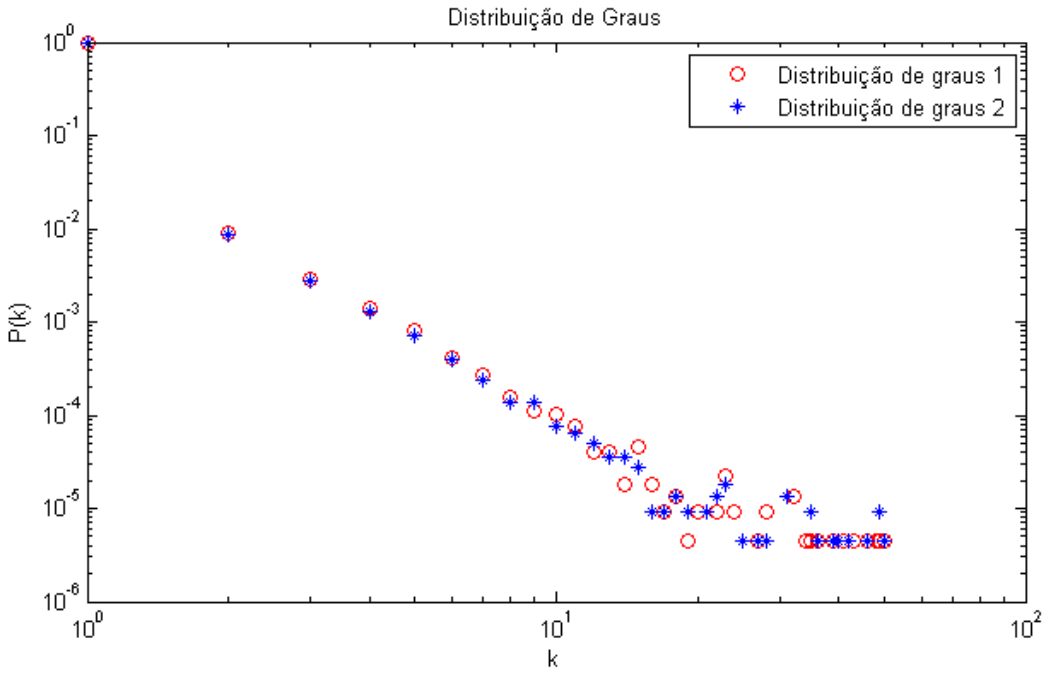


Figura 4.2: Lei de potência da base de dados *oshumed*

Na Fig. 4.2, o valor k de um nó foi calculado de forma diferente para as duas distribuições. Na distribuição em vermelho, o valor k é a quantidade de arestas que entram e saem de um nó. No caso da distribuição em azul, o k é obtido a partir da somatória de todas as arestas que entram e saem de um nó. A Equação (4.1) demonstra o cálculo do grau de um nó i.

$$k_i = \sum_j M_{i,j} + M_{j,i} \qquad \text{Para } i \neq j \tag{4.1}$$

A Fig. 4.2 mostra que a rede possui uma distribuição de graus conforme a equação da lei de potência (Equação (2.4)). Isso é um forte indício de que a rede se comporta de acordo com o modelo de rede livre de escala.

4.3 Cálculo do Coeficiente de Clusterização

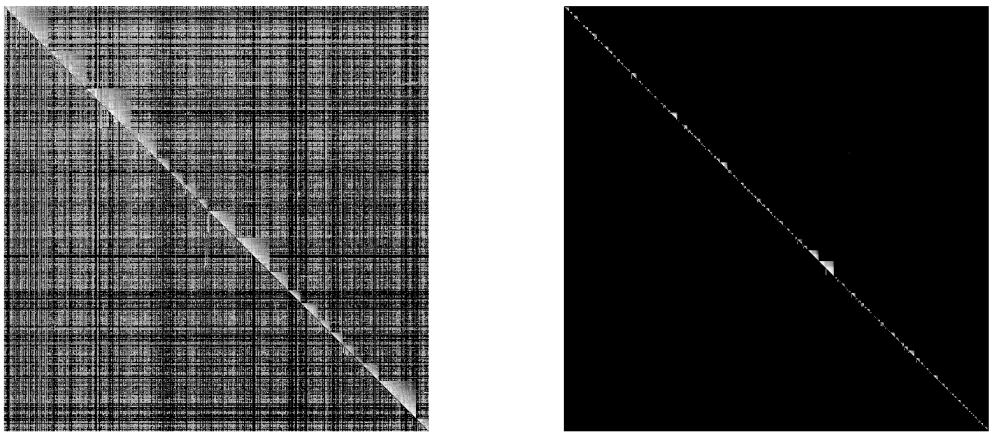
Nesta Seção, apresentamos e discutimos os resultados obtidos a partir dos dois algoritmos desenvolvidos na Seção 3.3.2. Na Seção 4.3.1, são apresentados os resultados do método

de Simplificação Baseada em Imagens (Seção 3.3.2.1). Na Seção 4.3.2 são apresentados os resultados obtidos a partir do método de Simplificação Estatística (Seção 3.3.2.2).

4.3.1 Experimentos com a Simplificação Baseada em Imagens

Com o objetivo de mensurar a precisão da técnica de redução de imagens e comparar os resultados com a técnica aleatória, conduzimos um experimento com matrizes pequenas ($N \leq 10000$), cujo coeficiente de clusterização real é possível de se calcular em tempo razoável e comparamos nossos resultados com o coeficiente de clusterização real.

Duas matrizes de nossa base foram montadas para o experimento. A primeira matriz (Veja Fig. 4.3(a)) foi extraída de uma partição densa da rede complexa e a segunda matriz (Fig. 4.3(b)) de uma partição esparsa da mesma rede. O tamanho de ambas as matrizes é $N = 8847$ nós. Porém, a primeira matriz tem probabilidade de conexão $p = 0.52$ e a segunda tem $p = 0.004$. A primeira matriz tem $CC = 0.46$ e a segunda matriz $CC = 0.34$, de acordo com a Equação (2.9).



