Desenvolvimento de novos modelos para reconhecimento de autoria com a utilização de redes complexas

Vanessa Queiroz Marinho

Vanessa Queiroz Marinho

Desenvolvimento de novos modelos para reconhecimento de autoria com a utilização de redes complexas

Monografia apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, para o Exame de Qualificação, como parte dos requisitos para obtenção do título de Mestra em Ciências – Ciências de Computação e Matemática Computacional.

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
Janeiro de 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

Marinho, Vanessa Queiroz
M634d Desenvolvimento de novos modelos para reconhecimento de autoria com a utilização de redes complexas / Vanessa Queiroz Marinho; orientador Diego Raphael Amancio. - São Carlos - SP, 2016.
85 p.

Monografia (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. Reconhecimento de Autoria. 2. Redes Complexas. 3. Teoria dos Grafos. 4. Processamento de Texto. I. Amancio, Diego Raphael, orient. II. Título.

Vanessa Queiroz Marinho

Development of new models for authorship recognition using complex networks

Monograph submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, as part of the qualifying exam requisites of the Master Program in Computer Science and Computational Mathematics.

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Diego Raphael Amancio

USP – São Carlos
January 2016

RESUMO

MARINHO, V. Q.. **Desenvolvimento de novos modelos para reconhecimento de autoria com a utilização de redes complexas**. 2016. 85 f. Monografia (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

A modelagem de grafos e redes complexas vem sendo aplicada com sucesso em diferentes domínios, sendo objeto de estudo de distintas áreas que incluem, por exemplo, a matemática e a computação. A descoberta de que métodos derivados do estudo de redes complexas podem ser utilizados para analisar textos em seus distintos níveis de complexidade proporcionou grandes avanços em tarefas de processamento de línguas naturais. Exemplos de aplicações analisadas com os métodos e ferramentas de redes complexas são a detecção de conceitos relevantes, a criação de sumarizadores extrativos automáticos e reconhecedores de autoria. Esta última tarefa, que é foco deste projeto de mestrado, tem sido estudada com certo sucesso através da representação de redes de adjacência de palavras que conectam apenas as palavras mais próximas. O objetivo deste projeto de mestrado é estender a modelagem tradicional, escolhendo-se a janela de conexão ótima para o problema, para um dado conjunto de treinamento. Além disso, pretende-se utilizar informação de conectividade de palavras funcionais para complementar a caracterização de estilo de autores. Finalmente, pretende-se criar classificadores híbridos que sejam capazes de combinar fatores tradicionais com as propriedades fornecidas pela análise topológica de redes complexas. Através da adaptação, combinação e aperfeiçoamento da modelagem, pretendemos não apenas melhorar o desempenho dos sistemas de caracterização estilística textual e reconhecimento de autoria, mas também entender melhor quais são os fatores quantitativos textuais (medidos via redes) que podem ser utilizados na área de estilometria. Os avanços obtidos durante este projeto podem ser úteis para estudar aplicações relacionadas, como é o caso de análise de inconsistências estilísticas e plagiarismos.

Palavras-chave: Reconhecimento de Autoria, Redes Complexas, Teoria dos Grafos, Processamento de Texto.

ABSTRACT

MARINHO, V. Q.. **Desenvolvimento de novos modelos para reconhecimento de autoria com a utilização de redes complexas**. 2016. 85 f. Monografia (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

The modelling of graphs and complex networks has been successfully applied in different fields, being the object of study in different areas including, for example, mathematics and computer science. The discovery that methods derived from the study of complex networks can be used to analyze texts in their different complexity levels provided great advances in natural language processing tasks. Examples of applications analyzed with the methods and tools of complex networks are the detection of relevant concepts, development of automatic summarizers and authorship recognition systems. The latter task, which is the focus of this Master's project, has been studied with some success through the representation of adjacency networks that connect only the closest words. The purpose of this Master's project is to extend the traditional modelling, choosing the optimal connection window to the problem, for a given training set. In addition, we intend to use the connectivity information of function words to complement the characterization of authors' style. Finally, we intend to create hybrid classifiers that are able to combine traditional factors with properties provided by the topological analysis of complex networks. By adapting, combining and improving the model, we aim not only to improve the performance of textual stylistic characterization and authorship recognition systems, but also better understand what are the textual quantitative factors (measured through networks) that can be used in stylometry. The advances obtained during this project may be useful to study related applications, such as the analysis of stylistic inconsistences and plagiarism.

Key-words: Authorship Recognition, Complex Networks, Graph Theory, Text Processing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A modelagem de sistemas reais por meio de redes complexas têm sido útil para descrever uma grande variedade de sistemas encontrados no mundo real (ALBERT; BARABáSI, 2002). Alguns exemplos incluem a célula, que pode ser descrita como uma rede de substâncias conectadas por reações químicas, e a Internet, uma rede de roteadores e computadores conectados por links físicos (MIHALCEA; RADEV, 2011). Muitos trabalhos científicos na área de redes complexas são interdisciplinares, beneficiando-se de ideias provenientes de diferentes disciplinas, como matemática, física, biologia, ciência da computação, ciências sociais e outras (NEWMAN, 2010).

Tradicionalmente, o estudo de redes limitava-se à teoria dos grafos aplicada a sistemas aleatórios. Um dos precursores da teoria dos grafos foi o matemático Leonhard Euler que, em 1736, solucionou o famoso problema das pontes de Kőnigsberg. Desde então, a teoria dos grafos tem se beneficiado de grandes avanços (BOCCALETTI *et al.*, 2006). Nos últimos anos, no entanto, a descoberta de que vários sistemas reais podem ser caracterizados por uma estrutura de rede não aleatória (NEWMAN, 2010) permitiu um rápido desenvolvimento da área. A informatização e disponibilização de dados de várias áreas do conhecimento e o aumento do poder computacional proporcionaram a análise e o desenvolvimento de algoritmos eficientes em um amplo conjunto de aplicações de naturezas distintas. De particular interesse para os objetivos deste projeto, podemos citar as redes textuais, que podem ser formadas por relações sintáticas (CANCHO; SOLÉ; KÖHLER, 2004; AMANCIO *et al.*, 2012), semânticas (LIU, 2009) ou empíricas (LUDUEñA; BEHZAD; GROS, 2014). Nestes três tipos de redes, as propriedades universais *small-world* e *scale-free* foram encontradas (CANCHO; SOLé, 2001).

Um caso particular das redes sintáticas são as redes de co-ocorrência de palavras (ou redes de adjacência de palavras) (CANCHO; SOLé, 2001). Em tais redes, as conexões sintáticas são aproximadas pelas conexões de palavras adjacentes, já que a maioria das conexões sintáticas ocorre entre palavras vizinhas (CANCHO; SOLÉ; KÖHLER, 2004). O sucesso da representação

de elementos textuais como redes de ocorrências pode ser comprovado a partir de aplicações de reconhecimento de movimentos literários (AMANCIO; OLIVEIRA JR; COSTA, 2012a), sumarização automática (AMANCIO *et al.*, 2012), entre outras. Estudos das propriedades de tais redes desenvolvidos em (AMANCIO *et al.*, 2011; AMANCIO *et al.*, 2013a) demonstraram que a maioria das medidas topológicas da representação por redes captura características da sintaxe e estilo da linguagem. Por este motivo, estas redes são mais adequadas para tratar problemas de estilo, embora uma dependência entre topologia e semântica de palavras tenha sido também comprovada (SILVA; AMANCIO, 2013). Neste contexto, pretende-se fazer uso do poder discriminativo gerado por redes de co-ocorrência de palavras para discriminar estilos na tarefa de reconhecimento de autoria.

Métodos de reconhecimento de autoria são relevantes pois estes podem ser aplicados para classificar obras literárias, resolver direitos autorais (GRANT, 2007) ou até mesmo para interceptar mensagens terroristas (ABBASI; CHEN, 2005). As primeiras técnicas de reconhecimento de autoria foram exploradas após o famoso trabalho de Mosteller e Wallace na análise dos artigos federalistas (MOSTELLER; WALLACE, 1964). Após este estudo, pesquisadores têm proposto novos atributos para caracterizar estilos (JUOLA, 2006). Alguns dos atributos tradicionalmente utilizados para a tarefa incluem a análise das propriedades estatísticas de palavras (por exemplo, o comprimento médio, a frequência, o *burstiness* e a riqueza de vocabulário) (STAMATATOS, 2009) e caracteres (frequências e correlações) (GRANT, 2007). Além dos atributos léxicos, atributos sintáticos (por exemplo, frequências de *chunks* específicos) e semânticos também têm sido utilizados como atributos úteis para o problema (STAMATATOS, 2009). Atualmente, novos atributos têm sido propostos para a geração de classificadores robustos (STAMATATOS, 2009).

A adequabilidade das medidas de redes complexas para a tarefa foi observada em alguns trabalhos (ANTIQUEIRA *et al.*, 2006; AMANCIO *et al.*, 2011), nos quais os autores mostraram que as características topológicas das redes podem ser utilizadas para identificar autores. Este projeto pretende estender a metodologia utilizada nesses trabalhos com a introdução de novas modelagens e novas formas de caracterizar estilos. Além disso, pretende-se analisar a robustez da modelagem com relação ao tamanho do texto, escolha de palavras específicas e outros fatores. Como a análise topológica não considera os rótulos dos vértices, essa informação será incluída através da implementação de classificadores híbridos. Desse modo, duas componentes serão consideradas, a componente topológica (medidas de redes complexas) e a componente proveniente de métodos tradicionais, como a frequência de palavras funcionais específicas.

De maneira geral, este projeto pretende combinar técnicas tradicionais com técnicas de redes. Espera-se, portanto, uma melhoria de performance da tarefa, já que as desvantagens de cada técnica devem ser superadas pela classificação híbrida. Além de contribuir para melhorar o desempenho da tarefa, espera-se criar classificadores mais robustos que os atuais, que são vulneráveis a ataques (BRENNAN; GREENSTADT, 2009). Em resumo, os principais objetivos desse trabalho podem ser descritos pelo seguinte parágrafo:

Este projeto de mestrado tem por objetivo desenvolver novos modelos para o reconhecimento de autoria com o uso de redes complexas. Especificamente, pretende-se adaptar o modelo de adjacência ao incluir palavras de pouco conteúdo semântico e conectar palavras em um maior contexto (não apenas palavras vizinhas, como nas redes de adjacência convencionais). No lado topológico, novas medidas serão introduzidas para aperfeiçoar a caracterização da topologia. Por fim, espera-se ainda otimizar os métodos atuais de reconhecimento de autoria combinando os atributos obtidos de medidas topológicas com os atributos tradicionais em um classificador híbrido.

1.1 Organização da Monografia

Este trabalho está organizado da seguinte maneira. No Capítulo 2 é apresentada a fundamentação teórica referente às duas áreas relacionadas a esse projeto, as redes complexas e o reconhecimento de autoria. No Capítulo 3 são apresentados os principais trabalhos encontrados na literatura que tratam a tarefa de reconhecimento de autoria através do uso de redes complexas. Por fim, no Capítulo 4 são apresentadas a motivação e os objetivos, a metodologia, as principais tarefas propostas, alguns resultados iniciais e o cronograma a ser seguido na execução deste projeto de Mestrado.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações Iniciais

Para alcançar os objetivos propostos no Capítulo 1, este trabalho aborda o problema do reconhecimento de autoria através do uso de redes complexas. Neste capítulo, são apresentados os conceitos fundamentais relacionados às duas grandes áreas nas quais este trabalho está inserido, as redes complexas e o reconhecimento de autoria. Na Seção 2.2, são apresentadas algumas definições, os principais modelos de redes e as medidas de redes complexas utilizadas neste trabalho. Na Seção 2.3, são apresentadas algumas definições, os primeiros trabalhos realizados em reconhecimento de autoria e como eles evoluíram de acordo com o tempo. Ainda na Seção 2.3, os principais atributos estilométricos e métodos de atribuição são descritos.

2.2 Redes Complexas

2.2.1 Definições

Redes complexas têm sido utilizadas como base para a representação matemática de uma grande variedade de sistemas complexos. De acordo com Barabási (BARABÁSI, 2014), as propriedades emergentes de um sistema complexo não são facilmente inferidas apenas pela análise dos componentes do sistema. Dentre vários exemplos de sistemas complexos, podemos citar sistemas socias e biológicos, a Internet, a sociedade humana e muitos outros. Neste contexto, diversas pesquisas com redes complexas visam estudar e explicar como comportamentos específicos emergem desses sistemas (NEWMAN, 2010; COSTA *et al.*, 2011). Tradicionalmente, o estudo de redes limitava-se à teoria dos grafos aplicada a diversos sistemas aleatórios. Um dos precursores da teoria dos grafos foi o matemático Leonhard Euler que, em 1736, solucionou o famoso problema das pontes de Kőnigsberg (BARABASI, 2003). Desde então, a teoria dos grafos tem se beneficiado de grandes avanços (BOCCALETTI *et al.*, 2006).

Matematicamente, uma rede é um par ordenado $G = \{V, E\}$ formado por um conjunto $V = \{v_1, v_2, ..., v_n\}$ de vértices e por um conjunto $E = \{e_1, e_2, ..., e_m\}$ de arestas. A conectividade da rede pode ser representada através de uma matriz de adjacência A. Nessa matriz, os elementos A_{ij} podem apresentar valor 0 ou 1, onde $A_{ij} = 1$ se o vértice i estiver conectado com o vértice j. As redes podem ser direcionadas ou não direcionadas. No caso das redes não direcionadas, a matriz de adjacência A é simétrica, ou seja $A_{ij} = A_{ji}$ para todo vértice i e j. Uma maneira de representar a intensidade de conexões é através da associação de pesos às arestas (COSTA *et al.*, 2011). Desse modo, uma rede ponderada é definida por $G = \{V, E, W\}$, que, além dos conjuntos V e E apresentados anteriormente, possui o conjunto $W = \{w_1, w_2, ..., w_m\}$ de pesos.

Nos últimos anos, a área de redes complexas está recebendo bastante atenção da comunidade acadêmica. O principal motivo foi a descoberta de que vários sistemas reais podem ser caracterizados por uma estrutura de rede não aleatória (COSTA *et al.*, 2007). Outros fatores foram responsáveis pelo aumento do interesse no estudo de redes complexas (ALBERT; BARABáSI, 2002), como a informatização e disponibilização de dados de diversas áreas do conhecimento que permitiram o surgimento de grandes bases de dados de várias redes reais. Além disso, o aumento do poder computacional proporcionou a análise de redes contendo milhões de vértices. Por fim, a expansão das fronteiras de diversas disciplinas, como a biologia e a sociologia, permitiu aos pesquisadores o acesso às diversas bases de dados.

De acordo com Newman (2010), as redes mais utilizadas e estudadas podem ser divididas em quatro categorias: as redes tecnológicas, sociais, de informação e biológicas. Algumas redes podem ser classificadas em mais de uma categoria. A descrição de cada uma das categorias é apresentada a seguir:

- Redes Tecnológicas: Estas redes são redes artificiais que fornecem serviços para as sociedades modernas. Exemplos típicos são as redes de energia elétrica e a Internet.
- Redes Sociais: Uma rede social pode ser caracterizada como uma rede na qual os vértices são pessoas (ou grupos de pessoas) e as arestas representam alguma forma de interação social, como a amizade ou relações de trabalho.
- Redes de Informação: Estas redes são caracterizadas por dados conectados através de alguma relação. Os principais exemplos de redes de informação são a *World Wide Web* e as redes de citações.
- Redes Biológicas: Uma rede biológica pode ser utilizada para representar diversos padrões de interação entre elementos biológicos. Um exemplo é a representação das conexões entre os neurônios ou as reações químicas nas células.

Motivados pela diversidade de redes reais, os pesquisadores analisam as redes na tentativa de modelá-las e compará-las através de suas propriedades. A partir desses estudos, propriedades comuns entre as redes foram descobertas. Muitas redes reais compartilham características como

a presença de comunidades (GIRVAN; NEWMAN, 2002), motivos (MILO *et al.*, 2002), a distribuição do grau segundo uma lei de potência (CLAUSET; SHALIZI; NEWMAN, 2009), entre outras.

2.2.2 Modelos de Rede

Uma das maneiras de compreender os efeitos de diversas propriedades das redes é através da construção de modelos matemáticos (NEWMAN, 2010). Os modelos de rede são cada vez mais utilizados durante a investigação de diversos tipos de fenômenos. Estes modelos simulam os padrões de conexão de redes reais em uma tentativa de entender quais são suas implicações. Os três principais modelos de rede são descritos a seguir.

2.2.2.1 Modelo Erdős-Rényi (ER)

O modelo proposto por Erdös e Rényi (1959) é considerado um dos modelos mais simples de redes. Neste modelo, existem n vértices conectados por m arestas escolhidas aleatoriamente de um total de $\frac{n(n-1)}{2}$ possíveis arestas. Uma definição equivalente é dada através de um modelo binomial. Nessa definição, o número de arestas não é fixo, o que existe é um valor de probabilidade p de conexão entre os pares de vértices. Para este caso, a distribuição do grau dos vértices segue uma distribuição de Poisson. Esta é a razão pela qual esses modelos são conhecidos como Grafos Aleatórios de Poisson (NEWMAN, 2003). Apesar de sua simplicidade, o modelo ER não é adequado para modelar redes reais, uma vez que não representa algumas características dessas redes (COSTA *et al.*, 2011), como a presença de vários ciclos de tamanho 3 (triângulos) e a distribuição do grau $P(k)$ segundo uma lei de potência, onde $P(k) \sim k^{-\gamma}$ (CLAUSET; SHALIZI; NEWMAN, 2009).

2.2.2.2 Modelo Watts-Strogatz (WS)

Uma solução para um dos problemas citados anteriormente (a baixa presença de ciclos no modelo ER) foi publicada por Watts e Strogatz (1998). Este modelo ficou conhecido como *Small-World Networks*. A propriedade *small-world* pode ser percebida em várias redes reais, nas quais muitos vértices podem ser acessados através de poucos passos na rede. Para gerar as redes *small-world*, inicia-se com um anel com N vértices, onde cada vértice é conectado com seus m vizinhos mais próximos à esquerda e seus m vizinhos mais próximos à direita. Em seguida, para cada vértice, cada aresta que conecta este vértice aos seus vizinhos mais próximos em sentido horário é reconectada com probabilidade p. O final da aresta é conectado à qualquer vértice escolhido de maneira uniforme, evitando-se auto conexões. O processo é realizado até que todos os vértices tenham sido analisados. O processo de reconexão de arestas permite que as redes *small-world* se posicionem entre redes regulares e redes semelhantes às aleatórias. Quando $p = 0$, nenhuma aresta é reconectada e a rede permanece regular com muitos ciclos de 3 vértices e grandes caminhos mínimos. Para $p = 1$, todas arestas são reconectadas e a rede é semelhante a

uma rede aleatória, apresentando poucos triângulos e pequenos caminhos mínimos. Através de simulações numéricas, Watts e Strogatz (1998) mostraram que existe uma região entre esses dois extremos na qual a rede *small-world* irá apresentar pequenos caminhos mínimos e grande número de ciclos. Por exemplo, para $p = 0.01$, a simulação realizada pelos autores exibiu significativa queda no tamanho dos caminhos mínimos e pouca alteração no número de ciclos, ao comparar com $p = 0$. Embora tenha solucionado o problema da baixa presença de ciclos do modelo ER, a distribuição do grau dos vértices segue uma distribuição de Poisson, e não uma lei de potência.

2.2.2.3 Modelo Barabási-Albert (BA)

A ligação preferencial é a possível explicação para a distribuição do grau de acordo com uma lei de potência observada em várias redes reais, como as redes de citações e de páginas Web (NEWMAN, 2003). A ligação preferencial é uma das regras que compõem o modelo de rede introduzido por Barabasi e Albert (1999). As redes criadas por esse modelo são conhecidas como *scale-free networks*. Para gerar redes com distribuição do grau do tipo lei de potência, Barabasi e Albert (1999) propuseram a criação de uma rede com duas regras:

- Crescimento: A cada intervalo de tempo, um novo vértice i com m arestas é adicionado à rede.
- Ligação Preferencial: A probabilidade de uma aresta do novo vértice i ligar-se a um vértice j (que já está na rede) é proporcional ao número de conexões de j.

$$P(i \rightarrow j) = \frac{K_j}{\sum_{u=1}^{N} K_u} \tag{2.1}$$

onde N indica o número total de vértices na rede.

Através da ligação preferencial, os vértices mais conectados apresentam maior probabilidade de receberem novas arestas. Este fenômeno é conhecido como *"the rich get richer"* ("os ricos ficam mais ricos"). Uma característica das redes *scale-free* é a existência de *hubs*, ou seja, vértices que contêm uma significativa fração do total de arestas de uma rede (COSTA *et al.*, 2007). Por fim, esse modelo gera redes com distribuição de grau de acordo com uma lei de potência com expoente $\gamma = 3$ (NEWMAN, 2003).

2.2.3 Redes aplicadas para a Análise da Linguagem

Nos últimos anos, redes complexas têm sido cada vez mais utilizadas para modelar e analisar a linguagem humana (CONG; LIU, 2014). A linguagem humana também é considerada um sistema complexo (LARSEN-FREEMAN; LYNNE, 2008; BECKNER *et al.*, 2009). Por isso, os modelos e ferramentas de redes complexas constituem uma importante metodologia para o estudo da linguagem. Uma rede N modelando a linguagem é dada por $N = (V, E)$, onde V é o conjunto de vértices representando as unidades linguísticas e E é o conjunto de arestas que

representa as relações entre as unidades linguísticas. As unidades linguísticas podem ser palavras, fonemas ou morfemas (CONG; LIU, 2014). As relações entre as unidades linguísticas podem ser extraídas de diferentes níveis do uso da linguagem. As principais são as relações de co-ocorrência, sintáticas ou semânticas. As relações de co-ocorrência representam a ordem linear das palavras em uma sentença. Nesse modelo, duas palavras são conectadas se forem adjacentes em pelo menos uma sentença. Nos modelos que utilizam relações sintáticas, duas palavras são conectadas por uma aresta se fizerem parte de uma relação sintática em pelo menos uma sentença. Por fim, as relações semânticas representam um nível de análise mais profundo (CONG; LIU, 2014), nas quais as palavras (ou conjunto de palavras) participam de relações do tipo predicado-argumento. Apesar das diferenças apresentadas, as redes complexas criadas com essas relações apresentam as propriedades universais *small-world* e *scale-free* (CANCHO; SOLé, 2001; CANCHO; SOLÉ; KÖHLER, 2004) que são encontradas em outras redes reais. Essas descobertas sugerem que a linguagem apresenta uma organização que pode ser descrita por padrões universais. Dorogovtsev e Mendes (2001) desenvolveram um modelo de rede para explicar a evolução da linguagem. Neste modelo, a linguagem é representada como uma rede evolutiva de interação de palavras.

