Seleção de sementes na identificação sistemática de artigos científicos com reduzido esforço do usuário

Matheus Vinícius Todescato, Jean Carlo Hilger, Guilherme Dal Bianco

***Resumo*—A *High-recall Information Retrieval* (HIRE) tem por finalidade a identificação de todos (ou quase todos) os documentos relevantes referente a uma consulta do usuário. Esta propriedade é vastamente aproveitada na revisão sistemática da literatura, onde buscam-se documentos acadêmicos (como por exemplo artigos científicos) que sejam pertinentes à um critério de busca. O número de documentos retornados ao usuário determina a maior parte do esforço que ele necessitará exercer para encontrar as informações que possui interesse. Desta forma, é imprescindível retornar o menor número de documentos possível. Tradicionalmente, no HIRE é empregado métodos de aprendizagem supervisionada, que depende de um treinamento para aprender os padrões dos documentos relevantes relacionados a consulta. No entanto, para isso é necessário encontrar boas sementes para iniciar o processo de aprendizado. Neste artigo, é proposta uma nova abordagem para a geração do treinamento inicial (semente) para se reduzir o número de documentos não relevantes enviados ao usuário. Dessa forma, foi desenvolvida uma abordagem que combina a seleção ativa de documentos informativos com a geração de ranqueamento. Os experimentos realizados demostram que a abordagem proposta pode reduzir em até 18% o custo do processo de rotulação.**

***Index Terms*—Recuperação de informação, High-recall Information Retrieval, Aprendizado ativo, Geração de treinamento inicial, Identificação de artigos científicos.**

I. Introdução

É Notório o fato de que cada vez mais dados são gerados. Na mesma medida que a capacidade de produzir dados elava-se, cresce a urgência em encontrar informações pertinentes em meio a estes grandes volumes de dados. Em linhas gerais, esta premissa integra o objeto central de estudo da *IR (Information Retrieval)*. Seu objetivo pode ser determinado como a busca, em grandes coleções de dados, por determinado material que atenda à uma necessidade informativa [1].

Sob certas circunstâncias, porém, a mera busca por informações relevantes não é suficiente para solucionar um problema. Há situações em que deseja-se obter todo o material informativo existente em uma base de dados. A *HIRE (High-Recall Information Retrieval)* [2] é uma subárea da IR que tem como propósito atender à esta exigência, fornecendo ferramentas apropriadas para tal. Técnicas de HIRE são aplicadas em diversos cenários. No contexto acadêmico, vale mencionar métodos que auxiliam na revisão auxiliada por tecnologia (TAR), cujo pressuposto é retornar à um usuário documentos acadêmicos (usualmente artigos científicos), de modo a atender às necessidades preestabelecidas, de modo comum, por meio de uma consulta textual.

Comumente, métodos HIRE retornam ao usuário um ranqueamento dos documentos (ordenados por relevância), fazendo uso de técnicas de aprendizado supervisionado para este fim [3]. No processo de identificação de documentos relevantes, exige-se que o usuário realize a rotulação de diversas instâncias presentes na base de dados, a fim de possibilitar o treinamento do algoritmo. É imprescindível requisitar o menor número de rotulações, mitigando o esforço por parte do usuário. Para alcançar tal objetivo, vale a utilização de aprendizado ativo, cuja eficácia mostra-se promissora [4]. A aprendizagem ativa tem como objetivo selecionar apenas documentos que possuem maior valor informativo, evitando a apresentação de documentos redundantes [5]. Assim o esforço se dá nos documentos que têm maior probabilidade de serem relevantes ou que podem aprimorar o método. O método S-CAL [6] usufrui desta técnica.

Dentre os diversos desafios que surgem na modelagem de sistemas HIRE, vale destacar o problema da geração do conjunto de treinamento inicial (semente). Este conjunto inicial, utilizado pelo algoritmo de aprendizado no inicio do processo, é primordial para a identificação de padrões que configuram um documento relevante. Esta etapa possui grande impacto no esforço despendido pelo usuário, posto que pode acelerar ou atrasar o aprendizado do algoritmo. O problema descrito é conhecido como *cold-start* [7], estando presente não apenas no campo do HIRE mas também em algoritmos de aprendizado alheios à este. Há conjuntos de dados em que a taxa de documentos relevantes (prevalência) é extremamente baixa (por exemplo, na base do CLEF 2017 [8] à cada 700 documentos apenas um é relevante), tornando a tarefa ainda mais desafiadora e vital.

Neste trabalho, é apresentada uma abordagem, denominada S-CAL++, que aprimora os resultados do método S-CAL [6]. O objetivo é buscado adicionando ao ciclo uma nova camada capaz de encontrar um conjunto de documentos informativos para uma semente relevante, e não mais sintética, como é no S-CAL original. Esta camada emprega o algoritmo BM25 para a geração de um ranque (em ordem decrescente de relevância) de documentos relevantes. Após calculado, é aplicado sobre o ranque o algoritmo SSARP [9], com o intuito de remover informações redundantes e desnecessárias, possibilitando assim um menor esforço de rotulação.

Os experimentos foram realizados com a base de dados do CLEF 2017[8], contendo artigos científicos da área da medicina. Tal experimentação permitiu identificar que S-CAL++

Matheus Vinícius Todescato, Universidade Federal da Fronteira Sul (UFFS), Campus Chapecó, Brasil, matheus.todescato@estudante.uffs.edu.br.

Jean Carlo Hilger, Universidade Federal da Fronteira Sul (UFFS), Campus Chapecó, Brasil, jean.hilger@estudante.uffs.edu.br.

Guilherme Dal Bianco, Universidade Federal da Fronteira Sul (UFFS), Campus Chapecó, Brasil, guilherme.dalbianco@uffs.edu.br.

reduziu o esforço inicial no processo de seleção de semente em até 18%, quando comparado ao método base.