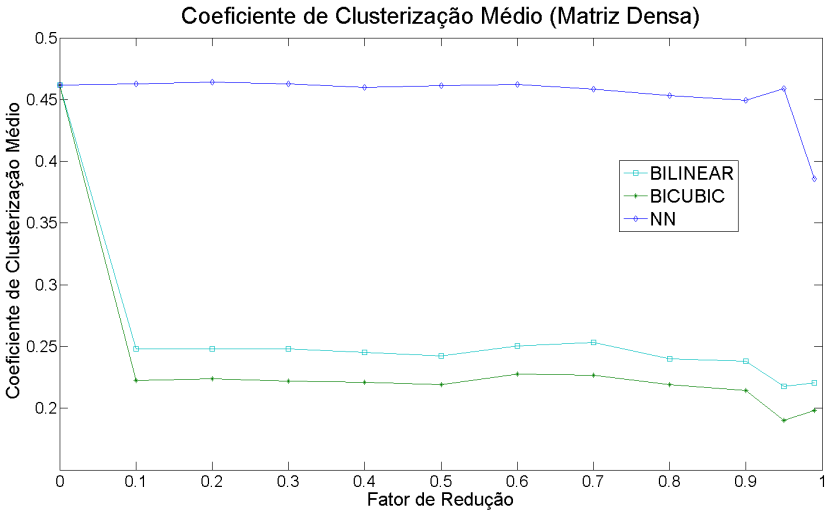
(a) Matriz de adjacência ponderada densa (b) Matriz de adjacência ponderada esparsa

Figura 4.3: Os dois tipos de matrizes estudados no experimento.

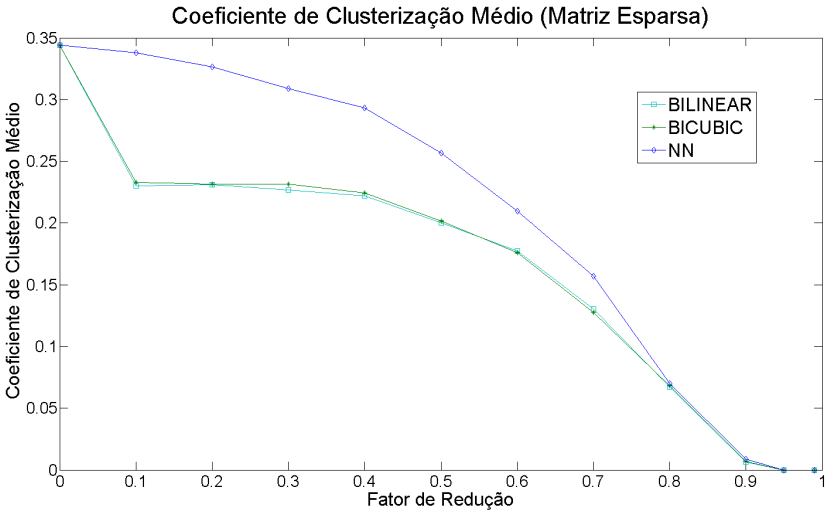
Os resultados são apresentados na Fig. 4.4 e claramente mostra que a redução NN teve

valores mais próximos do real (fator de redução zero). Para ambas as figuras, a o fator de redução 0 representa o coeficiente de clusterização real.

Podemos observar também que, quando a matriz é densa, um fator de redução de 90% tem pouco efeito sobre o coeficiente de clusterização neste caso. Além disso, o *NN* tem outra vantagem sobre as interpolações *bilinear* e *bicubic*: para uma matriz densa, o *CC* se torna constante (em torno de 0.46) até o fator de redução 0.9.



(a) Coeficiente de Clusterização em Matriz Densa

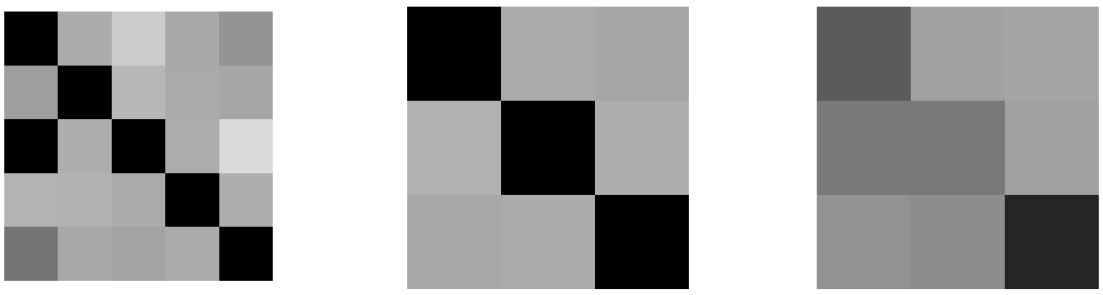


(b) Coeficiente de Clusterização em Matriz Esparsa

Figura 4.4: Resultado do experimento para os dois tipos de matrizes. Note que, nos dois casos, a redução por *NN* foi a que mais se aproximou do Coef. Clust. Médio real.

Comparado com *NN*, os métodos de redução *Bilinear* e *Bicubic* foram insatisfatórios uma vez que a distribuição de pesos da matriz de adjacência foi maciçamente alterada.

Metodos que utilizam funções de interpolação podem criar novas conexões, o que afeta diretamente o cálculo do coeficiente de clusterização, um vez que o número de vizinhos pondera o relacionamento entre eles. Por exemplo, se aplicarmos uma redução *Bilinear* em uma imagem, os pixels mudarão a medida que novas conexões são criadas na rede devido à esta interpolação (veja Figs. 4.5(c) e 4.6(c)).

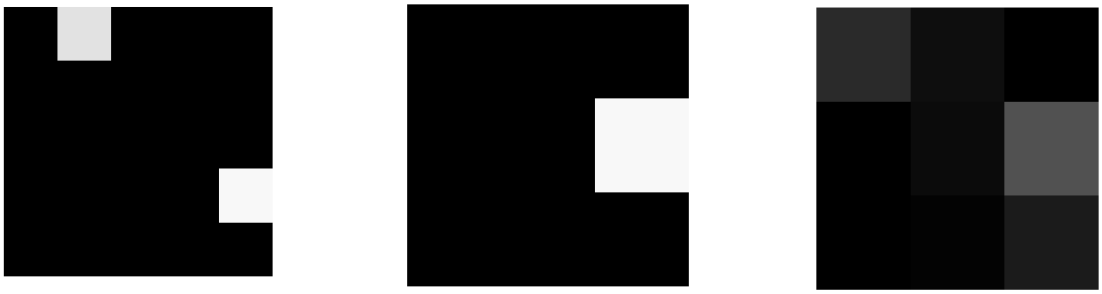


(a) Pequena Matriz Densa.

