Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Estabilidade de Métricas de Redes Complexas

Paulo R. A. Margarido
Maria das G. V. Nunes
Thiago A. S. Pardo
Osvaldo N. de Oliveira Jr.
Luciano da F. Costa

NILC-TR-08-03
Março, 2008

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

Resumo

Neste relatório será apresentado um experimento realizado com base em experimentos do NILC (Núcleo Interinstitucional de Linguística Computacional) que teve como finalidade obter algumas informações sobre a dinâmica das redes complexas, enquanto representação de textos para uso em atividades de PLN (Processamento de Língua Natural). Este trabalho é focado especificamente na estabilidade de diversas medidas que fornecem informações sobre uma dada rede, de maneira a desenvolver uma base de medidas confiáveis que possa ser utilizada em experimentos futuros, permitindo assim obter resultados mais expressivos e robustos. Outro objetivo do trabalho aqui apresentado é determinar um tamanho de texto que gerasse redes com comportamento razoavelmente previsível, de maneira a determinar um limite mínimo para o tamanho de textos a serem trabalhados usando redes complexas para PLN.

Índice

1 Introdução **2**

2 Redes complexas **3**
2.1 Definição de redes complexas 3
2.2 Medidas utilizadas 4

3 Definição do experimento **7**
3.1 Descrição do corpus 7
3.2 Metodologia 9

4 Resultados **10**

5 Conclusões **15**

Referências Bibliográficas **16**

Capítulo 1

Introdução

As redes complexas têm recentemente recebido grande visibilidade em várias áreas da computação devido à sua flexibilidade para representar diversos tipos de problema. Seguindo esta tendência de aumento de utilização, um método foi desenvolvido para representar textos como redes complexas. Com isso, permitiu-se que várias aplicações clássicas de PLN fossem realizadas utilizando as redes complexas como método de representação de textos [4], como, por exemplo, as áreas de atribuição de autoria [3], avaliação de qualidade [2], sumarização automática, tradução automática, entre muitas outras.

Para aumentar a efetividade deste método de representação dos textos é desejável que as redes construídas sejam tão robustas quanto possível, garantindo assim que a base utilizada para a realização de experimentos seja confiável. O experimento descrito neste relatório tem o intuito de analisar a dinâmica da construção das redes [1], de modo a determinar quais métricas são mais estáveis e gerar dados que sirvam de base para que experimentos futuros tenham melhores resultados.

Outro objetivo deste trabalho é determinar um limite inferior confiável para o tamanho dos textos modelados como redes para que as métricas sejam previsíveis o bastante. Este tipo de informação é importante para determinar quais aplicações são as mais recomendáveis para as redes complexas.

No capítulo 2 está uma definição mais detalhada das redes complexas e das métricas utilizadas neste experimento. O capítulo 3 contém detalhes do experimento, como corpus e métodos utilizados, e no capítulo 4 os resultados obtidos são exibidos. Finalmente, o capítulo 5 contém as conclusões obtidas deste trabalho.

Capítulo 2

Redes complexas

Neste capítulo são brevemente apresentados tanto o conceito de redes complexas quanto algumas das medidas existentes, utilizadas nestes experimentos.

2.1 Definição de redes complexas

Há algum tempo já se vem observando a viabilidade da utilização de grafos para representação e análise computacional de problemas em diversas áreas, com boa quantidade de resultados positivos. Porém, boa parte dos casos reais são complexos demais para serem estudados com métodos da teoria original de grafos, e daí deriva a opção de muitos pela utilização da teoria mais elaborada de redes complexas. Certos grafos com características especiais, conhecidos como redes complexas, vêm sendo aplicados em crescente escala ao PLN como uma forma de representação computacional de textos. Os experimentos deste relatório se baseiam nessa teoria e é, portanto, importante que seja feita uma definição mais formal de redes complexas.

O traço mais importante das redes complexas é que se trata de grafos com alto número de vértices (nós), o que também ocorre em problemas reais. Algumas outras características também são relevantes:

- Apesar do tamanho das redes, o número de nós entre um dado par de nós é relativamente pequeno, isto é, o caminho mínimo entre dois nós tem baixo custo. Tais redes são conhecidas como redes mundo pequeno ou *small world*.