2.2.3.1 Modelo Dorogovtsev-Mendes (DM)

A distribuição do número de conexões das palavras obtida por Cancho e Solé (2001) apresenta uma característica peculiar, a presença de duas regiões que seguem uma lei de potência com expoentes diferentes. Para capturar esse efeito, Dorogovtsev e Mendes (2001) propuseram uma teoria sobre a evolução da linguagem humana baseada em uma rede evolutiva de colaboração de palavras. A cada intervalo de tempo, uma nova palavra é adicionada à rede e o total de vértices é representado por t. O modelo utiliza a ligação preferencial, de modo que a nova palavra é conectada a uma palavra i já existente na rede com probabilidade proporcional ao grau k_i, assim como é feito no modelo de Barabasi e Albert (1999). Além disso, a cada intervalo de tempo, ct novas arestas são adicionadas entre as palavras que existiam anteriormente na rede (ou seja, desconsiderando apenas a palavra que acabou de ser adicionada), onde c representa uma constante. As novas arestas são adicionadas entre as palavras não conectadas i e j com probabilidades proporcionais ao produto $k_i k_j$.

A partir desse modelo, Dorogovtsev e Mendes (2001) conseguiram representar a rede de palavras descrita por Cancho e Solé (2001). As duas regiões com diferentes expoentes são obtidas através das duas taxas de crescimento de arestas empregadas nesse modelo: uma taxa constante envolvendo as arestas dos novos vértices e a taxa crescente para as arestas entre os vértices já existentes (BIEMANN, 2012).

2.2.4 Medidas de Redes Complexas

As principais medidas de caracterização das redes complexas utilizadas neste trabalho são descritas abaixo.

2.2.4.1 Grau

O grau de um vértice é uma das medidas mais simples que podem ser calculadas. Em redes não direcionadas, esta medida indica a quantidade de arestas incidentes no vértice i e pode ser obtida pela seguinte fórmula:

$$k_i = \sum_{j=1}^{N} A_{ij}. \tag{2.2}$$

Para as redes direcionadas, o grau de um vértice pode ser definido da seguinte maneira:

- k_i^{in}: Indica o número de arestas que chegam a um vértice i, também pode ser chamado de grau de entrada.
- k_i^{out}: Indica o número de arestas que saem de um vértice i, também pode ser chamado de grau de saída.

2.2.4.2 Assortatividade

Em algumas redes reais, há uma tendência de que vértices "parecidos"se conectem (NEWMAN, 2002). No caso do grau, é possível encontrar três configurações para o grau k_i e k_j dos vértices i e j, respectivamente:

- $k_i \sim k_j$ - *Hubs* conectam-se com *hubs*.
- $k_i \neq k_j$ - *Hubs* conectam-se com vértices pouco conectados.
- Não há qualquer relação entre os valores k_i e k_j.

A correlação de grau é determinada por meio da medida de assortatividade, que pode ser calculada através do coeficiente de Pearson do grau dos vértices presentes nas extremidades de cada uma das arestas. A fórmula dessa medida é apresentada a seguir (COSTA *et al.*, 2007):

$$r = \frac{\frac{1}{M}\sum_{j>i} k_i k_j A_{ij} - \left[\frac{1}{M}\sum_{j>i}\frac{1}{2}(k_i + k_j)A_{ij}\right]^2}{\frac{1}{M}\sum_{j>i}\frac{1}{2}(k_i^2 + k_j^2)A_{ij} - \left[\frac{1}{M}\sum_{j>i}\frac{1}{2}(k_i + k_j)A_{ij}\right]^2}, \tag{2.3}$$

onde M representa o número total de arestas e k_i e k_j representam o grau do vértice i e j, respectivamente.

Se $r > 0$, os vértices mais conectados tendem a se conectar com outros de grau semelhante e esta rede é dita assortativa (NEWMAN, 2002). Se $r < 0$, a rede é chamada de disassortativa. Nas redes disassortativas, os vértices mais conectados tendem a se conectar com aqueles que apresentam poucas ligações. Por fim, se $r = 0$, não há qualquer relação entre o grau dos vértices e a rede é dita não assortativa. Na maioria dos casos, as redes de adjacência de palavras são disassortativas.

2.2.4.3 Grau Médio dos Vizinhos

A medida $k_{nn}(i)$ (PASTOR-SATORRAS; VÁZQUEZ; VESPIGNANI, 2001) calcula o grau médio dos vizinhos de um vértice. Este valor pode ser obtido através da seguinte fórmula:

$$k_{nn}(i) = \frac{1}{k_i} \sum_{j=1}^{N} k_j A_{ij}, \tag{2.4}$$

onde k_i e k_j são os graus dos vértices i e j, respectivamente, e $A_{ij} = 1$ se existir uma aresta entre o vértices i e j.

2.2.4.4 Coeficiente de Aglomeração

A aglomeração ou transitividade é uma propriedade típica de redes de amizade, nas quais se duas pessoas A e B possuem um amigo em comum, a probabilidade de A e B também serem amigos é alta. A transitividade está relacionada com a presença de triângulos (ou ciclos) na rede. Diferentemente das redes aleatórias de Erdős-Rényi, as redes reais apresentam alta frequência de *loops* envolvendo poucos vértices (COSTA *et al.*, 2007).

O coeficiente de aglomeração local de um vértice i, $cc(i)$, indica a probabilidade de dois vizinhos desse vértice estarem conectados. Esta medida é obtida da seguinte maneira:

$$cc(i) = \frac{2 * e_i}{k_i * (k_i - 1)}, \tag{2.5}$$

onde e_i representa o número de arestas entre os vizinhos do vértice i e k_i é o grau do vértice i. Se $k_i = 1 \mapsto cc(i) = 0$.

Como caracterização global da rede, a medida $< CC >$ é obtida através da média do coeficiente de aglomeração local de todos os vértices:

$$< CC >= \frac{1}{N} \sum_{i=1}^{N} cc(i), \tag{2.6}$$

onde N corresponde à quantidade de vértices na rede. Para o cálculo dessa medida, cada vértice i possui o mesmo peso, independente do grau k_i.

Uma outra maneira de caracterizar globalmente a rede é através do coeficiente de aglomeração obtido pela fórmula da transitividade. Esta medida é calculada da seguinte maneira:

$$C = \frac{3 \times N_{\triangle}}{N_3}, \tag{2.7}$$

onde $N_{\triangle}$ corresponde ao número de triângulos presentes na rede e N_3 é o número total de triplas conectadas. Os valores possíveis para essa medida são $0 \leq C \leq 1$, sendo que $C = 0$ corresponde à uma rede sem triângulos e $C = 1$ é uma rede com transitividade perfeita. Diferentemente da medida anterior, a medida C dá o mesmo peso para cada triângulo na rede. Por esse motivo, $< CC >$ e C resultam em diferentes valores, uma vez que vértices com alto grau possivelmente

apresentam um maior número de triângulos se comparados com graus menores (COSTA *et al.*, 2007). Sobre essa medida, Cancho e Solé (2001) constataram que o coeficiente de aglomeração em redes representando textos é muito maior que o valor esperado para a correspondente rede aleatória de palavras.

2.2.4.5 Média dos Caminhos Mínimos Geodésicos

As distâncias entre os vértices têm um papel importante no transporte e na comunicação dentro de uma rede (BOCCALETTI *et al.*, 2006). O envio de uma pacote de um computador para outro através da Internet ocorre, na maioria das vezes, através do menor caminho entre esses dois computadores (PASTOR-SATORRAS; VÁZQUEZ; VESPIGNANI, 2001). Um caminho é definido por uma sequência de vértices sem repetição de arestas. O comprimento do caminho entre os vértices i e j é igual ao número de arestas entre esses vértices.

O caminho geodésico entre os vértices i e j é um dos caminhos existentes entre esses vértices que apresenta o menor comprimento (COSTA *et al.*, 2007). A distância geodésica entre i e j é denotada por l_{ij}. As distâncias geodésicas l_{ij} entre todos os pares i e j de uma rede podem ser representadas em uma matriz de distâncias D da seguinte maneira:

$$d_{ij} = l_{ij}. \tag{2.8}$$

Uma maneira de caracterizar a estrutura interna de uma rede é através da média dos caminhos mínimos geodésicos. Esta medida indica a separação média entre dois vértices e pode ser calculada da seguinte forma:

$$L = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} l_{ij}, \tag{2.9}$$

onde N é o número total de vértices na rede. Quando a rede apresenta mais de um componente, alguns pares i e j não estarão conectados por um caminho. Nesse caso, $d_{ij} = \infty$ e, por definição, $L = \infty$. Uma das soluções para evitar a divergência de L é considerar apenas o maior componente conexo para o cálculo desta medida. Em redes textuais, L quantifica a relevância de cada palavra de acordo com sua distância às palavras mais frequentes (AMANCIO *et al.*, 2011).

2.2.4.6 Coeficiente de Intermediação

O coeficiente de intermediação, também conhecido como *Betweenness Centrality*, é utilizado para medir o tráfego que passa por um vértice ou uma aresta. Assumindo que as mensagens em uma rede são trocadas entre pares de vértices e que estas trafegam pelos menores caminhos, é possível medir a relevância de um vértice (ou aresta) através da quantidade de menores caminhos que passam por esse vértice (ou aresta) (BOCCALETTI *et al.*, 2006). A medida de *Betweenness Centrality* é calculada da seguinte maneira:

$$B_i = \sum_{st} \frac{n_{st}^i}{g_{st}}, \tag{2.10}$$

onde n_{st}^{i} é o número de caminhos mínimos entre os vértices s e t que passam pelo vértice i e g_{st} indica o número total de menores caminhos entre os vértices s e t, o somatório é realizado para todos os pares de vértices s e t. Vértices com alto valor de *Betweenness Centrality* podem apresentar uma considerável influência na rede devido ao controle da informação que é passada para os outros vértices (NEWMAN, 2010).

Em uma rede representando texto, palavras muito frequentes costumam apresentar alto valor dessa medida. Entretanto, algumas palavras podem servir como "pontes"conectando conceitos de comunidades distintas e, assim, apresentar alto valor de *Betweenness Centrality*. Por isso, esta medida pode quantificar a diversidade de contextos nos quais uma palavra pode ser utilizada (AMANCIO *et al.*, 2011).

2.2.4.7 Motivos

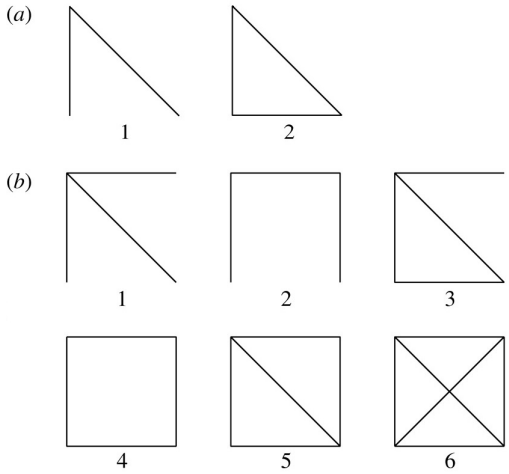
As redes complexas também podem ser caracterizadas através da extração de motivos. Motivos são padrões de interconexão significativos e recorrentes que ocorrem em redes reais em quantidades significativamente maiores do que em redes aleatórias (MILO *et al.*, 2002), expressos em forma de pequenos subgrafos. Milo *et al.* (2002) perceberam a presença de motivos em diversas redes reais. Algumas redes como a de transcrição genética e de conectividade neural possuíam o mesmo conjunto de motivos, sugerindo uma semelhança estrutural entre essas duas redes. Além disso, os motivos extraídos de um conjunto de redes de circuitos elétricos foram capazes de separá-los em duas classes distintas, sem utilizar nenhum conhecimento prévio (MILO *et al.*, 2002; KASHTAN *et al.*, 2004). Cada classe obtida representava uma funcionalidade diferente do conjunto de circuitos elétricos.

A extração de motivos para a análise de textos foi feita em (AMANCIO *et al.*, 2013a). Neste trabalho, Amancio *et al.* (2013a) perceberam a presença de alguns motivos significativos em textos. Além disso, os autores constataram que alguns subgrafos raramente ocorrem em textos reais. A extração de motivos é computacionalmente cara. Por isso, a análise de motivos realizada neste projeto ficou restrita aos subgrafos apresentados na Figura 1 e na Figura 2.

2.2.4.8 Acessibilidade

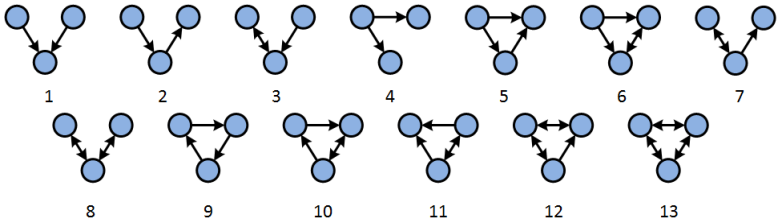
Além das medidas citadas anteriormente, neste projeto foram utilizadas medidas recentes de redes complexas para análise de fatores estilísticos de textos. Por exemplo, sabe-se que a frequência de palavras específicas representa um fator importante para a caracterização de estilos. Em redes tradicionais de co-ocorrência, a frequência pode ser medida pelo grau dos vértices. Sabendo disto, foram utilizadas extensões do conceito de grau que têm sido úteis para aperfeiçoar a caracterização de outros sistemas textuais (TRAVENçOLO; VIANA; COSTA, 2009). Uma das extensões do conceito de grau corresponde ao uso da versão hierárquica desta medida. As medidas hierárquicas são caracterizadas por analisar vizinhanças mais distantes. Portanto, o grau

Figura 1 – Todos os motivos não direcionados com 3 vértices (a) e 4 vértices (b).



Fonte: Adaptada de Silva e Stumpf (2005).

Figura 2 – Todos os motivos direcionados com 3 vértices.



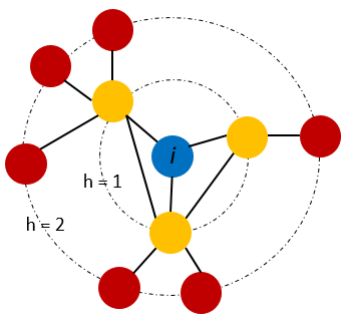
Fonte: Gabasova (2014).

hierárquico de um nó i à distância h, é definido como

$$k_i(h) = N_i(h), \tag{2.11}$$

onde $N_i(h)$ representa o número de nós que estão a uma distância h do nó i. A Figura 3 ilustra o cálculo do grau hierárquico de um vértice i para diferentes valores de h.

Figura 3 – Os três vértices amarelos representam aqueles que estão a uma distância h = 1 de i. Portanto, o grau hierárquico $k_i(1) = 3$. Analogamente, os seis vértices vermelhos estão a uma distância h=2 de i, logo $k_i(2) = 6$.



Fonte: Elaborada pelo autor.

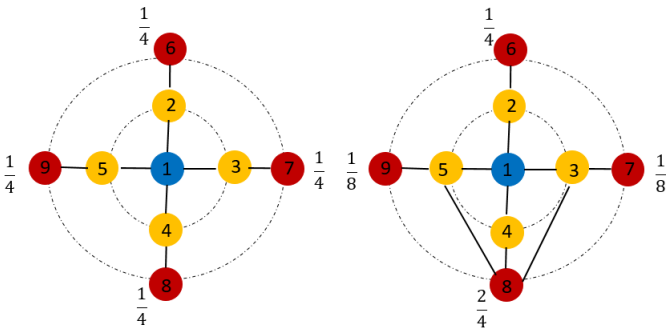
Embora a versão hierárquica forneça mais informação topológica local do que simplesmente a medida de grau, esta versão não considera que o acesso a vértices mais distantes pode ocorrer de maneira irregular. Desta forma, mesmo que um vértice possua grau alto, existe a possibilidade que apenas alguns de seus vários vizinhos sejam efetivamente acessados. Para considerar esta possibilidade, foi utilizada a medida de acessibilidade (TRAVENçOLO; COSTA, 2008).

A medida de acessibilidade quantifica o número de nós efetivamente acessíveis a partir de um nó inicial (VIANA; BATISTA; COSTA, 2012). Para calcular esta medida, considere que $P_h(i,j)$ representa a probabilidade de se alcançar um vértice j a partir de i através de uma caminhada aleatória auto excludente de comprimento h (TRAVENçOLO; VIANA; COSTA, 2009). Neste caso, são considerados os caminhos a partir do vértice i para cada um dos vértices situados no anel concêntrico de distância h. Considerando esta notação, a medida de acessibilidade é dada por

$$a^{(h)}(i) = \exp(-\sum P_h(i,j)\ln P_h(i,j)). \tag{2.12}$$

Com esta definição, $0 < a^{(h)}(i) \leq N_i(h)$, onde $N_i(h)$ é o número de vértices no anel h. O valor máximo é obtido quando todas as probabilidades de transição para um dado nível concêntrico são iguais. Nesse caso, todos os nós podem ser acessados igualmente. A Figura 4 ilustra as probabilidades de transição e a medida de acessibilidade para duas configurações de redes. Diferentemente de outras medidas apresentadas, a acessibilidade não apresenta correlação com a frequência das palavras. Outro importante aspecto dessa medida é a propriedade de detectar palavras-chave relevantes do conteúdo abordado (AMANCIO *et al.*, 2012). A fração de palavras-chave também se mostrou relevante para a tarefa de reconhecimento de autoria (AMANCIO *et al.*, 2011).

Figura 4 – No grafo à esquerda, a probabilidade de transição para cada um dos vértices do segundo nível concêntrico é a mesma, resultando no maior valor de acessibilidade $a^{(2)}(1) = 4$. No grafo à direita, com a adição de duas novas arestas, as probabilidades de transição se alteram e observa-se uma tendência aparente de acesso ao vértice 8, resultando em um número menor de vértices efetivamente acessados.



Fonte: Elaborada pelo autor.

2.3 Reconhecimento de Autoria

2.3.1 Definições e História

Em um típico problema de reconhecimento de autoria, um texto de autoria desconhecida é atribuído a um autor, de um conjunto de vários possíveis autores. As primeiras atividades de reconhecimento de autoria apoiadas por métodos estatísticos tiveram início no século XIX. Uma das primeiras abordagens e que ainda é utilizada em muitos estudos envolve o uso de modelos probabilísticos (MOSTELLER; WALLACE, 1964). Estes métodos têm por objetivo maximizar a probabilidade $P(x|a)$ de um texto x pertencer a um possível autor a. Em (MOSTELLER; WALLACE, 1964), os autores analisaram a autoria de centenas de ensaios políticos, conhecidos como *The Federalist Papers*. Neste estudo, os autores introduziram uma grande contribuição para a área ao mostrarem que a frequência de palavras comuns (como *and* e *to* na língua inglesa) produz resultados significativos para a distinção entre os possíveis autores.

Diferentemente dos métodos de reconhecimento de autoria tradicionais, utilizados por linguistas e peritos, o trabalho de Mosteller e Wallace (1964) iniciou os estudos de atribuição de autoria baseados em estatística. Desde então, muitos trabalhos se dedicaram a definir características que quantificassem o estilo de escrita do autor, conhecido como estilometria (HOLMES, 1994). De acordo com (STAMATATOS, 2009), os atributos estilométricos são divididos em algumas categorias, dentre elas destacam-se as características léxicas, baseadas em caracteres, sintáticas e semânticas.

Desde o fim da década de 90, a área de reconhecimento de autoria passou por algumas mudanças (STAMATATOS, 2009). A grande quantidade de textos disponíveis na *web* (como emails e mensagens em blogs e fóruns) aumentou a necessidade de análise desses dados de maneira eficiente e estimulou o desenvolvimento de diversas áreas, como recuperação de informação, aprendizado de máquina e Processamento de Linguagens Naturais (PLN). O reconhecimento de autoria se beneficiou dos avanços obtidos nas áreas citadas pelos seguintes motivos:

- Pesquisas em recuperação de informação desenvolveram técnicas eficientes para representar e classificar grandes volumes de texto.
- Diversos algoritmos de aprendizado de máquina foram criados para tratar dados esparsos e multi-dimensionais.
- Diversas ferramentas de PLN capazes de analisar textos de maneira eficiente foram desenvolvidas.

Além disso, o conjunto de textos disponíveis eletronicamente revelou o potencial do reconhecimento de autoria em aplicações de diversas áreas como:

- Inteligência (por exemplo, a análise de mensagens terroristas) (ABBASI; CHEN, 2005);

- Direito civil e criminal (GRANT, 2007; CAROLE, 2005);
- Computação forense (FRANTZESKOU *et al.*, 2006);

Além das áreas citadas acima, muitos trabalhos em reconhecimento de autoria são dedicados às aplicações tradicionais em textos literários (BURROWS, 2002). Além da tarefa de reconhecimento de autoria, outras tarefas relacionadas podem ser realizadas, como a detecção de plágio e inconsistências estilísticas e a verificação de autoria, que consiste em decidir se um texto foi escrito por um autor ou não.