(b) Fig. 4.5(a) reduzida 50% com NN e ampliada 2x para fins de visualização.

(c) Fig. 4.5(a) reduzida 50% com interpolação Bilinear e ampliada 2x para fins de visualização.

Figura 4.5: Efeitos da Redução em Matrizes Densas.



(a) Pequena Matriz Esparsa.

(b) Fig. 4.6(a) reduzida 50% com NN e ampliada 2x para fins de visualização.

(c) Fig. 4.6(a) reduzida 50% com interpolação Bilinear e ampliada 2x para fins de visualização.

Figura 4.6: Efeitos da Redução em Matrizes Esparsas.

As Figs. 4.5 e 4.6 mostram como a interpolação pode afetar a aparência das imagens reduzidas. Por exemplo, ambas interpolações *Bilinear* nas Figs. 4.5(c) e 4.6(c) modificam diversos pixels da imagem original após a aplicação da redução. Como a interpolação funciona como uma média dos pixels que estão próximos a amostra da imagem reduzida, os pixels que possuem valores zerados tendem a se tornar de maior valor. Isso significa que novas conexões são criadas quando se aplica tal método. Finalmente, podemos concluir que o método de interpolação não pode mudar os valores dos pixels, principalmente os pixels que têm valores zerados, uma vez que o número de vizinhos deve ser proporcionalmente (na potência de 2) reduzido.

Em contrapartida, o método *Nearest Neighborhood* (NN) escolhe o pixel que está mais próximo da grade de redução (veja Seção 3.3.2.1). O efeito disso do ponto de vista das redes complexas é que alguns nós são excluídos da rede. Contudo, o número de conexões é proporcional à rede original. Se a rede possui muitas conexões (Matriz Densa), o NN terá muitos nós para excluir, porém, se a matriz for esparsa, remover alguns nós resulta na perda de informações importantes, uma vez que a rede possui poucas arestas, o que explica o baixo desempenho desse algoritmo no caso de matriz esparsa (Fig. 4.4(b)).

4.3.2 Experimentos com a Simplificação Estatística

A *propriedade fundamental* do NN (definida na Seção 3.3.2.1) diz que a proporção dos valores dos vizinhos ao redor de um pixel i tende a se manter após a aplicação desse algoritmo (Fig. 4.4(a), curva NN). Este efeito parece ser atenuado quando a matriz é esparsa (Fig. 4.4(b), curva NN). Neste gráfico da redução NN para matriz densa (Fig. 4.4(a)), é também observado que a média dos coeficientes de clusterização se mantém estável em torno de 0.46 até uma redução extrema de quase 95%, o que significa que a imagem reduzida de 440 x 440 pixels de nossa base afeta a *propriedade fundamental*, e resulta em um comportamento equivalente para matrizes esparsas (Fig. 4.4(b)).

A ideia de se tomar pixels aleatoriamente e computar uma média aceitável do Coeficiente de Clusterização possibilita a seguinte especulação. De acordo com a matriz original (sem nenhuma redução) e um subconjunto aleatório de nós, é possível computar uma média aceitável do CC. No trabalho de [Schank and Wagner, 2004] foi apresentado um algoritmo baseado na mesma ideia aleatória, porém com amostras de triplas (conexões com forma de triângulos) com probabilidade uniforme. Os resultados foram satisfatórios, uma vez que a complexidade do algoritmo foi reduzida sem impactar significantemente a média do coeficiente de clusterização.

Dessa forma, propomos o estudo de como o subconjunto de nós, escolhidos aleatoriamente de uma rede complexa ponderada de palavras, influencia o cálculo do coeficiente de clusterização médio, para matrizes esparsas e densas, possibilitando o uso do método proposto numa ampla gama de aplicações.

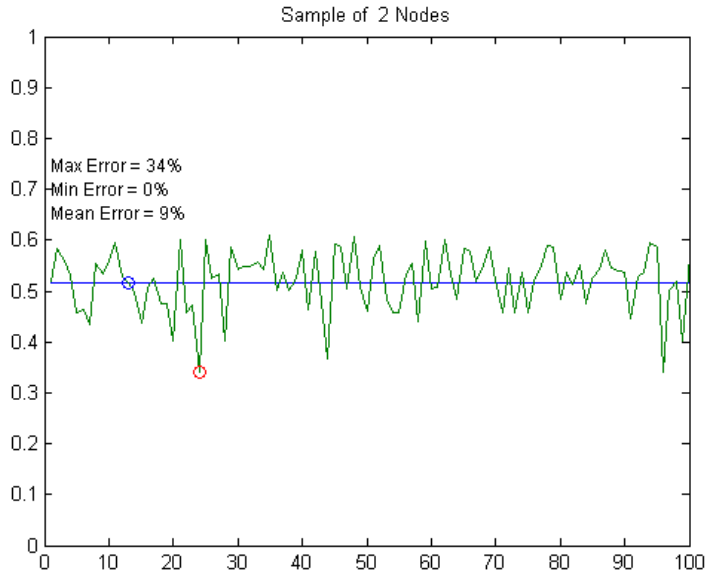
Com o objetivo de quantificar a tolerância dessa aproximação, conduzimos um experimento em uma matriz pequena ($N = 2000$). Primeiramente, para efeito de comparação, computamos o coeficiente de clusterização real utilizando a Equação (2.8) para todos os nós da rede. Alterando o número de nós no conjunto de amostras e fazendo 100 testes para cada simulação, foi possível medir experimentalmente a tolerância do cálculo (Veja Figs. 4.7, 4.8 e Tabela 4.1).