- Quando um grupo pequeno em relação ao tamanho da rede possui nós que são fortemente conectados entre si, trata-se de um aglomerado ou um *cluster*. As redes complexas têm uma tendência a ter um alto coeficiente de aglomeração (definido na seção 2.2), isto é, elas tendem a apresentar grupos muito coesos.

- Outra característica marcante das redes complexas é o fato de que elas são livres de escala ou *scale-free*. Isso significa que redes complexas têm poucos hubs (nós com grande quantidade de conexões) e muitos nós com baixa quantidade de conexões, e novos nós adicionados tendem a se ligar a nós já bem conectados.

Antes de modelar um texto como uma rede complexa [4], uma etapa de pré-processamento é realizada, na qual são primeiramente removidas as *stopwords* (que são palavras comuns que aumentam o tempo de processmento mas contribuem pouco para o resultado final), e em seguida as palavras remanescentes são lematizadas, isto é, flexionadas para uma forma comum (por

exemplo, todos os verbos são flexionados para o infinitivo), para evitar problemas com dados esparsos, e cada uma dessas palavras será um nó na rede.

As redes dos textos utilizados no experimento são então criadas da seguinte maneira: cada associação entre palavras no texto gera uma conexão entre dois nós na rede, e cada associação entre palavras aumenta em 1 o peso de sua aresta correspondente. Vale lembrar que trata-se de um grafo orientado, isto é, as arestas são unidirecionais. Deste modo, quanto mais vezes uma associação aparece (na mesma ordem) no texto, maior é o peso dessa aresta na rede. As arestas obtidas são então representadas como uma matriz NxN, onde N é o número de vértices. A partir dessa modelagem, o texto passa a ser tratado simplesmente como uma rede complexa e é portanto submetido à análise sob essa perspectiva.

2.2 Medidas utilizadas

A representação de textos por redes complexas permite vários focos de análise através de suas diferentes medidas. Algumas dessas medidas foram implementadas no NILC e utilizadas nos experimentos [5]. Segue uma breve descrição de cada uma delas:

- Outdegree (Grau): uma rede complexa é, por definição, um grafo direcionado com pesos nas arestas. Portanto, devemos levar em consideração ambos os graus de entrada e saída de cada nó. Entretanto, dado o método de formação da rede, cada aresta que incide em um nó tem uma aresta complementar que sai deste mesmo nó (exceto pelas primeira e última palavra, que se compensam). Por isso os resultados para grau de saída e grau de entrada são sempre os mesmos, e com isso, apenas o grau de saída foi levado em conta.

  O grau de cada nó i da rede é medido com a seguinte fórmula:

  $$OD(i) = \sum_{j=1}^{N} W(j,i) \tag{2.1}$$

  Onde i e j são nós da rede e $W(j,i)$ o peso da aresta $i \rightarrow j$. A medida final é a média aritmética dos graus de todos os nós da rede.

- Clustering Coefficient (Coeficiente de aglomeração): para um nó i, o coeficiente é dado da seguinte maneira: quanto mais arestas existem entre um vizinho específico de i e os outros vizinhos de i, maior é o coeficiente de aglomeração.

  Isto é, o CC verifica quão próximo o subgrafo dos vizinhos do nó i da rede está do grafo completo com o mesmo número de nós. O valor da medida é 1 quando todos os nós de um conjunto de vizinhos se conectam entre si. O valor para a rede toda é a média aritmética de todos os CCs.

  Matematicamente, temos: seja R o conjunto de todos os nós que se conectam ao nó i. Temos que $N_c = |R|$ e B é o número de arestas existentes entre os nós do conjunto R. O CC do nó R é calculado da seguinte maneira:

  $$CC(i) = \frac{B}{N_c(N_c - 1)} \tag{2.2}$$

  Se $N_c = 0$ ou $N_c = 1$, $CC(i) = 0$.

- Components Dynamics Deviation (Desvio na dinâmica dos componentes): o CDD é uma medida que, intuitivamente, indica a uniformidade com a qual novas palavras são inseridas em um texto. Seu cálculo é feito da seguinte maneira: inicialmente, tem-se todos os nós e nenhuma aresta. Então, cada aresta da rede é adicionada, e o número de componentes conexos é contado (isto é, é verificado se um novo nó foi adicionado à rede parcial). O desvio da dinâmica dos componentes é o resultado da comparação entre a curva do número de componentes obtida pela adição de arestas e a reta ótima, na qual os nós são adicionados uniformemente a cada nova aresta. O valor da medida para toda a rede é a média aritmética do desvio para todos os pontos da curva.

