UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

CAMILO AKIMUSHKIN VALENCIA

Dinâmica de redes complexas aplicada a reconhecimento de autoria

São Carlos

2015

CAMILO AKIMUSHKIN VALENCIA

Dinâmica de redes complexas aplicada a reconhecimento de autoria

Monografia apresentada ao Programa de Pós-Graduação em Física do Instituto de Física de São Carlos da Universidade de São Paulo, para o Exame de Qualificação como parte dos requisitos para obtenção do título de Doutor em Ciências.

Área de concentração: Física Básica
Orientador: Prof. Dr. Osvaldo Novais de Oliveira Jr.

São Carlos

2015

RESUMO

AKIMUSHKIN, C. *Dinâmica de redes complexas aplicado a reconhecimento de autoria*. Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, 2015.

Parte da complexidade implícita na linguagem se reflete na ordem das palavras, o que já foi usado para caracterizar linguagens, movimentos literários e autores por meio da criação de redes de co-ocorrência de palavras. O reconhecimento de autoria visa a separar textos em grupos que representam cada autor, tal que seja possível identificar o autor de um texto em disputa. Redes de co-ocorrência têm mostrado sucesso na tarefa de reconhecimento de autoria, mas pouco se tem estudado sobre a influência da dinâmica da rede. Isto é curioso, uma vez que a dinâmica é a responsável pelas propriedades estruturais da rede. Portanto, aprofundar no estudo da dinâmica, além do benefício prático de servir para o reconhecimento de autoria, pode trazer maior compreensão dos mecanismos de evolução de redes de textos. Um problema recorrente do reconhecimento de autoria é a escassez e heterogeneidade dos textos disponíveis. Neste projeto propõe-se uma metodologia para o reconhecimento de autoria baseada na dinâmica de redes de co-ocorrência. Para testar o método utiliza-se uma coleção de 300 textos de 27 autores na língua inglesa. Para cada texto são obtidas séries temporais para 6 medidas de rede. As séries temporais são estacionárias, permitindo usar os quatro primeiros momentos da distribuição para caracterizar a série. Os 24 atributos obtidos são usados em algoritmos de classificação e agrupamento. O desempenho da classificação é comparável ao de técnicas anteriores. Por outro lado, o agrupamento baseado em densidade mostra ótimos resultados, agrupando corretamente 296 dos 300 textos analisados. Os melhores resultados são alcançados com $\varepsilon = 1$, a qual parece ser a separação natural entre os grupos. As medidas introduzidas mostram ser características de cada autor.

Palavras-chave: Redes complexas. Séries temporais. Classificação e agrupamento de textos.

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

Os sistemas observados no dia a dia são quase sempre compostos de grande quantidade de elementos interagentes. A física, tradicionalmente caracterizada pelo reducionismo, poderia não ser considerada como a melhor ferramenta para abordar sistemas com tais propriedades. Porém, é fato que existem princípios básicos para descrever os fenômenos observados. Isto ocorre com os chamados sistemas complexos, os quais seguem regras bastante simples. O nome vem das propriedades não triviais como a perda de escala nestes sistemas típica, por exemplo, dos fractais. A linguagem é um destes sistemas complexos.

Um tipo especial de sistemas complexos vem sendo estudado com muito sucesso: as redes complexas. Redes podem construir-se partindo da maioria de sistemas disponíveis, sendo a representação mais natural de alguns deles. Muitos desses sistemas geram redes que satisfazem as condições requeridas para ser classificadas como redes complexas (1). A complexidade manifesta-se em redes como uma ou mais das seguintes características: distribuições em forma de leis de potência*, alta clusterização, efeito de mundo pequeno, e organização hierárquica.

Textos representados em forma de redes apresentaram as características anteriores (2), (3). Várias redes obtidas a partir da linguagem, além de textos escritos, são exemplos disto (4). Redes fonológicas de palavras têm propriedades modulares e perda de escala. Redes de similaridade semântica apresentam o fenômeno de mundo pequeno e perda de escala. Redes de dependência sintática mostram organização hierárquica, e redes de colocação apresentam o fenômeno de mundo pequeno e distribuições livres de escala. As redes de co-ocorrência são um caso particular de rede de colocação, construídas criando-se uma aresta entre duas

*As leis de potência são um exemplo de funções que apresentam perda de escala.

palavras (os nós da rede) cada vez que essas palavras aparecem consecutivamente num texto. Acredita-se que as linguagens evoluíram adquirindo estas qualidades para facilitar o aprendizado e a navegação por palavras.

O surgimento de complexidade é uma descoberta importante já que permite aplicar o conhecimento teórico dos modelos de redes, o que ajudaria a entender o comportamento das línguas a partir de mecanismos de evolução de redes. Igualmente, o fato de apresentar um comportamento conhecido permite caracterizar estas redes, e assim textos inteiros podem ser razoavelmente bem descritos em termos de uns poucos parâmetros. Considerando então que a representação por meio de redes consegue capturar aspectos relevantes dos textos, poderíamos usá-las para um fim prático.

A análise automática de textos vem adquirindo crescente importância. A disponibilidade de *corpora* (coleções de textos) é cada vez maior e os textos, que continuam sendo gerados diariamente, precisam ser analisados, organizados e acessados adequadamente. A mineração de textos é talvez a parte mais evoluída da área de mineração de dados, principalmente porque textos são processados mais facilmente do que, por exemplo, uma imagem. Na mineração de textos existem várias tarefas: recuperação de informação, indexação, extração de características, sumarização e classificação e agrupamento de textos.

As ferramentas mencionadas acima são usadas separadamente e em conjunto para diversos fins. A recuperação de informação está associada com a fase de preparação dos textos, necessária para a maioria das outras aplicações. A indexação permite converter um texto num conjunto de palavras-chave para otimizar a busca. A extração de características e a sumarização procuram os elementos relevantes dentro de um texto. A classificação e o agrupamento de textos podem ser aplicados à organização dos mesmos, por exemplo a categorização automática de mensagens de correio eletrônico é utilizada na deteção de spam. Outra aplicação consiste em organizar páginas de um domínio, por exemplo em seções num jornal de notícias.

Estas e outras ferramentas de processamento de língua natural têm evidentes aplicações, principalmente na internet. Neste trabalho utilizaremos duas delas, a classificação e o agrupamento. Elas se encaixam naturalmente no nosso objetivo: a predição automática do autor

de um texto.

As ferramentas tradicionais de classificação e agrupamento de textos usam uma representação do texto do tipo *bag of words*. A classificação de textos tem usado a maioria das técnicas de aprendizado supervisionado. Técnicas bem conhecidas como a TF-IDF e similares obtêm uma medida da relevância de cada texto para uma consulta (tal como na tarefa de indexação) baseando-se no número de vezes que aparece uma palavra-chave no documento. Do outro lado, no escopo de agrupamento se atribui um vetor cujas componentes são por exemplo as frequências de cada termo no documento. A dissimilaridade, isto é, a distância entre documentos, pode ser a diferença entre vetores ou ser o inverso do cosseno do ângulo entre eles. Nas técnicas anteriores o número de atributos que representa um documento é da ordem de grandeza do vocabulário do corpus. Uma forma de reduzir a dimensionalidade consiste em extrair os termos relevantes.

