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

Atribuição de Autoria usando 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-08-02
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 serão apresentados dois experimentos realizados com base em experimentos do NILC (Núcleo Interinstitucional de Linguística Computacional) que tiveram como finalidade avaliar a capacidade das redes complexas de realizar tarefas de *clustering* de autores e de atribuição de autoria. A teoria de redes complexas vem crescendo no ambiente de pesquisas, e é bastante recente no meio de PLN (Processamento de Língua Natural), e suas capacidades e limitações ainda estão sendo verificadas, o que foi a grande motivação para a realização destes experimentos.

Índice

1 Introdução **2**

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

3 Descrição do corpus **6**

4 Clustering de Autores **8**

5 Atribuição de Autoria **12**

6 Conclusões **14**

Referências Bibliográficas **15**

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, avaliação de qualidade [2], sumarização automática, tradução automática, entre muitas outras.

Este trabalho foi particularmente focado na área de atribuição automática de autoria. Uma atribuição de autoria, como o próprio nome indica, consiste em determinar quem é o autor de um dado texto cuja autoria é desconhecida ou disputada por dois ou mais pessoas diferentes. Esta é uma tarefa pouco trivial, uma vez que um sistema automático capaz de realizar esta atribuição deve ser capaz de detectar características de estilo de um determinado autor.

Várias abordagens automáticas já foram desenvolvidas, algumas com resultados bastante positivos, porém a representação de textos por redes complexas ainda não foi direcionada a este problema específico, e neste trabalho está descrita um método desenvolvido com base na teoria de redes complexas, e suas medidas.

No capítulo 2 está uma definição mais detalhada das redes complexas e das métricas utilizadas neste experimento. O capítulo 3 contém detalhes sobre o corpus utilizado, e os capítulos 4 e 5 contêm descrições dos experimentos realizados e seus resultados. Finalmente, o capítulo 6 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 [6]. 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, 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 [4]: 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 [5]. Algumas dessas medidas foram implementadas no NILC e utilizadas nos experimentos. 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.

Capítulo 3

Descrição do corpus

Neste capítulo está uma breve descrição do experimento, contendo informações sobre o corpus utilizado. 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**	**Palavras utilizadas**
Charles Dickens	A Tale of Two Cities	60000
	American Notes	46000
	David Copperfield	158000
	Great Expectations	80000
	Hard Times	48000
	Master Humphrey's Clock	20000
	Oliver Twist	76000
	The Old Curiosity Shop	102000
	The Seven Poor Travelers	4000
Ernest Hemingway	The Garden of Eden	28000
	Green Hills of Africa	28000
Lewis Carroll	Alice's Adventures in Wonderland	12000
	Sylvie and Bruno	30000
	Through the Looking-Glass	14000
Pelham G. Wodehouse	My Man Jeeves	22000
	Tales of St. Austin's	24000
	The Adventures of Sally	36000
	The Clicking of Cuthbert	28000
	The Gem Collector	16000
	The Man with Two Left Feet	32000
	The Pothunters	18000
	The Swoop	8000
	The White Feather	20000
Thomas Hardy	A Changed Man and Other Tales	38000
	A Group of Noble Dames	30000
	Desperate Remedies	64000
	Far from the Madding Crowd	64000
	The Hand of Ethelberta	60000
Virginia Woolf	Jacob's Room	28000
	Night and Day	76000
	The Voyage Out	64000

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.

Capítulo 4

Clustering de Autores

Dois experimentos foram realizados com o intuito de avaliar a capacidade das redes complexas para detectar características de estilo de um autor [3]: o primeiro consistiu num agrupamento, ou *clustering* dos autores descritos na seção 3. O *clustering* consiste em tentar diferenciar os autores de uma base de textos (sem necessariamente atribuir a autoria de textos não-pertencentes à base). A seguir está uma descrição do desenvolvimento deste experimento.

Para o *clustering* de autores, de cada um dos 31 textos descritos anteriormente na seção 3 foram extraídas as 2000 primeiras associações de palavras, e cada uma das medidas descritas em 2.2 foi extraída para textos de tamanho crescente, isto é, para cada quantidade de associações de palavras (de 1 a 2000), gerando uma curva de evolução das medidas conforme o tamanho dos textos. A figura 4.1 exemplifica esta evolução.