A Fig. 4.7 mostra os resultados da simulação para um número crescente de nós esco-

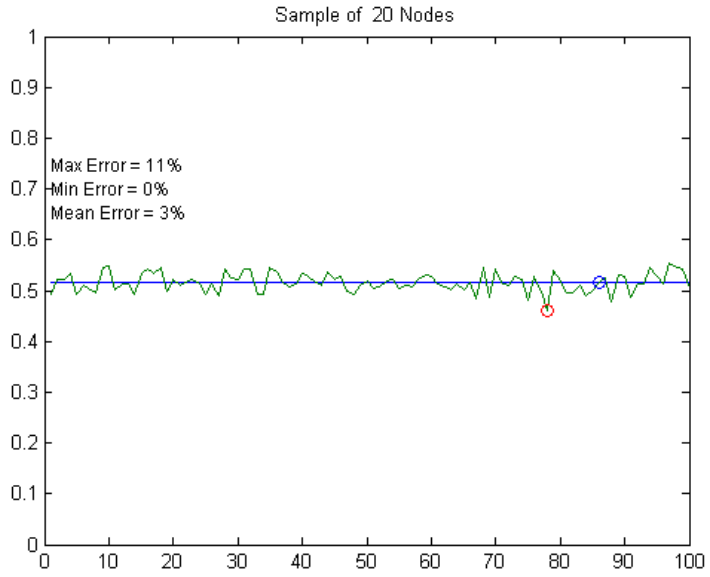
lhidos aleatoriamente. Em cada gráfico da Fig. 4.7, a linha sólida central ℓ representa o coeficiente de clusterização médio real computado de acordo com a Equação (2.9) para a matriz original de $N = 2000$ nós. Essa matriz é similar àquela utilizada na Seção 4.3.1 (Fig. 4.3(a)), que corresponde a uma matriz densa com $p = 0.421$. O eixo horizontal representa os 100 diferentes conjuntos de nós aleatórios e o eixo vertical o coeficiente de clusterização médio correspondente (CC).

A Fig. 4.8 mostra o experimento correspondente da Fig. 4.7, porém para uma matriz esparsa similar àquela utilizada na Seção 4.3.1 (Fig. 4.3(b)) com $N = 2000$ nós e $p = 0.0015$. Na Fig. 4.8, os eixos verticais e horizontais e a linha ℓ possuem a mesma interpretação da Fig. 4.7.

Na Fig. 4.7, conforme a amostra aumenta (gráficos da esquerda para direita e do topo para baixo), a estabilidade do método aleatório também aumenta, se aproximando do valor ótimo representado pela linha ℓ. Mesmo para 1% do total de nós que são considerados no cálculo, como mostrado na Fig. 4.7(b), a estabilidade é aceitável (erro médio de 3%), o que significa uma vantagem sobre o método baseado em redução de imagens. Para matrizes esparsas (Fig 4.8), o método estatístico também se comporta com boa estabilidade (erro médio de 3%) mesmo para uma amostra com somente 1% dos nós.

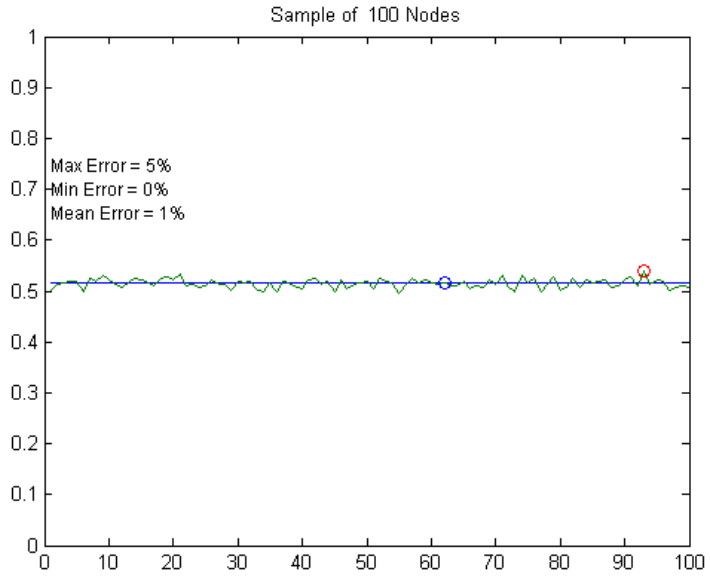


(a) Resultado dos experimentos para 0.1% dos nós

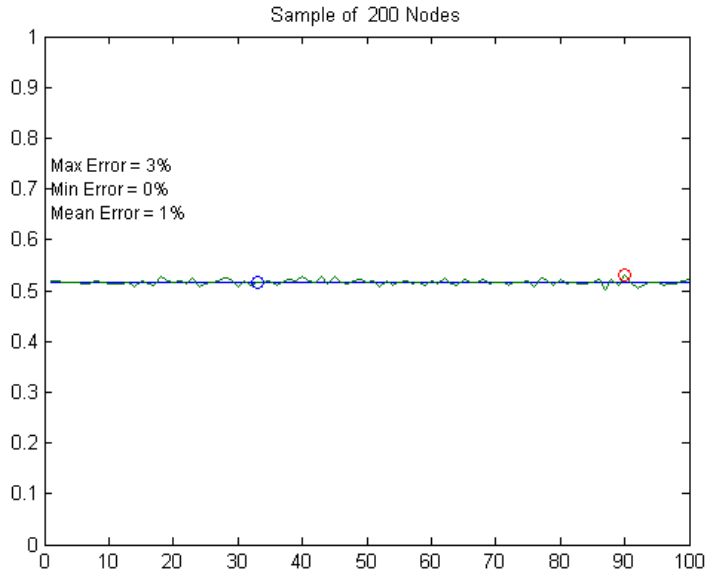


(b) Resultado dos experimentos para 1% dos nós

Figura 4.7: Resultados para 100 experimentos utilizando a Simplificação Estatística em uma Matriz Densa (Continua...).