  Sejam f_a a função da curva supracitada, f_s a função da reta, N o número de nós na rede e L o número total de associações de palavras no texto. O CDD é medido da seguinte maneira:

$$CDD = \frac{\sum_{x=1}^{L} |f_a(x) - f_s(x)|}{NL} \tag{2.3}$$

- Matching index (MI): o *matchingindex* é uma medida aplicada às arestas, e observa a similaridade entre a conectividade dos dois nós adjacentes à aresta em questão, isto é, quanto menor o valor do MI, menor é a semelhança entre as regiões do grafo conectadas pela aresta. O cálculo da medida é feito da seguinte maneira:

  O MI da aresta $i \rightarrow j$ é o número de conexões compartilhadas entre i e j, isto é, quantas conexões ambos têm em comum com um terceiro vértice k, dividido pelo número total de conexões entre i, k e j, k‘(exceto conexões entre i e j). Como o grafo de um texto é direcionado, são tomadas como arestas compartilhadas apenas as quem possuem a mesma direção. Matematicamente, temos:

$$MI(i,j) = \frac{\sum_{k \neq i,j} a_{ik} a_{jk}}{\sum_{k \neq i} a_{ik} + \sum_{k \neq j} a_{jk}} \tag{2.4}$$

- Hierarchical Degree (Grau Hierárquico): para definir o GH, é necessário primeiramente definir os conceitos de distância e anel. A distância entre os nós i e j será tomada neste relatório como o número de arestas necessárias para se chegar do nó i ao nó j. Por exemplo, os vizinhos (diretos) de um nó sempre têm distância 1 do mesmo. Já o anel $R_d(i)$ de distância d do nó i é o conjunto de todos os nós que distam d de i (os nós com distâncias menores do que d também são desconsiderados).

  O grau hierárquico de distância d de um nó i pode então ser definido como o número de arestas que conecta os anéis $R_d(i)$ e $R_{d-1}(i)$. Também é feita a diferenciação pela direção das arestas, resultando em graus hierárquicos de saída e de entrada. Nos experimentos que utilizaram o GH, a distância utilizada foi 2, com 8 variações: todas as combinações de entrada/saída, acumulado/não-acumulado, com peso/sem peso das arestas. O termo acumulado indica que, para calcular o grau hierárquico de distância d, os graus de distância 1 até $d-1$ devem ser somados ao valor final da medida.

- Degree Correlation (Correlação de Grau): A CG mede quão próximos estão os graus (foi utilizado apenas o grau de saída de cada nó) dos nós adjacentes a uma mesma aresta, averiguando assim quão uniforme a distribuição das arestas é na rede. Seu cálculo é feito da seguinte maneira: para cada aresta, os graus de seus dois nós adjacentes são obtidos, e então dois vetores podem ser formados, $X = (X_1, X_2, ..., X_n)$ e $Y = (Y_1, Y_2, ..., Y_n)$, onde n é o número de arestas na rede, X_i e Y_i são os graus dos nós de incidência e de partida da

i-ésima aresta. O valor da correlação de grau utilizado no experimento foi um coeficiente de Pearson entre os vetores X e Y, calculado pela seguinte fórmula:

$$CG(X,Y) = \frac{\sum_{t=1}^{n}(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \bar{X})^2 \sum_{t=1}^{n}(Y_t - \bar{Y})^2}} \tag{2.5}$$

onde $\bar{X}$ e $\bar{Y}$ são as médias dos valores de X e Y. Os valores obtidos para o CG variam entre -1 e 1, e quanto mais próximos forem (em módulo) de 1, maior é a correlação linear entre os valores dos graus. Se o CG, a correlação entre os graus definitivamente não é linear.

Capítulo 3

Definição do experimento

Neste capítulo está uma breve descrição do experimento, contendo informações sobre o corpus utilizado e sobre o procedimento realizado.

3.1 Descrição do corpus

O corpus deste experimento consistiu de 31 livros de 6 autores consagrados da língua inglesa. Os livros foram utilizados na sua versão original, e os tamanhos dos textos se encontram na tabela 3.1.