Na seção II são apresentados os fundamentos teóricos, utilizados no processo de confecção desta contribuição. A seção III trata de trabalhos correlatos à este, que visam solucionar problemas que surgem em técnicas HIRE. Na seção IV é apresentada e discutida a proposta de seleção de semente, aplicada ao método S-CAL ([6]). Em seguida, a seção V expõe a forma como os experimentos foram conduzidos, bem como exibe resultados e discussões relacionadas à eles. Por fim, a seção VI conclui o trabalho.

II. Referencial Teórico

Nesta seção, são apresentados os principais conceitos e métodos envolvendo a abordagem proposta. As informações apresentadas são importantes para o entendimento deste trabalho.

A. Extração de Características

Documentos são tradicionalmente apresentados em formato de texto livre. Tal representação deve ser modificada, a fim de possibilitar seu processamento por algoritmos de aprendizagem de máquina. Uma das técnicas básicas existentes com este propósito é o BoW (sigla para *Bag of Words*), que consiste em contabilizar a ocorrência de palavras em um dado documento [1]. Desta forma, o documento passa a ser representado por um vetor contendo o número de ocorrências de cada palavra que o compõe.

Alguns documentos podem apresentar tamanho elevado em comparação à outros em um mesmo corpus e consequentemente o número de ocorrências de suas palavras será maior. Neste cenário, a utilização do BoW fica comprometida, uma vez que sua representação torna-se tendenciosa para os documentos mais extensos. Outras técnicas de extração de características, mais sofisticadas, não possuem tal inconveniente.

É o caso do TF-IDF (*Term Frequency - Inverse Document Frequency*) que consiste em um algoritmo de extração de características que, semelhante ao BoW, utiliza a frequência das palavras para construir os vetores de característica. Porém, o TF-IDF considera também a frequência dos termos perante à base como um todo, fazendo com que o resultado seja independente do tamanho dos documentos. O valor TF-IDF de um termo é dado pelo produto do TF (*Term Frequency*) e do IDF (*Inverse Document Frequency*) deste mesmo termo [10]. Deste modo, tem-se:

$$tf_{td} = \frac{f_{t,d}}{|d|}, \tag{1}$$

onde $f_{t,d}$ denota a frequência do termo t no documento d e $|d|$ representa o tamanho do documento, ou seja, o total de termos presentes nele,

$$idf_t = log\frac{|D|}{|d \in D : t \in d|}, \tag{2}$$

onde D é a coleção completa de documentos, $|D|$ corresponde ao total de documentos na coleção e df_t configura o número de documentos que possuem o termo t.

Finalmente,

$$tf\text{-}idf = tf \times idf, \tag{3}$$

De modo geral, a natureza do método permite extrair um valor único para cada documento, determinado com base na consulta do usuário, o que permite utilizar o tf-idf como um algoritmo de ranqueamento de documentos. Dessa forma, é possível a geração de um ranque a partir da similaridade dos documentos com a consulta do usuário.

Outro método tradicional para geração do ranque é a partir do BM25, cuja principal inovação foi trazer à modelos probabilísticos os valores utilizados no tf-idf [1]. Embora hajam diversas variantes para o algoritmo, pode-se defini-lo, de maneira mais genérica, conforme a formula:

$$BM25 = \sum_{t \in q} idf_t \times \frac{tf_{td}(k_1 + 1)}{k_1((1-b) + b \times (L_d/L_a)) + tf_t d}, \tag{4}$$

onde k_1 e b são parâmetros arbitrários, tf_{td} e idf_t referem-se aos termos discutidos anteriormente, q é a consulta do usuário, L_d é o tamanho (em número de termos) do documento sendo processado e L_a é a média dos comprimentos dos documentos da coleção (também, em número de termos). Após calculado, o valor do BM25 pode ser utilizado para a geração de um ranque ou até mesmo como uma característica, a ser utilizada por outro algoritmo.

B. Aprendizado Supervisionado

Algoritmos de aprendizado supervisionado são aqueles aplicados sobre conjuntos de dados onde cada exemplo possui um rótulo conhecido [11] que será utilizado para permitir o aprendizado de padrões existentes. Desta forma, o propósito destes algoritmos é fornecer uma função que mapeia dados para classe [1].

Mais formalmente, seja um conjunto de dados X, definido por n características. A cada exemplar $x_i \in X$, com $x_i = (c_1, c_2, ..., c_n)$ onde c_j representa uma característica j, é associado um rótulo $y_i \in Y$. Desta forma, diz-se que um algoritmo de aprendizado supervisionado E tem como propósito encontrar valores $E(x_i) = \hat{y}_i$ tais que $\hat{y}_i - y_i \approx 0$.

Algoritmos supervisionados podem ser divididos em classificadores e regressores [11]. O primeiro tem como resultado um valor discreto, representando uma classe à qual o exemplo de entrada pertence. O segundo resultará em um valor contínuo, que representa uma possível resposta para a combinação de dados fornecida na entrada do algoritmo. Entre os algoritmos baseados em técnicas supervisionadas pode-se indicar [11]: *Naïve Bayes*, *Support Vector Machines* - SVM e Redes Neurais Artificiais.

Naïve Bayes é um algoritmo probabilístico, cuja fundamentação remete ao teorema de Bayes, por meio do qual é possível obter probabilidades condicionadas. Seu aprendizado é efetuado buscando encontrar quais termos em um documento fornecem mais evidencias de que ele pertence à determinada classe [1]. Para a aplicação do algoritmo, assume-se uma forte independência entre as características analisadas.

O algoritmo SVM (*Support Vector Machines*) visa dividir o conjunto de dados em dois. Esta divisão é feita encontrando um hiperplano de tal modo que, a distância entre este e os pontos mais próximos a ele seja máxima [1]. Embora o algoritmo seja primordialmente para classificação binária (onde há apenas duas classes), é possível utilizá-lo com mais classes.

Os algoritmos de aprendizado supracitados mostram-se eficazes para a maioria das tarefas, todavia apresentam dificuldade em aprender de maneira eficiente quando expostos à conjuntos de dados com uma grande dimensionalidade (grande número de *features*) [11].

