IDENTIFICAÇÃO SISTEMÁTICA DE ARTIGOS CIENTÍFICOS COM REDUZIDO ESFORÇO DO USUÁRIO

MATHEUS VINÍCIUS TODESCATO[1], JEAN CARLO HILGER[2], GUILHERME DAL BIANCO[3]

1 Introdução

O crescente volume de dados faz com que informações pertinentes se tornem cada vez mais necessárias e valiosas. Há circunstâncias em que almeja-se a obtenção de todas as informações relevantes disponíveis em um conjunto de dados, o que configura o objeto de investigação da *High Recall Information Retrieval* (HIRE). No âmbito da Revisão Sistemática da Literatura (RSL), por exemplo, encontrar todos (ou quase) documentos/artigos relevantes acerca dos sintomas de uma doença é imprescindível. Neste cenário, a perda de documentos relevantes pode acarretar em sérias complicações, como a não identificação de determinados sintomas referentes a uma doença.

Tradicionalmente métodos HIRE utilizam aprendizado supervisionado para a geração de um ranqueamento (Cormack *et al*., 2015). Os métodos supervisionados dependem de um conjunto de treinamento para o aprendizado dos padrões presentes na base de dados. O revisor decide se os documentos do topo do ranking (mais próximos da consulta do usuário, como um buscador *Web* tradicional) são relevantes ou não para determinado tópico. O método de ranqueamento será atualizado com as novas informações (de documentos avaliados e revisados pelo usuário) após cada rodada.

Em pesquisas realizadas anteriormente para redução do esforço do usuário, foi identificado que o aprendizado ativo é promissor na forma de se alcançar esse objetivo (Yu *et al.*, 2018). O aprendizado ativo visa identificar os documentos informativos evitando que documentos redundantes/desnecessários sejam apresentados ao usuário. Assim o esforço se dá nos documentos que têm maior probabilidade de serem relevantes ou que podem melhorar o classificador.

1 Acadêmico de Ciência da Computação, Universidade Federal da Fronteira Sul, *campus* Chapecó, **Bolsista**, contato: mvtodescato@hotmail.com.
2 Acadêmico de Ciência da Computação, Universidade Federal da Fronteira Sul, *campus* Chapecó.
3 Doutor, UFFS, **Orientador**.

O HIRE contém alguns desafios em relação ao seu desempenho e a quantidade de esforço aplicado na revisão. Entre eles temos a geração do treinamento inicial, pois o método depende de informações iniciais para a aprendizagem do padrão de informação requisitada pelo usuário. Em algumas situações, a quantidade de documentos relevantes (prevalência) de uma consulta pode ser muito baixa (por exemplo, na base do CLEF 2017 (Kanoulas *et al*., 2017) a cada 700 documentos somente 1 é relevante). Dessa forma, a geração de semente, se torna uma tarefa desafiadora mas vital para que o método aprenda inicialmente o que é um documento relevante no contexto de cada tópico.

2 Objetivos

Este trabalho tem por objetivo aprimorar um método de HIRE a partir do melhoramento da estratégia de seleção da semente. Para isso, será proposta uma abordagem utilizando a aprendizagem ativa juntamente com uma técnica para a geração de um ranqueamento.

3 Metodologia

Inicialmente identificou-se os principais trabalhos científicos que exploram técnicas de HIRE. As pesquisas relacionadas apresentam diversas estratégias para cada etapa desse processo. Dessa forma, os principais estudos foram selecionados para se entender o seu comportamento. O AutoTar (Cormack *et al*., 2015), representa um dos métodos mais simples para o HIRE, na qual o seu diferencial é operar sobre qualquer tópico, porém sem se preocupar com esforço do usuário e o ponto de parada. Já o S-CAL (Cormack *et al*., 2016) exige que o usuário apenas revise uma parcela menor dos documentos.

No estudo da bibliografia foi identificado que um dos principais métodos é o REVEAL (Bianco *et al.,* 2020) que combina o S-CAL com uma nova proposta de aprendizagem ativa. O REVEAL representa uma melhora em relação ao estado da arte, no entanto, a criação da semente ainda é feita usando um documento sintético. O documento sintético nada mais é do que os termos usados na consulta do usuário. Ou seja, termos não adequados podem impossibilitar a convergência do método, pois a consulta pode ser pouco informativa ou não representar a informação demandada pelo usuário. Dessa forma, o código fonte do REVEAL foi estudado em detalhes para se desenvolver uma proposta para a escolha do documento inicial a ser avaliado pelo usuário. Posteriormente, foi projetada uma nova abordagem para escolha da semente. Além disso, a base de dados utilizada (CLEF 2017) foi

pré-processada. Por fim, a abordagem proposta está sendo analisada experimentalmente para se avaliar seu comportamento em comparação com outros métodos.