Tabela 3.1: Autores e livros pertencentes ao corpus deste experimento, com os títulos originais em inglês

Autor	**Livros**
Charles Dickens	A Tale of Two Cities American Notes David Copperfield Great Expectations Hard Times Master Humphrey's Clock Oliver Twist The Old Curiosity Shop The Seven Poor Travelers
Ernest Hemingway	The Garden of Eden Green Hills of Africa
Lewis Carroll	Alice's Adventures in Wonderland Sylvie and Bruno Through the Looking-Glass
Pelham G. Wodehouse	My Man Jeeves Tales of St. Austin's The Adventures of Sally The Clicking of Cuthbert The Gem Collector The Man with Two Left Feet The Pothunters The Swoop The White Feather
Thomas Hardy	A Changed Man and Other Tales A Group of Noble Dames Desperate Remedies Far from the Madding Crowd The Hand of Ethelberta
Virginia Woolf	Jacob's Room Night and Day The Voyage Out

Este corpus, embora seja pequeno, contém um número razoável de livros, que embora pequenos têm tamanho o suficiente para a realização deste experimento, e 31 livros é uma quantidade que permite a obtenção dos resultados desejados, que serão descritos no capítulo 4.

3.2 Metodologia

Cada um dos textos sofreu a seguinte análise: as 2000 primeiras associações de palavras do texto foram obtidas, e os valores de todas as medidas foram obtidas para cada segmento destes de maneira cumulativa, isto é, as medidas foram primeiramente calculadas para textos com 1 palavra, em seguida com as 2 primeiras palavras, e assim por diante. Com isso, cada par texto-medida gerou 2000 valores, que foram organizados em curvas, como na figura 3.1, onde a curva da esquerda representa os valores das medidas (a curva em preto é uma aproximação por polinômio de grau 3 obtido pelo Método dos Mínimos Quadrados Generalizado ($MMQG$)), e a curva à direita representa a flutuação entre os valores da medida (curva azul) e a aproximação (curva preta).

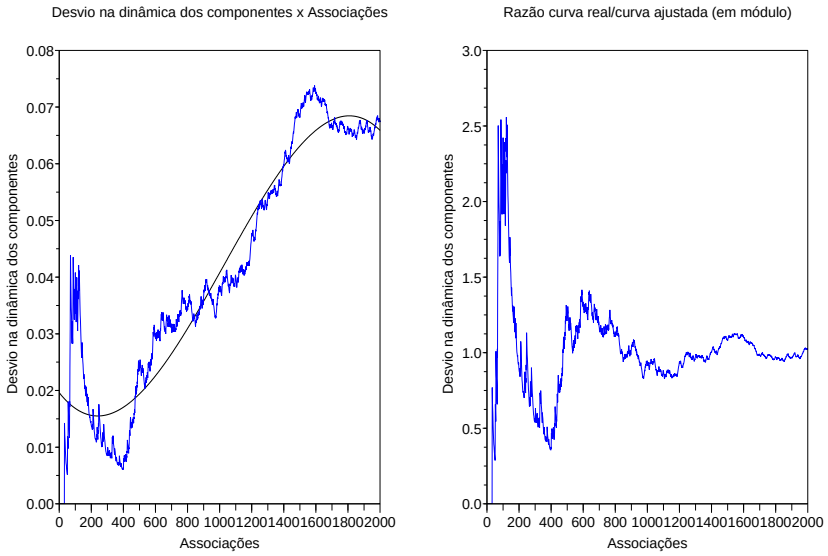

Figura 3.1: Gráfico da rede correspondente ao livro *American Notes*, de Dickens. À esquerda: cada valor no eixo das associações indica a quantidade de palavras selecionadas cumulativamente, isto é, o valor n no eixo indica que as n primeiras associações foram utilizadas. À direita: gráfico da flutuação dos valores das medidas para textos de diferentes tamanhos

Estas duas curvas nos permitem obter todas as informações desejadas, isto é, por elas é possível analisar a estabilidade de cada métrica, assim como é possível determinar um tamanho de texto para o qual a medida já se comporta de maneira razoavelmente uniforme, aumentando a confiabilidade do valor obtido.

Uma análise mais profunda sobre os resultados está localizada no capítulo 4.