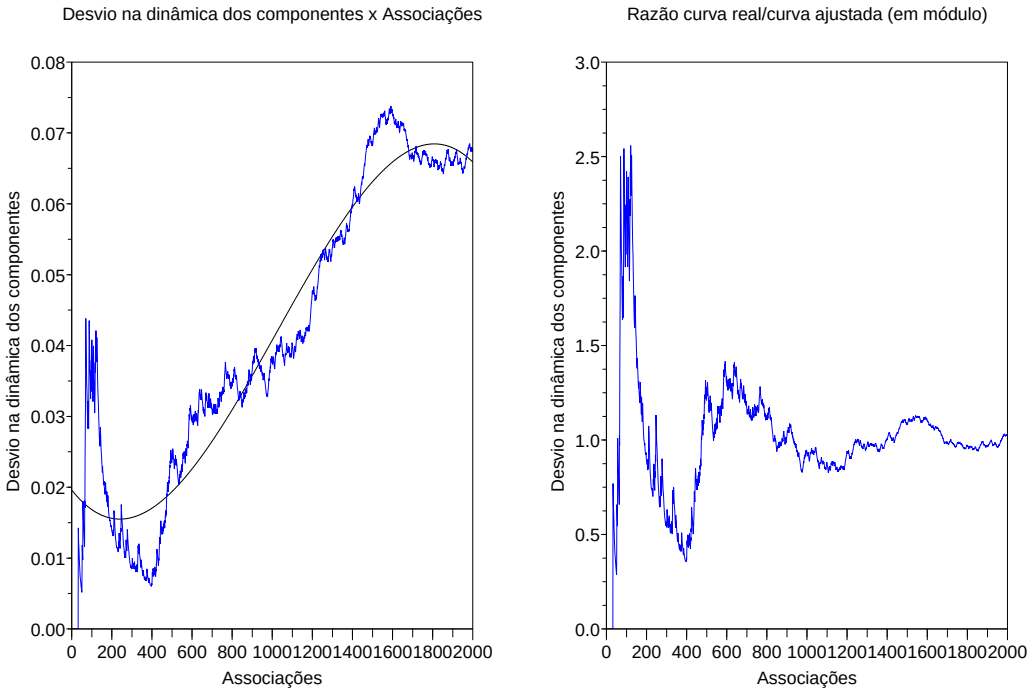

Figura 4.1: À esquerda, valores de CDD para as 2000 primeiras associações do livro $AmericanNotes$, de Dickens. À direita, flutuação relativa conforme a evolução de tamanho do texto

Após a extração dos valores de cada medida, foi feita uma aproximação da curva utilizando o Método dos Mínimos Quadrados Generalizado (MMQG) para obter um polinômio de grau 3 que condensa toda a evolução de uma dada medida em 4 coeficientes (curva em preto na figura 4.1, gráfico da esquerda).

A análise final foi feita a partir destes coeficientes, não levando em conta os valores reais das medidas. Os seguintes valores foram utilizados:

- OD_c0
- CDD_c0
- GHnns_c0
- GHnss_c0
- OD_c1
- CDD_c1
- GHnns_c1
- GHnss_c1
- OD_erro
- CDD_erro
- GHnns_erro
- GHnss_erro

onde GHnns significa Grau Hierárquico não acumulado (apenas o último anel é levado em conta, os anteriores são descartados), de saída e com pesos nas arestas, GHnss é o mesmo, mas os graus considerados são de entrada, c0 e c1 são os coeficientes de grau 0 e 1, respectivamente e erro é a somatória do módulo da diferença dos valores entre curva real e curva aproximada (azul e preta na figura 4.1) para os 2000 valores. O motivo para c2 e c3 terem sido descartados é o fato de que estes são muito pequenos (pode-se dizer que as curvas são muito "horizontais", pois o eixo das associações é maior, o que gera valores pequenos para garus maiores).

Uma Análise de Variáveis Canônicas (CVA) foi realizada sobre os valores descritos acima, e os resultados obtidos estão na figura 4.2, onde cada cor representa um autor diferente e cada ponto é um livro.