4 Experimentos e Resultados

Como supracitado, o método REVEAL utiliza como documento inicial (ou semente) a consulta realizada pelo usuário. No entanto, não há garantia de que uma semente sintética seja efetiva para encontrar os documentos relevantes. Assim sendo, neste trabalho é proposta uma nova abordagem para seleção da semente complementando a abordagem REVEAL. A Figura 1 ilustra a proposta na qual em vermelho é projetada a contribuição deste trabalho.

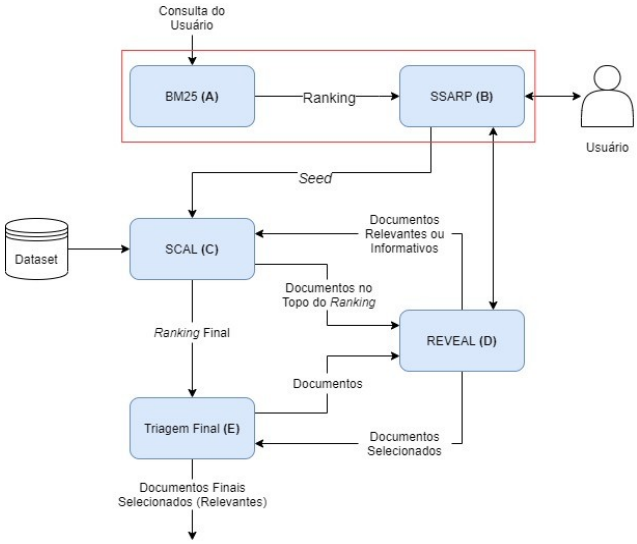

Figura 1. Estrutura de funcionamento da abordagem proposta.

Em um primeiro momento, uma técnica para geração de ranqueamento, BM25(A), é incorporada ao início do método para identificar os documentos mais promissores de acordo com a consulta do usuário. O método BM25 atribui a cada documento uma pontuação, utilizando para tal a frequência dos termos da consulta presentes no documento. Desta forma, com uma semente contendo as mesmas características dos demais documentos, a convergência do método REVEAL pode ser alcançada antes, elevando a revocação (*recall)* e diminuindo o esforço de rotulação. Após, o método de aprendizagem ativa, chamado de SSARP(B), é aplicado para remover documentos redundantes, ou seja, evitar que o usuário receba documentos similares e não relevantes. Dessa forma, é reduzido o esforço do usuário com documentos não relacionados à sua consulta.

Os experimentos parciais identificaram que a abordagem proposta se mostra promissora, no entanto, novos experimentos estão em fase final de elaboração.

5 Conclusão

Este trabalho apresentou uma abordagem para a tarefa do *High-Recall Information Retrieval*, aprimorando a geração de semente do método já existente REVEAL. A seleção proposta de uma semente mostrou-se eficaz mas ainda necessita de uma avaliação mais precisa, levando em consideração os valores de revocação e de esforço de rotulação. Isto posto, a investigação de técnicas ainda mais efetivas para a seleção da semente manifesta elevado potencial.

Referências

YU, Zhe; KRAFT, Nicholas A.; MENZIES, Tim. Finding better active learners for faster literature reviews. **Empirical Software Engineering**, v. 23, n. 6, p. 3161-3186, 2018.

CORMACK, Gordon V.; GROSSMAN, Maura R. Scalability of continuous active learning for reliable high-recall text classification. In: **Proceedings of the 25th ACM international on conference on information and knowledge management**. 2016. p. 1039-1048.

DAL BIANCO, Guilherme; DUARTE, Dênio; GONÇALVES Marcos. REVEAL-HIRE - A New Active Framework for the High Recall Task. **Em fase de avaliação**, 2020.

CORMACK, Gordon V.; GROSSMAN, Maura R. Autonomy and reliability of continuous active learning for technology-assisted review. **arXiv preprint arXiv:1504.06868**, 2015.

KANOULAS, Evangelos et al. CLEF 2017 technologically assisted reviews in empirical medicine overview. In: **CEUR Workshop Proceedings**. 2017. p. 1-29.

Palavras-chave: recuperação de informação; *High-recall Information Retrieval*; aprendizado ativo.

Financiamento: esse projeto foi financiado com recursos do Edital 459/GR/UFFS/2019.