UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE FÍSICA DE SÃO CARLOS

CAMILO AKIMUSHKIN VALENCIA

Dinâmica de redes complexas aplicada a reconhecimento de autoria

São Carlos

2015

CAMILO AKIMUSHKIN VALENCIA

Dinâmica de redes complexas aplicada a reconhecimento de autoria

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

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

São Carlos

2015

RESUMO

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

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

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

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

1.1 Proposta de pesquisa

1.2 Objetivos

CAPÍTULO 2

MATERIAIS E MÉTODOS

CAPÍTULO 3

RESULTADOS

CAPÍTULO 4

CONCLUSÕES

CAPÍTULO 5

CRONOGRAMA

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

Tabela 5.1 – Cronograma de atividades

5.1 Disciplinas Cursadas

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

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

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

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