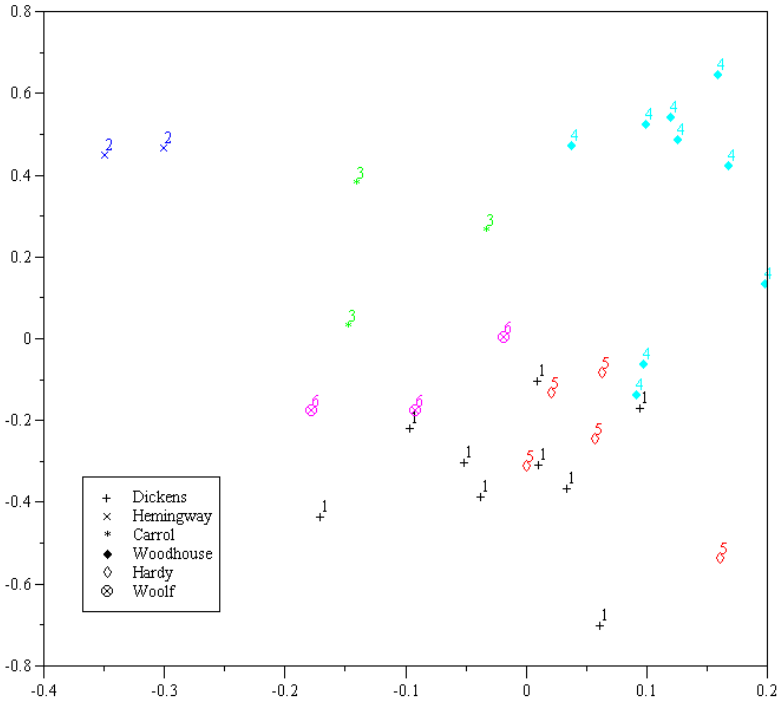

Figura 4.2: Resultados do *clustering* de autores

Os resultados obtidos neste experimento são muito positivos, pelos seguintes motivos: todos os autores ficaram em regiões bastante definidas do gráfico, exceto Dickens e Hardy (números 1 e 5, respectivamente), o que indica que até certo ponto as redes complexas são capazes de detectar informações de estilo de um autor bastante precisamente. Vale realçar aqui que o *clustering* na verdade separa estilos, e não autores, portanto se dois autores escreveram dois livros de estilos próximos, provavelmente esses livros resultariam próximos em uma análise estilística. Os autores Dickens e Hardy tinham estilos notadamente parecidos de escrita, o que explica parte da proximidade observada entre os livros dos dois autores no *clustering*.

Apenas como complemento deste experimento, para verificar a viabilidade deste método para textos menores, foram selecionados aleatoriamente (mas do mesmo período) textos de 4 autores do jornal *The London Telegraph* [1]: Grif Rhys Jones, Jenny McCartney, Matthew d'Ancona, Nigel Farndale. A quantidade de associações utilizada foi de 341 (máximo possível de modo a uniformizar o corpus) e o *scatter-plot* de 4.3 foi gerado pelo mesmo processo anterior. Novamente, os autores ficam em regiões definidas, com poucos pontos fugindo à regra. O tamanho dos textos é bem menor do que no experimento anterior, o que explica a maior ocorrência de erros, mas ainda assim é um resultado bastante positivo.

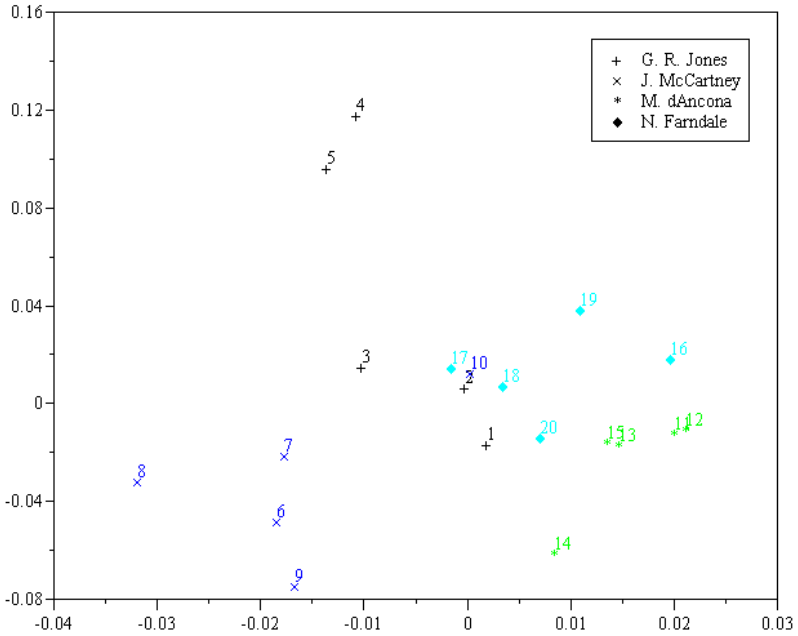

Figura 4.3: Novos resultados de classificação

Mais uma vez, fica demonstrado que as redes complexas são capazes de identificar características importantes de estilo de escrita, e portanto é possível tentar uma abordagem que se utilize de redes para atribuição de autoria, e ademais, as redes se mostraram capazes de boa performance mesmo quando os textos utilizados são bastante pequenos, o que pode definir um nicho de aplicação.

Capítulo 5

Atribuição de Autoria

O segundo experimento descrito neste relatório é uma atribuição de autoria propriamente dita [3]. Com os resultados obtidos no *clustering* dos autores, viu-se possível tentar realizar esta tarefa, que consiste em, tendo uma base de textos, encontrar o autor de um textos não-pertencente à base.

O procedimento realizado neste experimento é o seguinte: cada texto foi dividido em segmentos de 2000 associações de palavras, e em seguida as redes complexas foram criadas para cada segmento. Como o custo seria computacional seria muito grande para calcular as medidas de maneira cumulativa (como descrito no capítulo 4), calculou-se apenas as métricas sobre textos de 2000 associações, o que seria equivalente ao último ponto no eixo das associações da figura 4.1. Com isso, cada segmento de 2000 palavras proporcionou apenas um valor para cada métrica.