Capítulo 4

Resultados

Os resultados descritos neste capítulo foram obtidos após a realização das atividades descritas na seção 3.2. Alguns exemplos estão expostos nas figuras 4.1-4.6, que correspondem ao livro $DavidCopperfield$, de Dickens. Pelas figuras é possível observar que na maioria dos casos os valores da flutuação diminuem conforme o tamanho do texto aumenta, eventualmente se aproximando de 1, onde o comportamento dos valores pode ser aproximado mais precisamente com o $MMQG$. A figura 4.7 contém uma comparação entre os valores do CDD para 6 livros diferentes, com o intuito de demonstrar que os valores geralmente seguem um mesmo padrão de comportamento para vários livros diferentes.

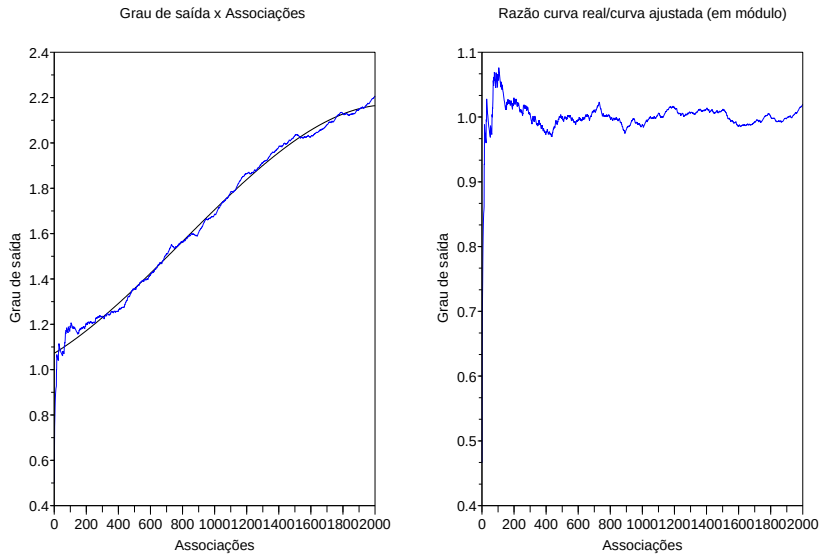

Figura 4.1: Gráficos para OD. É provavelmente a medida mais estável de todas, seguindo praticamente um comportamento monotonicamente crescente

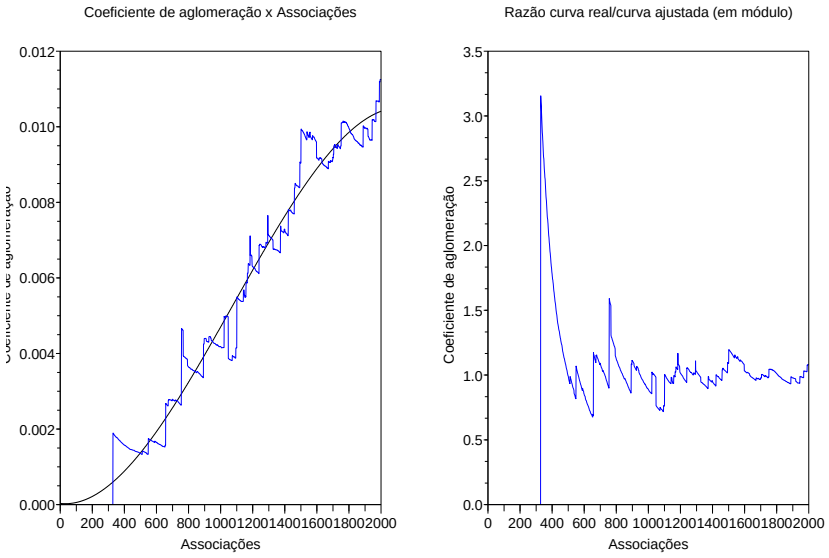

Figura 4.2: Gráficos para CC. Esta medida se caracteriza por somente crescer após um certo patamar de tamanho de textos