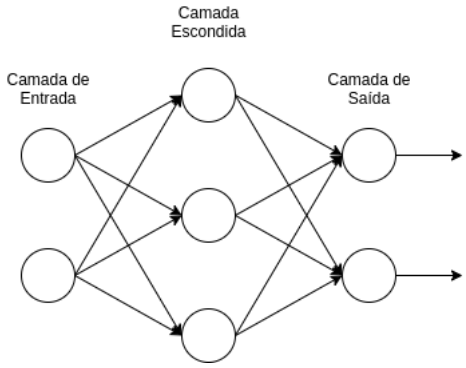

Figura 1. Estrutura básica de uma rede neural artificial.

Neste cenário, são necessários algoritmos mais eficazes, cujo aprendizado não seja comprometido pelo excesso de características. Redes Neurais Artificiais são exemplos destes algoritmos. Uma rede neural pode ser representada - graficamente - por um grafo, onde vértices representam neurônios e arestas representam as ligações entre eles. A Figura 1 exemplifica a estrutura. Os dados são fornecidos através da camada de entrada, são processados e por fim o resultado é exibido na camada de saída. Na estrutura, as arestas simbolizam pesos que definem os estados dos neurônios.

Redes neurais demonstram um grande potencial em extrair informações pertinentes de conjuntos de dados com dimensionalidade elevada. Todavia, este poder é acompanhado por um alto custo computacional na fase de treinamento o que para determinadas aplicações pode ser um fator inviabilizante.

C. Aprendizado não Supervisionado

Algoritmos não supervisionados não requerem dados rotulados para o treinamento do método. Nesta categoria, instâncias x_i pertencentes a um conjunto de dados X não estão atrelados à um rótulo y, e por consequência, a classificação (ou regressão) dá-se por meio de inferências acerca de padrões encontrados nos dados [11]. Algoritmos de clusterização são exemplos do emprego de aprendizado não supervisionado.

A clusterização tem como finalidade dividir o conjunto de dados em grupos (*clusters*) de modo que os exemplos dentro de um mesmo grupo sejam mais semelhantes entre si o possível [11]. O algoritmo *K-means* é um dos mais utilizados para a tarefa. Sua meta é minimizar média das distâncias (euclidiana) das instâncias de um *cluster* para o centroide deste *cluster* [1]. O centroide de um grupo é definido como a média de todos os documentos contidos nele.

D. Aprendizado Ativo

Em muitas aplicações, a utilização de técnicas de aprendizado de máquina - em essência, na modalidade supervisionada - requer um alto grau de qualidade dos dados de treinamento sobre os quais o algoritmo irá operar [5]. Sobretudo, a existência de dados rotulados é imprescindível. Não obstante, há cenários em que a obtenção de rótulos representa um custo elevado, inviabilizando a utilização de técnicas supervisionadas. Assim sendo, o emprego de técnicas alternativas mostra-se de fundamental importância, dentre as quais há o aprendizado ativo.

Técnicas de aprendizado ativo partem do pressuposto de que permitir que um algoritmo escolha de quais dados irá extrair o conhecimento resulta em uma maior eficácia à um menor custo de treinamento [5]. De maneira sucinta, o algoritmo selecionará instâncias de um conjunto de dados cuja rotulação é desconhecida, e irá solicitar para que um usuário aplique uma rotulação conveniente à elas. Em seguida tais instâncias passam a integrar o conjunto de treino (com rótulos) que será consumido por um algoritmo supervisionado, aperfeiçoando-o.

Silva, Gonçalves e Veloso [9] propuseram um método para ranqueamento baseado em regras de associação, chamado SSAR (*Selective Sampling using Association Rules*). Regras de associação permitem definir a relação existente entre um conjunto de valores arbitrário com um outro valor a definir-se. No caso dos autores, associou-se características dos documentos (como por exemplo o valor BM25 de cada documento) à níveis de relevância (por exemplo, relevante ou não relevante) [9], de modo a possibilitar a obtenção da relevância por meio da observação destas características. Vale ressaltar, que o aprendizado baseado em regras de associação visa inferir um conjunto de regras acerca da base de dados, e não tomar decisões com base em um conjunto de regras pré-estabelecido.

O algoritmo explora o fato de haver redundâncias nas informações dos documentos de uma coleção. Sejam U e D o conjunto de documentos não rotulados e o conjunto de treino, respectivamente. Para cada documento $u \in U$ são extraídas regras de associação, e caso possua menos regras do que qualquer outro documento já em D, é requisitada a rotulação de u que passa a integrar D [9]. Ao final do método, documentos que pertencem à D serão os que possuem mais riqueza de informação, e portanto, os mais relevantes.

Documento	**Características**
1	$a\ y\ c$
2	$x\ y\ z$
3	$x\ b\ z$

Tabela I
EXEMPLO DE UM CONJUNTO DE DOCUMENTOS NÃO ROTULADOS (U).

À título de exemplo, considera-se um conjunto de documentos como descrito na Tabela II-D. O algoritmo constrói uma projeção do conjunto não rotulado, para selecionar instâncias a integrarem o conjunto de treino. Isto é alcançado, selecionando

os documentos que sejam menos redundantes em relação à T. Inicialmente, então, seleciona-se o documento 2, em seguida o documento 1 e por fim o documento 3. Cada vez que um documento é adicionar à T, as características redundantes são omitidas. O resultado final da projeção é exibido na Tabela II-D. O intuito é obter o número de regras de associação geradas para cada entrada (coluna 3). Será requisitada a rotulação do documento que apresentar o menor número de regras.

Documento	Características	# de Regras
2	$x\ y\ z$	5
1	$a\ -\ c$	3
3	$-\ b\ -$	2

Tabela II
EXEMPLO DA PROJEÇÃO DO CONJUNTO NÃO ROTULADO.