2.3.2 Atributos Estilométricos

Características léxicas e baseadas em caracteres consideram o texto como uma sequência de palavras ou caracteres, respectivamente. Exemplos de características léxicas utilizadas para a tarefa de reconhecimento de autoria são o comprimento das palavras e sentenças, frequência de palavras, a riqueza do vocabulário utilizado (*vocabulary richness*), entre outros (GRAHAM; HIRST; MARTHI, 2005). Considerada uma das primeiras tentativas de atribuição de autoria, Mendenhall (1887) utilizou-se de medidas de comprimento das sentenças e das palavras para identificar a autoria de textos. Grande parte dos estudos em atribuição de autoria utilizam, pelo menos parcialmente, características léxicas para representar o estilo dos autores (STAMATATOS, 2009).

Uma importante descoberta com relação ao uso de atributos léxicos é o fato de que palavras comuns (artigos, preposições, pronomes), conhecidas como *function words*, estão entre as melhores características para diferenciar um conjunto de autores (BURROWS, 1987). Por serem independentes de tópico e utilizadas de maneira inconsciente, essas palavras são capazes de capturar escolhas estilísticas de cada um dos autores (STAMATATOS, 2009). Além disso, palavras frequentes são responsáveis pelo sucesso da abordagem proposta por físicos na qual a distância dos *ranks* de frequência de palavras é usada para computar similaridade entre textos (HAVLIN, 1995).

Outra possível abordagem utilizando características léxicas é extrair as palavras mais frequentes em um texto. Nessa abordagem, cada texto é representado como um vetor de frequências de palavras. Assim, técnicas de aprendizado de máquina podem ser aplicadas para distinguir esses vetores. Vários trabalhos utilizaram diferentes quantidades de palavras extraídas do conjunto de dados, desde 100 até 1000 palavras mais frequentes (BURROWS, 1987; STAMATATOS, 2006).

Para as medidas baseadas em caracteres, um texto é analisado como uma simples sequência de caracteres. Alguns exemplos de medidas são a contagem de caracteres alfabéticos ou numéricos, frequência de letras e sinais de pontuação, entre outros (STAMATATOS, 2009). Uma outra abordagem é extrair a frequência de n-gramas em nível de caracteres. N-gramas a nível de caracteres representam subconjuntos de n caracteres adjacentes. No contexto de reconhecimento de autoria, foi observado que a frequência dos caracteres mais comuns representa um bom

atributo para discriminar autores, apresentando melhores resultados do que com a utilização de características léxicas (JANKOWSKA; MILIOS; KESELJ, 2014). A vantagem da utilização de características baseadas em caracteres é que podem ser utilizadas em basicamente qualquer linguagem natural. Portanto, é um método independente de linguagem.

A extração de características sintáticas e semânticas requer uma análise mais elaborada do texto. De acordo com Stamatatos (2009), informações sintáticas são consideradas mais representativas do que características léxicas. Um indicativo desse poder de representação é o bom desempenho que as palavras comuns (*function words*) apresentam na caracterização do estilo de escrita, uma vez que essas palavras são encontradas em certas estruturas sintáticas.

Algumas abordagens sintáticas incluem a extração da frequência de regras de reescrita (*rewrite rules*). Regras de reescrita ou de produção indicam como uma sentença pode ser decomposta em sintagmas. Sintagmas são sequências de palavras que juntas formam uma unidade representativa. Cada sintagma possui uma palavra principal (denominada núcleo) e pode apresentar outras palavras dependentes. Uma sentença é considerada válida se for possível gerar todos os seus terminais (palavras) de acordo com algumas regras de reescrita (KUMAR, 2012). Um exemplo de uma regra de reescrita é $S \rightarrow NP\ \ VP$, na qual S significa sentença, NP significa sintagma nominal (*Noun Phrase*) e VP significa sintagma verbal (*Verb Phrase*). Essa regra de reescrita significa que uma sentença é constituída por um sintagma nominal e um sintagma verbal. Utilizando essa regra, a sentença *"O aluno realizou todos os experimentos"* é dividida em *O aluno* (sintagma nominal) e *realizou todos os experimentos* (sintagma verbal). Essas regras são importantes pois, além de descreverem a classe sintática, mostram como as palavras se combinam para formar frases.

Além das regras de reescrita, a análise de sentenças ou *chunks* (pedaços) e *part-of-speech tagging* também são utilizadas nas abordagens sintáticas. *Part-of-speech tagging* consiste na rotulação de palavras em determinadas classes gramaticais como substantivos, adjetivos, entre outros. Um dos primeiros trabalhos que utilizaram informações sintáticas para o reconhecimento de autoria foi realizado por Baayen, Halteren e Tweedie (1996). Eles extraíram as frequências das regras de reescrita e estas medidas proporcionaram melhores resultados do que aqueles obtidos com abordagens léxicas. Hirst e Feiguina (2007) combinaram as abordagens de frequência de bigramas com análise sintática. Os bigramas sintáticos utilizados foram úteis para distinguir diferentes autores, mesmo quando aplicados em textos curtos. Por fim, alguns exemplos de características semânticas que podem ser extraídas são os grafos de dependências semânticas (GAMON, 2004) e informações sobre sinônimos e hiperônimos de palavras (MCCARTHY *et al.*, 2006).

Quanto mais detalhada for a análise textual para extração das características estilométricas, menor será a acurácia obtida com as medidas (STAMATATOS, 2009). As ferramentas atuais de processamento de línguas naturais ainda não são capazes de representar adequadamente o estilo através de informações sintáticas e semânticas, apresentando inserção de ruídos. Assim,

poucos trabalhos tentam explorar estas características estilométricas, sendo mais utilizadas como complemento às outras medidas de caracterização, como léxicas ou a nível de caracteres.

2.3.3 Métodos de Atribuição

Em um problema de reconhecimento de autoria, há um conjunto de possíveis autores, amostras de textos cuja autoria é conhecida (conjunto de treinamento) e um ou mais textos de autoria desconhecida (conjunto de teste). Cada um dos textos do conjunto de teste deve ser atribuído a um dos autores. Considera-se a existência de duas abordagens diferentes para o reconhecimento de autoria (STAMATATOS, 2009). Uma delas trata cada texto de treinamento individualmente (abordagem baseada em instâncias) e na outra, os textos são acumulados por autor (abordagens baseadas em perfis).

2.3.3.1 Abordagens baseadas em Perfis

Nas abordagens baseadas em perfis, os textos de treinamento são combinados em um único texto t_p por autor. Este texto é então utilizado para extrair as propriedades de estilo. Estas propriedades caracterizam o *perfil* do autor. Cada novo texto, cuja autoria precisa ser atribuída, é comparado com cada um dos textos t_p dos autores. Uma das principais maneiras de realizar essa comparação é através de medidas de distância.

Uma vez que cada um dos textos do autor não é utilizado separadamente, as possíveis diferenças entre os textos de treinamento são descartadas. Além disso, os atributos estilométricos extraídos do arquivo combinado podem ser diferentes das que seriam extraídas de cada arquivo separadamente. Portanto, estas abordagens prezam por representar o estilo geral de escrita de cada autor.

O processo de treinamento utilizado pelas abordagens baseadas em perfis é simples. O perfil de cada um dos autores é extraído dos arquivos durante a fase de treinamento. Por fim, o modelo de atribuição é baseado em funções de distância que calculam as diferenças entre o texto de autoria desconhecida e os perfis de cada um dos autores. Alguns exemplos de abordagens baseadas em perfis são os modelos probabilísticos, modelos de compressão e o Common n-Grams (CGN) (STAMATATOS, 2009). Uma breve explicação de cada um dos modelos é apresentada a seguir:

- Modelos probabilísticos: Tais métodos tem por objetivo maximizar a probabilidade $P(x|a)$ de um texto x pertencer a um possível autor a.
- Modelos de compressão: Nesse modelo, um algoritmo de compressão é aplicado no texto concatenado de cada um dos autores, produzindo um arquivo F_c. Um texto de autoria desconhecida é adicionado ao texto concatenado de cada autor e o algoritmo de compressão é novamente aplicado, criando o arquivo F_c^*. As similaridades entre os autores e o novo texto são dadas pela diferença entre F_c^* e F_c.

- Common n-Grams (CGN): Nesse modelo, o perfil de um texto t é composto pelos n-gramas a nível de caracteres mais frequentes em t. Deste modo, algumas medidas de distância podem ser utilizadas para estimar a similaridade entre dois textos.

2.3.3.2 Abordagens baseadas em Instâncias

A maioria das abordagens de reconhecimento de autoria considera cada texto de treinamento separadamente (STAMATATOS, 2009). Desse modo, cada texto de autoria conhecida é considerado uma instância para o problema e pode ser representado por um conjunto de atributos. Algoritmos de classificação são treinados no conjunto de instâncias com o objetivo de desenvolver o modelo de atribuição. Por fim, este modelo criado será capaz de estimar o autor de textos com autoria desconhecida. Essas abordagens prezam por quantificar o estilo de escrita presente em cada um dos documentos.

Para que o modelo gerado seja capaz de reconhecer a autoria de novos textos com maior confiabilidade, é preciso fornecer aos algoritmos de classificação instâncias suficientes e representativas de cada uma das classes. Além disso, é preciso que as instâncias apresentem tamanhos iguais. Caso contrário, os textos devem ser segmentados para que fiquem com o mesmo tamanho (SANDERSON; GUENTER, 2006). Diversos tamanhos de texto podem ser utilizados. Hirst e Feiguina (2007) avaliaram a tarefa de reconhecimento de autoria em textos de tamanhos variados (de 200 até 1000 palavras) e perceberam que a acurácia reduziu drasticamente para os textos de tamanhos menores.

Alguns exemplos de abordagens baseadas em instâncias são Vector Space Models (VSM) ou modelos de espaço vetorial e modelos baseados em similaridade. Uma breve explicação de cada um dos modelos é apresentada abaixo:

- Modelos de espaço vetorial: Tais modelos compreendem a maioria das abordagens baseadas em instâncias. Nesses modelos, cada texto de treinamento é considerado como um vetor com diversos atributos. Desse modo, diversos algoritmos de aprendizado de máquina podem ser utilizados para construir um modelo de classificação a partir dos vetores.

- Modelos baseados em similaridade: Tais modelos extraem medidas de similaridade entre o texto de autoria desconhecida t com todos os outros textos do conjunto de treinamento. A autoria do texto t será a mesma do texto do conjunto de treinamento que apresentar a maior similaridade com t.

2.4 Considerações Finais

Neste capítulo foram apresentados os principais conceitos relacionados às duas grandes áreas nas quais este trabalho está inserido, as redes complexas e o reconhecimento de autoria. É importante ressaltar que apenas as medidas de redes complexas utilizadas nesse trabalho foram

descritas nesse Capítulo. Entretanto, estima-se que há mais de 100 medidas de redes complexas que podem ser utilizadas para caracterizar a estrutura topológica das redes (COSTA *et al.*, 2007).

No próximo capítulo são apresentados diversos trabalhos relacionados que utilizam redes complexas para a tarefa de reconhecimento de autoria. Ao fim do capítulo, alguns trabalhos que utilizam redes de co-ocorrência para outras tarefas linguísticas são descritos.

CAPÍTULO

3

TRABALHOS RELACIONADOS

3.1 Considerações Iniciais

Encontram-se na literatura diversas abordagens realizadas dentro da área de redes complexas com o objetivo de quantificar e caracterizar os estilos de escrita dos autores. Neste capítulo, são apresentados os principais trabalhos relacionados à proposta deste projeto. Na Seção 3.2.1, são apresentados alguns trabalhos que aplicam redes complexas para a tarefa de reconhecimento de autoria. Na Seção 3.2.2, são apresentados alguns trabalhos que utilizam redes de co-ocorrência para outras tarefas linguísticas, como a classificação e distinção de linguagens.

3.2 Trabalhos Relacionados

Nos últimos anos, a teoria de redes complexas tem sido amplamente utilizada para provar propriedades da linguagem. Redes podem ser utilizados para modelar objetos e as conexões existentes entre eles. Este tipo de modelagem é aplicado na representação de diferentes conjuntos de dados, sendo objeto de estudo de distintas áreas da matemática discreta e da computação em geral. Com relação a computação, técnicas baseadas em grafo têm sido aplicadas na análise e construção de arquiteturas de software (MOURA; LAI; MOTTER, 2003), filtros de spam (KONG *et al.*, 2006) e sistemas de processamento de línguas naturais (MIHALCEA; RADEV, 2011; CONG; LIU, 2014).

Com relação à analise textual, redes complexas demonstraram sua importância melhorando a performance de tarefas de processamento de línguas naturais (AMANCIO, 2015b). Modelos baseados em redes complexas têm sido aplicados para analisar diversos níveis de organização da linguagem, incluindo o nível sintático (CANCHO; SOLÉ; KÖHLER, 2004; AMANCIO *et al.*, 2012) e semântico (SILVA; AMANCIO, 2013; MASUCCI *et al.*, 2011). Um modelo muito utilizado é a rede de adjacência de palavras (também conhecido como rede de co-ocorrência), que realiza a conexão de palavras próximas. Uma vez que este modelo consegue representar fatores

sintáticos e estilísticos (AMANCIO *et al.*, 2013a), ele tem sido empregado em análises de complexidade sintática (AMANCIO *et al.*, 2013b), detecção de movimentos literários (AMANCIO; OLIVEIRA JR; COSTA, 2012a) e em estilometria (GRABSKA-GRADZINSKA A. KULIG; DROZDZ, 2012).

Uma importante descoberta com relação as redes de co-ocorrência foi reportada em (CANCHO; SOLé, 2001). Nesse trabalho, Cancho e Solé (2001) utilizaram a modelagem de co-ocorrência de palavras e provaram que tais redes apresentam duas importantes características encontradas em vários sistemas de redes complexas, a propriedade livre de escala (*scale free*) e o fenômeno do pequeno mundo (*small world*). A estrutura livre de escala nas redes textuais pode ser explicada como consequência de um processo de otimização, de modo que o esforço em transmitir e receber uma mensagem é minimizado (CANCHO; SOLé, 2003). Esse processo de otimização é também uma das hipóteses para a emergência da lei de Zipf (ZIPF, 1949), que afirma que a relação entre o *ranking* de frequência das palavras e a frequência absoluta segue uma lei de potência. Em um outro trabalho, Cancho, Solé e Köhler (2004) mostraram que a propriedade livre de escala e o fenômeno do pequeno mundo se mantiveram constantes nas redes de dependência sintática construídas para três línguas europeias.

3.2.1 Redes aplicadas ao Reconhecimento de Autoria

3.2.1.1 Trabalho de Antiqueira et al. (2006)

O trabalho de Antiqueira *et al.* (2006) modela diversos livros como redes complexas para a caracterização de autoria. Nesse trabalho, os autores investigam as correlações entre textos do mesmo autor e as medidas extraídas de suas respectivas redes.

Os autores utilizaram livros de 8 autores de diferentes gêneros literários, como ficção, poesia e textos científicos. O pré-processamento do texto consistiu na remoção das *function words* e lematização das palavras remanescentes. A etapa de lematização é aplicada com a utilização de um rotulador de classes gramaticais (*part-of-speech tagger*). Palavras no plural são transformadas para sua forma singular, verbos são transformados para sua forma infinitiva e nomes são convertidos para sua forma masculina. Assim, palavras referentes a um mesmo conceito são associadas a um mesmo vértice, independente das variações de flexão.

Após a etapa de pré-processamento do texto, foi construída uma rede de co-ocorrência direcionada, onde as palavras representam os vértices e as arestas direcionadas são estabelecidas entre palavras adjacentes. A aresta sai do vértice da primeira palavra e termina no vértice da palavra seguinte. Os pesos das arestas foram determinados pelo número de vezes que as palavras adjacentes aparecem no texto.

Além de uma medida desenvolvida pelos autores baseada na conectividade da rede, o grau médio de saída, o coeficiente de aglomeração e a medida de correlação do grau (assortatividade) foram extraídas das redes de cada um dos livros. Algumas dessas medidas apresentaram variações

consideráveis entre os diferentes autores. A partir da combinação de algumas medidas, foi possível agrupar diferentes autores e separar outros, caracterizando um estilo comum de escrita. A maior diferença de estilo percebida foi entre Charles Darwin e William Wordsworth, isso já era esperado uma vez que os textos de Charles Darwin apresentam uma escrita científica e os textos de Wordsworth são poesias.

De acordo com Antiqueira *et al.* (2006), as redes produzidas por cada autor apresentam características específicas e, portanto, não são úteis apenas para capturar características autorais mas também podem ser utilizadas para a tarefa de reconhecimento de autoria.

3.2.1.2 Trabalho de Amancio et al. (2011)

O trabalho de Amancio *et al.* (2011) investiga a dependência entre as medidas de redes complexas e de intermitência com as características de estilo de diferentes autores.

A base de dados utilizada consistia em 5 livros para cada um dos oito autores. O pré-processamento do texto foi realizado com a remoção das *function words* e lematização das palavras remanescentes. Embora a frequência de *function words* tenha sido utilizada em alguns trabalhos de reconhecimento de autoria, os autores preferiram retirar as *function words* para analisarem apenas a relação entre palavras com conteúdo semântico. Os autores também utilizaram redes de co-ocorrência para a modelagem do texto.

As medidas de redes complexas extraídas foram o coeficiente de aglomeração, a média dos caminhos mínimos geodésicos e o coeficiente de intermediação (ou *Betweenness*). A medida de intermitência quantifica a regularidade da distribuição de palavras ao longo do texto. Esta medida pode ser uma boa caracterização das palavras relacionadas aos tópicos, como nomes de personagens e localizações. Além das medidas citadas, a frequência das palavras também foi utilizada para caracterização de cada um dos livros. Essas medidas foram depois transformadas de modo a caracterizar globalmente cada livro.

Três algoritmos de aprendizado de máquina foram utilizados para avaliar a capacidade das medidas em distinguir os diferentes autores. Quando todos os atributos foram utilizados, um dos algoritmos de aprendizado de máquina acertou a autoria em 50% dos casos. Entretanto, com a seleção do melhor subconjunto de atributos, a taxa obtida subiu para 65%.

Nesse trabalho, os fatores com maior dependência com a autoria são a média dos caminhos mínimos e a medida de assimetria da distribuição de intermitência das palavras, $\gamma(I)$. A média dos caminhos mínimos quantifica a distância das palavras aos *hubs* da rede (palavras frequentes). A medida $\gamma(I)$ pode ser interpretada como a fração de todas as palavras-chave de um texto.

Por fim, foram realizados novos experimentos para ilustrar como as medidas utilizadas podem ser complementares aos métodos tradicionais. Em um dos experimentos, foram extraídas a frequência e a intermitência de um conjunto composto pelas 5 palavras mais frequentes de

cada livro. A taxa de acerto para a tarefa de reconhecimento de autoria alcançou 80%. Amancio *et al.* (2011) concluem que medidas de redes complexas e de intermitência são capazes de extrair diversas características relacionadas com a autoria do textos.

3.2.1.3 Trabalho de Mehri, Darooneh e Shariati (2012)

Neste trabalho, Mehri, Darooneh e Shariati (2012) utilizam redes complexas para analisar a autoria de 36 livros de 5 famosos escritores Persas. Inicialmente, todas as palavras foram mapeadas para letras minúsculas e os sinais de pontuação e números foram removidos. Os textos não passaram por nenhuma etapa de lematização. Uma rede de co-ocorrência de palavras não direcionada é criada para cada um dos textos.

Algumas medidas de redes complexas foram extraídas das redes, como o grau, o grau médio dos vizinhos e o coeficiente de aglomeração. Além dessas, os autores também definem duas medidas, o parâmetro q e o expoente α. O parâmetro q está relacionado com a distribuição do grau e pode ser considerado uma generalização da lei de potência. O expoente α relaciona o número de arestas (N_a) e o número de vértices (N_v) em uma rede de co-ocorrência da seguinte maneira, $N_a \approx N_v^{\alpha}$.

Diferentemente dos outros trabalhos que utilizam redes complexas para o reconhecimento de autoria, no trabalho de Mehri, Darooneh e Shariati (2012) não são utilizados algoritmos de reconhecimento de padrões para induzir classificadores. Após a extração das medidas de cada uma das redes, cada autor é representado por um vetor v_a com as médias obtidas para cada uma das medidas extraídas dos livros de sua autoria. Assim, um texto de autoria desconhecida X também é representado por um vetor v_x e são calculadas as probabilidades desse texto X pertencer a cada um dos autores. Em linhas gerais, essa probabilidade é calculada em termos da distância das médias obtidas para cada uma das medidas. Quanto maiores as distâncias, menor será a probabilidade do texto X pertencer a um certo autor.

Para testar a metodologia proposta, um livro é considerado a cada passo e os vetores de cada um dos autores são calculados utilizando os 35 livros remanescentes. A autoria do livro é atribuída ao autor que apresentar a maior probabilidade. Os autores reportam uma acurácia de 91%, ao utilizarem uma fórmula baseada no número de falsos (e verdadeiros) positivos e negativos. Entretanto, os outros trabalhos citados nesta seção utilizam a acurácia para se referir à porcentagem de atribuições de autoria corretas. Nesse caso, a taxa de acerto obtida neste trabalho seria de 77.7%, uma vez que 8 livros foram atribuídos erroneamente. Por fim, os autores analisaram o poder discriminativo de pares de atributos e concluíram que a combinação do parâmetro q e o expoente α trazem bons resultados para o reconhecimento de autoria. Além disso, esses atributos podem ser combinados com outros em diversas análises da linguagem.

3.2.1.4 Trabalho de Amancio, Oliveira Jr e Costa (2012b)

Com o objetivo de medir a similaridade entre textos, Amancio, Oliveira Jr e Costa (2012b) propõem métodos que combinam características semânticas com a topologia de redes complexas. Essa abordagem foi testada em diferentes aplicações, como o reconhecimento de autoria e a classificação e avaliação de traduções.