O reconhecimento de autoria também é problema central de uma nova área de pesquisa chamada estilometria. A estilometria junta técnicas, mesmo com focos distintos, para encontrar elementos característicos do estilo de cada autor. A estilometria pode precisar de conhecimento de especialista, isto é, aquele obtido da linguística.

1.1 Proposta de pesquisa

Mesmo que bem sucedidas, as técnicas tradicionais de reconhecimento de autoria têm problemas inerentes. A dimensionalidade tem papel importante já que mesmo usando técnicas para reduzir o número de palavras-chave, este continua sendo considerável. Pior ainda, este número depende do maior elemento presente na coleção. Este problema se trata com algoritmos otimizados para cálculos em muitas dimensões. Além disso, focar na semântica das palavras pode ser prejudicial no reconhecimento de autoria. Considere por exemplo dois livros de dois autores diferentes, ambos abordando temas sobre o mar. Considerar esses livros como similares pode ser útil para outros fins mas não necessariamente para descobrir o autor.

Dentre as técnicas usadas para o reconhecimento de autoria, redes de co-ocorrência apresentam várias vantagens. O uso de conhecimento de especialista limita-se à fase preparatória, a qual está totalmente automatizada, diferentemente das técnicas de estilometria. A dimensionalidade do problema fica a cargo do usuário, que define quantas e quais medidas quer usar. Sob certas condições esta dimensionalidade pode ser baixa. Diferentemente das técnicas tradicionais que funcionam extraindo-se palavras-chave, pode tratar facilmente diversos gêneros literários.

As propriedades das redes de co-ocorrência já serviram para identificar linguagens (2), (5), correntes literárias (6) e autores (7). No entanto não há estudos sobre caracterização de textos usando a evolução temporal, ou dinâmica, da rede de co-ocorrência. Acreditamos que a dinâmica de rede e o estilo da escrita estejam relacionados. Baseamo-nos no fato que a dinâmica define a topologia da rede (1), (3).

Gostaríamos de extrair medidas, em particular, métricas de rede à medida que ela evolui. No entanto, dois grandes problemas surgem. Primeiro, as métricas de uma rede dependem enormemente do tamanho da rede. Considere um exemplo extremo: a conectividade média de uma rede com distribuição livre de escala é infinita, se a rede for infinita. Portanto deve-se ter cuidado ao comparar métricas de redes de diferentes tamanhos, principalmente se a topologia não for conhecida. O segundo problema tem a ver com a independência das observações. Se medirmos uma métrica numa rede e a mesma métrica depois que a mesma rede evolui, as duas medidas com certeza não serão independentes. Este problema, embora tenha solução, precisa ser tratado com cuidado.

O método proposto neste trabalho evita estes e outros problemas para analisar redes. Em particular, o reconhecimento de autoria sofre pela heterogeneidade e escassez de textos que remete ao problema de comparar duas redes de tamanho distinto.

Se as medidas de rede dependerem do autor do texto, além das aplicações práticas haveria importantes questões sobre a linguagem. O resultado implicaria que o autor de um texto tem algum controle sobre a rede ao invés de ela estar governada por um mecanismo rigidamente fixo. Assim como sutis diferenças entre redes de textos sugerem características próprias de

linguagens específicas (no entanto possuindo todas um mecanismo comum (2), (3)), as medidas de rede poderiam dizer algo a respeito da forma de escrever de uma pessoa.

1.2 Objetivos

O objetivo geral do projeto é avançar na compreensão do processo de criação de textos escritos visando a entender a influência do autor na dinâmica de redes. Os objetivos específicos são:

- Observar os aspectos gerais da dinâmica das redes de co-ocorrência.
- Propor uma metodologia para testar a possível dependência da dinâmica de redes com respeito à autoria de textos escritos.
- Criar e disponibilizar uma ferramenta para reconhecimento de autoria baseada na metodologia proposta.

CAPÍTULO 2

FUNDAMENTAÇÃO TEÓRICA

A base teórica necessária para a realização do trabalho consiste em conceitos básicos e definições de redes complexas, séries temporais e conhecimento dos principais algoritmos de aprendizado de máquina.

2.1 Redes complexas

A linguagem pode ser representada mediante diferentes tipos de redes. No entanto, a caracterização de um texto individual beneficia-se particularmente de um tipo: as redes de co-ocorrência. Uma rede de co-ocorrência é gerada associando-se palavras consecutivas em pares. As redes de co-ocorrência apresentam distribuições em forma de lei de potência, isto é, perda de escala.

Muitos mecanismos geram perda de escala, como o paradoxo de São Petersburgo do jogo de moeda. No entanto, o paradigma atual em redes complexas é a ligação preferencial, proposta por Albert Barabási e Reka Albert (1). Em termos gerais, a ligação preferencial preconiza que quanto maior for o número de conexões de um nó numa rede, tanto maior é a chance de esse nó adquirir uma nova conexão, o que às vezes é chamado como fenômeno "*rich get richer*". O mecanismo de Barabási e Albert gera redes com distribuções de probabilidade em forma de lei de potência, $P(k) \sim k^{-\gamma}$ na faixa de valores $\gamma \in (2, 3]$. As redes de Barabási e Albert descrevem as redes de co-ocorrência entre outras. Um refinamento do modelo, proposto por

Dorogovtsev e Mendes, especificamente para redes de texto (3), gera distribuições com dois regimes de lei de potência.

As métricas extraídas das redes de co-ocorrência são bem conhecidas na literatura (8), (9). A primeira é conhecida como conectividade ou grau de um nó. A conectividade é o número de arestas que conectam o nó. Em redes com arestas dirigidas distingue-se entre conectividade de entrada e de saída. Pela construção da rede é facil ver que os nós das redes de co-ocorrência de textos têm a mesma conectividade de entrada e de saída exceto pelo primeiro e o último. Neste trabalho usamos a conectividade total obtida como a soma das conectividades de entrada e saída.

Uma outra métrica que define o efeito de mundo pequeno é a menor distância ou menor caminho (às vezes chamado de geodésica) medida como o número de arestas que se deve percorrer de um nó a outro. Em redes com efeito de mundo pequeno esta distância é pequena comparada com o tamanho da rede. Em princípio é facil calcular os caminhos de um certo tamanho usando a definição de matriz de adjacência. O elemento A_{ij} da matriz de adjacência é igual a um se existir uma aresta do i-ésimo ao j-ésimo nó e igual a zero caso contrário. A n-ésima potência da matriz de adjacência contém os caminhos de comprimento n entre os nós. Em particular, as conectividades dos nós são dadas pelos elementos diagonais do quadrado da matriz de adjacência, A^2. Na prática este método pode ter problemas se usado com uma rede muito grande.