III. TRABALHOS RELACIONADOS

A busca por todos ou quase todos os documentos relevantes configura o *High-Recall Information Retriveal*(HIRE). Em comum, os trabalhos que exploram HIRE utilizam técnicas de aprendizagem de máquina com base na revisão de documentos pelo usuário em ciclos [6], [3]. O alto esforço de rotulação do usuário e a dificuldade em alcançar uma alta revocação em tópicos com prevalência baixa de documentos, são, por exemplo, desafios de HIRE. Utilizando-se de aprendizado de máquina, diversos métodos propõem técnicas importantes a fim de melhorar o processo do HIRE e explorar os desafios do mesmo. Dentre estes, existem alguns, que serão descritos a seguir.

O AutoTAR [3] é um dos métodos mais simples para o problema TAR e sua estratégia é explorada em diversos outros métodos [6], [12], [13]. Seu diferencial é ser totalmente autônomo, não requerendo ajuste de parâmetros por conta de tópicos ou bases de dados específicas. No entanto, o método tem como ponto fraco a alta demanda de documentos rotulados pelo usuário.

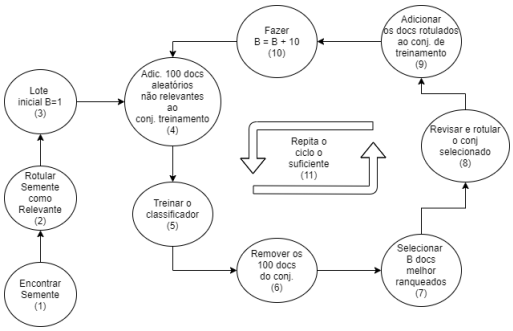

Figura 2. Estrutura de funcionamento AutoTAR. [14]

A estrutura de funcionamento do AutoTAR é descrita no diagrama da Figura 2. No Passo 1 é feito uma busca por um documento relevante, podendo-se utilizar uma pesquisa *ad-hoc* (pesquisa realizada com base na consulta do usuário). Outra alternativa, a partir da consulta, é criar uma semente sintética rotulada como "relevante"(Passo 2). A geração sintética é feita a fim de possibilitar a criação do conjunto de treinamento inicial do classificador. Em seguida, cada lote de documentos tem um tamanho B, iniciando primeiramente em 1 (Passo 3) e aumentando em 10 documentos a cada ciclo(Passo 10). A partir do Passo 4 se inicia um primeiro ciclo, na qual é adicionado documentos aleatórios ao conjunto de treinamento considerando-os como não relevantes para compor o treinamento. Em seguida, o modelo (Passo 5) é treinado e aplicado ao conjunto dos documentos não rotulados a fim de ranqueá-los. Após retirar os documentos aleatórios colocados no conjunto de treinamento e com o modelo treinado, é necessário selecionar B documentos que estão melhor ranqueados (Passo 7) e rotulá-los a fim de identificar os mesmo como "relevante" ou não "relevante" (Passo 8). Como dito anteriormente, esses documentos rotulados irão para o conjunto de treinamento (Passo 9). Esse ciclo iniciado no Passo 4 continua até que se tenha um número suficiente de documentos relevantes ou toda base de dados seja rotulada (Passo 11).

Já em [6] é proposto uma melhoria AutoTAR para melhorar a escalabilidade em grandes conjuntos de dados. O objetivo do S-CAL é oferecer uma alta revocação em grandes bases de dados com reduzido esforço do usuário. O trabalho tem como foco bases de dados voltadas para *eDiscovery(Eletronic Discovery)* que nada mais é do que a busca por documentos relevantes para processos judiciais. Para se atingir tal objetivo essa nova versão traz mudanças no funcionamento do *AutoTAR*. A principal diferença é que são rotuladas apenas pequenas amostras de cada lote sucessivo de documentos e o processo se faz escalar até que se esgote a coleção. Com isso se reduz o esforço do usuário, não sendo mais necessário rotular todo o lote de documentos a cada rodada.

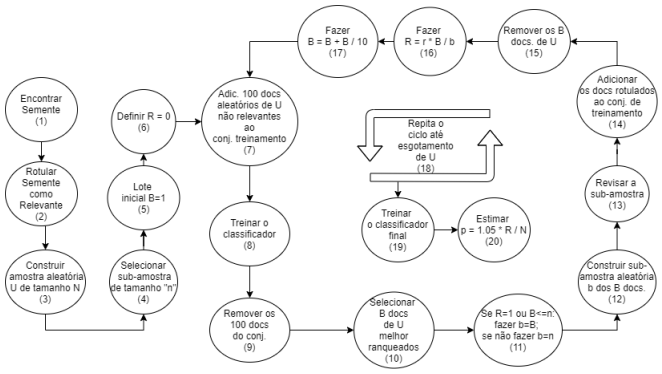

Figura 3. Estrutura de funcionamento S-CAL [6]

A estrutura de funcionamento do S-CAL é descrita na Figura 3). Primeiramente, é construído um documento relevante sintético para o processo de semente (Passo 1 e Passo2) e o tamanho do lote de documentos começa em 1 (Passo 5). Porém o *S-CAL* traz mudanças relacionadas principalmente ao tamanho do lote durante o ciclo. Como mostrado no Passo 3, é construída uma amostra aleatória "U"da população de documentos (ao invés de utilizar a base inteira), e dessa amostra é selecionado uma sub-amostra "n"(Passo 4). Ao conjunto de treinamento também se adiciona 100 documentos aleatórios, porém vindo da amostra "U". É então treinado o

classificador (Passo 7 e 8) e aplicado nele o conjunto dos documentos não rotulados a fim de ranqueá-los. Então são removidos os 100 documentos adicionados aleatoriamente no Passo 9.

O lote de revisão B aumenta em 10% a cada ciclo (Passo 17) e quando B fica maior do que ”n”ou ”R”(iniciado em 0 no Passo 6 e calculado novamente no Passo 16) é selecionado apenas uma sub-amostra aleatória de tamanho ”n”dos documentos de B. Dessa forma, não é mais necessário rotular todo o lote (Passos 10, 11, 12 e 13). Essa sub-amostra continua a ser adicionada ao conjunto de treinamento a cada ciclo (Passo 14). Os documentos ”B”de ”U”também são removidos mesmo não sendo completamente rotulados (Passo 15). Ao final do ciclo, onde ”U”estará esgotado (Passo 18) é então treinado o classificador pela última vez (Passo 19), estimando um valor ”p”(Passo 20) que corresponde ao limite ou ponto final, delimitando uma parte do ranque do classificador para mostrar ao usuário.