Os textos são pré-processados e modelados como redes de co-ocorrência. O grau, a média dos caminhos mínimos e o coeficiente de aglomeração e de intermediação foram extraídos dessas redes. Além dessas medidas, os autores extraíram as frequências de motivos com 3 vértices. Três grupos de índices de similaridade são então definidos para avaliar e classificar as traduções automáticas e o reconhecimento de autoria. O primeiro e o segundo grupo utilizam apenas informações semânticas e informações topológicas, respectivamente. Por fim, o último grupo é considerado híbrido por combinar essas informações.

Em uma das aplicações, Amancio, Oliveira Jr e Costa (2012b) investigaram a adequabilidade de informações semânticas e topológicas para a tarefa de reconhecimento de autoria. Dois conjuntos de textos foram utilizados nesse experimento, um de textos poéticos e outro em prosa. Para ambos os conjuntos de texto, foi constatado que tanto a característica semântica quanto a topológica são relevantes para caracterizar a autoria. Sobre os textos em prosa, foi observado que alguns autores compartilham os mesmos conteúdos semânticos, porém utilizam estilos de escrita diferentes. Do mesmo modo, outros apresentam estilos de escrita semelhantes, embora escrevam sobre assuntos completamente diferentes.

Por fim, os autores concluem que os métodos mais adequados para a classificação de textos dependem do propósito de cada aplicação, seja distinguir o estilo de escrita do autor ou os diferentes tópicos utilizados por eles. Autores com estilos de escrita semelhantes podem ser diferenciados pelo conteúdo semântico de seus textos. Por outro lado, autores que escrevem sobre o mesmo tópico podem ser separados pelo estilo individual de escrita. Além disso, estas informações podem ser combinadas para o reconhecimento de autoria quando muitos autores precisam ser diferenciados.

3.2.1.5 Trabalho de Amancio (2015c)

Em contraste com os trabalhos citados anteriormente que fazem uso de livros inteiros ou textos longos, Amancio (2015c) utiliza amostras de textos de tamanho relativamente curto para verificar a manutenção das propriedades topológicas das redes complexas. O objetivo é investigar se a topologia de pequenas amostras de texto fornece atributos relevantes para a análise textual. Além disso, as amostras obtidas foram analisadas com relação à tarefa de reconhecimento de autoria.

Cada uma das amostras passa por etapas de pré-processamento e depois é modelada através do uso de redes de co-ocorrência. Diversas medidas, incluindo o coeficiente de aglomeração,

a intermitência e a acessibilidade, foram utilizadas. A primeira análise realizada foi investigar o comportamento das propriedades topológicas em um total de 50 livros. Cada livro foi dividido em amostras com W *tokens* (no caso, palavras), com W variando de 300 a 2100. Os resultados revelaram que a maioria das medidas topológicas são estáveis e apresentam baixa variabilidade nas amostras de textos com tamanhos pequenos. De acordo com Amancio (2015c), os autores tendem a manter seus estilos em diferentes porções do mesmo livro.

Para verificar a aplicabilidade em outras tarefas, o autor investigou como a amostragem de textos afeta a tarefa de reconhecimento de autoria. Foram utilizados cinco livros para cada um dos quatro autores analisados. O tamanho das amostras variou entre 500 e 21400 *tokens*. O tamanho de 21400 *tokens* corresponde ao caso no qual não é feita nenhuma amostragem e o texto está completo. Foram utilizados quatro métodos de classificação supervisionados para quantificar os efeitos da amostragem na classificação de diferentes estilos. Ao analisar as taxas de acerto obtidas, confirmou-se que as características topológicas das amostras são capazes de diferenciar os autores. As taxas de acerto mais baixas correspondem às amostras com 500 *tokens*.

Na maioria dos casos, as maiores taxas de acerto não ocorreram quando livros inteiros (21400 *tokens*) foram utilizados e sim com amostras menores, algumas representando menos de 8% do tamanho de um livro original. Isso sugere que quando a amostragem é feita adequadamente, a performance de classificadores pode ser melhorada. Com esse trabalho, Amancio (2015c) mostrou que textos curtos podem ser analisados com métodos e conceitos de redes complexas. Por fim, uma desvantagem apresentada pelo autor é que o método só pode ser aplicado em textos longos. A amostragem de documentos curtos gera textos com grande variabilidade topológica.

3.2.1.6 Trabalho de Amancio, Silva e Costa (2015)

Neste trabalho, Amancio, Silva e Costa (2015) utilizam o conceito de simetria para analisar os padrões de conectividade em redes de co-ocorrência (adjacência) de palavras. A medida de simetria é depois avaliada na tarefa de reconhecimento de autoria. Essa aplicação é escolhida devido ao fato de que autores tendem a utilizar diversas restrições (ou escolhas) em seus textos que afetam o estilo de escrita. Por exemplo, repetir uma palavra ou trocá-la por diversos sinônimos. Portanto, espera-se que isso se reflita em características simétricas na estrutura da rede.

As medidas de simetria concêntrica utilizadas baseiam-se na acessibilidade (VIANA; BATISTA; COSTA, 2012). Estas medidas são denominadas simetria *backbone* e *merged* e ambas realizam transformações nas redes analisadas. Para o cálculo da simetria *backbone*, todas as arestas existentes entre vértices do mesmo nível concêntrico h são removidas. Diferentemente, na simetria *merged*, todos os vértices com arestas no mesmo nível concêntrico h são combinados em um super vértice. Após essas transformações, as probabilidades de transição são extraídas para cada um dos vértices (como na Seção 2.2.4.8) e utilizadas no cálculo da simetria. Os autores mostraram que as medidas de simetria não apresentam correlação significativa com

outras medidas tradicionais, como o coeficiente de aglomeração e de intermediação.

Uma outra análise foi realizada para investigar se as medidas de simetria são capazes de caracterizar diversos estilos de escrita. Para a tarefa de reconhecimento de autoria, foram utilizados 40 livros de 8 autores distintos. As medidas de simetria *merged* e *backbone* foram calculadas para as 229 palavras que aparecem em comum em todos os livros. Foram utilizados quatro métodos de classificação supervisionados para mensurar o poder discriminativo das medidas de simetria. As taxas de acerto variam de 20% a 82.5% para as diversas combinações de algoritmos de reconhecimento de padrões, simetria *merged* e *backbone* e níveis concêntricos h.

Os autores concluem que as medidas de simetria são capazes de caracterizar as marcas estilísticas deixadas pelos autores, uma vez que cada autor apresenta um viés na escrita com relação ao uso de diferentes padrões e isso é quantificado através da homogeneidade de acesso aos vértices. Além disso, por não apresentarem correlação significativa com outras medidas tradicionais, as medidas de simetria podem ser utilizadas para complementar a caracterização das redes em diversas atividades, como o reconhecimento de autoria.

3.2.1.7 Trabalho de Amancio (2015a)

Embora existam vários trabalhos sobre reconhecimento de padrões em texto, poucos se dedicaram à análise de flutuações estilísticas. Padrões informativos podem estar escondidos nessas flutuações. Por exemplo, ao considerar a distribuição da frequência de palavras, as flutuações em torno da média podem ser úteis para detectar os conceitos mais relevantes. Neste contexto, Amancio (2015a) analisa a variabilidade estilística entre textos através da análise de redes complexas e da medida de intermitência.

No primeiro experimento, Amancio (2015a) investiga se as variações estilísticas entre textos fornecem atributos úteis para a tarefa de reconhecimento de autoria. A evolução estilística foi analisada através da divisão de cada texto em textos menores com W *tokens* (palavras), com W entre 500 e 1300. Cada subtexto foi modelado como uma rede de co-ocorrência e foram extraídas diversas medidas dessas redes. As medidas extraídas foram a acessibilidade, a média dos caminhos mínimos e o coeficiente de aglomeração e de intermediação. Desse modo, cada medida X gera uma série temporal onde cada elemento da série x_i representa o valor obtido para a medida X no subtexto i. As séries temporais de cada livro foram decompostas através da transformação de Fourier e alguns componentes dessa transformação foram utilizados como atributos para os algoritmos de reconhecimento de padrões.

A partir da análise da variação das medidas, diversas características interessantes foram encontradas. A menor taxa de acerto para o reconhecimento de autoria ocorreu para $W = 500$ e a maior com $W = 1300$, 35% e 45%, respectivamente. O atributo mais relevante para discriminar diferentes autores foi a variabilidade da média dos caminhos mínimos. Os resultados obtidos confirmam que as variações estilísticas podem ser utilizadas para distinguir o estilo de escrita dos autores. Além disso, a metodologia proposta pode ser utilizada como um atributo estilístico

complementar e combinada com atributos tradicionais.

Em outro experimento, Amancio (2015a) investiga se apenas a medida de intermitência de algumas palavras pode ser utilizada na tarefa de reconhecimento de autoria. Os textos originais (sem pré-processamento) foram utilizados para esse experimento. Os valores de intermitência das 100 palavras mais frequentes no conjunto de livros foram utilizados como atributos para os classificadores. A taxa de acerto obtida para a tarefa de reconhecimento de autoria foi de 65%. Com a remoção dos três autores responsáveis pela maioria dos erros de classificação, a porcentagem de acertos subiu para 90%. Apesar do grande número de atributos utilizados, o poder discriminativo concentrou-se em apenas algumas *function words*, como *but*, *and*, *I*, *who* e *as*. Este resultado sugere que a distribuição não uniforme de algumas palavras também fornece atributos úteis para o reconhecimento de autoria.

3.2.1.8 Trabalho de Amancio (2015b)

Apesar da grande quantidade de trabalhos que utilizam redes complexas para analisar fenômenos da linguagem, apenas uma pequena parcela deles investiga como as redes complexas podem ser utilizadas para melhorar a performance de diversas tarefas de PLN e, assim, melhorar o estado da arte. Em alguns casos, os melhores resultados ainda são obtidos com métodos tradicionais de PLN. Neste contexto, Amancio (2015b) trata esse problema com a definição de métodos que utilizam redes complexas para melhorar a performance de tarefas de classificação de texto, como o reconhecimento de autoria e a identificação de estilos.

Antes de modelar os textos como redes de adjacência de palavras, algumas etapas de pré-processamento são aplicadas aos textos. Com relação às medidas topológicas, o grau médio, a acessibilidade, a assortatividade, entre outras medidas, foram extraídas das redes. A frequência de algumas *function words* e dos bigramas de caracteres são algumas medidas tradicionais e que foram utilizadas nesse trabalho. Dois métodos diferentes foram utilizados para combinar o componente tradicional e o baseado nas medidas de redes complexas. Em um dos métodos, denominado *Hybrid*, as características topológicas e tradicionais são utilizadas através de uma combinação linear. No outro método, denominado *Tiebreaker*, os atributos topológicos são utilizados apenas quando a classificação realizada pelos métodos tradicionais não é satisfatória.

Para a realização dos experimentos, diversas combinações de atributos tradicionais e topológicos foram consideradas. Em termos gerais, ao combinar características tradicionais com medidas topológicas, os dois métodos melhoraram a performance para a tarefa de reconhecimento de autoria. Além disso, a combinação fornecida pelo método *Hybrid* superou, em muitos casos, as acurácias obtidas quando apenas características tradicionais ou topológicas são utilizadas separadamente.

Tanto para a tarefa de reconhecimento de autoria quanto para a identificação de estilos, foram constatados ganhos de performance nos classificadores quando os métodos tradicionais foram combinados com as técnicas baseadas na topologia das redes. Desse modo, as medidas

topológicas são úteis para complementar a caracterização do estilo de escrita dos autores. As técnicas desenvolvidas nesse trabalho são de grande importância, uma vez que melhoram as atuais estratégias de classificação.

3.2.2 Outros Trabalhos

Esta seção apresenta alguns trabalhos que não estão relacionados com o reconhecimento de autoria mas que utilizaram a topologia de redes de co-ocorrência para analisar o estilo de escrita, classificar e distinguir linguagens ou movimentos literários.

3.2.2.1 Trabalho de Amancio, Oliveira Jr e Costa (2012a)

Com o objetivo de investigar as variações no estilo de escrita dos autores, Amancio, Oliveira Jr e Costa (2012a) utilizam 77 livros publicados entre os anos de 1590 e 1922 e identificaram diversos movimentos literários. O período de publicação dos livros coincide com importantes movimentos literários dos últimos cinco séculos. Os livros são pré-processados e modelados como redes direcionadas de co-ocorrência de palavras. Diversas medidas de redes complexas foram extraídas dessas redes, como o coeficiente de aglomeração, a média dos caminhos mínimos e a assortatividade.

Através dos resultados obtidos com essas medidas, os autores procuraram o melhor particionamento dos dados. Isso foi realizado através da análise da qualidade do agrupamento utilizando a restrição de que livros com datas de publicação consecutivas deveriam pertencer ao mesmo grupo ou estar nas fronteiras de grupos consecutivos. Para isso, os autores variaram alguns parâmetros, como o número de grupos, e mediram a qualidade do agrupamento. A melhor partição encontrada obteve seis grupos (*clusters*) de livros, com quase nenhuma sobreposição entre os grupos. Cada um dos *clusters* representava corretamente um movimento literário clássico.

Além do resultado citado anteriormente, Amancio, Oliveira Jr e Costa (2012a) fizeram outras importantes observações. Grupos de períodos de tempo consecutivos são normalmente posicionados lado a lado, o que indica a ocorrência de mudanças suaves no estilo de escrita de um movimento literário para outro. Essa conclusão foi também verificada através da análise hierárquica dos grupos obtidos. Além disso, os autores concluem que as mudanças de estilo entre dois grupos consecutivos parecem ser movidas pela oposição, no sentido de que o movimento literário atual pode ser visto como uma oposição aos movimentos literários anteriores. A medida mais importante para a distinção entre os diferentes estilos literários foi o comprimento médio do caminho mínimo. Por fim, os autores sugerem que a abordagem utilizada neste trabalho pode ser útil para o estudo da evolução temporal de outros sistemas.

3.2.2.2 Trabalho de Liu e Cong (2013)

Neste trabalho, Liu e Cong (2013) analisaram a viabilidade de classificar linguagens com o uso de redes de co-ocorrência de palavras. A classificação de linguagens era usualmente feita através de redes de dependência sintática. Entretanto, a construção dessas redes é custosa e depende da existência de ferramentas para cada uma das linguagens.

Como forma de manter constante a semântica e gênero dos textos analisados, os autores utilizaram 14 traduções de uma novela russa e criaram uma rede de co-ocorrência não direcionada para cada uma delas. As *function words* não foram removidas nesse trabalho. Dentre as linguagens das traduções, 12 consistem em diferentes línguas eslavas e 2 não eslavas, o mandarim e o inglês. Foram extraídas diversas medidas de redes complexas, como o grau médio, o coeficiente de aglomeração, entre outras.

A partir da combinação de diversos parâmetros extraídos das redes, algoritmos de agrupamento puderam distinguir as linguagens eslavas das não eslavas. Além disso, também foi possível agrupar corretamente as linguagens eslavas dentro de suas sub-divisões. Portanto, redes de co-ocorrência também podem ser utilizadas com sucesso para a tarefa de classificação topológica de linguagens e constituem uma alternativa ao uso das redes de dependência sintática.

Uma outra observação feita em Liu e Cong (2013) é que a rede de co-ocorrência de palavras tende a ser similar à rede de dependência sintática extraída do mesmo texto. Isso ocorre porque existe uma grande sobreposição de arestas entre as duas redes, uma vez que muitas das relações sintáticas acontecem entre palavras adjacentes (LIU, 2008).

3.2.2.3 Trabalho de Amancio et al. (2013a)

De acordo com (AMANCIO *et al.*, 2013a), poucos estudos dedicaram-se à investigação das medidas estatísticas com a análise de diferentes linguagens e textos. Neste trabalho, Amancio *et al.* (2013a) propõem um *framework* capaz de determinar se um texto é compatível com uma linguagem natural e de quais linguagens o texto é mais próximo, sem utilizar nenhum conhecimento sobre o significado das palavras.

A abordagem desenvolvida baseia-se na utilização de três classes de medidas, as medidas de primeira ordem (como o tamanho do vocabulário), medidas extraídas da rede (como o coeficiente de aglomeração e a quantidade de motivos) e a medida de intermitência. Para as análises comparativas, as propriedades foram obtidas de um mesmo livro (O Novo Testamento) em 15 linguagens e diferentes livros escritos em Inglês e Português. O objetivo dessa comparação é identificar os atributos capazes de distinguir um texto de sua versão aleatória e determinar a proximidade entre textos. Além disso, foi possível identificar medidas que estão mais relacionadas com a linguagem (sintaxe) do que com o conteúdo do texto (semântica).

Para demonstrar a relevância do *framework* em investigações de textos desconhecidos, os autores analisaram um manuscrito conhecido como Manuscrito de Voynich, até hoje não

decifrado. Os autores chegaram à conclusão que o manuscrito difere-se de uma sequência aleatória de palavras, sendo compatível com linguagens naturais. Isso indica a possível existência de uma verdadeira mensagem no manuscrito. Mesmo sem ter o objetivo de decifrar o manuscrito, as abordagens definidas nesse trabalho foram capazes de extrair palavras chaves do texto, as quais podem ser úteis para tentativas futuras de analisá-lo.

3.2.2.4 Trabalho de Arruda, Costa e Amancio (2015)

Com relação à tarefa de classificação de documentos, muitas abordagens existentes utilizam o conteúdo semântico dos textos. Entretanto, algumas características como a estrutura textual são utilizadas em apenas uma pequena parcela desses trabalhos. Neste contexto, Arruda, Costa e Amancio (2015) utilizam redes complexas para modelar e classificar textos em prosa em duas categorias, os informativos e os imaginativos. O conjunto de textos informativos é formado principalmente por manuscritos científicos e biografias. O conjunto de textos imaginativos apresenta romances, ficções em geral, entre outros.

Após passarem por etapas de pré-processamento, os textos em prosa são modelados como redes de co-ocorrência de palavras. Diversas medidas de redes complexas foram extraídas dos textos, como a acessibilidade e a simetria, medidas de centralidade, entre outras. Diferentes estratégias foram utilizadas nesse trabalho, algumas utilizaram as *function words* e em outras, essas palavras foram removidas na etapa de pré-processamento. Além disso, algumas estratégias consideraram as medidas globais (obtidas para toda a rede) e outras utilizaram as locais (obtidas para cada uma das palavras) como atributos para os classificadores. Três algoritmos de aprendizado de máquina supervisionados foram utilizados.

A partir dos experimentos realizados, a maior taxa de acerto obtida foi de 95%. As taxas obtidas nas estratégias locais foram maiores do que as obtidas com as estratégias globais, o que sugere que apenas algumas palavras contribuem para a tarefa de classificação. Uma análise sistemática da relevância dos diversos atributos revelou que a simetria e a acessibilidade de algumas palavras (como *the*, *by*, e *have*) foram os atributos mais relevantes para a tarefa de classificação de textos em prosa. A partir dos resultados obtidos, os autores sugerem que as medidas de acessibilidade e simetria podem ser utilizadas em aplicações relacionadas, uma vez que apresentam um papel complementar à caracterização dos textos.

3.3 Considerações Finais

Neste capítulo foi apresentado o estado da arte em reconhecimento de autoria através do uso de redes complexas. A maioria dos trabalhos apresentados utiliza apenas medidas topológicas das redes. Poucos são os estudos que combinam essas medidas com técnicas tradicionais de processamento de linguagens naturais. Como mencionado no Capítulo 1, um dos objetivos desse trabalho é combinar o uso de técnicas tradicionais de reconhecimento de autoria com as

medidas extraídas da topologia das redes. A Tabela 1 apresenta uma comparação entre todos os trabalhos descritos na Seção 3.2.1. É importante ressaltar que alguns fatores como as etapas de pré-processamento realizadas, o número de autores, a quantidade e o tamanho dos textos disponíveis influenciam na taxa de acerto obtida em cada método.

No próximo capítulo são definidos formalmente o objetivo e a motivação deste trabalho, a metodologia, as principais atividades planejadas e alguns resultados prévios.

Tabela 1 – Resumo dos trabalhos relacionados apresentados em 3.2.1

Trabalho	**Tamanho do Corpus**	**Língua**	**Número de Autores**	**Máxima Taxa de Acerto do Método Proposto**
Antiqueira *et al.* (2006)	44	Inglês	8	Não se aplica
Amancio *et al.* (2011)	40	Inglês	8	65%
Mehri, Darooneh e Shariati (2012)	36	Persa	5	77.7%
Amancio, Oliveira Jr e Costa (2012b)	20 Vários poemas	Inglês Inglês	4 4	Não se aplica Não se aplica
Amancio (2015c)	20	Inglês	4	86.67%
Amancio, Silva e Costa (2015)	40	Inglês	8	82.5%
Amancio (2015a)	40	Inglês	8	65%
Amancio (2015b)	40	Inglês	8	Não se aplica

CAPÍTULO

4

PROPOSTA DE TRABALHO

4.1 Considerações Iniciais

No Capítulo 3 foram apresentados diversos trabalhos que utilizam redes complexas para resolver problemas de processamento de linguagens naturais, em especial, o problema de reconhecimento de autoria. A partir dos trabalhos apresentados, pode-se perceber que a principal diferença entre eles encontra-se nas medidas de redes complexas utilizadas. Entretanto, quase todos utilizam a mesma metodologia de criação da rede, com redes de co-ocorrência de palavras que conectam apenas as palavras adjacentes. Além disso, poucos deles beneficiam-se das técnicas tradicionais de reconhecimento de autoria.

Neste capítulo, a proposta de pesquisa deste trabalho é apresentada. Na Seção 4.2, são apresentadas a motivação e o objetivo do trabalho; na Seção 4.3, é descrita a metodologia utilizada; os resultados preliminares obtidos são apresentados na Seção 4.5; e, por fim, na Seção 4.6, são definidos o cronograma e o plano de atividades planejado para este trabalho de Mestrado.