Outra métrica a ser considerada é chamada de centralidade, que mostra quão importante é um nó na rede. Existem muitas medidas de centralidade; a conectividade de um nó, por exemplo, pode ser considerada como medida de centralidade. No entanto usamos uma métrica que considera as propriedades globais da rede inteira. A centralidade de intermediação (do inglês *betweenness centrality*) mede quanto um nó possui de menores caminhos, isto é quão importantes são as conexões que possui. A centralidade de intermediação do nó i é definida como:

$$x_i = \sum_{s \neq i \neq t} \frac{g^i_{st}}{g_{st}} \tag{2.1}$$

onde g_{st} é o número total de menores caminhos entre os nós s e t e g^i_{st} é o número destes

caminhos que passam por i. Um nó tem portanto alta centralidade se para ir de uma parte da rede a outra é preciso passar por ele. Usamos também outra medida que tem sido útil na caracterização de textos, a intermitência média. A intermitência é tão maior quanto mais irregular for a distância entre duas aparições da mesma palavra. Assim, se uma palavra aparecer muito mais em intervalos regulares ela terá intermitência baixa. Nas redes de co-ocorrência a intermitência tem relação com o comprimento dos ciclos ou caminhos fechados. Também definimos o peso de uma aresta como sendo o número de vezes que o respectivo par de palavras aparece no texto. As outras medidas consideradas neste trabalho são o número total de nós e de arestas (às vezes chamadas de ordem e tamanho da rede). Como fixaremos o peso total nas arestas da rede, estas duas grandezas são variáveis.

2.2 Séries temporais

É preciso definir alguns conceitos simples de séries temporais (10), (11). Considere um conjunto de realizações $\{x_1, x_2, \ldots x_T\}$ de uma variável X. A série satisfaz a forma estrita de ser estacionária se a distribuição estatística conjunta de qualquer grandeza da série (tal como a média ou a variância) nunca varia no tempo. Formalmente, se $F_X(x_{t_1+\tau}, \ldots x_{t_k+\tau})$ é a função de distribuição cumulativa da série $\{X_t\}$ nos instantes $t_1 + \tau, \ldots t_k + \tau$, diz-se que $\{X_t\}$ é estritamente, ou fortemente estacionária, se

$$F_X(x_{t_1+\tau}, \ldots x_{t_k+\tau}) = F_X(x_{t_1}, \ldots x_{t_k}) \tag{2.2}$$

para todos k e τ e para todos os instantes $t_1, \ldots t_k$. Portanto, F_X não depende do tempo, isto é, a série não apresenta fenômenos como tendência (*trending*) ou sazonalidade (*seasonality*). A forma fraca da estacionariedade ou estacionariedade de segunda ordem exige unicamente que a média e a variância sejam constantes e que a autocovariância não dependa do tempo. Existem vários testes de estacionariedade, incluindo: o teste aumentado de Dickey-Fuller, de Elliott-Rothenberg-Stock, teste KPSS de raiz unitária, o de Phillips-Perron, de Schmidt-Phillips, de Zivot-Andrews e o de Priestley-Subba Rao. Eles vêm inclusos nos principais pacotes de estatística. O resultado dos testes é um p-value que aceita ou rejeita com certo grau de

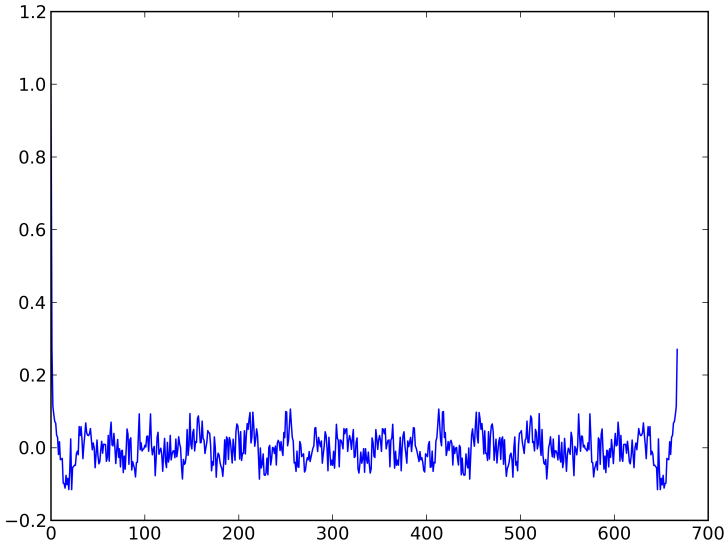


Figura 2.1 – Autocorrelação de Pearson para um dos textos da coleção.

confiança a hipótese nula de que a série seja estacionária.

Muitas vezes uma simples análise visual permite determinar se uma série pode ser estacionária. Uma grandeza a conferir é a autocorrelação. O parâmetro de correlação de Pearson entre duas séries finitas se define como:

$$r_{xy} = \frac{\sum_{i=1}^{T}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{T}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{T}(y_i - \bar{y})^2}} \tag{2.3}$$

Onde $\bar{x}$ e $\bar{y}$ são as médias das distribuições. A autocorrelação é a correlação entre uma série e ela mesma deslocada um número τ de passos temporais. Para $\tau = 0$ a autocorrelação tem o máximo valor igual à unidade. Para τ diferente de zero a autocorrelação deve cair rapidamente a zero se a série for estacionária tal como se mostra na figura 2.1 obtida da série temporal de um dos textos da coleção. Observe a simetria da figura. A autocorrelação também é igual à transformada de Fourier do quadrado absoluto da série devido ao teorema de Wiener-Khinchin.

2.3 Reconhecimento de autoria

Para reconhecer automaticamente a classe ou grupo (no nosso caso o autor) à qual pertence um determinado elemento é preciso realizar alguma tarefa de aprendizado de máquina. O aprendizado de máquina divide-se em duas grandes vertentes. No aprendizado supervisionado a máquina utiliza um conjunto de dados de treino para predizer a classe de um conjunto menor de dados de teste. No aprendizado não-supervisionado não existem dados de treino e portanto a solução, isto é, o agrupamento dos dados se faz usando alguma medida de similaridade. Em aprendizado de máquina os dados, no caso as 24 medidas introduzidas, são chamados de atributos.

2.3.1 Classificação

No aprendizado supervisionado ou classificação usamos cinco algoritmos, sendo que cada um é representativo de uma classe geral de algoritmos. O algoritmo mais simples é o ZeroR. Este algoritmo identifica a classe com mais elementos e atribui a esta classe todos os elementos da coleção, assim, a taxa de acerto está predefinida como a razão entre o número de elementos da maior classe e o número total. O algoritmo OneR, abreviação do inglês *one rule*, cria uma regra a partir de cada atributo e testa a taxa de acerto de cada regra, posteriormente utiliza a regra com a melhor taxa de acerto com todos os atributos.