Em [12] é proposto um método baseado em aprendizado ativo, chamado de *Fast2*, explorando técnicas para a construção de semente. São propostas novas estratégias com o intuito de evoluir o trabalho anterior [4] para a busca sistemática na literatura no contexto do HIRE. O diferencial deste trabalho é que ele apresenta novas formas de abordar os desafios relacionados a geração de treinamento inicial, o erro humano no processo de revisão e o ponto de parada para o método. O FAST2 traz uma abordagem baseada em aprendizado ativo com utilização de métodos para construção da semente inicial e, semelhante ao AutoTar, utiliza uma abordagem incremental na qual a cada ciclo um lote de documentos é avaliado. Tal lote é treinado utilizando o algoritmo SVM, empregando uma estratégia de predição de erros humanos e um estimador de revocação. Na versão utilizada neste trabalho ao invés do estimador é utilizado o método do joelho [13]. Esse método é utilizado para detectar quando a curva de revocação estabiliza, possibilitando que o processo possa ser finalizado sem desperdiçar esforço do usuário.

Já o método proposto em [14] tem como foco a geração de semente inicial para a construção do treinamento. A abordagem tem como base dois módulos principais: (A) um gerador de pseudo-documento que leva em consideração a informação da semente para pré-treinar uma rede neural;(B) um módulo de autotreinamento com documentos reais não rotulados utilizando a rede treinada pelos pseudo-documentos. Nos experimentos realizados foi identificado que a abordagem tem uma performance significativamente melhor do que os métodos bases(TF-IDF, LDA e etc [10]). No entanto um problema é a falta de integração entre as informações diferentes das sementes. Se isso for resolvido pode impulsionar ainda mais os resultados positivos.

O método aqui proposto explora a geração de treinamento inicial (semente) utilizando BM25 e adicionando a aprendizagem ativa com regras de associação como diferencial.

IV. Proposta S-CAL++

O método S-CAL utiliza como documento inicial relevante (ou semente), a consulta realizada pelo usuário. No entanto, não há garantia de que uma semente sintética contenha termos similar aos documentos relevantes presentes na base de dados. Neste trabalho, é proposta uma nova abordagem para seleção da semente combinando a criação de ranque e a aprendizagem ativa.

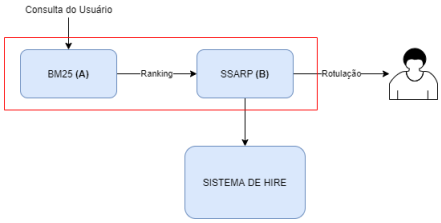

Figura 4. Proposta gerador de semente

A Figura 4 ilustra uma visão geral da proposta (destacando em vermelho a contribuição). A 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. Assim, o método SSAR (B) é aplicado para remover os documentos redundantes do ranque, ou seja, evitar que o usuário receba documentos similares e com informações não relevantes. Desta forma, com uma semente contendo características similares aos demais documentos relevantes, a convergência do método S-CAL pode ser alcançada antes, diminuindo o esforço de rotulação.

O SSAR é essencial para que não se desperdice esforço do usuário ao analisar o ranking do BM25 com documentos não relevantes. Dado que a consulta do usuário pode ser pouco informativa, pode ser necessário percorrer um substancial número de documentos até que se encontre relevantes. No entanto, o SSAR é um método custoso em termos computacionais, devido a necessidade de recalcular a cada documento rotulado a informatividade daqueles que ainda não foram rotulados (conforme descrito na seção II). Devido a isto, o método proposto fornece ao SSAR apenas lotes de N documentos do ranque, diminuindo consideravelmente o custo de processamento do mesmo. Caso dentro de um lote não seja encontrado um documento relevante, é então selecionado os próximos N, usando a ordem do ranque, até encontrar um documento relevante.

O funcionamento detalhado do S-CAL++ é ilustrado a partir de um fluxograma na Figura 5. Inicialmente é feito o TF-IDF da consulta do usuário para transformar o texto em um vetor de características (Passo 1). Em seguida, o método BM25 é aplicado em toda a base base de dados (Passo 2) para se criar um ranque a partir da similaridade com a consulta do usuário (Passo 3).

Para que seja possível utilizar o SSAR, é necessário discretizar as *features* (Passo 4). Após, é então selecionado os últimos N documentos do ranking e rotulados como negativos (Passo 5). Esses N documentos são adicionados ao conjunto de treinamento para que o SSAR saiba quais documentos tem menos chance de serem relevantes e evitar enviar para o usuário (Passo 6). Em seguida, se inicia a seleção de documentos

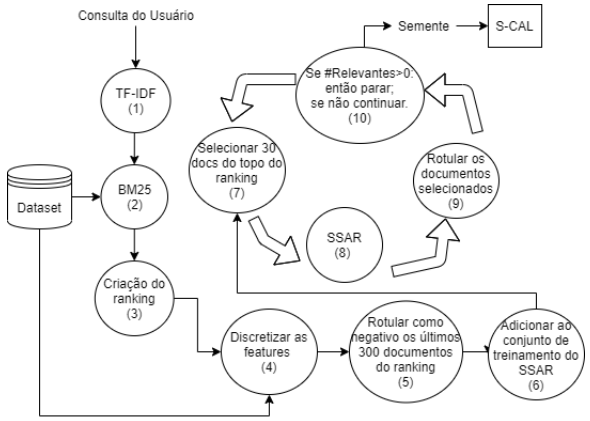

Figura 5. Estrutura de funcionamento do gerador de semente

para análise pelo SSAR, sendo selecionado os N documentos do topo do ranque (Passo 7). Esses documentos são processados pelo SSAR para que seja identificado os documentos não redundantes (Passo 8). Os documentos selecionados são então rotulados pelo usuário (Passo 9). Se pelo menos um dos documentos rotulados pelo usuário for relevante, então o processo de seleção de semente finaliza; caso contrário, será executado novamente o Passo 7 escolhendo os próximos N documentos do ranking para ser processado pelo SSAR. O ciclo será executado até que se encontre um documento relevante. É importante salientar que o SSAR tem como função evitar que todos os N documentos sejam submetidos para o usuário reduzindo assim o esforço para se encontrar a semente inicial.