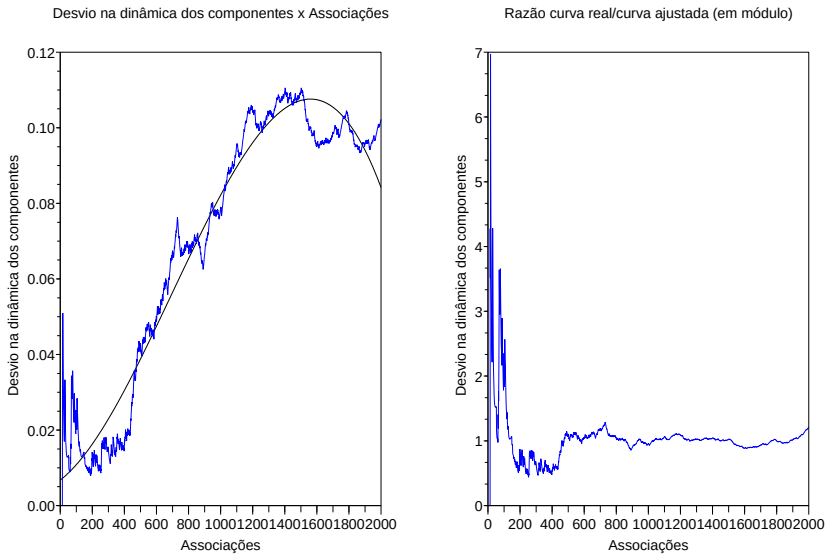

Figura 4.3: Gráficos para CDD. Para textos extremamente pequenos, esta medida é altamente instável, mas seu comportamento se altera para textos maiores

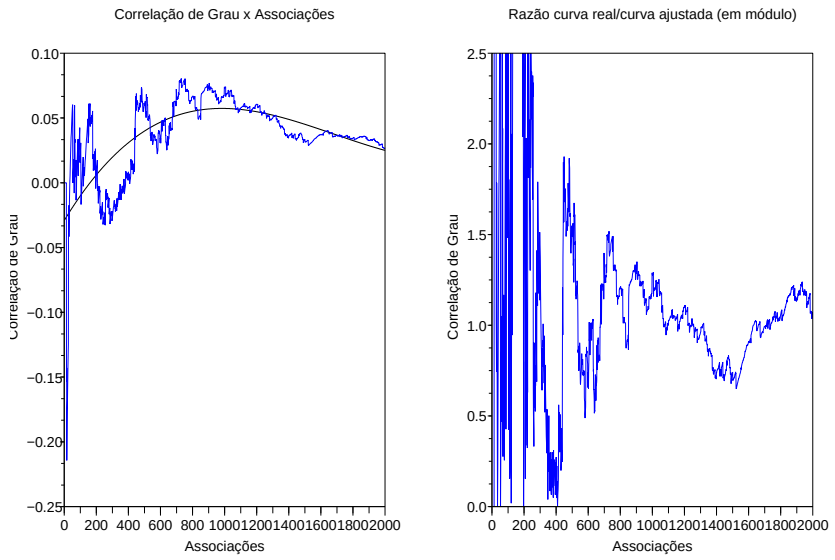

Figura 4.4: Gráficos para CG. Os valores desta medida tendem a ficar sempre próximos de 0, diminuindo assim a precisão da medição da flutuação

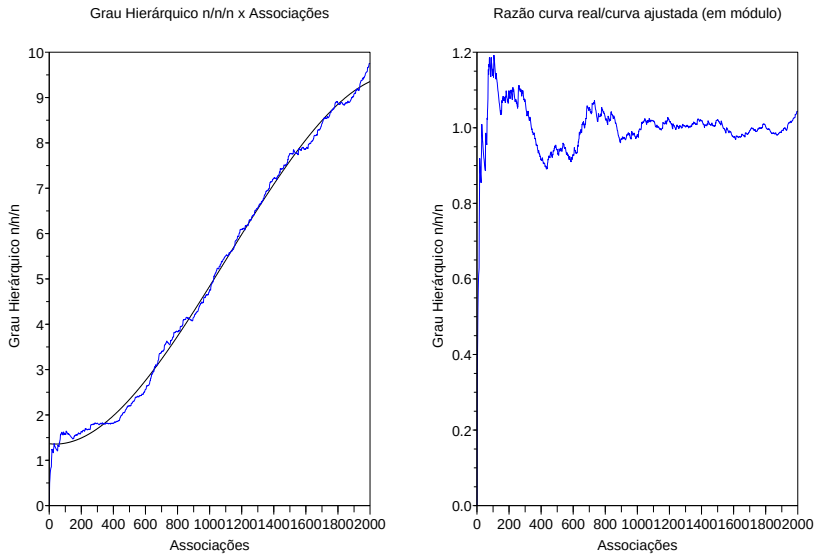

Figura 4.5: Gráficos para GH. Resultados bastante próximos aos do OD. Entre as diversas variações do GH, os resultados são geralmente muito próximos