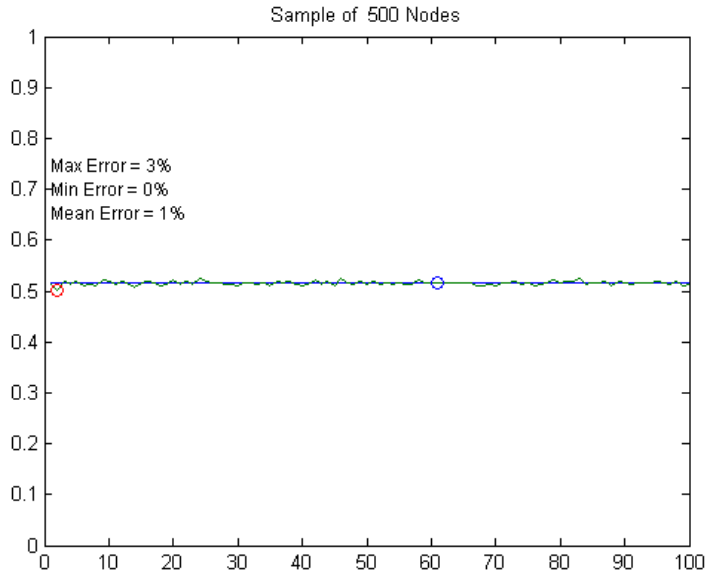


(c) Resultado dos experimentos para 5% dos nós

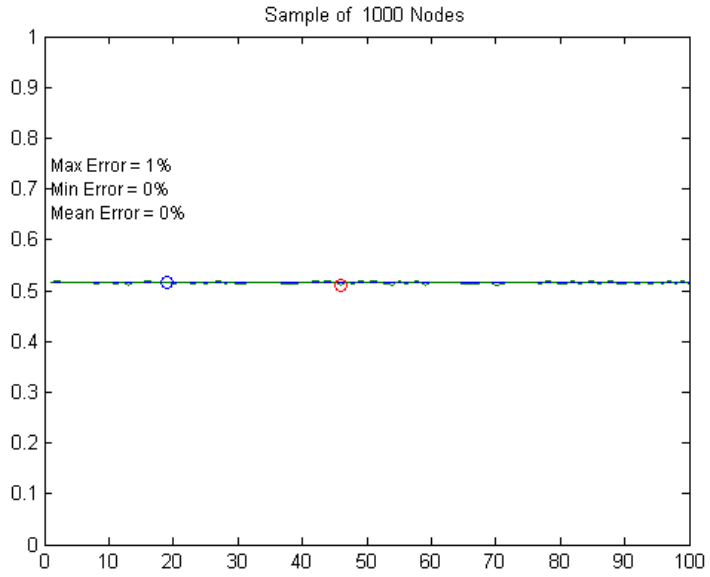


(d) Resultado dos experimentos para 10% dos nós

Figura 4.7: Resultados para 100 experimentos utilizando a Simplificação Estatística em uma Matriz Densa (Continua...).

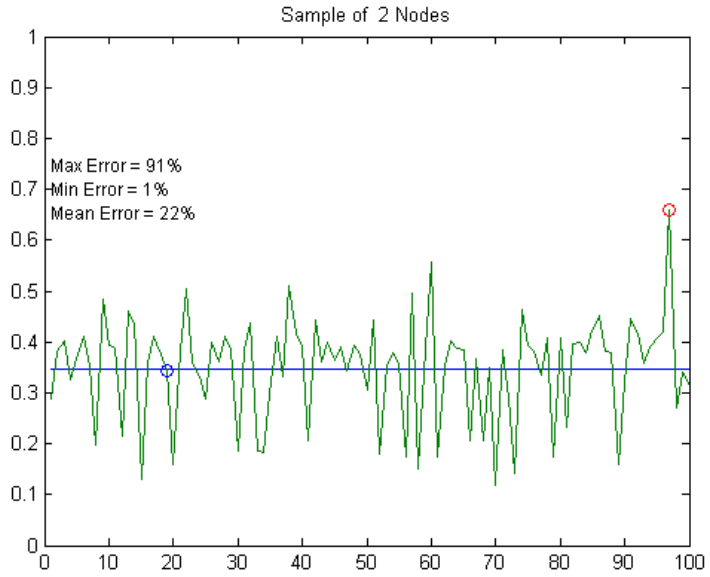


(e) Resultado dos experimentos para 25% dos nós

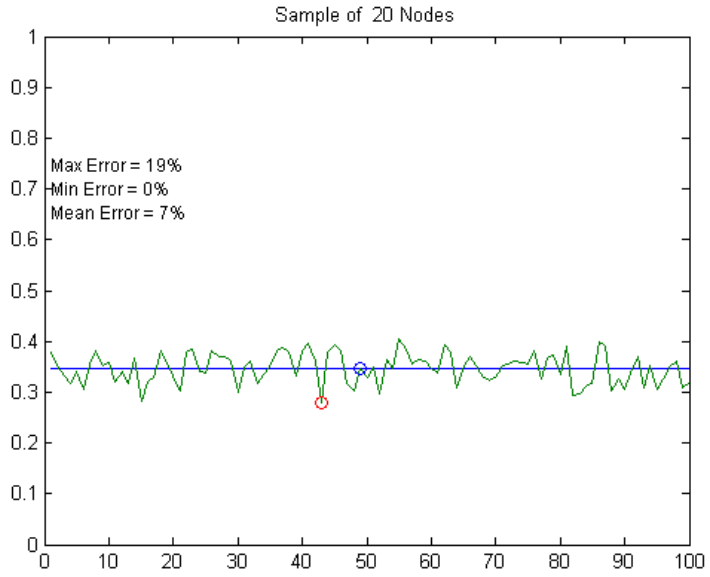


(f) Resultado dos experimentos para 50% dos nós

Figura 4.7: Resultados para 100 experimentos utilizando a Simplificação Estatística em uma Matriz Densa.