V. Experimentos

Nesta seção, serão descritos aos experimentos realizados com objetivo de avaliar e comparar a abordagem proposta. Primeiramente será apresentado a base de dados e como ela é configurada, quais são os tópicos e suas consultas. Em seguida, será apresentado como foi implementada extração de características, a configuração dos experimentos, e as métricas aplicadas. Em Por fim, são apresentados e discutidos os resultados encontrados nos experimentos.

A. Base de Dados

A coleção de documentos utilizada é originaria da Tarefa 2 do "CLEF 2017 e-Health" [15]. O conjunto de dados é composto por um total de 125.464 documentos, contendo 20 consulta do usuário.

Na tabela V-A encontra-se uma descrição dos tópicos com o número de documentos relevantes e a prevalência (taxa de documentos relevantes perante à base completa). A prevalência dos tópicos é baixa e observa-se que alguns tópicos apresentam uma quantidade extremamente baixa de documentos relevantes, como, por exemplo, os tópicos 3, 4 e 17 com 11, 7 e 2 documentos relevantes respectivamente, impondo uma dificuldade ainda maior na tarefa de HIRE.

Além dos documentos, o conjunto de dados necessita de outros dois arquivos adicionais. O primeiro contém os julgamentos que detém a informação sobre quais documentos são

Tópico	# Relevantes	Prevalência
$tr0$	95	0.076 %
$tr1$	123	0.098 %
$tr2$	274	0.218 %
$tr3$	11	0.009 %
$tr4$	7	0.006 %
$tr5$	73	0.058 %
$tr6$	162	0.129 %
$tr7$	122	0.097 %
$tr8$	144	0.115 %
$tr9$	78	0.062 %
$tr10$	117	0.093 %
$tr11$	76	0.061 %
$tr12$	39	0.031 %
$tr13$	32	0.026 %
$tr14$	48	0.038 %
$tr15$	215	0.171 %
$tr16$	113	0.090 %
$tr17$	2	0.002 %
$tr18$	619	0.493 %
$tr19$	454	0.362 %

Tabela III
DESCRIÇÃO DE CADA TÓPICO DA BASE DE DADOS CLEF 2017.

positivos para um determinado tópico. Este arquivo segue o mesmo formato como descrito em [**?**]. O segundo abrange as consultas para seus respectivos tópicos. Por exemplo, na tabela V-A é descrito um exemplo de consulta referente ao tópico 19. Nesse caso *"Urine tests for Down syndrome screening"* será a primeira informação utilizada para a busca por documentos relevantes.

Tópico	Consulta
$tr19$	Urine tests for Down syndrome screening.

Tabela IV
EXEMPLO DE CONSULTA PARA O TÓPICO $tr0$

B. Extração de características

A base de dados utilizada encontra-se em formato texto, consequentemente faz-se necessário o uso de técnicas de extração de características que permitam usufruir ao máximo dos algoritmos de aprendizado e ranqueamento empregados. Vale considerar que devido ao fato do algoritmo S-CAL compreender o núcleo do algoritmo S-CAL++, os métodos de extração de características de ambos são muito similares, como descrito a seguir.

Primeiramente, utilizou-se da técnica *Bag of Words* (BoW) para uma conversão de texto para números. Para aumentar a eficácia e versatilidade das características, recorreu-se ao uso de outra técnica: *Term Frequency-Inverse Document Frequency* (TF-IDF). Também, empregou-se o algoritmo SVD (*Singular Value Decomposition*) cujo objetivo é reduzir a dimensão dos atributos, para possibilitar permitir a discretização dos mesmos.

C. Configuração dos experimentos

Os experimentos foram realizados com objetivo de avaliar o comportamento do método proposto S-CAL++ em relação

ao método S-CAL. A intuição é avaliar se a nova abordagem para a geração do treinamento inicial causa impactos positivos no processo.

Os algoritmos foram executado 5 vezes, considerando sempre todos os tópicos. Com exceção dos tópicos $tr3$, $tr4$ e $tr17$ cujo número de relevantes demonstrou-se ser substancialmente baixo (apenas 2 documentos relevantes no tópico $tr17$ e ambos os métodos não convergiram. Assim sendo, os resultados foram tomados como sendo a média de todas estas execuções.

D. Métricas

No intuito de validar o desempenho do método proposto, buscou-se por métricas que resultassem em informações conclusivas no contexto da tarefa de HIRE. A principal delas, a revocação ou *recall*, consiste na relação entre os documentos relevantes encontrados e o total de documentos relevantes presentes na base [1], cuja representação matemática é exposta na Equação 5.

$$\text{Revocação} = \frac{\text{\# relevantes retornados}}{\text{\# total de relevantes}} \tag{5}$$

Nos experimentos realizados aplicou-se as métricas revocação e esforço de rotulação. O esforço de rotulação trata do valor absoluto de documentos manualmente rotulados pelo usuário do sistema, durante todo o processo.

Analisando a revocação e o esforço, é possível inferir a qualidade do ranque final retornado para usuário - no sentido de quantos documentos retornados que são de fato relevantes. De modo geral, almeja-se atingir uma revocação elevada e ao passo que o esforço de rotulação deve ser mínimo.

Para comparar estatisticamente os valores de revocação foi utilizado testes de significância estatística (teste-t) com um intervalo de confiança de 95%.

