APRENDIZADO DE MÁQUINA PARA ANÁLISE SEMÂNTICA DE ARTIGOS CIENTÍFICOS

Aluno: Éverton Cardoso Acchetta,
Orientador: Prof. Dr. Paulo Sergio Silva Rodrigues[2]
[1,2] *Ciência da Computação, Centro Universitário FEI*
unieeacchetta@fei.edu.br e psergio@fei.edu.br

Resumo: A obtenção na internet de artigos científicos de um assunto específico ainda é uma tarefa trabalhosa, uma vez que a quantidade de base de dados tem crescido de maneira volumosa, refletindo a importância e o interesse da produção científica, mas ao mesmo tempo dificultando a recuperação e organização de informações. Neste projeto, é apresentada uma proposta para indexação, recuperação, organização e seleção automática de informação de artigos acadêmicos-científicos utilizando como fonte o *Semantic Scholar*.

1. Introdução

Tradicionalmente, quando deseja-se pesquisar em uma determinada área acadêmica, uma das primeiras tarefas que deve ser feita é o levantamento bibliográfico para o estudo atual do estado-da-arte. Nesse ponto, uma boa fonte de obtenção desses recursos é o site *Semantic Scholar*, por possuir um banco de mais de 175.000.000 de artigos, permitindo a aquisição de informação relacionada, atual, bem estruturada, de maneira rápida e eficaz.

Além da busca por artigos científicos, o gerenciamento é outra tarefa fundamental, que de forma geral deve ser capaz de organizar as principais informações de cada artigo de tal maneira que seja fácil compará-lo a outros relacionados ao mesmo assunto. Nesse contexto, há diversos programas que podem realizar esse gerenciamento, tais como: Zotero, Docear, Mendeley, JabRef, Qiqqa, entre muitos outros [1].

Apesar desses gerenciadores de artigos facilitarem o processo de organização de bibliografias, eles ainda estão longe de analisar de forma automática o conteúdo semântico dos mesmos. Por esse motivo, hoje há uma demanda crescente por sistemas que possam reconhecer o conteúdo de um artigo científico avançando, ainda mais esse tipo de tarefa.

Uma área bem conhecida da computação que lida com a compreensão textual é a de Processamento de Linguagem Natural [2]. Essa área é um ramo da Inteligência Artificial que agrega um extenso conjunto de ferramentas estatísticas e modelagem computacional. Um desses modelos computacionais é chamado Modelos Baysianos. Trabalhos recentes, tais como [7, 4] têm demonstrado como essa teoria pode ser aplicada na área de Processamento de Linguagem Natural.

Por outro lado, a rígida formalização encontrada em artigos acadêmicos, sobretudo científicos, em que divide sua informação entre tópicos, tais como: objetivos, metodologias, revisão bibliográfica, conceitos, metrologia, experimentos, resultados e discussão, pode facilitar a busca baseadas em modelos estatísticos como as Redes Bayesianas dentro de processos de aprendizado de máquina. Assim, é possível criar modelos matemático-estatístico-computacionais que aprendam a reconhecer essas estruturas formais de artigos de modo a sustentar a construção de arquiteturas de softwares que sejam capazes de organizar essas informações.

Assim, a ideia geral desse projeto é explorar esse tipo de informação estruturada a fim de facilitar a extração de conteúdo textual de cada uma dessas partes, de modo a viabilizar o estudo da revisão bibliográfica de novos artigos, de modo a produzir com mais clareza suas afirmações semânticas.

2. Metodologia

A metodologia proposta neste trabalho é desenvolver um sistema automatizado que recebe um termo-chave do usuário e retorna uma tabela com resultados obtidos do site *Semantic Scholar* relacionados a este termo. Esta tabela possui os seguintes metadados dos artigos: título, autores, data de publicação, velocidade de citação, fator de influência, um índice criado com a intenção de priorizar o que o usuário julgar mais importante durante a ordenação, BibTeX, e por fim, o objetivo de cada artigo.

Para que este objetivo seja atingido, um compilador de informações (*crawler*) foi proposto para uma tabela *Excel*. Também foram desenvolvidos dois sistemas de suporte: um responsável por extrair o texto de arquivos .pdf e salva-los em formato .txt para análise futura, e um que será usado para a geração da base de dados, extraindo informações desejadas de tabelas de metadados de artigos que foram preenchidas manualmente.

O modelo matemático-estatístico-computacional proposto aqui segue a linha epistemológica idealizada originalmente em [5] e utilizada com sucesso em trabalhos como [3, 6] e [7] para recuperação de informação textual, e em [4] para recuperação de imagens com base no conteúdo visual.

Esse modelo prevê que o cálculo da probabilidade de um evento, considerando a ocorrência de eventos prévios, possa ser computado com a combinação logística do operador XOR entre as probabilidades dos eventos internos. No trabalho proposto aqui, cada evento interno é também ponderado potencialmente por constantes calculadas com um modelo de otimização.

Essa ideia é explanada pela relação condicional mostrada na Equação (1), onde calcula-se a probabilidade de um texto $F_i = \{P_1, P_2, ..., P_n\}$, que é uma sequência de palavras que formam o texto F_i de n palavras, sendo P_i cada palavra contida em F_i, ser uma sentença considerada como o objetivo do artigo, considerando que os eventos sequenciais de palavras ocorreram previamente.