Outro algoritmo usado é o K-*Nearest Neighbors*, no qual a classe de um elemento é dada pela classe dos K elementos mais próximos. Uma variante atribui mais peso aos elementos mais próximos. O quarto algoritmo é o *Naive Bayes* ou Bayesiano ingênuo. Este algoritmo assume independência entre os atributos. Assim, para cada atributo se obtém a média e a variância para caracterizá-lo como sendo distribuído normalmente. Para a nova instância a ser classificada se calcula a probabilidade de pertencer a uma classe dado o valor do seu atributo. A probabilidade de pertencer a uma classe é dada pelo produto das probabilidades condicio-

nais de todos os atributos. O último algoritmo de classificação usado é o de árvore de decisão J48. Este é a versão livre em Java do algoritmo C4.5 (que por sua vez é uma extensão do algoritmo ID3). O algoritmo cria uma árvore de decisão onde cada bifurcação é um critério para classificar as instâncias como pertencendo a uma ou outra classe. O critério de decisão é o valor de um dos atributos. Escolhem-se primeiro os atributos que geram um maior ganho da informação.

2.3.2 Agrupamento

Os algoritmos de aprendizado de máquina não-supervisionado ou agrupamento procuram os grupos que formam os elementos naturalmente. Para isto alguma definição de distância é requerida. Quando os atributos são numéricos a escolha mais razoável é geralmente a distância Euclidiana. Três tipos de algoritmos foram implementados: baseados em densidade, hierárquicos e maximização da expectativa.

O algoritmo baseado em densidade padrão é o DBSCAN (*Density-Based Spacial Clustering of Applications with Noise*), que precisa de dois parâmetros. O raio ε define uma esfera no espaço de atributos. $MinPoints$ é o número mínimo de pontos que deve encontrar-se dentro da espera para considerá-la densa. Um ponto é considerado pertencente a um grupo se a esfera centrada nele é densa, caso contrário considera-se o ponto como ruído. Uma generalização do DBSCAN é o algoritmo OPTICS (*Ordering Points To Identify the Clustering Structure*) que fornece uma ideia visual do agrupamento de pontos. Este algoritmo retorna uma lista ordenada dos elementos junto com duas distâncias características do algoritmo. O primeiro elemento da lista é arbitrário, o seguinte é o elemento mais próximo ao primeiro, o terceiro é o mais próximo ao segundo, e assim sucessivamente. É possível ver os grupos formados já que os elementos pertencentes a um mesmo grupo têm as menores distâncias enquanto ao passar de um grupo para o seguinte têm-se as maiores.

Outro tipo de algoritmos de agrupamento é o chamado agrupamento hierárquico aglome-

rativo. Nesta classe de algoritmos uma matriz contendo as distâncias entre todos os elementos é definida. Segundo um certo critério os pontos se unem sucessivamente, reduzindo o tamanho da matriz e formando os grupos. Depois de unir todos os elementos tem-se uma hierarquia que pode ser representada em forma de dendrograma. Os diferentes critérios de ligação definem os algoritmos e podem ser mais ou menos úteis dependendo dos dados a serem agrupados. A ligação simples junta dois grupos se a menor distância entre qualquer par de elementos for pequena. Na ligação média a distância a ser considerada é a média das distâncias dos elementos de ambos grupos. A ligação centroide considera a distância entre os centroides dos grupos igual à média, que é mais efetivo com grupos regulares. Uma ligação similar é a mediana. Na ligação completa considera-se a máxima distância entre pares de elementos dos grupos, o que pode ser ruim na presença de elementos muito afastados. A ligação com peso atribui mais importância aos grupos maiores. Na ligação Ward a distância entre dois grupos é a soma dos desvios padrão dos elementos com relação aos centroides visando a minimizar a soma de quadrados dentro dos grupos.

O último tipo de algoritmo de agrupamento é chamado de maximização da expectativa. Este pressupõe que os elementos estão posicionados segundo uma distribuição gaussiana. Mais do que um agrupamento, o algoritmo fornece as probabilidades de que um elemento pertença a um determinado grupo. Este algoritmo é especialmente útil quando existe superposição dos grupos. No entanto, a suposição de que os grupos são distribuídos segundo uma Gaussiana pode ser muito forte.

CAPÍTULO 3

METODOLOGIA

Neste trabalho propõe-se uma metodologia para obter medidas da evolução temporal, isto é, da dinâmica das redes de co-ocorrência de textos escritos. As medidas devem caracterizar o texto, já que são usadas posteriormente para identificar a autoria do texto sob análise. Como primeiro passo devemos escolher uma coleção de textos para aplicar a metodologia.

3.1 Corpus

O corpus foi obtido da base de dados aberta gutenberg.org, contendo 300 textos de 27 autores na língua inglesa criados desde meados do século XIX e no século XX (porém, três deles escreveram suas obras em outras línguas, posteriormente traduzidas ao inglês). Os textos incluem principalmente contos, novelas e ensaios. A coleção foi propositalmente construída com grande heterogeneidade no tamanho das obras como mostra a figura 3.1. O detalhe da figura 3.1 também mostra que o número de obras por autor é variável. Os dois autores mais prolíficos possuem respectivamente 52 e 49 textos enquanto 5 autores têm um texto cada. Esta heterogeneidade simula as características das coleções de textos reais. Dispondo de uma coleção de textos o primeiro passo numa análise computacional é a preparação dos textos.

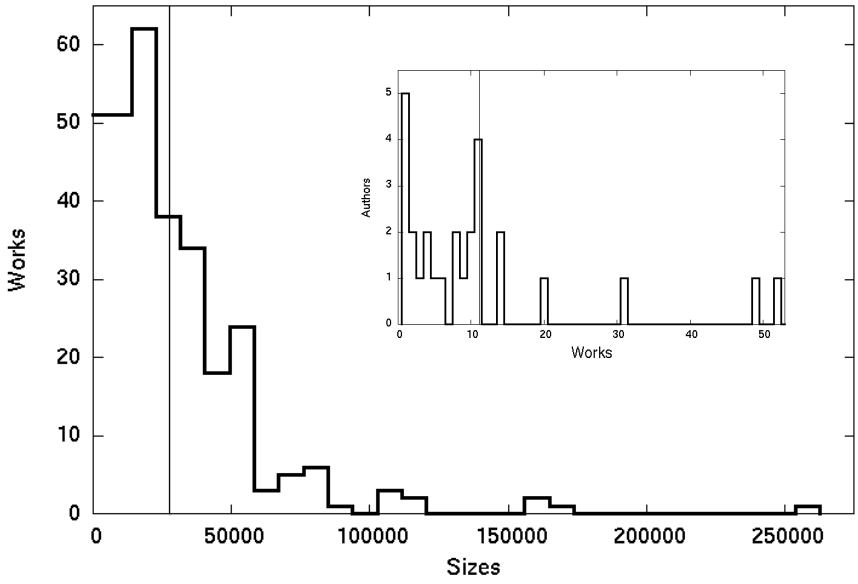


Figura 3.1 – Corpus utilizado. Distribuição do tamanho dos textos em número de palavras depois do pré-processamento. A linha fina assinala a média, 27755.8 palavras. Número de bins $= 30$. Detalhe: distribuição do número de livros por autor. A linha fina assinala a média, 11.1 livros por autor.