E. Seleção da semente inicial

Neste experimento o objetivo é avaliar o custo da identificação da semente do S-CAL++ em comparação com a geração sintética (S-CAL). A Figura 6 ilustra o esforço de rotulação para se identificar documentos positivos. Pode-se notar que são utilizadas menos de 30 documentos rotulados em todos os tópicos para encontrar a semente correta. Com exceção dos tópicos $tr3$, $tr4$ e $tr17$ na qual ambos os métodos não foram capazes de encontrar uma semente positiva (documento relevante). Isso é devido ao baixo número de documentos relevantes em tais tópicos, apenas 2, 7 e 11 de um total de 125.464 documentos, respectivamente.

Em média o custo de rotulação foi de 22 documentos para se produzir a semente. Isso demonstra uma redução de cerca de 28% no número de documentos para se identificar a semente sem prejudicar a qualidade do processo em relação ao S-CAL. No tópico $tr8$, por exemplo, são rotulados cerca de 80 documentos até encontrar o primeiro relevante no S-CAL, já a abordagem proposta demanda de somente 19 documentos rotulados. Dessa forma, além do beneficio de encontrar o primeiro documento relevante com menor esforço de rotulação, o processo de aprendizagem do modelo não fica comprometido conforme veremos nas próximas seções, garantindo a eficiência no processo.

	S-CAL	S-CAL ++
tr0	0.93	0.93
tr1	0.70	0.75
tr2	0.65	0.66
tr5	0.92	0.93
tr6	0.70	0.76
tr7	0.75	0.77
tr8	0.63	0.86
tr9	0.94	0.96
tr10	0.83	0.87
tr11	0.96	0.96
tr12	0.84	0.82
tr13	0.93	0.96
tr14	0.98	0.98
tr15	0.76	0.77
tr16	0.58	0.67
tr18	0.38	0.42
tr19	0.46	0.50

Tabela V
REVOCAÇÃO NO PROCESSO DE APRENDIZAGEM (TREINAMENTO) DO S-CAL.

F. Análise da revocação no processo de treinamento

Um ponto importante que deve ser salientado é a evolução da revocação ao longo do processo de treinamento. Ou seja, quanto antes o usuário obtiver os documentos relevantes menos tempo o mesmo irá gastar para realizar a sua consulta. Além disso, se o usuário só acessar documentos não relevantes no início do processo pode acabar desistindo e desperdiçando o seu esforço inicial.

Neste experimento o objetivo é avaliar se o método proposto de seleção de semente auxilia na recuperação de documentos relevantes durante o processo de treinamento do S-CAL. Dessa forma, foi executado o método S-CAL com a semente sintética e e o S-CAL com a semente gerada pela abordagem proposta (S-CAL++).

A tabela V-F detalha a revocação de cada tópico do S-CAL e o S-CAL++. Como pode ser observado o S-CAL++ é equivalente ou superior na grande maioria dos tópicos. De acordo com a tabela, pode-se perceber que semente gerada pelo S-CAL++ auxilia a recuperar mais documentos relevantes em relação a sementes sintética, chegando a um ganho de até 9%. A única exceção foi o tópico $tr12$ que apresentou uma perda de 2%. Futuros experimentos serão realizados nos próximos trabalhos para identificar o motivo de tal perda.

A Figura 7 consolida os resultados, mostrando a diferença na evolução da curva da revocação média e do número de documentos rotulados ao longo do processo de treinamento do S-CAL de todos os tópicos. Como pode ser observado, o ganho de revocação acontece antes, assim como o joelho da curva, onde o método já não encontra mais nenhum documento relevante. Dessa forma, o SCAL++ tem uma revocação superior na grande maioria dos tópicos após a finalização do processo de aprendizagem.

Por fim, este experimento demostrou que a geração de semente do S-CAL++ foi capaz de auxiliar no processo de treinamento, recuperando em média mais documentos relevantes, que o método comparado.

Figura 6. Seleção da semente

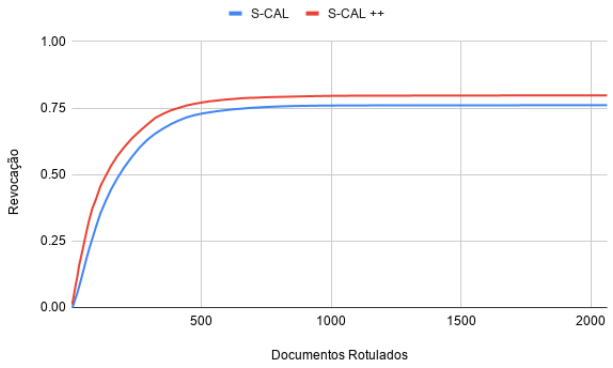

Figura 7. Curvas de crescimento médio da revocação no processo de treinamento

G. Custo de rotulação vs. revocação

Neste experimento o objetivo é de avaliar o custo de rotulação final (após o processo de treinamento e da geração do ranque final) e a revocação ao acoplar o método proposto, S-CAL++, em relação ao S-CAL.

Na Figura 8 são consolidados o esforço de rotulação (Eixo Y esquerda), a revocação (Eixo Y direita) e o tópico com com o respectivo número de documentos relevantes no Eixo X. A figura é interessante para compararmos o custo de rotulação em comparação com a revocação. O S-CAL e o S-CAL++ apresentam uma revocação equivalente na grande maioria dos tópicos, no entanto, o custo de rotulação varia consideravelmente. O S-CAL é melhor em 7 tópicos já o S-CAL++ em 9, tendo apenas 1 com empate. Um ponto a destacar é a substancial redução no custo de rotulação do tópico TR18 (619 relevantes) e o TR9 (78 relevantes), onde o S-CAL ++ tem uma eficiência maior.

Para se ter uma visão geral do custo do processo, na Tabela V-G é apresentada a média entre todos os tópicos em relação

	ESFORÇO	Revocação
S-CAL	5846	96%
S-CAL ++	**4777**	96%

Tabela VI
MÉDIA DE ESFORÇO DE ROTULAÇÃO E REVOCAÇÃO

a revocação e o custo de rotulação. Os dois métodos tem a mesma média de revocação final de 96%. Estatisticamente, conforme o teste-t, os dois métodos são iguais em relação a revocação em grande maioria dos tópicos, comprovando a premissa de que os dois métodos, S-CAL e S-CAL++, buscam sempre atingir 95% de revocação, não importando o quão antes se encontre o primeiro documento relevante. Ou seja o método tem como objetivo principal obter uma revocação miníma conforme indicado pelo usuário, deixando em segundo plano o custo de rotulação.

