Atribuição de autoria utilizando redes complexas

Paulo Rodrigues Alves Margarido[+]
Maria das Graças Volpe Nunes[+]
Luciano da Fontoura Costa[*]
Osvaldo Novais de Oliveira Jr.[*]
Thiago Alexandre Salgueiro Pardo[+]

[+]Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos
pauloram@grad.icmc.usp.br, {gracan, taspardo}@icmc.usp.br

[*]Instituto de Física de São Carlos (IFSC), USP/São Carlos
{luciano, chu}@if.sc.usp.br

1. Introdução

A escrita é algo que sempre fascinou a humanidade, e esta é uma atividade que sempre cresceu em interesse entre os homens. Porém, com este aumento, naturalmente emergem certos problemas, como, por exemplo, quando mais de um autor reivindica autoria sobre um determinado livro, e como atribuir corretamente o crédito por tal autoria.

Mais recentemente, pesquisadores da área de Processamento de Língua Natural (PLN) começaram a dar atenção a este problema, e diversos métodos foram desenvolvidos com este propósito, já com resultados expressivos.

2. Objetivos

O propósito deste trabalho é desenvolver um método que possibilite determinar sem o auxílio de um ser humano quem é de fato o escritor que merece os créditos pela autoria de um determinado livro, isto é, realizar uma atribuição automática de autoria.

Para tanto, o conceito de redes complexas [1], adaptado para aplicações de PLN [2], permitiu a utilização de diversos recursos matematicamente embasados inerentes a esta forma de representação na realização dos experimentos desenvolvidos.

3. Métodos

As redes complexas são uma classe de grafos com características especiais provenientes de uma organização complexa subjacente aos dados modelados. Elas vêm sendo aplicadas em um número crescente de áreas de pesquisa, incluindo Mecânica Física e PLN. Na área de PLN, as redes já foram testadas para sumarização automática, avaliação automática de qualidade textual, e até em tradução automática. A razão que motivou a utilização de redes complexas foi o forte embasamento teórico e matemático já existente [3].

A partir de medidas extraídas da rede (muitas oriundas da Teoria dos Grafos), tentou-se agrupar e classificar (usando técnicas de Aprendizado de Máquina, como regras e árvores de decisão, métodos bayesianos, etc.) os resultados para diferentes autores, para classificação dos mesmos.

4. Resultados

Os resultados do agrupamento, ainda preliminares, mostraram que é possível separar autores em um córpus contendo 31 livros de diversos autores consagrados da língua inglesa. Em relação à classificação

automática, a maior precisão verificada até o momento foi de 62%, ainda baixo devido a pouca quantidade de exemplos para o aprendizado.

Embora estes sejam resultados modestos quando comparados a outros trabalhos da área [4, 5], eles comprovam a aplicabilidade das redes complexas em mais esta área do conhecimento.

Em relação à continuidade deste trabalho, mais exemplos de aprendizado estão sendo coletados. Pretende-se, também, comparar os resultados obtidos com as redes com os resultados das técnicas apresentadas em [4] e [5], as quais estão sendo reproduzidas.

5. Conclusões

A teoria de redes complexas ainda é relativamente recente no meio de pesquisas, mas vem cada vez mais se firmando como uma forma de representação de dados confiável, além de altamente adaptável, rapidamente expandindo sua gama de aplicações.

A partir da base que recentes experimentos têm formado sobre as redes complexas, novas possibilidades podem ser abertas, e resultados cada vez melhores podem ser obtidos.

6. Referências Bibliográficas

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

[2] L. Antiqueira, M. G. V. Nunes, O. N. Oliveira Jr., & L. da F. Costa: 2005, 'Modelando textos como redes complexas'. Technical report, Instituto de Ciências Matemáticas e de Computação.

[3] R. Ferrer i Cancho & R. V. Solé: 2001, 'The Small World of Human Language'. In: P. Roy. Soc. Lond. B Bio., 268, 2261.

[4] E. Stamatatos, N. Fakotakis & G. Kokkinakis: 1999, 'Automatic Authorship Attribution'. In: Proceedings of EACL 1999.

[5] F. Peng, D. Schuurmans, V. Keselj, & S. Wang, 2003: Language Independent 'Authorship Attribution using Character Level Language Models'. In: Proceedings of EACL 2003.