4.2 Motivação e Objetivos

Desde o fim da década de 90, os estudos de atribuição de autoria passaram por grandes mudanças. Com a popularização e facilidade de acesso à rede mundial de computadores, uma grande quantidade de textos eletrônicos tornou-se disponível na Internet (por exemplo e-mails, posts em blogs e fóruns) e aumentou a necessidade por métodos eficientes de tratamento dessas informações. Devido ao aumento do poder computacional, os métodos de atribuição de autoria, que antes eram assistidos por computador, passaram a ser baseados em computador, com o desenvolvimento de sistemas totalmente automatizados (STAMATATOS, 2009). Essas mudanças trouxeram um grande e positivo impacto em áreas como aprendizado de máquina e processamento de línguas naturais.

Nos últimos anos, a quantidade de textos disponíveis e de fácil acesso na Web revelou o potencial da análise de autoria em diferentes aplicações de diversas áreas. Esta tarefa é bastante relevante dentro da área de processamento de línguas naturais contribuindo para diversos avanços na literatura (MATTHEWS; MERRIAM, 1993), história (TWEEDIE; SINGH; HOLMES, 1996), serviços de inteligência (ABBASI; CHEN, 2005), computação forense (JUOLA, 2006; FRANTZESKOU *et al.*, 2006) e também em investigações criminais (CHASKI, 2005). Outra aplicação importante desta tarefa se dá no contexto do plagiarismo, pois é possível identificar trechos de plágios e inconsistências estilísticas a partir da tarefa de verificação de autoria. A importância desta tarefa se torna ainda mais evidente também quando se deseja estimar a similaridade de textos (AMANCIO *et al.*, 2013a).

Embora redes complexas já tenham sido utilizadas para reconhecer autoria, o estado da arte ainda não foi atingido (AMANCIO *et al.*, 2011). Por este motivo, o objetivo deste projeto de mestrado é tentar aperfeiçoar os atuais modelos baseados em redes através da extensão do modelo de co-ocorrência atual e concepção/uso de novas medidas de análise que sejam mais dependentes do estilo. Este projeto visa também estudar métodos de criação de classificadores híbridos que forneçam a combinação mais adequada para o problema. Além da contribuição esperada em termos de desempenho, os métodos de redes serão adaptados para situações em que existe a possibilidade de ataque (autores disfarçando o estilo) e em casos onde os textos obtidos sejam curtos. Os avanços obtidos neste projeto devem ser úteis não apenas para o reconhecimento de autoria, mas também para identificar plagiarismos caracterizados por textos com estrutura similar. Em resumo, os principais objetivos desse trabalho podem ser descritos pelo seguinte parágrafo:

Este projeto de mestrado tem por objetivo desenvolver novos modelos para o reconhecimento de autoria com o uso de redes complexas. Especificamente, pretende-se adaptar o modelo de adjacência ao incluir palavras de pouco conteúdo semântico e conectar palavras em um maior contexto (não apenas palavras vizinhas, como nas redes de adjacência convencionais). No lado topológico, novas medidas serão introduzidas para aperfeiçoar a caracterização da topologia. Por fim, espera-se ainda otimizar os métodos atuais de reconhecimento de autoria combinando os atributos obtidos de medidas topológicas com os atributos tradicionais em um classificador híbrido.

Para alcançar estes objetivos, a metodologia de modelagem de textos através de redes de co-ocorrência de palavras será utilizada e adaptada. Na próxima seção, são detalhadas a metodologia de modelagem de textos como redes complexas e os demais aspectos que envolvem a análise de textos para o reconhecimento de autoria.

4.3 Metodologia

Como apresentado no Capítulo 3, os diversos trabalhos que utilizam redes complexas para abordar a tarefa de reconhecimento de autoria diferenciam-se, na maioria dos casos, pelas medidas de redes complexas que utilizam. Em geral, as abordagens realizadas por esses trabalhos utilizam a mesma modelagem de co-ocorrência de palavras. Para atingir os objetivos propostos, pretende-se alterar essa modelagem, definindo novas maneiras de conexões entre os vértices. A partir desta nova estrutura, o desempenho de algumas tarefas de processamento de linguagens naturais será avaliado, em especial, o reconhecimento de autoria.

Nesta seção, são descritos o processo de construção das redes de co-ocorrência e algumas modificações planejadas para essa modelagem. Depois, são apresentados os passos necessários para, a partir da rede, avaliar a tarefa de reconhecimento de autoria. A seguir, é apresentada a base de dados utilizada nas análises iniciais desse projeto de mestrado.

4.3.1 Base de Dados

A base de dados utilizada para as análises de reconhecimento de autoria é composta por 40 livros (em inglês), 5 de cada um dos 8 autores selecionados. Os livros foram publicados entre 1835 e 1922 e estão disponíveis no repositório online denominado Project Gutemberg [1]. A lista com os 40 livros utilizados está disponível no Apêndice A. Para evitar a influência de textos com tamanhos diferentes, cada livro foi limitado às suas primeiras 22.444 palavras, representando o tamanho do menor livro. No restante dessa seção, alguns exemplos são ilustrados utilizando-se o livro *"The Adventures of Sherlock Holmes"* ("As Aventuras de Sherlock Holmes") do autor Arthur Conan Doyle.

4.3.2 Pré-processamento e Conexão de Palavras

A modelagem de textos como redes complexas pode ser dividida em duas etapas, o pré-processamento do texto e a conexão de conceitos. Uma das atividades iniciais de pré-processamento é a retirada de sinais de pontuação. Nesse trabalho, palavras contendo apóstrofes (como *he's* e *isn't*) não foram alteradas. Acredita-se que a escolha de utilizar *isn't* em vez de sua forma equivalente *is not* está relacionada com o estilo de escrita do autor e, por isso, deve ser representada de maneira diferente na rede.

Um procedimento tipicamente adotado em trabalhos de literatura (AMANCIO; OLIVEIRA JR; COSTA, 2012a) é a extração das *stopwords* ou *function words*, por apresentarem pouco conteúdo semântico. A lista de *stopwords* utilizada neste trabalho é apresentada no Apêndice B. Porém, em uma das atividades desse projeto, as *stopwords* serão incluídas na construção da rede de palavras. O objetivo é analisar o aumento do poder descritivo da rede para a tarefa de reconhecimento de autoria, visto que tais palavras são úteis em estratégias tradicionais de

[1] http://www.gutenberg.org/

reconhecimento de autoria não baseadas em grafos. Outra etapa do pré-processamento é a lematização. Essa etapa é aplicada nas palavras remanescentes com a utilização de um rotulador de classes gramaticais (*part-of-speech tagger*). Palavras no plural são transformadas para sua forma singular, verbos são transformados para sua forma infinitiva e nomes são convertidos para sua forma masculina. Assim, palavras referentes a um mesmo conceito serão associadas a um mesmo vértice, independente das variações de flexão. O rotulador de classes gramaticais utilizado neste trabalho foi o NLTK (*Natural Language Toolkit*) (BIRD; KLEIN; LOPER, 2009). Para exemplificar os passos de pré-processamento que são utilizados neste trabalho, a Tabela 2 ilustra suas aplicações em um extrato em inglês do livro *"The Adventures of Sherlock Holmes"*, de Arthur Conan Doyle.

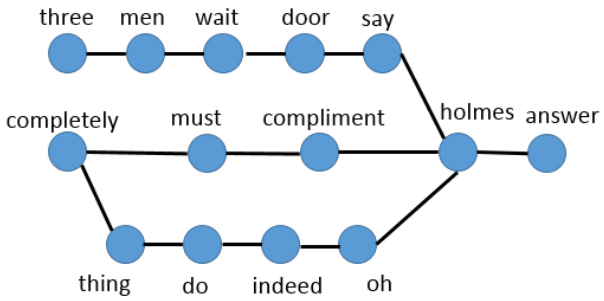
Tabela 2 – Exemplo de aplicação do pré-processamento para modelagem de textos como redes. Um extrato obtido do livro *"The Adventures of Sherlock Holmes"*, de Arthur Conan Doyle, está ilustrado após a remoção das *stopwords* e subsequente lematização

Original	**Sem *stopwords***	**Após lematização**
There are three men waiting for	*three men waiting*	*three men wait*
him at the door, said Holmes.	*door said holmes*	*door say holmes*
Oh, indeed! You seem to have	*oh indeed seem*	*oh indeed seem*
done the thing very completely.	*done thing completely*	*do thing completely*
I must compliment you.	*must compliment*	*must compliment*
And I you, Holmes answered.	*holmes answered*	*holmes answer*

Após o pré-processamento, é necessário realizar a conexão de conceitos. Por ser formada por cadeias lineares de palavras, o modo mais simples de representar a linguagem escrita é conectar palavras adjacentes. Este tipo de rede, conhecido como rede de co-ocorrência, é amplamente utilizado na literatura (CANCHO; SOLé, 2001; AMANCIO *et al.*, 2011; ROXAS; TAPANG, 2010). Como a maioria das relações sintáticas acontece na primeira vizinhança, esta modelagem pode ser vista como uma aproximação das ligações sintáticas (CANCHO; SOLÉ; KÖHLER, 2004). Em uma rede de co-ocorrência representando textos, os vértices representam palavras e as arestas são estabelecidas entre palavras vizinhas. A rede obtida com o exemplo da Tabela 2 está ilustrada na Figura 5. Esta rede foi criada conectando-se cada palavra ao seu vizinho mais próximo.

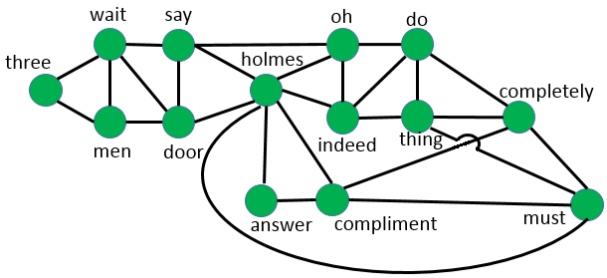
Outro fator importante na modelagem de co-ocorrência refere-se à escolha do tamanho da janela de conexão de palavras. A abordagem de conexão de palavras imediatamente vizinhas, como na Figura 5, representa uma simplificação considerável, já que palavras distantes podem estar relacionadas sintaticamente ou semanticamente (ALVAREZ-LACALLE *et al.*, 2006). Um dos objetivos deste trabalho é analisar a efetividade da modelagem para diferentes janelas de conexão. Além disso, em uma modelagem diferente, todas as palavras de uma sentença serão conectadas. O objetivo é que, com essas alterações na modelagem, seja possível capturar uma quantidade maior de *links* relevantes entre palavras e caracterizar melhor o contexto. A Figura 6 ilustra o subgrafo construído, para a mesma sentença do livro *"The Adventures of Sherlock Holmes"*, utilizando uma janela de conexão de tamanho 2.

Figura 5 – Subgrafo obtido para as sentenças apresentadas na Tabela 2. Nesse subgrafo, cada palavra foi conectada ao seu vizinho mais próximo.



Fonte: Elaborada pelo autor.

Figura 6 – Subgrafo obtido para as sentenças apresentadas na Tabela 2. Nesse subgrafo, cada palavra foi conectada ao seus dois vizinhos mais próximos.



Fonte: Elaborada pelo autor.

Vários são os exemplos da utilidade das redes de co-ocorrência para modelagem de textos (AMANCIO; OLIVEIRA JR; COSTA, 2012a; AMANCIO *et al.*, 2012; AMANCIO; OLIVEIRA JR; COSTA, 2012b; AMANCIO, 2015c). É importante ressaltar que a representação por redes e os métodos de classificação utilizados são genéricos, sendo aplicáveis em várias tarefas de processamento de línguas naturais, não apenas em reconhecimento de autoria.

Após cada livro ser mapeado em uma rede complexa, as medidas descritas na Seção 2.2.4 podem ser extraídas de cada uma das redes. A seguir, é explicado o processo pelo qual medidas locais (calculadas para cada uma das palavras) são transformadas em medidas globais (representam toda a rede).

4.3.3 Extraindo Propriedades Globais a partir de Propriedades Locais

Na Seção 2.2.4, foram apresentadas as medidas de redes complexas utilizadas nesse trabalho. A quantidade de motivos extraída é referente à toda a rede. A medida de assortatividade é global, porém as outras medidas são extraídas para cada uma das palavras. O objetivo é obter valores que possam ser usados como medidas globais de cada um dos livros (e não das palavras).

A escolha mais natural é calcular a média $\langle X \rangle$, que corresponde à média da medida X sobre todas as M palavras distintas. Além disso, o desvio de cada medida também foi calculado. Por fim, a *skewness*, $\gamma(X)$, foi utilizada para quantificar a assimetria da distribuição da medida X. Em resumo, os três atributos que foram utilizados para cada uma das medidas são descritos a seguir:

$$\text{Média:} \qquad \langle X \rangle = \frac{1}{M} \sum_{i=1}^{M} X_i \tag{4.1}$$

$$\text{Desvio:} \qquad \sigma(X) = \sqrt{\frac{\sum_{i=1}^{M} (X_i - \langle X \rangle)^2}{M-1}} \tag{4.2}$$

$$\textit{Skewness:} \qquad \gamma(X) = \left\langle \left(\frac{X - \langle X \rangle}{\sigma(X)} \right)^3 \right\rangle \tag{4.3}$$

onde X representa cada uma das medidas descritas na Seção 2.2.4, com exceção da assortatividade e dos motivos.

4.3.4 Técnicas de Reconhecimento de Padrão

Com o objetivo de mensurar a habilidade de diferenciação de autores através das medidas extraídas das redes, foram utilizados algoritmos de aprendizado de máquina que induzem classificadores a partir de uma base de treinamento. A robustez dos resultados foi testada utilizando-se quatro diferentes algoritmos de aprendizado supervisionado.

Os algoritmos utilizados para avaliar os modelos de reconhecimento de autoria são *Support Vector Machines*, kNN, *Naive Bayes* e C4.5 (DUDA; HART; STORK, 2000). Antes de descrever esses métodos, considere as seguintes definições. O conjunto de treinamento $X_{training} = \{(x_1, y_1), ..., (x_l, y_l)\}$ apresenta l tuplas, onde o primeiro componente da i-ésima tupla $x_i = (f_1, ...f_d)$ representa os atributos do livro. O segundo componente, y_i, é o nome do autor do livro. O objetivo de um sistema de classificação supervisionada é aprender o mapeamento $x \mapsto y$ a partir do conjunto de treinamento. Para verificar a acurácia do mapeamento, um conjunto de teste $X_{test} = \{x_{l+1}, ...x_{l+u}\}$ é utilizado. As técnicas utilizadas neste trabalho são descritas a seguir:

- ***Support Vector Machines***: Nesta técnica, os exemplos de treinamento são divididos em diversas regiões do espaço de acordo com suas categorias. Esta divisão é realizada por funções específicas que objetivam maximizar a margem de separação (FACELI *et al.*, 2011). Desse modo, novas instâncias são então classificadas de acordo com seu mapeamento em uma das regiões de separação.

- **kNN (*k-Nearest Neighbors*)**: Esta técnica é baseada em um processo de votação realizado sobre as k instâncias do conjunto de treinamento mais próximas (k_{nn}), em um espaço normalizado que envolve todos os atributos (AHA *et al.*, 1991). Se a maioria das instâncias

no conjunto k_{nn} são classificadas com a classe y', então esta classe também é atribuída à instância desconhecida. O valor de k utilizado neste trabalho é $k = 1$.

- ***Naive Bayes***: Este método é baseado no Teorema de Bayes e usa a regra de decisão que afirma que a correta classe y' de uma instância satisfaz a condição

$$P(y'|f_1, \ldots f_d) > P(y_k|f_1, \ldots f_d) \tag{4.4}$$

para cada $y_k \neq y'$, onde $P(y_k|f_1, \ldots f_d)$ é a probabilidade de uma instância ser classificada com a classe y_k dadas as características $F = \{f_1, \ldots f_d\}$ (JOHN; LANGLEY, 1995).

- **C4.5**: Este método cria uma árvore de decisão baseada na informação adicionada por cada um dos atributos (QUINLAN, 1993).

Neste trabalho, os algoritmos foram aplicados a um conjunto de treinamento independente do conjunto de teste, utilizando a técnica de validação cruzada *leave-one-out* (FACELI *et al.*, 2011). Com o uso dessa técnica, a cada ciclo exatamente um livro de cada autor é utilizado como teste, enquanto os 4 livros restantes são utilizados no treinamento do preditor. O desempenho de cada algoritmo foi dado pela porcentagem de atribuições corretas ao testar a autoria de cada livro individualmente.

4.4 Principais Atividades

Nesta seção, as principais atividades propostas para este projeto são detalhadas.

4.4.1 Extensão da Modelagem

Ao longo deste projeto, serão propostos modelos alternativos ao modelo de co-ocorrência tradicional. Uma das mudanças refere-se à inclusão de *stopwords* (ou *function words*) na construção da rede, pois tais palavras são úteis em estratégias tradicionais de reconhecimento não baseadas em grafos (STAMATATOS, 2009). Também investigaremos como a conexão de palavras em um maior contexto (não apenas palavras vizinhas como nas redes de adjacência convencionais) afeta a tarefa. Em especial, acreditamos que haverá um tamanho de janela J ótimo onde a classificação será maximizada. Obviamente, o valor de J não deve ser muito alto já que a conexão de palavras mais distantes tornará a modelagem capaz de capturar as características semânticas do texto (AMANCIO *et al.*, 2013a), que a princípio não são úteis nas tarefas de estilometria. Uma vez que a introdução de novos modelos pode mudar a interpretação das medidas, para todos os modelos propostos será realizada uma análise das características das medidas referentes à sua informatividade (capacidade de identificação de estruturas em textos) e natureza (ou seja, se a medida captura fatores sintáticos ou semânticos).

Um problema comum referente ao tratamento de textos com redes complexas que pode afetar negativamente a capacidade de captura de características estilométricas é a dependência de

algumas medidas topológicas com o tamanho do vocabulário (ou seja, o número de vértices da rede). Em tarefas onde esta questão é crucial, propomos a normalização das medidas pelo valor esperado em textos aleatórios, fazendo uma analogia direta com o conceito de modularidade em redes, onde o número de inter- ou intra-conexões é normalizado pelo número esperado em redes aleatórias (FORTUNATO, 2010). Para definir esta normalização, considere a seguinte notação. X representa o valor de uma medida (Seção 2.2.4) obtida em um dado texto. $\mu(X_R)$ representa a média obtida para X em vários textos aleatorizados (ou seja, textos em que a ordem das palavras é estabelecida aleatoriamente mantendo-se as distribuições de frequências originais). Desta forma, a medida normalizada $\tilde{X}$ é definida como:

$$\tilde{X} = \frac{X}{\mu(X_R)}. \tag{4.5}$$

A normalização efetuada na equação 4.5 é útil por permitir comparar cada medida com um modelo nulo. Dessa forma, uma medida fornece informação significativa somente se seu valor $\tilde{X}$ não é próximo de 1. Além do uso da equação 4.5, avaliaremos outras formas de normalização que têm sido utilizadas nas pesquisas de redes complexas (WIJK; STAM; DAFFERTSHOFER, 2010).

4.4.2 Uso e Concepção de Novas Medidas

Além da proposição dos novos modelos, neste projeto utilizaremos medidas recentes de redes complexas para análise de fatores estilísticos de textos. Por exemplo, sabe-se que a frequência de palavras específicas representa um fator importante para a caracterização de estilos. Em redes tradicionais de co-ocorrência, a frequência pode ser medida pelo grau dos vértices. Sabendo disto, serão utilizadas extensões do conceito de grau que têm sido úteis para aperfeiçoar a caracterização de outros sistemas textuais (TRAVENçOLO; VIANA; COSTA, 2009). As extensões do conceito do grau a serem utilizadas são o grau hierárquico e a medida de acessibilidade, apresentadas na Seção 2.2.4.8. Acredita-se que a introdução destas medidas seja capaz de aperfeiçoar a tarefa de reconhecimento de autoria.

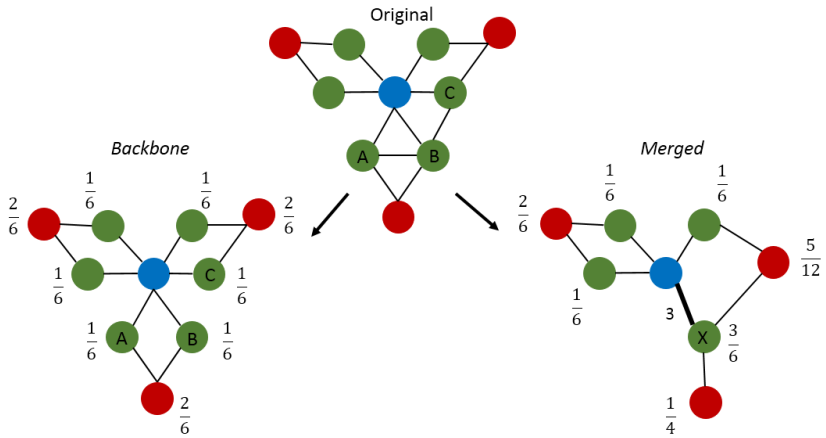
Um segundo fator que tem se mostrado importante para a tarefa de reconhecimento de autoria é a heterogeneidade da distribuição de certas quantidades textuais. Por exemplo, em (AMANCIO, 2015a), foi mostrado que a heterogeneidade da distribuição de palavras específicas é capaz de identificar estilos de forma significativa. Neste contexto, iremos estudar a heterogeneidade de acesso aos vizinhos de um vértice através de medidas de simetria local récem-propostas. Estas medidas serão empregadas nas novas modelagens desenvolvidas neste projeto. Silva *et al.* (2014) definem duas medidas que quantificam a simetria local de um nó, a simetria *backbone* e simetria *merged*. Estas medidas de simetria são versões normalizadas da acessibilidade (TRAVENçOLO; COSTA, 2008), onde o número de nós acessíveis é utilizado como fator de normalização. Para calcular essas medidas, as caminhadas aleatórias concêntricas

são utilizadas como forma de evitar transições à níveis concêntricos anteriores (SILVA *et al.*, 2014). Portanto, alterações devem ser feitas na rede de modo que as transições não utilizem arestas dentro de um mesmo nível concêntrico. Na medida de simetria *backbone*, as arestas que conectam nós pertencentes ao mesmo nível concêntrico são eliminadas. Com uma abordagem diferente, a medida *merged* considera essas arestas com custo 0 e os nós conectados por elas são colapsados. A partir dessas alterações na rede, as probabilidades de transição $P_h(i,j)$ são calculadas. A Figura 7 ilustra as alterações realizadas na rede para o cálculo de cada medida e as respectivas probabilidades de transição. Após o cálculo das probabilidades, a simetria *backbone* pode ser calculada da seguinte maneira:

$$Sb_i^{(h)} = \frac{\exp(-\sum P_h(i,j)\ln P_h(i,j))}{|\xi_i^{(h)}|} \tag{4.6}$$

onde $\xi_i^{(h)}$ representa o conjunto de nós acessíveis que estão a uma distância h do nó i. Para a simetria *merged*, o cálculo a ser realizado é semelhante, porém os pesos das arestas são considerados no cálculo das probabilidades de transição. Essas medidas de simetria conseguem capturar distintos padrões de conectividade (SILVA *et al.*, 2014). Por enriquecerem a caracterização das redes complexas, essas medidas também são capazes de aperfeiçoar a tarefa de reconhecimento de autoria. Como apresentado nos trabalhos relacionados, Amancio, Silva e Costa (2015) utilizaram essas medidas para o reconhecimento de autoria e as altas taxas de acerto confirmam o potencial da metodologia proposta.