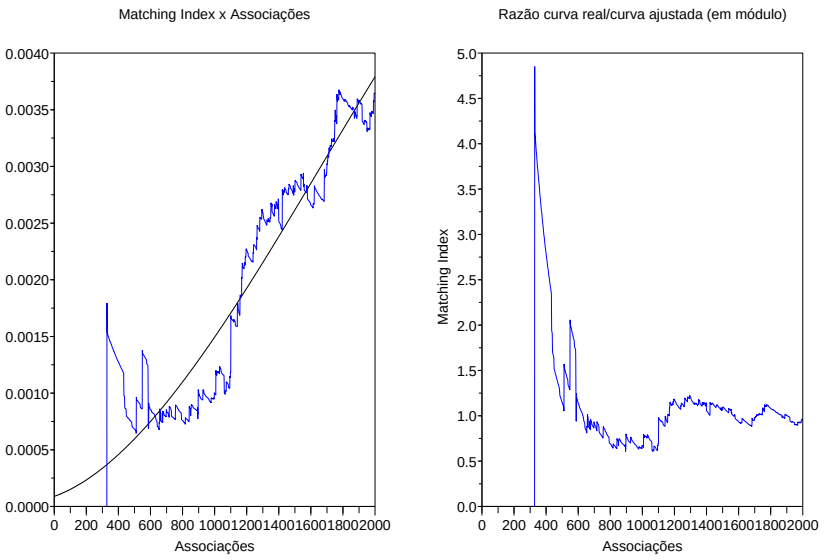

Figura 4.6: Gráficos para MI. Como é uma medida bastante parecida com CC, suas evoluções se comportam de modo bastante parecido

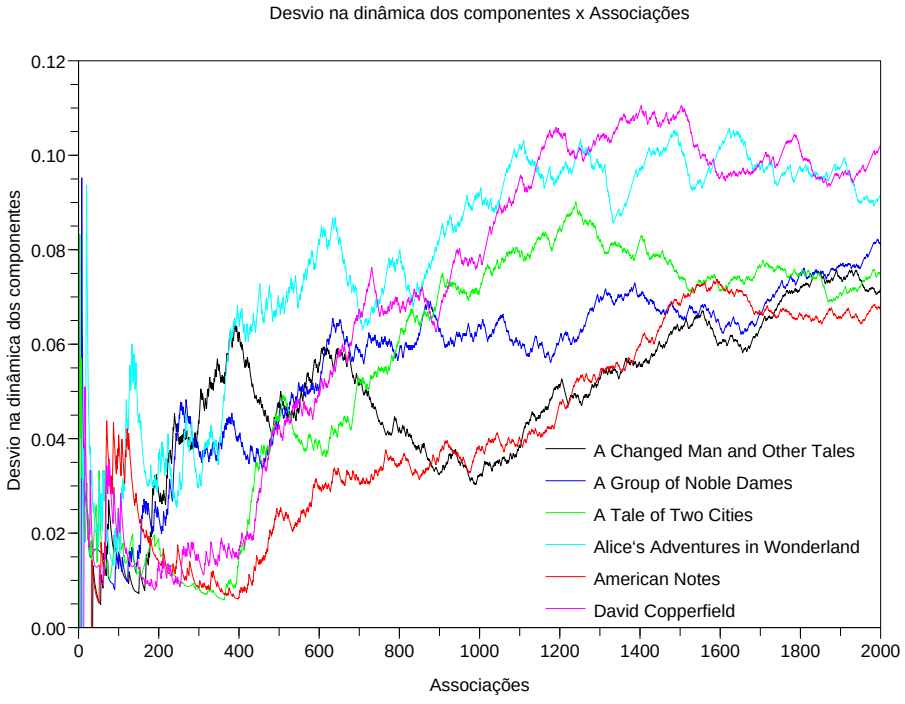

Figura 4.7: Comparação entre valores do CDD para vários livros diferentes

Como pode ser visto nas figuras 4.1-4.6, todas as medidas têm a mesma tendência de diminuir a flutuação conforme o tamanho do texto aumenta. Apesar disso, para o caso do CG, o tamanho de 2000 palavras não é o bastante para que os valores sejam previsíveis o bastante, e um destes casos de falha pode ser observado na figura 4.8.