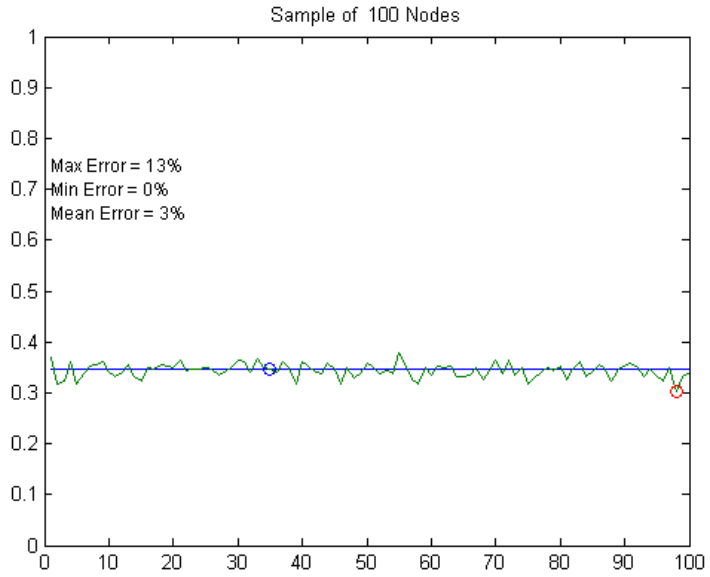


(a) Resultado dos experimentos para 0.1% dos nós

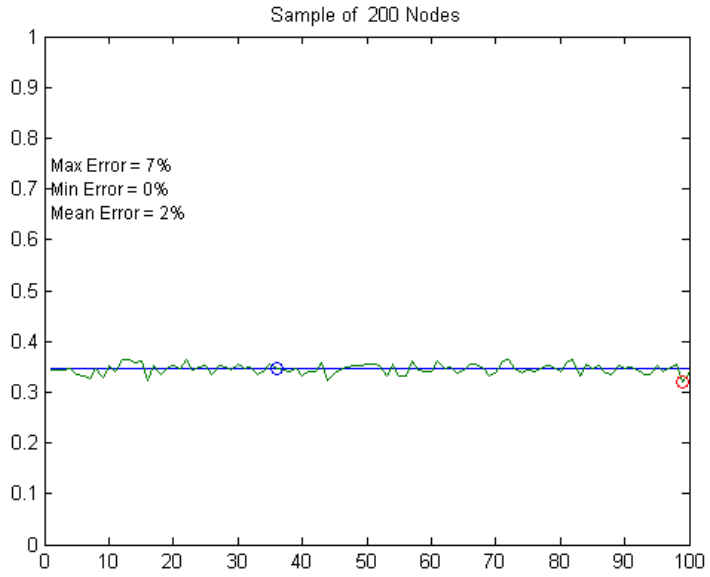


(b) Resultado dos experimentos para 1% dos nós

Figura 4.8: Resultados para 100 experimentos utilizando a Simplificação Estatística em uma Matriz Esparsa. (Continua...)

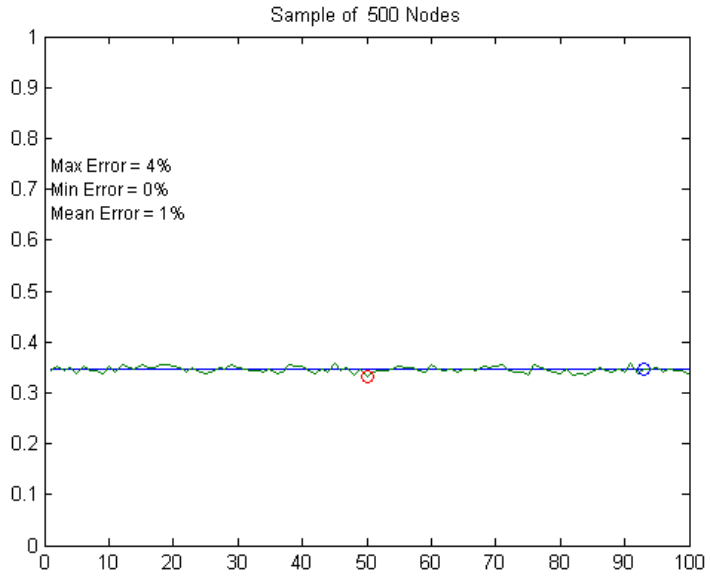


(c) Resultado dos experimentos para 5% dos nós

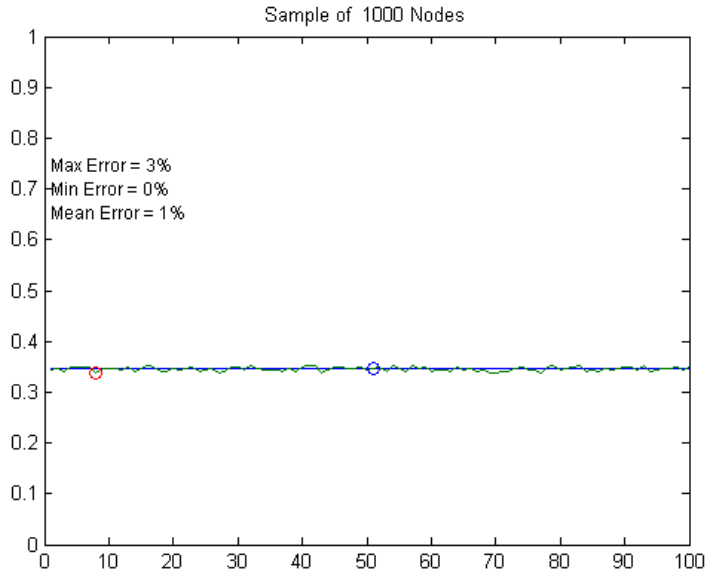


(d) Resultado dos experimentos para 10% dos nós

Figura 4.8: Resultados para 100 experimentos utilizando a Simplificação Estatística em uma Matriz Esparsa. (Continua...)



(e) Resultado dos experimentos para 25% dos nós



(f) Resultado dos experimentos para 50% dos nós

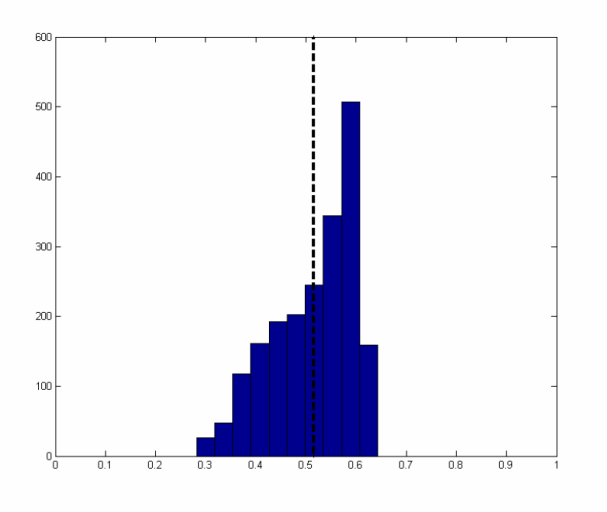
Figura 4.8: Resultados para 100 experimentos utilizando a Simplificação Estatística em uma Matriz Esparsa

Erros						
	Matriz Densa			Matriz Esparsa		
Amostra	Min	Max	Média	Min	Max	Média
0.1%	0%	34%	9%	1%	91%	22%
1%	0%	11%	3%	0%	19%	7%
5%	0%	5%	1%	0%	13%	3%
10%	0%	3%	1%	0%	7%	2%
25%	0%	3%	1%	0%	4%	1%
50%	0%	1%	0%	0%	3%	1%

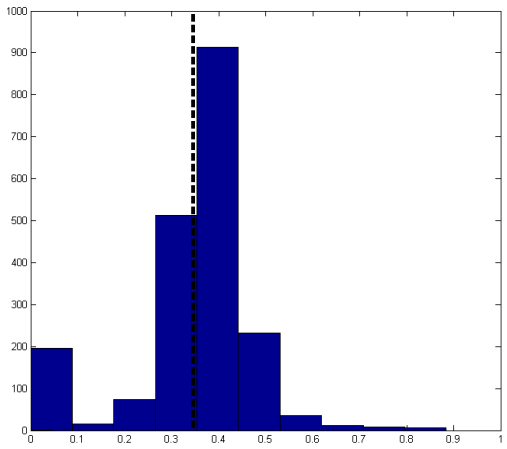
Tabela 4.1: Erros no Método de Simplificação Estatístico

A Tabela 4.1 mostra os percentuais do erros mínimo, máximo e médio para matrizes densas e esparsas. Os percentuais de amostras são mostrados na primeira coluna. Esta tabela indica que o erro aumenta à medida que as matrizes se tornam esparsas, porém com uma tolerância ainda aceitável.