Figura 7 – No grafo original, as arestas entre A e B e entre B e C conectam vértices pertencentes ao mesmo nível concêntrico. Essas arestas são removidas para realizar o cálculo da simetria *backbone*. Por outro lado, para o cálculo da simetria *merged*, os vértices A, B e C são colapsados em um único vértice X. Pode-se perceber que as diferentes alterações na estrutura do grafo modificam as probabilidades de transição para os vértices no segundo nível concêntrico.



Uma medida utilizada em (AMANCIO *et al.*, 2013a) e que apresenta bons resultados para a caracterização da topologia de redes é o motivo (*motif*) (MILO *et al.*, 2002). Neste trabalho, Amancio *et al.* (2013a) encontraram motivos significativos em textos, um atributo nunca utilizado para a tarefa de caracterização de estilo. Entretanto, os motivos ainda não foram utilizados para a tarefa de reconhecimento de autoria. Motivos são padrões de conectividade expressos

em pequenos blocos de construção (subgrafos). Os motivos foram descritos na Seção 2.2.4.7. O estilo de escrita de cada autor pode gerar diferentes padrões de conectividade na rede textual, por isso essa medida será utilizada neste projeto para auxilar a tarefa de reconhecimento de autoria.

4.4.3 Combinação com Técnicas Tradicionais

Os métodos de reconhecimento de autoria baseados em medidas topológicas de redes complexas têm gerado resultados significativos. No entanto, quando estes resultados são comparados com as técnicas tradicionais em Estilometria, geralmente as taxas de acertos obtidas usando apenas topologia são menores do que as obtidas com técnicas tradicionais. Por este motivo, pretendemos combinar as técnicas tradicionais com as baseadas em redes para aperfeiçoar o reconhecimento de estilo em textos. Uma possibilidade de combinação de métodos corresponde ao uso de classificadores híbridos. Neste caso, a classificação pode ser realizada levando em consideração a probabilidade de uma instância pertencer a uma dada classe considerando cada uma das abordagens. Seja $u_{ij}^{(R)}$ a probabilidade da instância i pertencer à classe j de acordo com a abordagem topológica e $u_{ij}^{(T)}$ a probabilidade da mesma instância pertencer à classe j segundo alguma técnica tradicional. Tais probabilidades podem ser combinadas através de um combinação convexa:

$$u_{ij}^{(H)} = \lambda u_{ij}^{(R)} + (1-\lambda) u_{ij}^{(T)}, \tag{4.7}$$

onde $u_{ij}^{(H)}$ representa a probabilidade da instância pertencer à class j de acordo com a combinação de classificadores e λ representa o peso dado à estratégia topológica e está restrito ao intervalo $0 \leq \lambda \leq 1$. Como na equação 4.7 temos uma probabilidade (por exemplo, $u_{ij}^{(R)}$) como resultado da classificação, será necessário utilizar classificadores *fuzzy* (KUNCHEVA, 2000). A princípio, o melhor valor a ser escolhido para λ não está claro. Por este motivo, também estudaremos formas de prever a melhor configuração para o parâmetro λ. Uma abordagem similar à proposta na equação 4.7 foi implementada com sucesso em (AMANCIO; OLIVEIRA JR; COSTA, 2015; AMANCIO, 2015b). Tal abordagem deve ser utilizada como base para este projeto.

Além da combinação fornecida na equação 4.7, existe também a possibilidade de se classificar inicialmente com a abordagem tradicional. Em seguida, a abordagem topológica poderia ser utilizada para classificar as instâncias cujas classes inferidas pelo método tradicional apresentaram alto grau de incerteza. Formas complementares de combinação de classificadores supervisionados também serão investigadas neste projeto.

A seguir, listamos alguns exemplos de atributos que podem ser utilizados para compor a classificação baseada em técnicas estatísticas tradicionais.

- **Palavras funcionais**: a frequência de palavras funcionais específicas representa uma das principais marcas estilísticas utilizadas para realizar o reconhecimento de autoria. Portanto, a frequência de tais palavras pode ser utilizada como sendo um atributo relevante para a abordagem topológica.

- **Caracteres**: a frequência de caracteres (ou bigramas) específicos também representa um fator relevante para a classificação estilométrica de textos. A consideração deste tipo de atributo nos classificadores tradicionais também é direta, já que a frequência de cada caractere (ou bigrama) pode ser utilizada como atributo. A vantagem do uso deste tipo de atributo é que ele não necessita de nenhuma informação textual profunda. Além disso, existe a possibilidade de criar redes de bigramas de caracteres.

- **Estrutura textual**: em vários estudos linguísticos comprova-se a existência de certos padrões organizacionais de textos. Sabe-se, por exemplo, que correlações de longo alcance ocorrem devido ao mapeamento de ideias de forma linear (ALTMANN; CRISTADORO; ESPOSTI, 2012). Outra descoberta relevante refere-se à entropia da distribuição de palavras, que comprova que certas palavras possuem distribuições espaciais de acordo com a função exercida (AMANCIO *et al.*, 2013a).

4.4.4 Aplicações Adicionais

Além das atividades apresentadas anteriormente, a concepção de novos métodos, medidas e combinações de classificadores podem também auxiliar o desenvolvimento de tarefas correlatas à tarefa de reconhecimento de autoria. A seguir, listamos algumas dessas tarefas:

- **Inconsistência estilística:** quando duas ou mais pessoas escrevem um documento, sendo cada um responsável por uma parte do texto, podem surgir inconsistências estilísticas devido à múltipla autoria, o que pode gerar textos com baixa coerência. A identificação de inconsistência estilística também pode ser utilizada para detectar fraudes em documentos que deveriam ser escritos por um único autor. Este problema pode ser identificado, por exemplo, dividindo-se o texto em pedaços e transformado cada pedaço em uma rede. Este tipo de divisão mostrou-se viável no estudo realizado em (AMANCIO, 2015c).

- **Detecção de plágio:** as técnicas de detecção de plágio tradicionais baseiam-se principalmente no conteúdo semântico dos textos, de uma maneira similar às medidas de avaliação de qualidade de traduções automáticas (PAPINENI *et al.*, 2002). As técnicas baseadas em análise de topologia de redes complexas poderiam ser úteis, por exemplo, para identificar tentativas de mascaramento de plágios. Se uma ou mais palavras da fonte original são substituídas por sinônimos, o *overlapping* com a fonte diminui. Este tipo de fraude não afetaria nossa técnica pois uma mudança global de rótulos de palavras específicas não afetaria a organização dos textos.

- **Identificação de fraudes em reconhecimento**: estudos recentes mostram que os métodos de detecção de autoria tradicionais não são robustos à ataques de autores que tentam mascarar estilos (BRENNAN; GREENSTADT, 2009). Neste sentido, pretendemos analisar a robustez das redes complexas com relação a estes ataques.

- **Análise de fenômenos cognitivos**: A teoria de redes complexas têm sido utilizada como ferramenta complementar para analisar diversos fenômenos cognitivos, como a doença de Alzheimer, transtornos bipolares e a esquizofrenia (HIRST; FENG, 2012; BORGE-HOLTHOEFER; MORENO; ARENAS, 2011; MOTA *et al.*, 2014). Pretende-se aplicar os modelos desenvolvidos nesse trabalho nas bases de dados utilizadas nesses artigos. Desse modo, é possível verificar se as novas modelagens melhoram a performance das análises realizadas e, assim, auxiliar diversos trabalhos dedicados a distinguir diversos pacientes com ou sem a patologia.

4.5 Resultados Preliminares

As atividades relacionadas a este projeto de mestrado que foram realizadas são descritas nesta seção. Na Seção 4.5.1 são exibidos os resultados relacionados com a extensão da modelagem. Em seguida, na Seção 4.5.2 são apresentados os resultados ao aplicar a extração de motivos para a tarefa de reconhecimento de autoria. Por fim, na Seção 4.5.3 são apresentadas algumas técnicas utilizadas para a adição de *links* entre as palavras.

4.5.1 Extensão da Modelagem

O modelo tradicional de co-ocorrência captura propriedades estilísticas ao conectar apenas as palavras adjacentes em um texto. Entretanto, este modelo é falho ao não capturar possíveis relações entre palavras mais distantes. Nesse contexto, com o objetivo de considerar possíveis *links* entre palavras não-adjacentes, a modelagem denominada *Further Neighborhoods* foi desenvolvida. Nessa modelagem, são conectados todos os pares de palavras que estão separados por no máximo $W-1$ palavras adjacentes, onde $W = 1, 2, 3$. Vale a pena ressaltar que $W = 1$ corresponde a modelagem tradicional de co-ocorrência.

Em uma outra modelagem, todas as palavras pertencentes a uma mesma sentença são conectadas. Para essa análise, uma sentença foi considerada como um trecho de texto separado por vírgula, ponto de exclamação ou de interrogação. Não foi encontrada na literatura nenhuma outra tentativa de modelar os textos dessa maneira para extrair informações estilísticas. Essa modelagem foi denominada *Sentence based*.

As medidas apresentadas na Seção 2.2.4 foram extraídas das diferentes modelagens apresentadas, com exceção dos motivos que não foram utilizados nessa análise. Os quatro algoritmos de reconhecimento de padrão foram utilizados. O desempenho de cada algoritmo é dado pela porcentagem de atribuições corretas ao testar a autoria de cada livro individualmente. Os resultados obtidos são apresentados na Tabela 3. Para cada algoritmo, a primeira coluna indica os resultados obtidos quando todos os atributos de entrada são utilizados (TA). Por outro lado, a segunda coluna indica os resultados após realizar uma seleção de atributos (baseada em filtro) e utilizar a melhor combinação dos atributos de entrada (SA).

Tabela 3 – Porcentagem de livros corretamente classificados para cada método de reconhecimento de padrão ao testar as novas metodologias propostas. Para cada método, dois conjuntos de atributos foram utilizados: (i) todos os atributos (TA); e (ii) os atributos obtidos após a seleção de atributos (SA)

Modelagem	**C4.5**		**kNN**		**SVM**		**Naive Bayes**	
	TA	SA	TA	SA	TA	SA	TA	SA
$W = 1$	27.5%	50.0%	50.0%	50.0%	52.5%	37.5%	47.5%	47.5%
$W = 2$	45.0%	47.5%	62.5%	67.5%	55.0%	52.5%	50.0%	60.0%
$W = 3$	45.0%	55.0%	55.0%	60.0%	55.0%	50.0%	50.0%	62.5%
Sentence based	32.5%	27.5%	42.5%	32.5%	55.0%	35.0%	40.0%	30.0%

A partir dos resultados da Tabela 3, podemos perceber que ao utilizar todos os atributos, as modelagens *Further Neighborhoods* com $W = 2$ e $W = 3$ resultaram em maiores taxas de acerto do que a modelagem de co-ocorrência tradicional ($W = 1$). Com a seleção de atributos, as modelagens $W = 2$ e $W = 3$ também resultaram em maiores taxas de acerto em 7 dos 8 casos. Esses resultados confirmam a hipótese inicial de que a conexão de palavras em um maior contexto melhora a performance da tarefa de reconhecimento de autoria. A maior taxa de acerto passou de 52.5% (modelagem tradicional, algoritmo SVM e todos os atributos) para 67.5% (modelagem $W = 2$, algoritmo kNN e seleção de atributos).

Entretanto, a modelagem *Sentence based* não proporcionou aumento de performance em todos os casos analisados. Uma vez que os textos podem apresentar sentenças muito longas, essa modelagem acaba adicionando conexões que não são relevantes para a caracterização de escrita do autor.

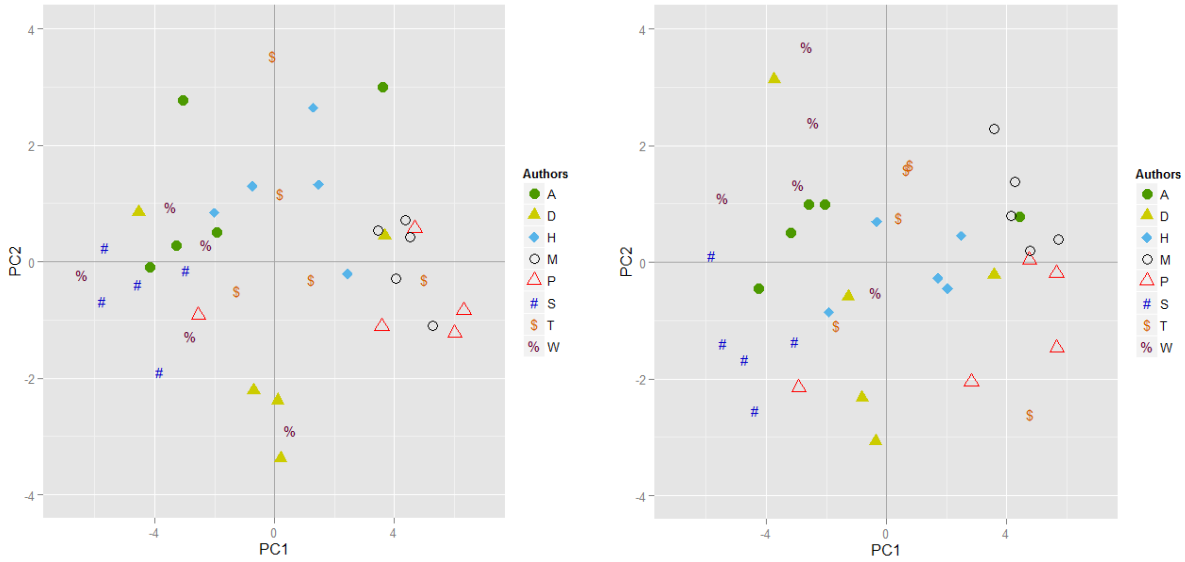
Uma outra forma de visualizar a caracterização fornecida pelas novas modelagens é através da análise de componentes principais (ou *Principal Component Analysis*, PCA). A análise de componentes principais é uma transformação matemática que mapeia um conjunto de exemplos (observações) de variáveis possivelmente correlacionadas em variáveis fracamente correlacionadas. Sabe-se, por exemplo, que algumas das medidas extraídas apresentam correlação com o grau. A Figura 8 apresenta a análise PCA quando todos os atributos de entrada são utilizados, nas diferentes modelagens. A Figura 9 apresenta os resultados da análise PCA para apenas os atributos selecionados. Cada um dos oito autores é representado por um símbolo e uma cor nos gráficos. Cada um dos elementos no gráfico representa um livro distinto.

A partir das Figuras 8 e 9 é possível perceber que, para alguns autores, os livros de sua autoria ficam mais próximos uns dos outros nas modelagens propostas nesse trabalho.

4.5.2 Extração de Motivos

Para a extração dos motivos, foram inicialmente consideradas duas redes para cada livro, uma não direcionada e outra direcionada. Na modelagem de co-ocorrência tradicional, a direção das arestas tem origem na primeira palavra e destino na palavra seguinte. Cada livro foi caracterizado por diversos conjuntos de atributos. Os conjuntos de atributos apresentavam

(a) *Further Neighborhoods* $W = 1$ (b) *Further Neighborhoods* $W = 2$



(c) *Further Neighborhoods* $W = 3$ (d) *Sentence Based*

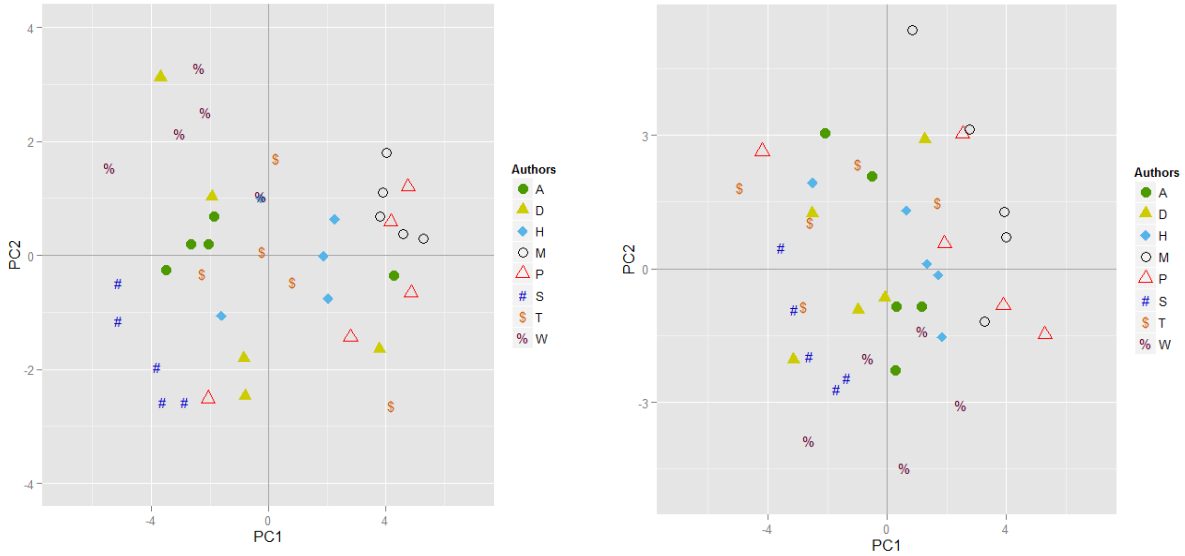
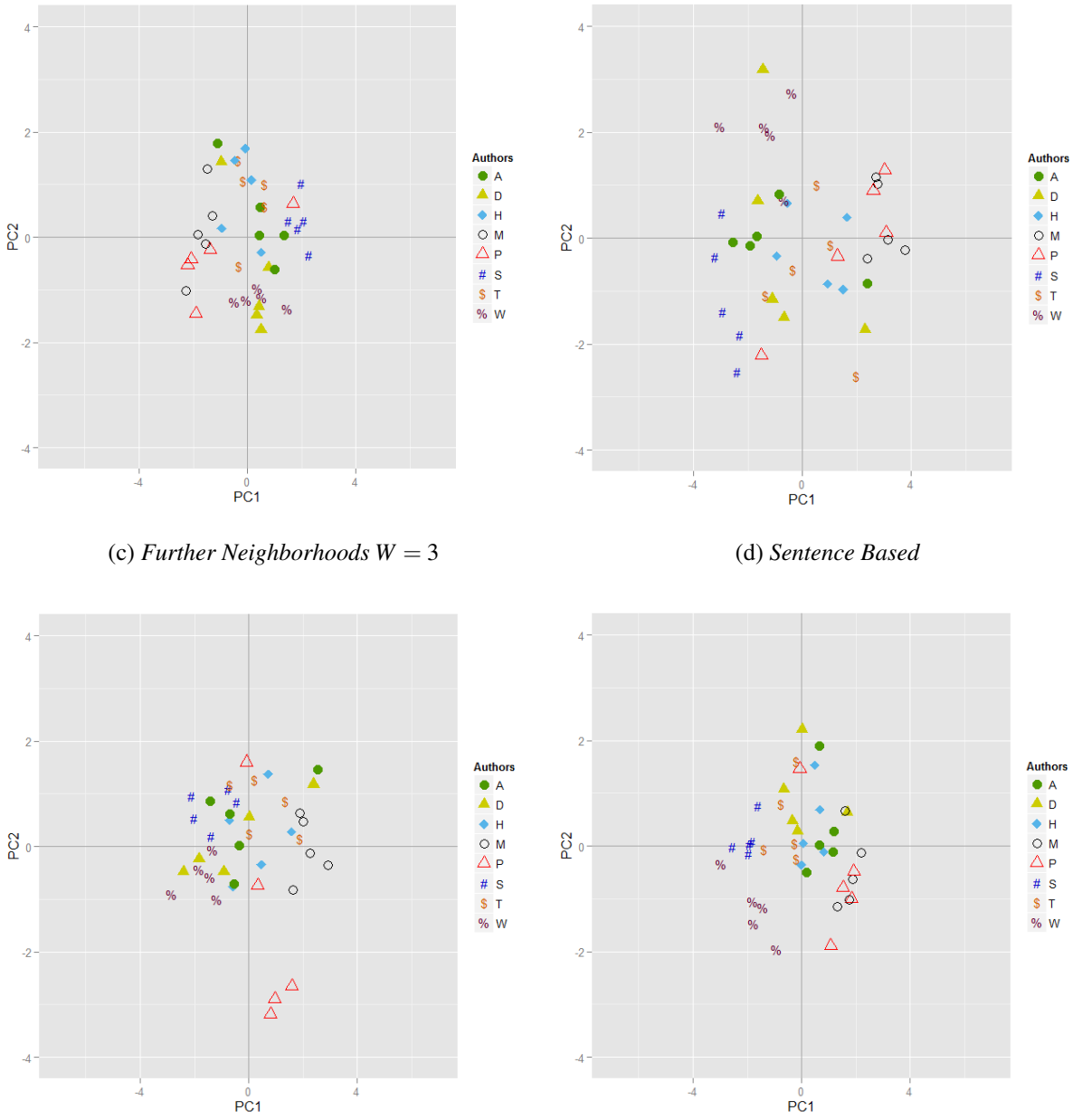


Figura 8 – Análise do componente principal quando todos os atributos são utilizados. As letras na legenda representam os seguintes autores, Arthur Conan Doyle (A), Bram Stoker (S), Charles Dickens (D), Edgar Allan Poe (P), Hector Hugh Munro (M), Pelham Grenville Wodehouse (W), Thomas Hardy (H), William Makepeace Thackeray (T).

tamanho 13 (número de motivos direcionados de tamanho $n = 3$) e 8 (número de motivos não direcionados de tamanho $n = 3$ e $n = 4$). A atribuição de autoria foi realizada utilizando os quatro algoritmos de aprendizado supervisionado descritos na Seção 4.3.4. O desempenho de cada algoritmo é dado pela porcentagem de atribuições corretas ao testar a autoria de cada livro individualmente.