O S-CAL++ tem uma redução de cerca de 1000 documentos no esforço de rotulação, o que representa 18% de diminuição no custo de rotulação. Isso pode ser explicado devido ao ranque final gerado, que é maior no caso do S-CAL.

Por fim, estes experimentos demostraram que a seleção de uma semente mais informativa auxilia substancialmente na redução do esforço do usuário. Além disso, foi possível mensurar na experimentação que o S-CAL++ é capaz de aprimorar o processo de treinamento, recuperando antes documentos relevantes se comparado ao S-CAL.

VI. CONCLUSÃO

O método de geração de semente proposto, chamado de S-CAL++, mostrou-se promissor na tarefa de reduzir o esforço de rotulação do usuário. Foi possível uma redução de até 18% no esforço empenhado pelo usuário. Além disso, foi possível constatar que a geração de uma semente informativa auxilia no treinamento a recuperar documentos relevantes antes se comparado ao método base. Todavia, a investigação de técnicas

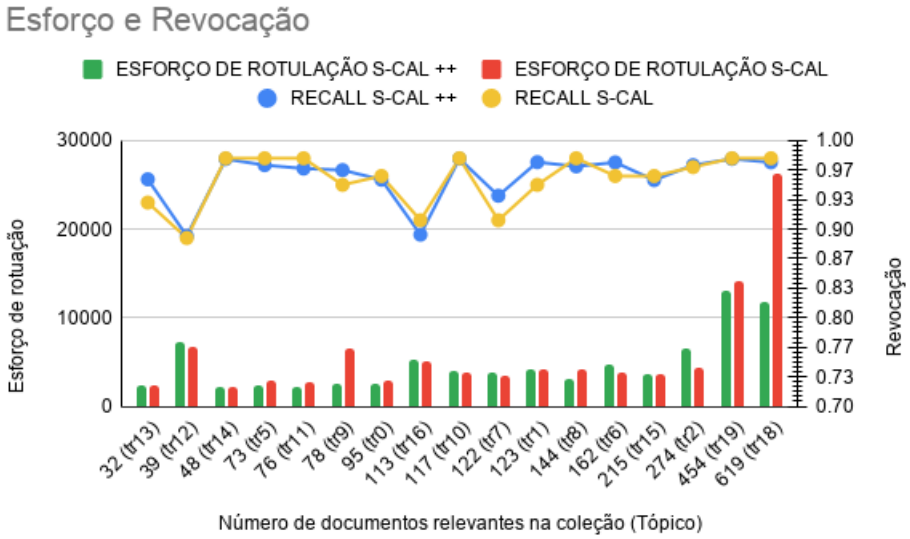

Figura 8. Revocação vs. custo de rotulação.

mais efetivas para geração da semente faz-se necessária, de modo à refinar os resultados obtidos e possibilitar a busca pela semente em todos na qual as sementes são raras.

Nos próximos trabalhos pretende-se explorar técnicas de aprendizado profundo, que mostram-se poderosas para o processamento de dados não estruturados. Assim, compete examinar o impacto de tais técnicas para o método.

REFERÊNCIAS

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2018.

[2] A. Roegiest, "On design and evaluation of high-recall retrieval systems for electronic discovery," 2017.

[3] M. R. G. G. V. Cormack, "Autonomy and reliability of continuous active learning for technology-assisted review," 2015.

[4] Z. Yu, N. A. Kraft, and T. Menzies, "Finding better active learners for faster literature reviews," *Empirical Software Engineering*, 2018. [Online]. Available: https://doi.org/10.1007/s10664-017-9587-0

[5] B. Settles, "Active learning literature survey," University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009.

[6] M. G. G. V. Cormack, "Scalability of continuous active learning for reliable high-recall text classification," 2016. [Online]. Available: http://dx.doi.org/10.1145/2983323.2983776

[7] M. Fisichella, R. Kawase, and U. Gadiraju, "Automatic classification of documents in cold-start scenarios," 2009.

[8] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker, "Clef 2017 technologically assisted reviews in empirical medicine overview," in *CEUR Workshop Proceedings*, vol. 1866, 2017, pp. 1–29.

[9] R. Silva, M. A. Gonçalves, and A. Veloso, "Rule-based active sampling for learning to rank," *Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science*, p. 240–255, 2011.

[10] S. Qaiser and R. Ali, "Text mining: Use of tf-idf to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, no. 1, p. 25–29, 2018.

[11] Y. Goodfellow and A. Courville, *Machine Learning Basics*. The MIT Press, 2016, p. 95–160.

[12] Z. Yu and T. Menzies, "Fast2: An intelligent assistant for finding relevant papers," *Expert Systems with Applications*, vol. 120, pp. 57 – 71, 2019.

[13] G. V. Cormack and M. R. Grossman, "Engineering quality and reliability in technology-assisted review," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 75–84.

[14] Y. Meng, J. Shen, C. Zhang, and J. Han, "Weakly-supervised neural text classification," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 983–992.

[15] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker, "Clef 2017 technologically assisted reviews in empirical medicine overview," in *CEUR Workshop Proceedings*, vol. 1866, 2017, pp. 1–29.

Matheus Vinícius Todescato é discente do curso de Ciência da Computação pela Universidade Federal da Fronteira Sul.

Jean Carlo Hilger é discente do curso de Ciência da Computação pela Universidade Federal da Fronteira Sul.

Guilherme Dal Bianco é doutor pela Universidade Federal do Rio Grande do Sul (2012) e, atualmente, é professor adjunto da Universidade Federal da Fronteira Sul (UFFS). Seus interesses em pesquisa são: extração de informações, tratamento de consultas, e integração de dados.