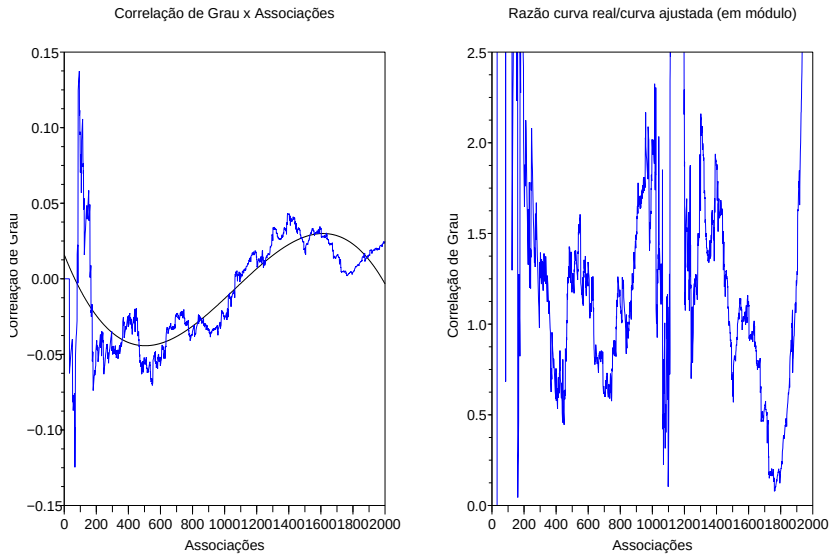

Figura 4.8: Exemplo de caso em que 2000 palavras não são suficientes para que a medida se estabilize

Para todas as outras medidas, os resultados foram bastante positivos, com todos os textos seguindo aproximadamente o padrão definido pelos gráficos 4.1-4.6. É plausível assumir que os problemas encontrados com o CG provenham do fato do alcance de valores da medida ser muito próximo de 0, o que aumenta a flutuação relativa da curva (e pode implicar em valores menos estáveis).

Com estes resultados, é possível afirmar que todas as medidas estudadas, exceto o CG, são razoavelmente previsíveis, e portanto recomendáveis para utilização em futuros experimentos utilizando redes complexas (embora isso não signifique que o CG deva ser descartado, mas apenas utilizado com mais cuidados). Além disso, foi possível se observar que textos com 2000 palavras já são de tamanho suficiente para que as redes complexas tenham um bom desempenho, o que é um resultado bastante positivo, uma vez que os populares métodos estatísticos para problemas de PLN normalmente requerem uma quantidade muito maior de dados para fornecer resultados satisfatórios.

Capítulo 5

Conclusões

Os resultados obtidos ao longo deste experimento comprovam que as diversas medidas para redes complexas estudadas, salvo a exceção da Correlação de Grau, são bastante previsíveis e podem, portanto, ser utilizadas em maior de escala para fornecer resultados cada vez mais expressivos na área. Com isso, a possibilidade de se desenvolver sistemas de maior precisão usando redes complexas aumenta, permitindo assim avanços tanto no estudo das redes quanto na área de PLN.

Outro resultado importante é a observação de um tamanho ideal para aplicações de PLN envolvendo redes complexas, em torno de 2000 palavras, que implica em textos bastante pequenos (poucas páginas), mas ainda assim fornecendo bons resultados. Esta informação é importante porque as aplicações de redes complexas normalmente são computacionalmente caras, exigindo boa quantidade de recursos e tempo e, como esperado, textos tão pequenos quanto possível colaboram para a construção de sistemas mais eficientes. Saber um tamanho de texto que seja previsível porém computacionalmente factível é de suma importância para o avanço da área.

Referências Bibliográficas

[1] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[2] L. Antiqueira, M.G.V. Nunes, O.N. Oliveira Jr., and L. da F. Costa. Strong correlations between text quality and complex networks features. *Physica A*, 373:811–820, 2006.

[3] L. Antiqueira, T.A.S. Salgueiro, M.G.V. Nunes, and O.N. Oliveira Jr. Some issues on complex networks for author characterization. *Revista Iberoamericana de IA, n. 36*, 11:51–58, 2007.

[4] Lucas Antiqueira, Maria das Graças V. Nunes, Osvaldo N. de Oliveira Jr., and Luciano da F. Costa. Modelando textos como redes complexas. Technical report, Instituto de Ciências Matemáticas e de Computação, 2005.

[5] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *arXiv:cond- mat/0505185*, 2006.