As modelagens *Further Neighborhoods* definidas anteriormente foram utilizadas na análise de extração de motivos, com $W = 1, 2, 3$. Os resultados obtidos são apresentados nas

(a) *Further Neighborhoods* $W = 1$ (b) *Further Neighborhoods* $W = 2$



(c) *Further Neighborhoods* $W = 3$ (d) *Sentence Based*

Figura 9 – Análise do componente principal após a seleção de atributos. As letras na legenda representam os seguintes autores, Arthur Conan Doyle (A), Bram Stoker (S), Charles Dickens (D), Edgar Allan Poe (P), Hector Hugh Munro (M), Pelham Grenville Wodehouse (W), Thomas Hardy (H), William Makepeace Thackeray (T)

Tabelas 4 e 5. A Tabela 4 apresenta as taxas de acerto obtidas ao utilizar os 13 motivos direcionados. Os valores obtidos variam entre 27.5% e 55%. A Tabela 5 apresenta as taxas de acerto obtidas ao utilizar os 8 motivos não direcionados. Os valores obtidos variam entre 27.5% e 45%.

A partir dos resultados da Tabela 4 e Tabela 5, podemos perceber que as taxas de acerto obtidas apenas utilizando os motivos são altas (55% nas redes direcionadas e 45% nas redes não

Tabela 4 – Porcentagem de livros corretamente classificados para os quatro algoritmos de aprendizado supervisionado utilizando todos os 13 motivos direcionados

Modelagem	**C4.5**	**kNN**	**SVM**	**Naive Bayes**
$W = 1$	30.0%	32.5%	27.5%	27.5%
$W = 2$	55.0%	40.0%	35.0%	30.0%
$W = 3$	42.5%	35.0%	32.5%	32.5%

Tabela 5 – Porcentagem de livros corretamente classificados para os quatro algoritmos de aprendizado supervisionado utilizando todos os 8 motivos não direcionados

Modelagem	**C4.5**	**kNN**	**SVM**	**Naive Bayes**
$W = 1$	30.0%	35.0%	27.5%	35.0%
$W = 2$	42.5%	40.0%	37.5%	42.5%
$W = 3$	40.0%	40.0%	40.0%	45.0%

direcionadas), confirmando a capacidade dos motivos de capturar o estilo de escrita dos diversos autores. Além disso, os resultados obtidos são consideravelmente maiores que a *baseline* para este problema, 12.5%. Esse valor representa a probabilidade de um livro X pertencer a um dos 8 autores, uma vez que cada autor possui a mesma quantidade de livros. Além disso, as maiores taxas obtidas foram para as modelagens propostas neste trabalho ($W = 2$ e $W = 3$), confirmando novamente a capacidade de modelagens de maior alcance em caracterizar o estilo de escrita dos autores.

4.5.3 Novas Abordagens para Adicionar Conexões

Por fim, foram analisadas duas novas abordagens para adicionar novas conexões entre palavras. A primeira delas consistiu em conectar todas as palavras não adjacentes que possuíam alguma relação sintática. O objetivo principal em adicionar conexões sintáticas (como arestas complementares) é baseado no fato de que palavras sintaticamente relacionadas também possuem uma relação semântica (ROMACKER; MARKERT; HAHN, 1999). Como mencionado anteriormente, redes sintáticas apresentam características em comum com diversos sistemas complexos. Mais especificamente, redes sintáticas têm sido aplicadas com sucesso para a identificação de conceitos para a tarefa de sumarização automática (AMANCIO *et al.*, 2012).

O *parser* disponibilizado pela Universidade de Stanford (CHEN; MANNING, 2014) foi utilizado para extrair as relações sintáticas entre todas as sentenças de cada livro. Um *parser* é um programa que extrai a estrutura gramatical das sentenças, por exemplo, quais palavras estão relacionadas e quais são o sujeito ou objeto de um verbo. As relações extraídas foram combinadas com as diferentes abordagens denominadas *Further Neighborhoods*. Por exemplo, ao criar a rede de co-ocorrência utilizando a abordagem tradicional ($W = 1$), todas as palavras adjacentes são conectadas. Além disso, nessa nova abordagem, foram conectadas as palavras w_i e w_j de cada par sintático (w_i, w_j) obtido com o parser. As medidas apresentadas na Seção 2.2.4 foram extraídas das redes, com exceção dos motivos que não foram utilizados nessa análise. Os quatro

algoritmos de reconhecimento de padrão foram utilizados. O desempenho de cada algoritmo é dado pela porcentagem de atribuições corretas ao testar a autoria de cada livro individualmente. Os resultados obtidos com essa nova abordagem são exibidos na Tabela 6. Para cada algoritmo, a primeira coluna indica os resultados obtidos quando todos os atributos de entrada são utilizados (TA). A segunda coluna indica os resultados após realizar uma seleção de atributos e utilizar a melhor combinação dos atributos de entrada (SA).

Tabela 6 – Porcentagem de livros corretamente classificados para cada método de reconhecimento de padrão ao adicionar *links* sintáticos. Para cada método, dois conjuntos de atributos foram utilizados: (i) todos os atributos (TA); e (ii) os atributos obtidos após a seleção de atributos (SA)

Modelagem	**C4.5**		**kNN**		**SVM**		**Naive Bayes**	
	TA	SA	TA	SA	TA	SA	TA	SA
$W = 1$	22.5%	37.5%	27.5%	30.0%	20.0%	27.5%	20.0%	27.5%
$W = 2$	35.0%	37.5%	25.0%	42.5%	30.0%	32.5%	27.5%	35.0%
$W = 3$	27.5%	30.0%	27.5%	45.0%	37.5%	30.0%	35.0%	42.5%

Ao comparar os resultados obtidos na Tabela 6 com os obtidos na Seção 4.5.1, podemos perceber que a adição de *links* sintáticos piorou a performance de classificação. Entretanto, na maioria dos casos, as modelagens de *Further Neighborhoods* com $W = 2$ e $W = 3$ apresentaram melhores resultados do que a modelagem de co-ocorrência tradicional com a adição dos *links* sintáticos.

A segunda abordagem, baseada no trabalho de Martinez-Romo *et al.* (2011), consistiu na conexão de palavras relevantes em um contexto. Nessa abordagem, são considerados todos os *links* relevantes dentro de um parágrafo. A conexão de todas as palavras pertencentes a um mesmo parágrafo iria gerar grandes comunidades na rede e poderia adicionar mais ruídos do que informações relevantes (como é o caso da abordagem *Sentence Based* apresentada anteriormente). Para resolver esse problema, primeiramente são adicionadas todas as palavras adjacentes, ou seja, é utilizada a abordagem *Further Neighborhoods* com $W = 1$. Depois, procura-se todos os *links* relevantes entre palavras não adjacentes em um mesmo parágrafo. Mais especificamente, um *link* é considerado relevante se a probabilidade de acontecer ao acaso em uma rede aleatória é baixa. Desse modo, duas palavras v_i e v_j serão conectadas se a frequência de ocorrência f_{ij} for maior do que a frequência esperada em um modelo aleatório.

Seja $p(k)$ a probabilidade de duas palavras aparecem no mesmo contexto (parágrafo) k vezes em um texto aleatório. Para calcular $p(k)$, considere que n_1 e n_2 são o número de partições (no caso, parágrafos) distintas nas quais v_i e v_j aparecem, respectivamente. O valor de $p(k)$ pode ser calculado através da seguinte expressão:

$$p(k) = \frac{(N;k,n_1-k,n_2-k)}{(N;n_1)(N;n_2)}, \text{ onde } (x;y_1,\ldots,y_n) \equiv \frac{x!}{x_1!\ldots x_n!}\frac{1}{(x-y_1-\ldots y_n)!}.$$

A expressão acima pode ser reescrita utilizando a notação $\{a\}_b$, definida como

$$\{a\}_b \equiv \prod_{i=0}^{b-1}(a-i) \tag{4.8}$$

onde $a \geq b$. Nesse caso, a probabilidade $p(k)$ pode ser reescrita como:

$$\begin{aligned} p(k) &= \frac{\{n_1\}_k\{n_2\}_k\{N-n_1\}_{n_2-k}}{\{N\}_{n_2}\{k\}_k} = \frac{\{n_1\}_k\{n_2\}_k\{N-n_1\}_{n_2-k}}{\{N\}_{n_2-k}\{N-n_2+k\}_k\{k\}_k} \\ &= \prod_{j=0}^{n_2-k-1}\left[\frac{N-j-n_1}{N-j}\right]\prod_{j=0}^{k-1}\frac{(n_1-j)(n_2-j)}{(N-n_2+k-j)(k-j)}. \end{aligned}$$

Se, em um dado texto, o número de co-ocorrências de duas palavras é r, o p-value associado a r pode ser calculado como

$$p(k \geq r) = \sum_{k \geq r} p(r) = \sum_{k \geq r} \prod_{j=0}^{n_2-k-1}\left(1-\frac{n_1}{N-j}\right)\prod_{j=0}^{k-1}\frac{(n_1-j)(n_2-j)}{(N-n_2+k-j)(k-j)}. \quad (4.9)$$

Desse modo, o valor de $p(k \geq r)$ definido na Equação 4.9 pode ser utilizado para estabelecer *links* entre palavras com frequência significativa. O *threshold* de $p-value$ utilizado nesse trabalho foi 0.000001. Ou seja, qualquer par cujo valor de p-value $p(k \geq r) < 0.000001$ foi considerado relevante.

A análise desenvolvida nesse trabalho utilizou 2 parâmetros diferentes, o número máximo de arestas relevantes adicionais (M_e) e o tamanho mínimo de um parágrafo (m_p). Para o parâmetro m_p, foram utilizados os valores 20, 50 e 100 *tokens*. O número máximo de arestas significativas, M_e, representou de 5% a 200% do número total de arestas da rede obtidas após a modelagem tradicional ($W = 1$). Por exemplo, suponha uma rede de adjacência tradicional com mil vértices, $M_e = 5\%$ e $m_p = 100$. Nessa rede, parágrafos com tamanhos menores que 100 palavras são combinados com os parágrafos seguintes até que o tamanho mínimo de 100 palavras seja alcançado. Como $M_e = 5\%$, no máximo 50 arestas mais significativas serão adicionadas (ou seja, os 50 pares com menores valores de $p-value$).

Após a criação das redes, as medidas apresentadas na Seção 2.2.4 foram extraídas, com exceção dos motivos. Os quatro algoritmos de reconhecimento de padrão foram utilizados. O desempenho de cada algoritmo é dado pela porcentagem de atribuições corretas ao testar a autoria de cada livro individualmente. Como foram utilizados vários valores para o parâmetro (m_p), apenas os resultados obtidos para o tamanho mínimo de parágrafo $m_p = 20$ são apresentados na Tabela 7. Esse foi o tamanho de parágrafo responsável pelos melhores resultados. Para cada algoritmo, a primeira coluna indica os resultados obtidos quando todos os atributos de entrada são utilizados. Por outro lado, a segunda coluna indica os resultados após realizar uma seleção de atributos (baseada em filtro) e utilizar a melhor combinação dos atributos de entrada.

Com esse último resultado, pode-se perceber que é possível melhorar a performance da modelagem tradicional ($W = 1$) apenas adicionando 10% de arestas referentes aos *links* mais significativos. Nesse caso, a taxa de acerto passa de 50% (Tabela 3 para $W = 1$ e algoritmo C4.5) para 60% (algoritmo kNN e seleção de atributos). A maior taxa de acerto obtida, 65%, ocorre quando 200% do total de arestas da rede é adicionado, para o algoritmo kNN com seleção de atributos.

Tabela 7 – Porcentagem de livros corretamente classificados para cada método de reconhecimento de padrão ao adicionar *links* significativos, com o tamanho mínimo de parágrafo igual a 20. Para cada método de reconhecimento de padrão, dois conjuntos de atributos foram utilizados: (i) todos os atributos (TA); e (ii) os atributos obtidos após a seleção de atributos (SA)

M_e	**C4.5**		**kNN**		**SVM**		**Naive Bayes**	
	TA	SA	TA	SA	TA	SA	TA	SA
5%	42.5%	50.0%	37.5%	57.5%	55.0%	52.5%	35.0%	55.0%
10%	47.5%	45.0%	37.5%	60.0%	50.0%	55.0%	32.5%	50.0%
20%	30.0%	35.0%	37.5%	60.0%	50.0%	42.5%	47.5%	60.0%
50%	35.0%	35.0%	37.5%	52.5%	47.5%	35.0%	35.0%	57.5%
100%	27.5%	27.5%	30.0%	40.0%	40.0%	40.0%	35.0%	42.5%
150%	30.0%	45.0%	35.0%	40.0%	42.5%	42.5%	37.5%	42.5%
200%	32.5%	42.5%	35.0%	65.0%	45.0%	35.0%	45.0%	50.0%

4.6 Cronograma Previsto

As seguintes atividades estão previstas para os 24 meses de execução deste trabalho de mestrado, com início em fevereiro de 2015 (data da matrícula da candidata como aluna regular de mestrado do Programa CCMC do ICMC/USP) e defesa prevista para março de 2017. O cronograma prevê a obtenção dos créditos de aula obrigatórios do programa, a realização do projeto de pesquisa, a realização de um estágio de pesquisa no exterior e os exames de qualificação e defesa.

Tabela 8 – Cronograma de Atividades.

	2015		2016		2017
Atividade	1º S.	2º S.	1º S.	2º S.	1º S.
1					
2					
3					
4					
5					
6					
7					
8					

1. **Disciplinas da pós-graduação:** Esta tarefa foi concluída no segundo semestre como aluna regular do programa. Os 51 créditos de disciplinas requeridos para obtenção do título de mestre em Ciência da Computação e Matemática Computacional foram concluídos, representando oito disciplinas do programa. Em todas as disciplinas cursadas (Preparação Pedagógica, Metodologia de Pesquisa Científica em Computação, Projeto de Algoritmos, Tópicos em Computação e Matemática Computacional I, Introdução ao Aprendizado de Máquina, Práticas em Preparação Pedagógica para a Computação, Mineração de Redes Complexas, Processos Dinâmicos em Redes Complexas), a aluna foi aprovada com conceito máximo 'A'.

2. **Revisão bibliográfica:** Revisão do estado da arte em redes complexas aplicadas à diversas tarefas de processamento de línguas naturais, principalmente para o reconhecimento de autoria. Em uma segunda etapa desse projeto, inclui-se a revisão do estado da arte em técnicas tradicionais de reconhecimento de autoria;

3. **Aplicabilidade e Obtenção de dados:** Levantamento de possíveis aplicações que podem ser melhoradas com a utilização dos modelos alternativos a serem propostos. Além do estudo de técnicas e ferramentas tradicionais em reconhecimento de autoria;

4. **Qualificação:** Preparação e apresentação do exame de qualificação, conforme determinam as normas do Programa de Pós-graduação;

5. **Estágio de Pesquisa no Exterior:** A aluna irá realizar um estágio de pesquisa no exterior com duração de 6 meses, início em 01/03/2016 e término em 31/08/2016. O supervisor no exterior será o professor Graeme Hirst, da Universidade de Toronto, Canadá. O professor Graeme Hirst possui uma ampla experiência de pesquisa na área de Processamento de Linguagens Naturais, incluindo os conceitos de reconhecimento de autoria e análises estilísticas;

6. **Desenvolvimento e Implementação:** Desenvolvimento e implementação da proposta do projeto;

7. **Avaliação de Resultados:** Avaliação dos resultados e comparação com ferramentas e técnicas existentes;

8. **Trabalhos Científicos:** Redação de artigos científicos, relatórios, participação de congressos e escrita da dissertação final de mestrado bem como sua apresentação para uma banca avaliadora.

4.7 Considerações Finais

Neste capítulo foram apresentados o objetivo deste trabalho e a metodologia proposta para a obtenção dos resultados esperados. Esta metodologia busca desenvolver novos modelos para a tarefa de reconhecimento de autoria através do uso de redes complexas, além de adaptar o modelo de adjacência tradicional. Também foram apresentadas as principais tarefas planejadas para esse projeto e alguns resultados iniciais. Muitos dos resultados obtidos foram satisfatórios, como a extensão da modelagem de adjacência, que proporcionou aumento de performance em quase todos os casos analisados. Alguns resultados foram inesperados, como as taxas de acerto consideravelmente altas obtidas com a extração de motivos. Pretende-se analisar a combinação dos atributos de redes complexas com os motivos. O último resultado satisfatório foi a constatação que, ao adicionar *links* relevantes, é possível melhorar a modelagem tradicional, o que não ocorreu com a adição de *links* sintáticos.

Acredita-se que o trabalho proposto soluciona problemas em aberto da área de reconhecimento de autoria. Além disso, a adaptação dos modelos de adjacência pode trazer benefícios à outras aplicações que utilizam redes complexas para a análise da linguagem. É importante ressaltar que a maioria das medidas de redes complexas utilizadas não foram implementadas durante o decorrer desse projeto. A maioria dessas medidas está disponível no pacote Igraph[2].

[2] http://igraph.org/c/

REFERÊNCIAS

ABBASI, A.; CHEN, H. Applying authorship analysis to extremist group web forum messages. **IEEE Intelligent Systems**, IEEE Educational Activities Department, v. 20, n. 5, 2005. Citado 3 vezes nas páginas 18, 32 e 54.

AHA, D. W.; KIBLER, D.; ALBERT; K., M. Instance-based learning algorithms. **Mach. Learn.**, Kluwer Academic Publishers, v. 6, n. 1, jan. 1991. Citado na página 58.

ALBERT, R.; BARABáSI, A.-l. Statistical mechanics of complex networks. **Rev. Mod. Phys**, 2002. Citado 2 vezes nas páginas 17 e 22.

ALTMANN, E. G.; CRISTADORO, G.; ESPOSTI, M. D. On the origin of long-range correlations in texts. **Proceedings of the National Academy of Sciences**, v. 109, n. 29, p. 11582–11587, 2012. Citado na página 63.

ALVAREZ-LACALLE, E.; DOROW, B.; ECKMANN, J. P.; MOSES, E. Hierarchical structures induce long-range dynamical correlations in written texts. **PNAS**, v. 103, n. 21, p. 7956–7961, maio 2006. Citado na página 56.

AMANCIO, D. R. Authorship recognition via fluctuation analysis of network topology and word intermittency. **Journal of Statistical Mechanics: Theory and Experiment**, 2015. Citado 5 vezes nas páginas 14, 45, 46, 51 e 60.

____. A complex network approach to stylometry. **PloS One**, 2015. Citado 5 vezes nas páginas 14, 39, 46, 51 e 62.

____. Probing the topological properties of complex networks modeling short written texts. **PLoS ONE**, 2015. Citado 6 vezes nas páginas 14, 43, 44, 51, 57 e 63.

AMANCIO, D. R.; ALTMANN, E. G.; OLIVEIRA JR, O. N.; COSTA, L. F. Comparing intermittency and network measurements of words and their dependence on authorship. **New Journal of Physics**, v. 13, n. 12, p. 123024, 2011. Citado 10 vezes nas páginas 14, 18, 28, 29, 31, 41, 42, 51, 54 e 56.

AMANCIO, D. R.; ALTMANN, E. G.; RYBSKI, D.; OLIVEIRA JR, O. N.; COSTA, L. F. Probing the statistical properties of unknown texts: Application to the voynich manuscript. **PLoS ONE**, Public Library of Science, v. 8, p. e67310, 07 2013. Citado 9 vezes nas páginas 14, 18, 29, 40, 48, 54, 59, 61 e 63.

AMANCIO, D. R.; ALUISIO, S. M.; OLIVEIRA JR, O. N.; DA, L. Complex networks analysis of language complexity. **EPL (Europhysics Letters)**, v. 100, n. 5, p. 58002+, fev. 2013. Citado na página 40.

AMANCIO, D. R.; NUNES, M. G. V.; OLIVEIRA JR, O. N.; COSTA, L. F. Extractive summarization using complex networks and syntactic dependency. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 4, p. 1855 – 1864, 2012. ISSN 0378-4371. Citado 6 vezes nas páginas 17, 18, 31, 39, 57 e 68.

AMANCIO, D. R.; OLIVEIRA JR, O. N.; COSTA, L. F. Identification of literary movements using complex networks to represent texts. **New Journal of Physics**, v. 14, n. 4, p. 043029, 2012. Citado 6 vezes nas páginas 14, 18, 40, 47, 55 e 57.

____. Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts. **Physica A: Statistical Mechanics and its Applications**, 2012. Citado 4 vezes nas páginas 14, 43, 51 e 57.