Para matrizes densas, este efeito é menor uma vez que elas tendem a ter melhores distribuições dos valores de CC_i pelos nós. A Fig. 4.9 compara a distribuição dos valores de CC_is nas matrizes densas e esparsas. Notavelmente, a matriz esparsa tem um histograma com um vasto leque em torno da média do CC, deferentemente da matriz densa (O histograma da Fig 4.9(a) tem uma distribuição variando de 0.3 à 0.65 e a Fig. 4.9(b) um histograma entre 0.0 e 0.9).



(a) Histograma dos coeficientes de clusterização para uma matriz densa.



(b) Histograma do coeficientes de clusterização para uma matriz esparsa.

Figura 4.9: Histogramas dos coeficientes de clusterização.

Conforme dito anteriormente, ambos os tipos de matrizes geram um coeficiente de clusterização médio com erros maiores a medida que o tamanho da amostra diminui. Uma

vez que uma amostra pequena representa pouco sua distribuição. Então, o coeficiente de clusterização médio para matrizes esparsas tende a ser mais instável. A razão para esta baixa estabilidade para matrizes esparsas é que uma baixa densidade pode resultar, como dito anteriormente, em uma faixa maior em torno da média do coeficiente de clusterização, dificultando a escolha por nós aleatórios que representam melhor a distribuição.

Finalmente, comparando o método aleatório com a Redução de Imagem por NN, a maior diferença é que, no caso do NN, o cálculo do CC_i para cada nó é feito usando uma matriz reduzida da original. Porém, o CC_i, para cada método estatístico, é calculado na matriz original, o que resulta em um coeficiente de clusterização médio mais próximo ao real.

4.3.3 Conclusão Preliminar

Os resultados dos experimentos relacionados com a distribuição de pesos (Seção 4.1) indicam que a rede complexa proposta nesse trabalho é esparsa. Quanto a distribuição de pesos, os resultados indicam que a rede possui características de redes livres de escala, uma vez que apresentou o comportamento da lei de potência. Finalmente, o método do cálculo do coeficiente de clusterização aleatório mostrou-se mais preciso que o método com redução de imagens, devido não apresentar mudanças significativas entre matrizes densas e esparsas.

Capítulo 5

Resultados e Discussão

Capítulo 6

Conclusões

Bibliografia

[Gov, 2000] (2000). *Heuristics for Internet map discovery*, volume 3.

[Ben, 2005] (2005). A fast approach for dimensionality reduction with image data. *Pattern Recognition*, 38.

[Albert and Barabási, 2002] Albert, R. and Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

[Albert et al., 1999] Albert, R., Jeong, H., and Barabasi, A. L. (1999). The diameter of the world wide web. *Nature*, 401:130–131.

[Albert et al., 2000] Albert, R., Jeong, H., and Barabási, A.-L. (2000). Attack and error tolerance of complex networks. *Nature*, (406):376–382.

[Baeza-Yates and Ribeiro-Neto, 1999] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley, 1st edition.

[Barabasi and Bonabeau, 2003] Barabasi, A.-L. and Bonabeau, E. (2003). Scale-free networks. *Scientific American*, pages 50–59.

[Boghosian, 1995] Boghosian, B. M. R. (1995). *J. Mol. Liq.* Number 4754. 53 edition.

[Boltzmann, 1864] Boltzmann, L. (1864). *Lectures on Gas Theory (English Translation)*. Berkeley.

[Borland et al., 1998] Borland, L., Plastino, A. R., and Tsallis, C. (1998). Information gain within nonextensive thermostatistics. *Journal of Mathematical Physics*, 39:6490–6501.

[Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Comput. Netw.*, 33(1-6):309–320.

[Cohen et al., 2000] Cohen, R., Erez, K., Ben-Avraham, D., and Gavlin, S. (2000). Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, (85):3682–3685.

[Cormen et al., 2001] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. The MIT Press, 2nd revised edition edition.

[de Pinho Tavares, 2003] de Pinho Tavares, A. H. M. (2003). Aspectos matemáticos da entropia. Master's thesis, Universidade de Aveiro.

[Duda et al., 2000] Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2 edition.

[Eakins, 2002] Eakins, J. P. (2002). Towards intelligent image retrieval. *Pattern Recognition*, 35(1):3–14.

[Erdős and Rényi, 1959] Erdős, P. and Rényi, A. (1959). On random graphs. I. *Publ. Math. Debrecen*, 6:290–297.

[Faloutsos et al., 1999] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262, New York, NY, USA. ACM.

[Farkas et al., 2002] Farkas, I. J., Jeong, H., Vicsek, T., Barabási, A. L., and Oltvai, Z. N. (2002). The topology of the transcription regulatory network in the yeast, s. cerevisiae. *Physica A*, 318(cond-mat/0205181):3–4. 18 p.

[Ferrer et al., 2001] Ferrer, R., Cancho, and Solé, R. V. (2001). The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268:2261–2266.

[G. W. Furnas and Lochbaum, 1988] G. W. Furnas, S. Deerwester, S. T. D. T. K. L. R. A. H. L. A. S. and Lochbaum, K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure. In *SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–480. ACM.

[Gonzalez and Woods, 2002] Gonzalez, R. and Woods, R. (2002). *Digital Image Processing*. Prentice-Hall, New Jersey, 2nd edition.

[Guelzim et al.,] Guelzim, N., Bottani, S., Bourgine, P., and Kepes, F.

[Harary, 1969] Harary, F. (1969). *Graph Theory*. Addison-Wesley.

[Hersh et al., 1994] Hersh, W., Buckley, C., Leone, T. J., and Hickam, D. (1994). Ohsumed: an interactive retrieval evaluation and new large test collection for research. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–201, New York, NY, USA. Springer-Verlag New York, Inc.