3.2 Pré-processamento

O conhecimento de especialistas só é usado neste passo preparatório da metodologia. Primeiro, é preciso remover as chamadas *stopwords*, como palavras funcionais consideradas pouco informativas, tais como os artigos e preposições. O método mais simples é usar uma lista contendo aquelas palavras que se deseja retirar. O seguinte passo é a lematização ou *stemming* na qual uma ou várias palavras são substituídas por um único conceito que descreve todas elas, como um verbo no infinitivo. Esta tarefa pode ser crítica em outras aplicações; por exemplo, se existirem vários significados para uma mesma palavra. No terceiro passo se reduz o número de palavras (conceitos) com a ajuda de um *thesaurus* onde as palavras estão ligadas segundo a dependência semântica, a mais simples das quais é a existente entre sinônimos.

Estas tarefas são feitas usando a base de dados léxica WordNet. Os três passos anteriores convertem a vasta quantidade de palavras de um texto escrito a uma lista ordenada de conceitos ou *tokens* que podem se repetir dentro da lista. No nosso caso os possíveis erros que estes possam causar são compensados de sobra pelo benefício. Como veremos, as redes construídas são extremamente esparsas e portanto diminuir o número de nós ajuda a ter maior

diversidade de comportamentos.

3.3 Criação de redes de co-ocorrência

A partir da lista ordenada de tokens, criamos redes de co-ocorrência. Primeiro criamos uma rede sem arestas cujos nós são os tokens (sem repetição). Em seguida cada conceito (token) é associado com o seguinte da lista. Se a aresta já estiver presente na rede aumentamos o valor de seu peso em uma unidade (uma nova aresta possui peso igual a um). Observe-se que na passagem por um dado token ele vai adquirir duas arestas, uma com o token anterior e outra com o seguinte (exceto pelo primeiro e o último token).

3.4 Extração de atributos dinâmicos

O objetivo do trabalho é extrair atributos que refletem a evolução temporal da rede. Como se comentou anteriormente isto pode trazer problemas. A abordagem proposta é a seguinte. Primeiro deve-se dividir a lista ordenada de tokens em janelas de igual tamanho (igual número de tokens e incluindo os repetidos). Os tokens restantes do final da lista não se consideram. Posteriormente se constrói uma rede de co-ocorrência para cada uma das janelas. De cada rede construída se extrai a média aritmética sobre todos os nós da rede das métricas. Neste trabalho utilizam-se: a conectividade, a centralidade de intermediação, o menor caminho e a intermitência média. Junto com eles se guardam o número total de nós e o de arestas da rede. Pela construção anterior o peso total nas arestas é fixo, sendo o único parâmetro do modelo. Teremos então para cada métrica uma lista ordenada de valores ou série temporal.

O resto do procedimento depende do resultado deste passo. Se todas as séries temporais

das métricas podem ser consideradas estacionárias (ver capítulo 2) continuamos a calcular os primeiros momentos da série. Caso contrário, técnicas mais sofisticadas de análise de séries temporais devem ser usadas. Para cada série $\{x_1, x_2, \ldots x_T\}$ de cada métrica calculamos a versão linearizada do i-ésimo momento como $\mu_i = \left[(T-1)^{-1}\Sigma_{j=1}^{T}(x_j - \mu_i)^i\right]^{1/i}$, onde $i > 1$ e μ_1 é o valor médio da série. Neste trabalho usamos os quatro primeiros momentos de cada série, portanto, havendo 6 séries das métricas de um texto teremos 24 atributos por texto.

3.5 Reconhecimento de autoria

Com os 24 atributos dinâmicos de rede que representam cada texto obtemos uma matriz de ordem 300 (número de elementos ou instâncias) por 24, chamada de matriz de atributos. Salvamos também um vetor de cadeias de caracteres de comprimento igual a 300 de tal forma que para o i-ésimo texto, a i-ésima fila da matriz de atributos contém os valores dos atributos do texto e o i-ésimo elemento do vetor contém a etiqueta da classe (autor) correspondente ao texto. Com esta informação criamos um arquivo de entrada para algum pacote de aprendizado de máquina. Usamos o pacote WEKA disponibilizado pela Universidade de Waikato.

Neste trabalho usam-se 5 algoritmos de classificação, cada um representativo de uma classe geral de algoritmos: ZeroR, OneR, K-Nearest Neighbors, Naive Bayes e J48. Na parte de agrupamento usam-se os seguintes algoritmos: Algoritmos baseados em densidade, DBSCAN e OPTICS; Algoritmos Aglomerativos Hierárquicos, entre os quais se usam as ligações: centroide, completa, média, mediana, ponderada, singular e Ward; e o algoritmo de Maximização da Expectativa.

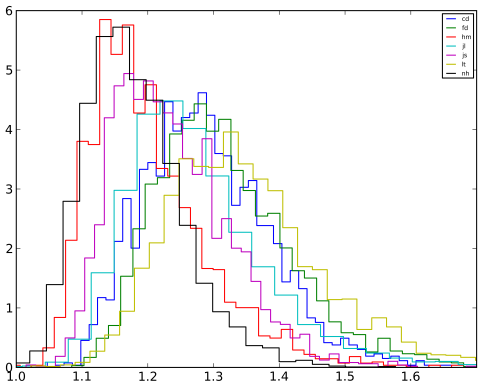
CAPÍTULO 4

RESULTADOS PRELIMINARES

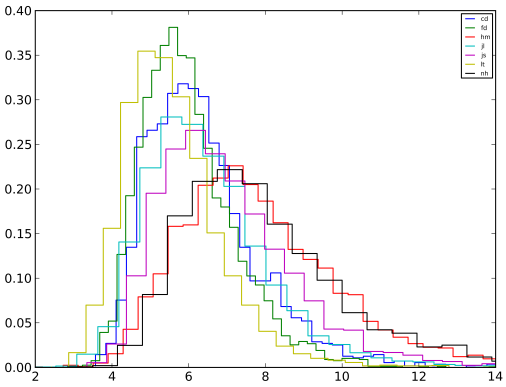
Conseguiu-se observar um comportamento da dinâmica das redes consistente entre os textos usados. Em geral verifica-se que a dinâmica é estacionária, pelo menos no que se refere ao comportamento da rede como um todo. As medidas consideradas oscilam, mas sempre ao redor de um valor médio fixo e estas oscilações mantêm sempre o mesmo formato. Podemos observar isto na figura 4.1 que mostra os histogramas normalizados, gerados a partir das séries temporais da conectividade e do menor caminho para os 7 autores com maior volume total de textos da coleção. Para efeitos de visualização os histogramas são obtidos juntando-se as séries temporais de toda a produção bibliográfica de cada autor e criando-se uma única série. Além de estacionárias, as séries temporais são representativas de cada autor. Observe que cada histograma tem um valor médio, uma variância e uma assimetria (*skewness*) bem diferenciadas. Para o menor caminho, o valor médio das distribuições parece seguir uma tendência na parte (b) da figura 4.1; porém, no caso da conectividade a ordem dos picos se altera (parte (a) da figura) quase dividindo-se em dois conjuntos de histogramas. Este comportamento garante que nossa escolha dos primeiros momentos dos histogramas como atributos para o aprendizado de máquina diferencie um autor de outro.