$P(F_i|P_1, P_2, ..., P_n)=1 - [(1 - f_0)^{\alpha 0}$ x...x $(1 - f_n)^{\alpha n}]$ (1)

3. Resultados Parciais

A primeira etapa do desenvolvimento deste sistema deu-se como a criação de um *Crawler* desenvolvido em *Python*, utilizando o *framework Selenium*, que permite automatizar os cliques em navegadores de internet (neste caso, foi utilizado o *Google Chrome*), e obter metadados das páginas visitadas. Com isso, o termo desejado pelo usuário é inserido no site *Semantic Scholar* e os resultados da pesquisa são salvos com todos os seus metadados disponíveis. A pesquisa é realizada de três maneiras diferentes: de maneira padrão, sem alterar nenhum filtro disponível; utilizando o filtro *Lit Reviews* e pesquisando artigos de no máximo 5 anos de idade. O usuário então escolhe com quantas páginas gostaria que a pesquisa seja realizada, o que resulta em aproximadamente 30 resultados por página pesquisada Após todo o processo de aquisição de metadados, os resultados são salvos utilizando o framework *Pickle* em formato .pkl, de maneira orientada a objetos e de fácil manipulação para processos futuros.

Após todos os metadados terem sido salvos pelo *crawler*, estes precisam ser apresentados de maneira que o usuário possa entender e caso necessário, edita-los. Para isso, um software também desenvolvido em *Python* foi criado. Usando o framework *xlsxwriter*, todos os metadados que foram salvos no formato .pkl pelo *crawler* são lidos e dispostos de maneira que todos os metadados relevantes estejam acessíveis em uma planilha do *Microsoft Excel*. Porém esses metadados possuem pesos diferentes para cada usuário, já que para uma pessoa, um artigo mais novo pode ser muito mais relevante que um artigo de 30 anos, porém com milhares de citações. Pensando nisso, foi aplicado um algoritmo (Equação 2).

$$P(P_{ot\_fat} \mid P_{vel}, P_{inf}, P_{ano}) = \alpha_1 P_{vel} + \alpha_2 P_{inf} + \alpha_3 P_{ano} \quad (2)$$

Onde α_1 é o peso desejado para a velocidade de citação, α_2 para o fator de influência, α_3 para a data de publicação do artigo, P_{vel} a velocidade de citação, P_{inf} o fator de influência e P_{ano} o ano de publicação.

Para melhorar a experiência final do usuário, o *crawler* e o compilador de resultados foram integrados em um único sistema com interface gráfica multi-plataforma (Linux, Windows e MacOS), desenvolvido em *Python* utilizando um *framework* chamado *appJar*, que envelopa a biblioteca *TKinter* (Figura 1).

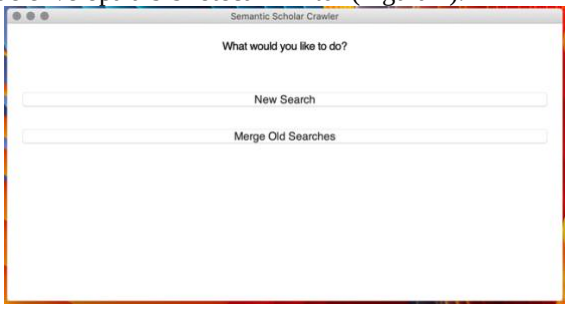

Figura 1 – Tela inicial da GUI

Para que os artigos em .pdf possam ser interpretados pelo modelo Baysiano proposto remediar este problema, foi desenvolvido um sistema que utiliza o *framework Google Tesseract OCR*, que utiliza reconhecimento ótico de caracteres para extrair o texto de um arquivo .pdf. Após isso, este texto extraído é salvo em formato .txt, o que permite que qualquer outro sistema tenha acesso a esses textos. Todo processo é realizado usando multiprocessos, permitindo que a atividade seja dividida para quantas *threads* estejam disponíveis.

4. Conclusões

O presente artigo apresenta um modelo estatístico-computacional para busca e análise de artigos científicos fornecidos no site *Semantic Scholar*. O trabalho ainda está em andamento, mas já possui uma interface em formato de Crawler que é capaz de obter, organizar e priorizar metadados de artigos científicos. Como resultado, tem ajudado a revisão bibliográfica de mais de 20 grupos de trabalhos de conclusão de curso do curso de ciência da computação da FEI, além de diversos mestrados e doutorados na área de Processamento de Sinais e Imagens. Como próximos passos, o modelo pretende fornecer automaticamente como informação as principais características dos artigos como objetivo, metodologias, métricas, base de dados resultados obtidos.

5. Referências

[1] José Luis Ortega. Chapter 4 - Reference Management Tools. In Social Network Sites for Scientists, pages 65–99. Chandos Publishing, 2016. DOI: 10.1016/B978-0-08-100592-7.00004-6.

[2] V Gudivada, Dhana Rao, and V Raghavan. Big data driven natural language processing research e applications. Big Data Analytics, 33:203, 2015.

[3] Paulo Sergio Silva Rodrigues. Um modelo bayesiano combinando análise semântica latente e atributos espaciais para recuperação de informação visual. 2003.

[4] Diego R Amancio. Probing the topological properties of complex networks modeling short written texts. PloS one, 10(2):e0118394, 2015.

[5] Baeza-Yates Ricardo et al. Modern information retrieval. Pearson Education India, 1999.

[6] Berthier AN Ribeiro and Richard Muntz. A belief network model for ir. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pages 253–260. ACM, 1996.

[7] Guilherme Alberto Wachs-Lopes and Paulo Sergio Rodrigues. Analyzing natural human language from the point of view of dynamic of a complex network. Expert Systems with Applications, 45:8–22, 2016.

Agradecimentos

À CNPq, CAPES e FAPESP, além do departamento de Ciencia da Computação da FEI pela realização das medidas ou empréstimo de equipamentos.

1 Aluno de IC do Centro Universitário FEI. Projeto com vigência de 03/18 a 02/19.