O procedimento anterior resultou em diversos valores separados. Para relacioná-los, o ambiente Weka [9] foi usado para executar diversos métodos de aprendizado de máquina, usando os valores das medidas como instâncias de aprendizado. Os resultados estão dispostos na tabela 5.1.

Tabela 5.1: Resultados da atribuição de autoria. Como esperado, conforme a quantidade de instâncias aumenta, também aumenta a precisão dos métodos

Método / Acertos em %	**31 segmentos**	**92 segmentos**	**669 segmentos**
Regras	26	43	54
Árvores de decisão	29	40	50
Redes bayesianas	32	42	42
Naive-bayes	35	38	40
Redes Neurais	45	54	62

Como era esperado, conforme aumenta a quantidade de instâncias de aprendizado, melhor é a precisão dos métodos. A maior precisão atingida é de 62%, mas acredita-se que isso seja devido à ausência de mais instâncias de aprendizado, uma vez que os métodos de aprendizado normalmente precisam de milhares de instâncias para atingirem sua precisão máxima. Estima-se que com um córpus maior se consiga atingir a precisão dos métodos de estado da arte, que chega a mais de 90% em alguns casos [7, 8].

Numa tentativa de melhorar a precisão deste método, a seguinte comparação foi feita: para

cada segmento de 2000 palavras, os dois métodos descritos anteriormente (o que utiliza os polinômios aproximadores e o que usa apenas um valor por segmento) foram comparados como instâncias de aprendizado para os métodos de aprendizado de máquina. Os resultados se encontram na tabela 5.2.

Tabela 5.2: Comparação entre valores dos coeficientes e das medidas

Método / Acertos em %	**Coeficientes dos polinômios**	**Valores das medidas**
Regras	35	26
Árvores de decisão	32	29
Redes bayesianas	29	32
Naive-bayes	39	35
Redes Neurais	42	45

Pode ser visto que o método dos coeficientes gera resultados mais precisos na maioria dos casos, e tem performance semelhante nos outros casos. Isto era até certo ponto esperado, uma vez que este método avalia toda a evolução de um texto, ao invés de olhar para um único ponto. O problema deste método é que o consumo de tempo é muito grande, pelo menos com os recursos computacionais disponíveis hoje. Provavelmente, este método será viável conforme os computadores avancem, e apenas então seria possível averiguar se o método dos coeficientes é de fato melhor.

Para validar a precisão deste método, tentou-se comparar com o método descrito em [7], que se baseia na teoria de modelos de língua. Modelos de língua são basicamente um método de associar a cada seqüência natural de palavras uma probabilidade. Neste trabalho, a idéia é associar a cada seqüência de palavras que de fato ocorre em um córpus uma probabilidade maior do que a de uma seqüência que nunca ocorre.

Porém, os resultados não correspondem aos descritos em [7], uma vez que, na grande maioria dos casos, o sistema não foi capaz de determinar qual era o autor do texto. Acredita-se que isso seja devido ao tamanho limitado do córpus, não sendo suficiente para treinar devidamente o sistema, o que reforça ainda mais a hipótese de que as redes complexas são melhores quando os textos são menores.

Capítulo 6

Conclusões

Os resultados obtidos nos experimentos descritos neste relatório são muito positivos, não apenas porque foi desenvolvido um sistema de atribuição de autoria de boa precisão usando recursos simples e com forte embasamento matemático, mas também porque demonstram uma extensão da aplicabilidade de redes complexas na área de PLN. A teoria de redes complexas é bastante recente, daí a importância de se buscar sempre por novas tarefas que possam ser resolvidas.

Embora o método desenvolvido não seja tão preciso quanto alguns existentes, um nicho dessa aplicação pode ser em aplicações onde a quantidade de textos é limitada, ou os textos são pequenos, sem que haja a necessidade de grandes corpus para que o método seja eficiente.

Mas talvez o resultado mais importante aqui descrito seja a demonstração da capacidade das métricas de redes complexas em detectar características de estilo de um texto, o que torna plausível a continuação da busca por novas métricas ou novas áreas de aplicação das redes em PLN.

Referências Bibliográficas

[1] http://www.telegraph.co.uk/, Setembro 2007.

[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.

[6] R. Ferrer i Cancho and R. V. Solé. The small world of human language. *P. Roy. Soc. Lond. B Bio.*, 268:2261, 2001.

[7] F. Peng, D. Schuurmans, V. Keselj, and S. Wang. Language independent authorship attribution using character level language models. In *Proceedings of EACL*, 2003.

[8] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic authorship attribution. In *Proceedings of EACL*, 1999.

[9] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, second edition, June 2005.