4.1 Classificação

O aprendizado de máquina supervisionado ou classificação dos textos foi realizado usando 5 classificadores. Um deles, o ZeroR, foi usado como limiar inferior para comparar o desem-



(a) Histogramas da conectividade.



(b) Histogramas do menor caminho.

Figura 4.1 – Histogramas normalizados de séries temporais da conectividade (a) e do menor caminho (b) para os 7 autores com maior volume total de texto.

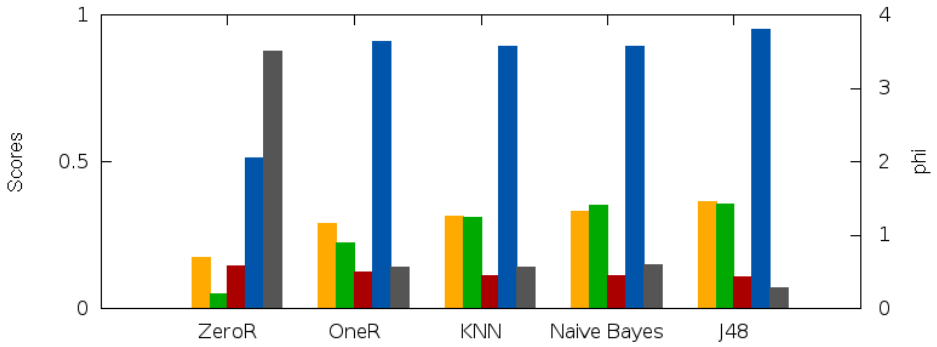


Figura 4.2 – Resultados dos algoritmos de classificação. Para a coleção completa: valores F micro- (amarelo) e macro- (verde) ponderadas e $\phi = \sqrt{\chi^2/N}$ (vermelho). Para os dois autores mais prolíficos valores F micro-ponderadas (azul) e $\phi = \sqrt{\chi^2/N}$ (cinza).

penho dos demais. Para a coleção completa de textos, e considerando que o autor com mais textos tem 52, ZeroR fornece uma taxa de acerto de 17.3%. Os outros quatro algoritmos praticamente dobram tal valor, com taxas similares entre si. O melhor desempenho foi uma taxa de acerto de 36.3% com o algoritmo de árvore de decisão $\mathrm{J48}$. Também consideramos só os dois autores mais prolíficos, com 52 e 49 obras respectivamente, para testar a capacidade de desambiguação do método. ZeroR tem uma taxa de acerto de 51.5%, ou seja, uma chance de aproximadamente um meio de acertar o autor de um texto. Os outros algoritmos superam de longe este limiar, o melhor sendo novamente o $\mathrm{J48}$ com uma taxa de acerto de 95%.

Estes resultados estão na figura 4.2. As barras amarelas e azuis correspondem às taxas de acerto para a coleção inteira e os dois autores mais prolíficos respectivamente. Elas são denominadas valores F micro-ponderadas (do inglês *micro-averaged F-scores*) já que são as

médias aritméticas das taxas de acerto de cada uma das classes (autores). Em contraste, para coleções que contêm várias classes com diversidade no número de elementos em cada classe (que é o nosso caso), uma média aritmética pode ser enganosa, sendo mais conveniente uma média ponderada pelo número de elementos por classe. Esta é chamada de valor F macro-ponderada ou *macro-averaged F-score*. Para a coleção completa, os F macro-ponderadas aparecem como barras verdes. Neste caso a diferença é mais marcante passando de 5% com o ZeroR para 35% com $\mathrm{J48}$. Para os dois autores mais prolíficos, as ponderações micro- e macro- têm pouca diferença, já que o número de textos dos dois autores é quase igual. A medida χ^2 ou chi-quadrado quantifica a dispersão dos elementos incorretamente classificados. Uma medida normalizada, definida como $\phi = \sqrt{\chi^2/N}$, mostra leve queda (do ZeroR para o $\mathrm{J48}$) de 0.58 a 0.43 para a coleção inteira (barras vermelhas) e uma bem mais visível, de 3.5 a 0.28 para os dois autores mais prolíficos (barras cinza).

As taxas de acerto dependem fortemente do corpus utilizado, mas uma comparação grosseira mostra que a metodologia proposta apresenta desempenho muito similar, tanto quanto às técnicas tradicionais como às mais recentes que usam redes de co-ocorrência. Por exemplo, em (12), um trabalho de reconhecimento autoral de poemas baseado em sequências de letras atinge 70% de sucesso ao comparar 3 autores, o que seria o dobro do desempenho do ZeroR se aplicado ao corpus.

4.2 Agrupamento

O agrupamento dos textos desconsidera as etiquetas previamente conhecidas e procura os grupos naturalmente formados. Devido às características dos algoritmos, a avaliação do desempenho dos algoritmos deve ser feita manualmente comparando-se as classes (os autores predefinidos) com os grupos gerados*. Os algoritmos baseados em densidade mostraram os resultados mais promissores. No DBSCAN fixamos o parâmetro $MinPoints = 2$ para encontrar todos os grupos, mesmo os menores. É claro que para este algoritmo os 5 autores

*Conhecendo as classes, medidas de avaliação internas perdem sentido.

que tenham um só texto devem ser interpretados como ruído[†]. Usando o valor padrão para o raio, $\varepsilon = 0.9$, obtemos somente 6 textos (2% da coleção) mal agrupados, sendo um grupo de dois elementos e outros quatro elementos erroneamente considerados como ruído. Fixando o raio em $\varepsilon = 1$, obtemos apenas quatro erros no agrupamento: novamente um grupo de dois elementos, e dois outros elementos se afastam dos seus grupos, e portanto são considerados como ruído. Notavelmente, nenhum elemento ficou em um grupo que não correspondesse ao do seu autor.

O algoritmo OPTICS é uma generalização do DBSCAN que permite ter uma ideia visual do agrupamento, cujo resultado é mostrado na figura 4.3. Observe-se que os poços, isto é, os grupos gerados, têm aproximadamente a mesma profundidade, o que significa uma densidade constante dentro de cada grupo. A altura da separação entre poços é dada pela menor distância desde o último elemento de um grupo até um elemento do grupo seguinte. Estas alturas são quase constantes, o que significa que existe uma distância característica entre grupos. Todas as distâncias entre grupos ficam numa estreita faixa, entre 1.00175 e 1.0285, exceto pelos 3 últimos elementos listados. Isto explica o sucesso obtido com $\varepsilon = 1$ no DBSCAN. Igualmente importante é o fato de não se ter encontrado nenhum elemento num grupo diferente ao do seu mesmo autor.