AMANCIO, D. R.; OLIVEIRA JR, O. N.; COSTA, L. F. da. Topological-collaborative approach for disambiguating authors' names in collaborative networks. **Scientometrics**, Springer Netherlands, v. 102, n. 1, p. 465–485, 2015. ISSN 0138-9130. Citado na página 62.

AMANCIO, D. R.; SILVA, F. N.; COSTA, L. F. Concentric network symmetry grasps authors' styles in word adjacency networks. 2015. Citado 4 vezes nas páginas 14, 44, 51 e 61.

ANTIQUEIRA, L.; PARDO, T. A. S.; NUNES, M. G. V.; OLIVEIRA JR, O. N.; COSTA, L. F. Some issues on complex networks for author characterization. In: **Fourth Workshop in Information and Human Language Technology (TIL'06) in the Proceedings of International Joint Conference IBERAMIA-SBIA-SBRN**. Ribeirão Preto, Brazil: ICMC-USP, 2006. Citado 5 vezes nas páginas 14, 18, 40, 41 e 51.

ARRUDA, H. F. de; COSTA, L. F.; AMANCIO, D. R. Classifying informative and imaginative prose using complex networks. 2015. Citado 2 vezes nas páginas 14 e 49.

BAAYEN, H.; HALTEREN, H. van; TWEEDIE, F. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. **Lit Linguist Computing**, v. 11, n. 3, p. 121–132, set. 1996. Citado na página 34.

BARABASI, A.-L. **Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life**. [S.l.]: Plume Books, 2003. Paperback. Citado na página 21.

BARABÁSI, A.-L. **Network Science**. [s.n.], 2014. Disponível em: <http://barabasi.com/networksciencebook/>. Citado na página 21.

BARABASI, A.-L.; ALBERT, R. Emergence of scaling in random networks. **Science**, v. 286, n. 5439, p. 509–512, 1999. Citado 2 vezes nas páginas 24 e 25.

BECKNER, C.; BLYTHE, R.; BYBEE, J.; CHRISTIANSEN, M. H.; CROFT, W.; ELLIS, N. C.; HOLLAND, J.; KE, J.; LARSEN-FREEMAN, D.; SCHOENEMANN, T. Language is a complex adaptive system: Position paper. **Language Learning**, v. 59, 2009. Citado na página 24.

BIEMANN, C. **Structure Discovery in Natural Language**. Heidelberg: Springer, 2012. (Theory and Applications of Natural Language Processing). ISSN 2192-032X. ISBN 978-3-642-25922-7. Citado na página 25.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural Language Processing with Python**. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2009. ISBN 0596516495, 9780596516499. Citado na página 56.

BOCCALETTI, S.; LATORA, V.; MORENO, Y.; CHAVEZ, M.; HWANG, D.-U. Complex networks : Structure and dynamics. **Phys. Rep.**, v. 424, n. 4-5, p. 175–308, Fervier 2006. Citado 3 vezes nas páginas 17, 21 e 28.

BORGE-HOLTHOEFER, J.; MORENO, Y.; ARENAS, A. Modeling abnormal priming in alzheimer's patients with a free association network. **PLoS ONE**, Public Library of Science, v. 6, 08 2011. Citado na página 64.

BRENNAN, M. R.; GREENSTADT, R. Practical attacks against authorship recognition techniques. In: HAIGH, K. Z.; RYCHTYCKYJ, N. (Ed.). **IAAI**. [S.l.]: AAAI, 2009. Citado 2 vezes nas páginas 18 e 63.

BURROWS, J. 'delta': a measure of stylistic difference and a guide to likely authorship. **Literary and Linguistic Computing**, v. 17, n. 3, p. 267–287, 2002. Citado na página 33.

BURROWS, J. F. Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. **Literary and Linguistic Computing**, v. 2, p. 61–70, 1987. Citado na página 33.

CANCHO, R. Ferrer i; SOLé, R. V. The small world of human language. **Proceedings of The Royal Society of London. Series B, Biological Sciences**, v. 268, p. 2261–2266, 2001. Citado 5 vezes nas páginas 17, 25, 28, 40 e 56.

CANCHO, R. Ferrer i; SOLé, R. V. Least Effort and the Origins of Scaling in the Human Language. **Proceedings of the National Academy of Science (USA)**, v. 100, 2003. Citado na página 40.

CANCHO, R. Ferrer i; SOLÉ, R. V.; KÖHLER, R. Patterns in syntactic dependency networks. **Phys. Rev. E**, American Physical Society, v. 69, p. 051915, May 2004. Citado 5 vezes nas páginas 17, 25, 39, 40 e 56.

CAROLE, E. C. Who's at the keyboard: Authorship attribution in digital evidence investigations. In: **Presented at the 8th Biennial Conference on Forensic Linguistics/Language and Law**. [S.l.: s.n.], 2005. Citado na página 33.

CHASKI, C. E. Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. **International Journal of Digital Evidence**, v. 4, 2005. Citado na página 54.

CHEN, D.; MANNING, C. A Fast and Accurate Dependency Parser using Neural Networks. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 740–750. Citado na página 68.

CLAUSET, A.; SHALIZI, C. R.; NEWMAN, M. E. J. Power-law distributions in empirical data. **SIAM Rev.**, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, v. 51, n. 4, p. 661–703, nov. 2009. ISSN 0036-1445. Citado na página 23.

CONG, J.; LIU, H. Approaching human language with complex networks. **Physics of life reviews**, Elsevier, v. 11, n. 4, p. 598–618, 2014. Citado 3 vezes nas páginas 24, 25 e 39.

COSTA, L. F.; OLIVEIRA JR, O. N.; TRAVIESO, G.; RODRIGUES, F. A.; Villas Boas, P. R.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, v. 60, n. 3, p. 329–412, 2011. Citado 3 vezes nas páginas 21, 22 e 23.

COSTA, L. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. **Advances in Physics**, v. 56, n. 1, p. 167–242, January 2007. Citado 6 vezes nas páginas 22, 24, 26, 27, 28 e 37.

DOROGOVTSEV, S. N.; MENDES, J. F. F. Language as an evolving word web. **Proceedings of the Royal Society of London. Series B: Biological Sciences**, v. 268, n. 1485, p. 2603–2606, 2001. Citado na página 25.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification (2Nd Edition)**. [S.l.]: Wiley-Interscience, 2000. ISBN 0471056693. Citado na página 58.

ERDöS, P.; RéNYI, A. On random graphs i. **Publicationes Mathematicae Debrecen**, v. 6, p. 290, 1959. Citado na página 23.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina.** [S.l.]: LTC, 2011. Citado 2 vezes nas páginas 58 e 59.

FORTUNATO, S. Community detection in graphs. **Physics Reports**, v. 486, n. 3–5, p. 75 – 174, 2010. ISSN 0370-1573. Citado na página 60.

FRANTZESKOU, G.; STAMATATOS, E.; GRITZALIS, S.; KATSIKAS, S. Effective identification of source code authors using byte-level information. In: **Proceedings of the 28th International Conference on Software Engineering**. New York, NY, USA: ACM, 2006. (ICSE '06), p. 893–896. Citado 2 vezes nas páginas 33 e 54.

GABASOVA, E. **Analysing programming languages using dependency networks**. 2014. Disponível em: <http://evelinag.com/blog/2014/06-09-comparing-dependency-networks/index.html#.VlnfD3qwelM>. Citado na página 30.

GAMON, M. **Linguistic correlates of style: authorship classification with deep linguistic analysis features**. 2004. Citado na página 34.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. **Proceedings of the National Academy of Sciences**, v. 99, n. 12, p. 7821–7826, 2002. Citado na página 23.

GRABSKA-GRADZINSKA A. KULIG, J. K. I.; DROZDZ, S. Complex network analysis of literary and scientific texts. **International Journal of Modern Physics C**, v. 23, 2012. Citado na página 40.

GRAHAM, N.; HIRST, G.; MARTHI, B. Segmenting documents by stylistic character. **Natural Language Engineering**, v. 11, n. 4, p. 397–415, December 2005. Citado na página 33.

GRANT, T. D. Quantifying evidence for forensic authorship analysis. **International journal of speech, language and the law**, 2007. First publication by 'International Journal of Speech, Language and the Law' and Equinox. Citado 2 vezes nas páginas 18 e 33.

HAVLIN, S. The distance between zipf plots. **Physica A: Statistical Mechanics and its Applications**, v. 216, n. 1, p. 148–150, 1995. Citado na página 33.

HIRST, G.; FEIGUINA, O. Bigrams of syntactic labels for authorship discrimination of short texts. **Literary and Linguistic Computing**, v. 22, n. 4, p. 405–417, 2007. Citado 2 vezes nas páginas 34 e 36.

HIRST, G.; FENG, V. W. Changes in style in authors with alzheimer's disease. **English Studies**, v. 93, n. 3, p. 357–370, 2012. Citado na página 64.

HOLMES, D. I. Authorship attribution. **Computers and the Humanities**, v. 28, n. 2, p. 87–106, 1994. Citado na página 32.

JANKOWSKA, M.; MILIOS, E.; KESELJ, V. Author verification using common n-gram profiles of text documents. In: **Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers**. [S.l.]: Dublin City University and Association for Computational Linguistics, 2014. Citado na página 34.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in bayesian classifiers. In: **Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (UAI'95), p. 338–345. ISBN 1-55860-385-9. Citado na página 59.

JUOLA, P. Authorship attribution. **Found. Trends Inf. Retr.**, Now Publishers Inc., Hanover, MA, USA, v. 1, n. 3, p. 233–334, dez. 2006. ISSN 1554-0669. Citado 2 vezes nas páginas 18 e 54.

KASHTAN, N.; ITZKOVITZ, S.; MILO, R.; ALON, U. Topological generalizations of network motifs. **Phys. Rev. E**, American Physical Society, v. 70, p. 031909, Sep 2004. Citado na página 29.

KONG, J. S.; REZAEI, B. A.; SARSHAR, N.; ROYCHOWDHURY, V. P.; BOYKIN, P. O. Collaborative spam filtering using e-mail networks. **Computer**, IEEE Computer Society, Los Alamitos, CA, USA, v. 39, n. 8, p. 67–73, 2006. ISSN 0018-9162. Citado na página 39.

KUMAR, E. **Natural language processing**. New Delhi: I K International, 2012. Citado na página 34.

KUNCHEVA, L. I. **Fuzzy Classifier Design**. [S.l.]: Springer, 2000. v. 49. 1-270 p. (Studies in Fuzziness and Soft Computing, v. 49). ISBN 978-3-7908-2472-8. Citado na página 62.

LARSEN-FREEMAN, D.; LYNNE, C. Complex systems and applied linguistics. v. 92, n. 4, 2008. Citado na página 24.

LIU, H. Dependency distance as a metric of language comprehension difficulty. **Journal of Cognitive Science**, v. 9, n. 2, p. 159–191, 2008. Citado na página 48.

____. Statistical properties of chinese semantic networks. **Chinese Science Bulletin**, SP Science in China Press, v. 54, n. 16, p. 2781–2785, 2009. ISSN 1001-6538. Citado na página 17.

LIU, H.; CONG, J. Language clustering with word co-occurrence networks based on parallel texts. **Science Bulletin**, v. 58, n. 10, p. 1139–1144, 2013. Citado 2 vezes nas páginas 14 e 48.

LUDUEñA, G.; BEHZAD, M.; GROS, C. Exploration in free word association networks: models and experiment. **Cognitive Processing**, Springer Berlin Heidelberg, v. 15, n. 2, p. 195–200, 2014. ISSN 1612-4782. Citado na página 17.

MARTINEZ-ROMO, J.; ARAUJO, L.; BORGE-HOLTHOEFER, J.; ARENAS, A.; CAPITáN, J. A.; CUESTA, J. A. Disentangling categorical relationships through a graph of co-occurrences. **Phys. Rev. E**, American Physical Society, v. 84, p. 046108, out. 2011. Citado na página 69.

MASUCCI, A. P.; KALAMPOKIS, A.; EGUÍLUZ, V. M.; HERNáNDEZ-GARCÍA, E. Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. **PLoS ONE**, Public Library of Science, v. 6, 2011. Citado na página 39.

MATTHEWS, R. A. J.; MERRIAM, T. V. N. Neural computation in stylometry i: An application to the works of shakespeare and fletcher. **Literary and Linguistic Computing**, v. 8, n. 4, p. 203–209, 1993. Citado na página 54.

MCCARTHY, P. M.; LEWIS, G. A.; DUFTY, D. F.; MCNAMARA, D. S. Analyzing writing styles with coh-metrix. In: SUTCLIFFE, G.; GOEBEL, R. (Ed.). **FLAIRS Conference**. [S.l.]: AAAI Press, 2006. p. 764–769. Citado na página 34.

MEHRI, A.; DAROONEH, A. H.; SHARIATI, A. The complex networks approach for authorship attribution of books. **Physica A: Statistical Mechanics and its Applications**, v. 391, n. 7, p. 2429 – 2437, 2012. Citado 3 vezes nas páginas 14, 42 e 51.

MENDENHALL, T. C. The characteristic curves of composition. **Science**, ns-9, n. 214S, p. 237–246, 1887. Citado na página 33.

MIHALCEA, R.; RADEV, D. **Graph-based natural language processing and information retrieval**. Cambridge; New York: Cambridge University Press, 2011. ISBN 9780521896139 0521896134. Citado 2 vezes nas páginas 17 e 39.

MILO, R.; SHEN-ORR, S.; ITZKOVITZ, S.; KASHTAN, N.; CHKLOVSKII, D.; ALON, U. Network motifs: simple building blocks of complex networks. **Science**, v. 298, n. 5594, p. 824–827, October 2002. Citado 3 vezes nas páginas 23, 29 e 61.

MOSTELLER, F.; WALLACE, D. L. **Inference and Disputed Authorship: The Federalist Papers**. Reading, Mass.: Addison-Wesley, 1964. Citado 2 vezes nas páginas 18 e 32.

MOTA, N.; FURTADO, R.; MAIA, P.; COPELLI, M.; RIBEIRO, S. Graph analysis of dream reports is especially informative about psychosis. **Science Reports**, 2014. Citado na página 64.

MOURA, A. P. de; LAI, Y.-C.; MOTTER, A. E. Signatures of small-world and scale-free properties in large computer programs. **Physical Review E**, v. 68, 2003. Citado na página 39.

NEWMAN, M. **Networks: An Introduction**. New York, NY, USA: Oxford University Press, Inc., 2010. Citado 5 vezes nas páginas 17, 21, 22, 23 e 29.

NEWMAN, M. E. Assortative mixing in networks. **Phys. Rev. Lett.**, v. 89, n. 20, p. 208701, 2002. Citado na página 26.

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM REVIEW**, v. 45, p. 167–256, 2003. Citado 2 vezes nas páginas 23 e 24.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: A method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (ACL '02), p. 311–318. Citado na página 63.

PASTOR-SATORRAS, R.; VÁZQUEZ, A.; VESPIGNANI, A. Dynamical and correlation properties of the Internet. **Physical Review Letters**, APS, v. 87, n. 25, p. 258701, 2001. Citado 2 vezes nas páginas 27 e 28.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0. Citado na página 59.

ROMACKER, M.; MARKERT, K.; HAHN, U. Lean semantic interpretation. In: **Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. (IJCAI'99), p. 868–875. Citado na página 68.

ROXAS, R. M.; TAPANG, G. Prose and Poetry Classification and Boundary Detection Using Word Adjacency Network Analysis. **International Journal of Modern Physics C**, v. 21, p. 503–512, 2010. Citado na página 56.

SANDERSON, C.; GUENTER, S. Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation. In: **Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (EMNLP '06), p. 482–491. Citado na página 36.

SILVA, E.; STUMPF, M. P. H. Complex networks and simple models in biology. **Journal of the Royal Society Interface**, 2005. Citado na página 30.

SILVA, F. N.; COMIN, C. H.; PERON, T. K. D. M.; RODRIGUES, F. A.; YE, C.; WILSON, R. C.; HANCOCK, E.; COSTA, L. F. Concentric Network Symmetry. p. 1–15, 2014. Citado 2 vezes nas páginas 60 e 61.

SILVA, T. C.; AMANCIO, D. R. Discriminating word senses with tourist walks in complex networks. **The European Physical Journal**, 2013. Citado 2 vezes nas páginas 18 e 39.

STAMATATOS, E. Authorship attribution based on feature set subspacing ensembles. **International Journal on Artificial Intelligence Tools**, n. 5, p. 823–838, 2006. Citado na página 33.

____. A survey of modern authorship attribution methods. **J. Am. Soc. Inf. Sci. Technol.**, John Wiley & Sons, Inc., New York, NY, USA, v. 60, n. 3, p. 538–556, mar. 2009. ISSN 1532-2882. Citado 8 vezes nas páginas 18, 32, 33, 34, 35, 36, 53 e 59.

TRAVENçOLO, B.; COSTA, L. F. Accessibility in complex networks. **Physics Letters A**, v. 373, n. 1, p. 89 – 95, 2008. ISSN 0375-9601. Citado 2 vezes nas páginas 31 e 60.

TRAVENçOLO, B. A. N.; VIANA, M. P.; COSTA, L. F. Border detection in complex networks. **New Journal of Physics**, v. 11, n. 6, p. 063019, 2009. Citado 3 vezes nas páginas 29, 31 e 60.

TWEEDIE, F. J.; SINGH, S.; HOLMES, D. I. Neural network applications in stylometry: The federalist papers. **Computers and the Humanities**, v. 30, n. 1, p. 1–10, 1996. Citado na página 54.

VIANA, M. P.; BATISTA, J. L. B.; COSTA, L. F. Effective number of accessed nodes in complex networks. **Phys. Rev. E**, American Physical Society, v. 85, p. 036105, Mar 2012. Citado 2 vezes nas páginas 31 e 44.

WATTS, D.; STROGATZ, S. Collective dynamics of 'small-world' networks. **Nature**, n. 393, p. 440–442, 1998. Citado 2 vezes nas páginas 23 e 24.

WIJK, B. C. M. van; STAM, C. J.; DAFFERTSHOFER, A. Comparing brain networks of different size and connectivity density using graph theory. **PLoS ONE**, Public Library of Science, v. 5, n. 10, p. e13701, 10 2010. Citado na página 60.

ZIPF, G. Human behaviour and the principle of least-effort. In: . Cambridge, MA: Addison-Wesley, 1949. Citado na página 40.

APÊNDICE

A

LIVROS UTILIZADOS PARA O RECONHECIMENTO DE AUTORIA

As tabelas abaixo exibem os 40 livros utilizados para a tarefa de reconhecimento de autoria.

Autor	**Livro**	**Ano**
Arthur Conan Doyle	The Adventures of Sherlock Holmes	1892
	The Tragedy of the Korosko	1897
	The Valley of Fear	1914
	Through the Magic Door	1907
	Uncle Bernac - A Memory of the Empire	1896
Bram Stoker	Dracula's Guest	1914
	Lair of the White Worm	1911
	The Jewel Of Seven Stars	1903
	The Man	1905
	The Mystery of the sea	1902
Charles Dickens	A Tale of Two Cities	1859
	American Notes	1842
	Barnaby Rudge: A Tale of the Riots of Eighty	1841
	Great Expectations	1861
	Hard Times	1854
Edgar Allan Poe	The Works of E. A. P (Volume 1 - 5)	1835
Hector Hugh Munro (Saki)	Beasts and Super Beasts	1914
	The Chronicles of Clovis	1912
	The Toys of Peace	1919

	When William Came	1913
	The Unbearable Bassington	1912
Pelham Grenville Wodehouse	Girl on the Boat	1920
	My Man Jeeves	1919
	Something New	1915
	The Adventures of Sally	1922
	The Clicking of Cuthbert	1922
Thomas Hardy (1840-1928)	A Pair of Blue Eyes	1873
	Far from the Madding Crowd	1874
	Jude the Obscure	1895
	Mayor Casterbridge	1886
	The Hand of Ethelberta	1875
William Makepeace Thackeray	Barry Lyndon	1844
	The Book of Snobs	1848
	The History of Pendennis	1848
	The Virginians	1859
	Vanity Fair	1848

Tabela 9 – Lista com os 40 livros utilizados

APÊNDICE

B

LISTA DE STOPWORDS PARA O INGLÊS

A lista abaixo ilustra as 127 stopwords em Inglês que foram utilizadas no pré-processamento dos textos. Todas estas palavras foram desconsideradas durante a análise de reconhecimento de autoria.

’a’, ’about’, ’above’, ’after’, ’again’, ’against’, ’all’, ’am’, ’an’, ’and’, ’any’, ’are’, ’as’, ’at’, ’be’, ’because’, ’been’, ’before’, ’being’, ’below’, ’between’, ’both’, ’but’, ’by’, ’can’, ’did’, ’do’, ’does’, ’doing’, ’don’, ’down’, ’during’, ’each’, ’few’, ’for’, ’from’, ’further’, ’had’, ’has’, ’have’, ’having’, ’he’, ’her’, ’here’, ’hers’, ’herself’, ’him’, ’himself’, ’his’, ’how’, ’i’, ’if’, ’in’, ’into’, ’is’, ’it’, ’its’, ’itself’, ’just’, ’me’, ’more’, ’most’, ’my’, ’myself’, ’no’, ’nor’, ’not’, ’now’, ’of’, ’off’, ’on’, ’once’, ’only’, ’or’, ’other’, ’our’, ’ours’, ’ourselves’, ’out’, ’over’, ’own’, ’s’, ’same’, ’she’, ’should’, ’so’, ’some’, ’such’, ’t’, ’than’, ’that’, ’the’, ’their’, ’theirs’, ’them’, ’themselves’, ’then’, ’there’, ’these’, ’they’, ’this’, ’those’, ’through’, ’to’, ’too’, ’under’, ’until’, ’up’, ’very’, ’was’, ’we’, ’were’, ’what’, ’when’, ’where’, ’which’, ’while’, ’who’, ’whom’, ’why’, ’will’, ’with’, ’you’, ’your’, ’yours’, ’yourself’, ’yourselves’