[Hidalgo et al., 2007] Hidalgo, C. A., Klinger, B., Barabási, A.-L., and Hausmann, R. (2007). The product space conditions the development of nations. *Science*, 317:482.

[Hilgetag et al., 2000] Hilgetag, C. C., Burns, G. A., O'Neill, M. A., Scannell, J. W., and Young, M. P. (2000). Anatomical connectivity defines the organization of clusters of cortical areas in the macaque monkey and the cat. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 355(1393):91–110.

[Jain, 2010] Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

[Jeffreys, 1939] Jeffreys, H. (1939). *Theory of probability*. Oxford University Press, second edition.

[Legány et al., 2006] Legány, C., Juhász, S., and Babos, A. (2006). Cluster validity measurement techniques. In *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases*, pages 388–393, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS).

[Li et al., 2008] Li, Y., Chung, S. M., and Holt, J. D. (2008). Text document clustering based on frequent word meaning sequences. *Data Knowl. Eng.*, 64(1):381–404.

[Liljeros et al., 2001] Liljeros, F., Edling, C. R., Amaral, L. A., Stanley, E. H., and åberg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907–908.

[Lloyd, 2003] Lloyd, S. (2003). Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137.

[Macqueen, 1967] Macqueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Procedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability*, volume 1, pages 281–297. University of California Press.

[Newman et al., 2006] Newman, M., Barabasi, A.-L., and Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton University Press.

[Newman, M. E. J. and Girvan, M., 2004] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113+.

[Nicosia et al., 2010] Nicosia, V., Mangioni, G., Carchiolo, V., and Malgeri, M. (2010). Extending the definition of modularity to directed graphs with overlapping communities. In *II International Workshop on Complex Networks*, Rio de Janeiro, Brazil.

[Noel and Jajodia, 2005] Noel, S. and Jajodia, S. (2005). Understanding complex network attack graphs through clustered adjacency matrices. In *ACSAC '05: Proceedings of the 21st Annual Computer Security Applications Conference*, pages 160–169, Washington, DC, USA. IEEE Computer Society.

[Parker et al., 1983] Parker, J. A., Kenyon, R. V., and Troxel, D. E. (1983). Comparison of interpolating methods for image resampling. *IEEE Transactions on Medical Imaging*, MI-2:31–39.

[Price, 1965] Price, D. d. S. (1965). Network of scientific papers. *Science*, (149):510–515.

[Redner, 1998] Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 4(2):131–134.

[Ribeiro and Muntz, 1996] Ribeiro, B. A. N. and Muntz, R. (1996). A belief network model for ir. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 253–260, New York, NY, USA. ACM.

[Ribeiro et al., 2000] Ribeiro, B. A. N., Silva, I., and Muntz, R. (2000). Bayesian network models for ir. *Soft Computing in Information Retrieval Techniques and Applications*, pages 1259–291.

[Rodrigues and Giraldi, 2009] Rodrigues, P. and Giraldi, G. (2009). Computing the q-index for tsallis nonextensive image segmentation. In *XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI)*, pages 232–237, Rio de Janeiro, RJ, Brazil.

[Rodrigues and Giraldi, 2010] Rodrigues, P. S. and Giraldi, G. A. (to appear, 2010). Improving the non-extensive medical image segmentation based on tsallis entropy. *Pattern Analysis and Applications*.

[Santos, 1997] Santos, R. J. V. (1997). Generalization of shannon's theorem for tsallis entropy. *J. Math. Phys.*, 38:4104–4107.

[Schaeffer, 2007] Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27 – 64.

[Schank and Wagner, 2004] Schank, T. and Wagner, D. (2004). Approximating clustering-coefficient and transitivity. *Journal of Graph Algorithms and Applications*, 9(2):265–275.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

[Shannon and Weaver, 1948] Shannon, C. and Weaver, W. (1948). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

[Shen-Orr et al., 2002] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature genetics*, 31(1):64–68.

[Solomonoff and Rapoport, 1951] Solomonoff, R. and Rapoport, A. (1951). Connectivity of random nets. *Bulletin of Mathematical Biophysics*, (13):107–117.

[Sporns, 2002] Sporns, O. (2002). Graph theory methods for the analysis of neural connectivity patterns. *Complex.*, 8(1):56–60.

[Stauffer and Aharony, 1992] Stauffer, D. and Aharony, A. (1992). *Introduction to percolation theory / Dietrich Stauffer and Amnon Aharony*. Taylor and Francis, London, 2nd ed. edition.

[Steinhaus, 1956] Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, 1:801–804.

[Tsallis, 1988] Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479–487. 10.1007/BF01016429.

[Tsallis, 1999] Tsallis, C. (1999). Nonextensive statistics: Theoretical, experimental and computational evidences and connections. *Brazilian Journal of Physics*, 29(1).

[Tsallis, 2001] Tsallis, C. (Springer, Berlin, 2001). Nonextensive statistical mechanics and its applications. *Series Lecture Notes in Physics*. The bibliography of the subjects is regularly updated at http://tsallis.cat.cbpf.br/biblio.htm.

[Urbain et al., 2009] Urbain, J., Goharian, N., and Frieder, O. (2009). A dimensional retrieval model for integrating semantics and statistical evidence in context for genomics literature search. *Computers in Biology and Medicine*, 39:61–68.

[van Dongen, 2000] van Dongen, S. M. (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, The Netherlands.

[Walsh, 1999] Walsh, T. (1999). Search in a small world. In *IJCAI '99: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1172–1177, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of /'small-world/' networks. *Nature*, 393(6684):440–442.

[Wigner, 1955] Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Ann. Math.*, 62:548–564.

[Wigner, 1957] Wigner, E. P. (1957). Characteristic vectors of bordered matrices with infinite dimensions ii. *Ann. Math.*, 65:203–207.

[Wigner, 1958] Wigner, E. P. (1958). On the distributions of the roots of certain symmetric matrices. *Ann. Math.*, 67:325–327.

[Witten et al., 1999] Witten, I. H., Moffat, A., and Bell, T. C. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, San Francisco, CA.

[Yan et al., 2009] Yan, D., Huang, L., and Jordan, M. I. (2009). Fast approximate spectral clustering. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 907–916, New York, NY, USA. ACM.

[Ziviani, 1996] Ziviani, N. (1996). *Projeto de Algoritmos*. Pioneira, 3 edition.

Apêndice A

Apêndice: Artigo submetido para revista "Information Sciences" em 14/09/2010 (aguardando resposta)