Outra classe de algoritmos muito usada é o agrupamento aglomerativo hierárquico. Dos vários critérios de ligação conhecidos, a ligação singular é a que apresentou os melhores resultados. Se fixarmos o número de grupos na nossa hierarquia em 27, isto é, o número original de autores, dois grupos dividem-se deixando um grupo com 8 elementos e quatro grupos com um elemento cada. Para manter o número de grupos requerido, os dois maiores grupos juntam-se a outros três. Para evitar esta fusão de grupos aumentamos o número de grupos requerido a 30 e obtemos uma concordância perfeita com os autores, exceto pelos três textos que agora são considerados grupos independentes.

A maior fraqueza da ligação singular são os elementos inseridos entre dois grupos, os quais podem criar uma ponte fazendo com que os grupos se juntem. Concluímos que os poucos elementos perdidos da nossa coleção estão tão longe de seus respectivos grupos quanto dos

[†]Uma vantagem do agrupamento sobre a classificação é que para o segundo é preciso dividir o corpus num conjunto de treino e um de teste, fazendo com que autores com poucos textos sejam classificados erroneamente.

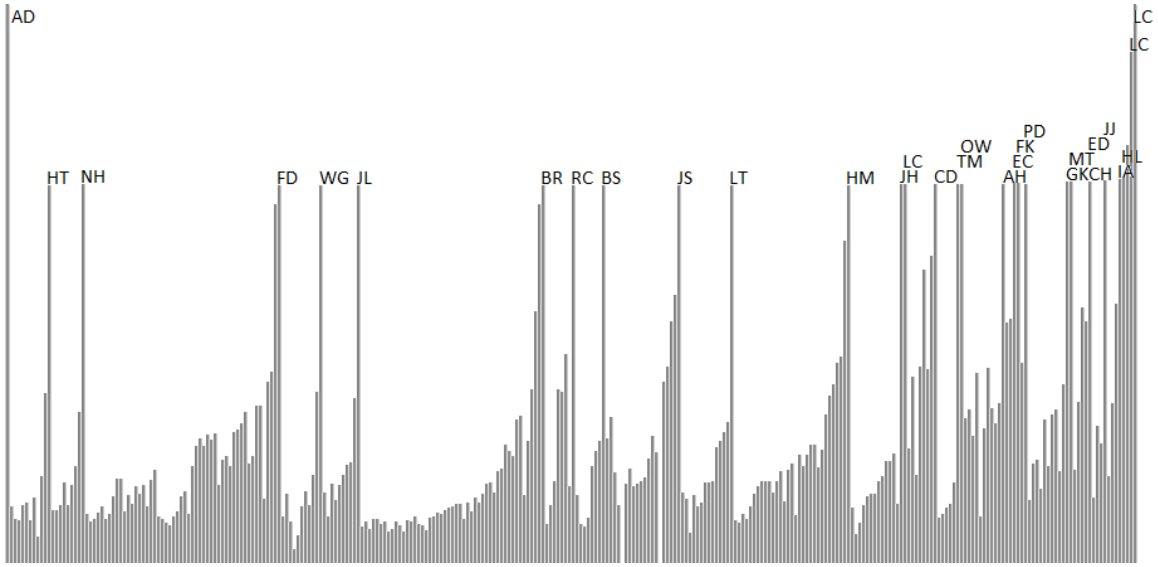


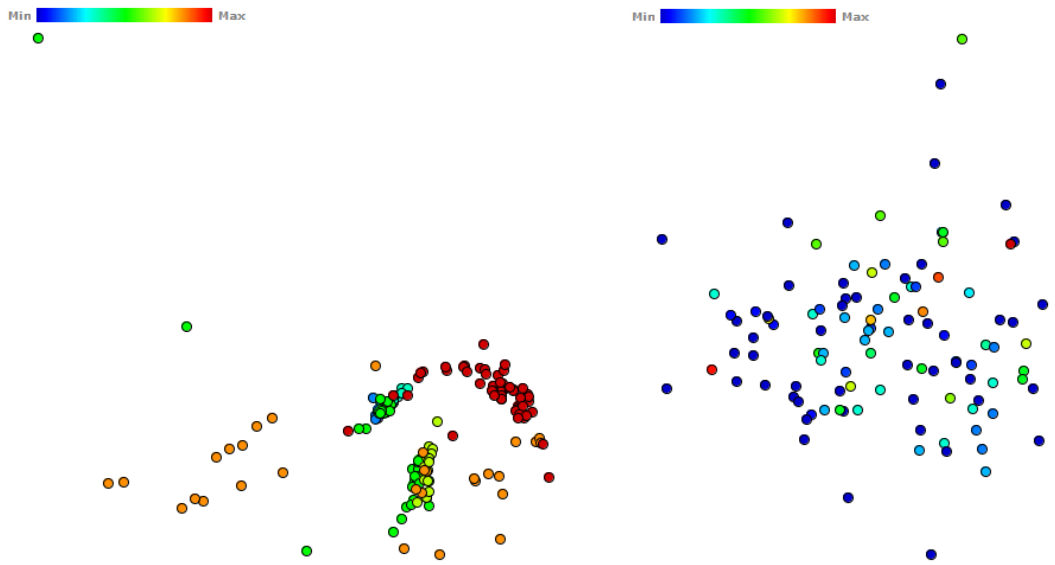
Figura 4.3 – Resultados do agrupamento OPTICS. As barras que possuem etiquetas correspondem ao autor dado pela etiqueta, as barras sem etiquetas correspondem ao mesmo autor dado pela etiqueta mais próxima à esquerda. A altura da maior distância plotada é $\varepsilon = 1.4$, $MinPoints = 2$. Figura adaptada do pacote WEKA.

outros. Além disso, o desempenho ruim com outros critérios de ligação significa que os grupos têm formas irregulares já que centroides e médias não representam corretamente os elementos do grupo.

Por último, usamos o algoritmo de maximização da expectativa. Neste caso não se obteve um bom agrupamento. O algoritmo de maximização da expectativa serve para lidar com grupos misturados e de elementos normalmente distribuídos de forma elíptica. Este algoritmo confirma que os grupos não se misturam consideravelmente e que têm formas irregulares e de densidade de elementos constante.

4.3 Visualização

Os algoritmos de redução da dimensionalidade ou projeção permitem visualizar os elementos em gráficos de dispersão bidimensionais. Em geral se observa superposição dos grupos. Na figura 4.4 é mostrada a melhor projeção, obtida usando *Isometric Feature Mapping* junto com uma redução a 15 dimensões usando Análise de Componentes Principais. Para comparar



(a) Coleção atual. (b) Coleção do Machado de Assis.

Figura 4.4 – Projeção Isometric Feature Mapping. (a): Os sete autores com maior volume de produção da coleção utilizada, contendo 177 textos. (b): coleção do Machado de Assis, contendo 106 textos.

os resultados, na parte (b) da figura mostramos a mesma técnica aplicada a um outro corpus correspondente ao autor Brasileiro Joaquim Machado de Assis. Machado assinou suas obras com diferentes pseudônimos, representados como as diferentes cores na parte (b) da figura. Os grupos (pseudônimos) no caso do Machado ficam totalmente misturados, enquanto os textos da coleção ficam razoavelmente separados.

CAPÍTULO 5

CONCLUSÕES GERAIS

A pesquisa realizada permite inferir aspectos gerais sobre a dinâmica de redes de co-ocorrência, em que medidas de topologia e de intermitência capturam a identidade do autor de um texto. Não somente a média de cada série, que poderia se associar com a média da distribuição sobre os nós da rede de um texto inteiro, mas os momentos de ordem superior são relevantes. Isto se verifica por exemplo observando a organização espacial dos elementos nos gráficos com um corte correspondente a estes momentos de ordem superior. Outra prova disto é a árvore de decisão obtida com o algoritmo J48 onde vários dos nós principais correspondem a estes momentos, o que significa que a sua utilização resulta num alto ganho de informação.

A utilidade das medidas introduzidas baseia-se no fato de as séries serem estacionárias, pois caso contrário o método não seria viável. Por exemplo, numa série na qual o valor esperado cresça com o tempo não faz sentido obter a média já que ela vai depender, entre outras coisas, do tamanho da série. Observe que só é preciso garantir uma forma fraca de estacionariedade na qual os momentos considerados sejam constantes. Entre as medidas usadas encontra-se o número de nós da rede, cujos momentos foram notórios ao longo do trabalho. O número de nós reflete a importância do peso nas arestas, que ficou fixo. É razoável pensar que o número de vezes no qual duas palavras aparecem juntas no texto possa ser caraterístico de um autor, mais do que a simples presença ou ausência do par.

Com respeito ao desempenho das medidas propostas, observamos: Os algoritmos de classificação apresentam um comportamento comparável, às vezes superior, ao obtido com outras técnicas bem estabelecidas de classificação de textos. Os resultados dependem fortemente do número de obras por autor assim como do número de autores, o que deve ser considerado

para uma aplicação prática. No caso do agrupamento, o melhor desempenho foi obtido com algoritmos baseados em densidade. A ferramenta OPTICS revelou que existe uma distância característica entre os grupos. Consistentemente, observou-se que os textos de dois autores não estão mais perto do que um raio $\varepsilon = 1$ (no espaço de atributos normalizado) ficando na estreita faixa entre 1.001 e 1.02. Além de uma distância característica inter-grupos, as distâncias intra-grupos são baixas e razoavelmente constantes como mostra a figura 4.3. A metodologia apresentada mostrou ser efetiva no tratamento de uma coleção heterogênea, e no caso do agrupamento, consegue juntar os textos de autores com poucas obras, mesmo aqueles com duas obras só.

O método introduzido apresenta uma vantagem sobre o outro método de redes de co-ocorrência para reconhecimento de autoria (7): a independência com relação aos elementos individuais da coleção. O método em (7) corta todos os textos pelo texto de menor tamanho para poder fazer a comparação das redes. Além de modificar drasticamente os objetos de estudo, todos os cálculos vão depender deste menor tamanho e deverão ser refeitos caso se introduza um novo elemento na coleção. Com a nova metodologia apresentada aqui, a coleção pode ser construída gradualmente e os possíveis problemas de um texto somente afetarão a classificação desse mesmo texto.

Uma fraqueza que em princípio poderia ser observada é o tratamento de textos curtos (fraqueza comum a todos os métodos de análise de textos). No método apresentado, um texto curto gera uma rede curta. Por sua vez, os momentos calculados a partir de uma rede curta podem não ser confiáveis fazendo com que a posição do elemento no espaço de atributos esteja longe dos outros elementos. No entanto, este não parece ser o caso e quando elementos ficaram afastados foi devido à formatação dos textos, como no exemplo da figura 4.4.

CAPÍTULO 6

CRONOGRAMA

Ano	Semestre	Atividade
2012	II	Revisão e estudo da bibliografia.
2013	I	Implementação computacional.
	II	Cursar disciplinas.
2014	I	Implementação computacional: refinamento do código e corpus.
	II	Apresentação de resultados.
2015	I	Exame de qualificação. Escrita de artigo.
	II	Cursar disciplina. Monitoria PAE.
2016	I	Defesa do doutorado. Escrita de artigo.

Tabela 6.1 – Cronograma de atividades

6.1 Disciplinas Cursadas

As disciplinas foram escolhidas visando a aperfeiçoar os conhecimentos gerais da física e adquirir os necessários na área de aprendizado de máquina e mineração de dados. As três disciplinas cursadas até agora são:

Tópicos especiais em teoria de muitos corpos É uma das disciplinas requeridas pelo instituto. O foco são as teorias de campo de partículas elementares.

Mineração de dados não estruturados Apresenta uma visão geral das diferentes áreas de mineração de dados na atualidade, incluindo uma revisão das técnicas para mineração de textos. É uma disciplina útil para conhecer o estado da arte em reconhecimento de autoria.

Análise de agrupamento de dados Concentra-se no aprendizado de máquina não-supervisionado detalhando nos conceitos e contas. Serve para aprender os principais algoritmos usados na atualidade.

REFERÊNCIAS BIBLIOGRÁFICAS

1 BARABÁSI, A.-L.; ALBERT, R. Emergence of scaling in random networks. *science*, American Association for the Advancement of Science, v. 286, n. 5439, p. 509–512, 1999.

2 CANCHO, R. Ferrer i; SOLÉ, R. V. Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, Taylor & Francis, v. 8, n. 3, p. 165–173, 2001.

3 DOROGOVTSEV, S. N.; MENDES, J. F. F. Language as an evolving word web. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, The Royal Society, v. 268, n. 1485, p. 2603–2606, 2001.

4 CHOUDHURY, M.; MUKHERJEE, A. The structure and dynamics of linguistic networks. In: *Dynamics on and of Complex Networks*. [S.l.]: Springer, 2009. p. 145–166.

5 BIEMANN, C.; QUASTHOFF, U. Networks generated from natural language text. In: *Dynamics on and of Complex Networks*. [S.l.]: Springer, 2009. p. 167–185.

6 AMANCIO, D. R.; JR, O. N. O.; COSTA, L. da F. Identification of literary movements using complex networks to represent texts. *New Journal of Physics*, IOP Publishing, v. 14, n. 4, p. 043029, 2012.

7 AMANCIO, D. R. et al. Comparing intermittency and network measurements of words and their dependence on authorship. *New Journal of Physics*, IOP Publishing, v. 13, n. 12, p. 123024, 2011.

8 BOCCARA, N. *Modeling complex systems*. [S.l.]: Springer Science & Business Media, 2010.

9 NEWMAN, M. *Networks: an introduction*. [S.l.]: Oxford University Press, 2010.

10 CHATFIELD, C. *The analysis of time series: an introduction*. [S.l.]: CRC press, 2013.

11 HAMILTON, J. D. *Time series analysis*. [S.l.]: Princeton university press Princeton, 1994.

12 HOORN, J. F. et al. Neural network identification of poets using letter sequences. *Literary and Linguistic Computing*, ALLC, v. 14, n. 3, p. 311–338, 1999.