UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

MYLLENE FERREIRA DA SILVA
Orientador: Guilherme Tavares de Assis

GERAÇÃO SEMIAUTOMÁTICA DOS CONJUNTOS INICIAIS DE TERMOS UTILIZADOS EM PROCESSOS DE COLETA TEMÁTICA DE PÁGINAS DA *WEB* BASEADA EM GÊNERO E CONTEÚDO

Ouro Preto, MG
2023

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

MYLLENE FERREIRA DA SILVA

GERAÇÃO SEMIAUTOMÁTICA DOS CONJUNTOS INICIAIS DE TERMOS UTILIZADOS EM PROCESSOS DE COLETA TEMÁTICA DE PÁGINAS DA *WEB* BASEADA EM GÊNERO E CONTEÚDO

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Guilherme Tavares de Assis

Ouro Preto, MG
2023

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Myllene Ferreira da Silva

Geração semiautomática dos conjuntos iniciais de termos utilizados em processos de coleta temática de páginas da Web baseada em gênero e conteúdo

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 25 de Agosto de 2023.

Membros da banca

Guilherme Tavares de Assis (Orientador) - Doutor - Universidade Federal de Ouro Preto
Andrea Gomes Campos Bianchi (Examinadora) - Doutora - Universidade Federal de Ouro Preto
Jadson Castro Gertrudes (Examinador) - Doutor - Universidade Federal de Ouro Preto

Guilherme Tavares de Assis, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 25/08/2023.

Documento assinado eletronicamente por **Guilherme Tavares de Assis**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 25/08/2023, às 12:25, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0577031** e o código CRC **55F97814**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho à minha mãe que sempre esteve comigo, independente da dificuldade.

Agradecimentos

Agradeço imensamente a minha mãe, por ser essa mulher batalhadora e que nunca deixou de acreditar em mim; minhas pequeninhas Ana Vitória e Pretinha por serem meu anjo da guarda e por todo amor. Em especial, faço um agradecimento ao meu pai, mesmo que não esteja mais aqui, sempre apoiou e incentivou meus estudos.

Agradeço também à vocês que foram peças importante em toda minha trajetória: João Vitor, Júlia Eduarda, Sabrina Suellen e Vitor Hugo - sem vocês nada disso estaria acontecendo.

Agradeço ao meu orientador, Guilherme Tavares de Assis, por toda a paciência, dedicação e todo carinho até aqui. E por fim, agradeço todos meus amigos que estiveram presentes ao longo dessa caminhada e minha amada casa - República Flor de Liz. Com certeza, todos foram essenciais e que levarei todos em meu coração.

Resumo

Os coletores temáticos apresentam o propósito maior de coletar páginas da *Web* que sejam relevantes a um tópico de interesse específico do usuário, sendo, com isso, importantes para uma grande variedade de aplicações. Nesse contexto, foi proposta e desenvolvida proposta em ASSIS et al.; JÚNIOR et al.(2009; 2008; 2007; 2021) uma abordagem para coleta temática em que o tópico de interesse do usuário pode ser expresso por termos que descrevem o conteúdo e o gênero das páginas da *Web* desejadas. Tal abordagem possibilita a construção de coletores temáticos que realizam processos de coleta eficazes, eficientes e escaláveis, caso tais termos de conteúdo e gênero, fornecidos como dados de entrada, sejam bem significativos ao tópico de interesse em questão. Durante a experimentação conduzida para validar a abordagem proposta, para cada tópico de interesse considerado, especialistas definiram os termos de gênero e conteúdo utilizados nos processos relativos de coleta. Assim, devido à importância de se definir termos significativos para a realização de processos de coleta relativos ao tópico de interesse desejado, faz-se necessário especificar os conjuntos iniciais de termos de gênero e conteúdo de uma forma mais precisa, segura e ágil. Neste contexto, este trabalho propõe desenvolver uma estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo a serem usados em processos de coleta temática baseada em gênero e conteúdo. Por meio de experimentos realizados, considerando a estratégia proposta e desenvolvida, foram obtidos resultados de precisão satisfatórios quanto aos termos gerados para distintos tópicos de interesse , chegando a níveis médios de precisão de 75% e 80% ao considerar 12 termos de gênero e conteúdo, respectivamente, retornados pela estratégia. No que tange às limitações dessa estratégia, é perceptível que a precisão ponderada é menor quando os documentos são inseridos no formato de *URL*. Isso acontece devido ao fato de que cada *URL* possui uma configuração específica em relação às tags *HTML* utilizadas em sua construção.

Palavras-chave: Coleta temática de páginas da web. Geração semiautomática de termos. Termos de gênero. Termos de conteúdo.

Abstract

Thematic collectors serve the overarching purpose of gathering web pages that are relevant to a user's specific topic of interest. In doing so, they prove valuable for a wide range of applications. In this context, an approach for thematic collection was proposed and developed in ASSIS et al.; JÚNIOR et al.(2009; 2008; 2007; 2021). This approach enables the construction of thematic collectors that conduct efficient, effective, and scalable collection processes. It allows users to express their topic of interest using terms that describe the content and genre of the desired web pages. During the experimentation carried out to validate the proposed approach, experts defined the genre and content terms used in the collection processes for each considered topic of interest. Consequently, due to the significance of defining meaningful terms for conducting collection processes relevant to the desired topic of interest, it becomes necessary to specify the initial sets of content and genre terms in a more precise, secure, and expedient manner. In this context, this work proposes the development of a strategy for semi-automatically generating initial sets of genre and content terms to be used in thematic collection processes based on genre and content. Through conducted experiments, considering the proposed and developed strategy, satisfactory precision results were achieved for the generated terms across various topics of interest. This included average precision levels of 75% and 80% when considering 12 genre and content terms, respectively, returned by the strategy. Regarding the limitations of this strategy, it is evident that the weighted precision is lower when documents are input in the URL format. This is due to the fact that each URL possesses a specific configuration with respect to the HTML tags used in its construction.

Keywords: Thematic collection of web pages. Semi-automatic generation of terms. Genre terms. Content terms.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

DECOM	Departamento de Computação
UFOP	Universidade Federal de Ouro Preto
XML	eXtensible Markup Language
EMP	Processo de Extração de Metadados
PDF	Portable Document Format
HTML	Linguagem de Marcação de HiperTexto
URL	Uniform Resource Locator

Sumário

1 Introdução

No atual momento, de acordo com AHLGREN (2023), 8.5 bilhões de consultas de pesquisa são realizadas todos os dias em todo o mundo. Ademais, houve em janeiro de 2023, cerca de 5.6 bilhões de usuários da internet sendo que o tempo médio global de uso da internet por usuário é de sete horas diárias, e existiam na internet, no final de 2022, mais de 1.97 bilhão de sites. Em virtude desses fatores, novas técnicas de recuperação de informação tornam-se fundamentais para facilitar a coleta de conteúdo da *Web* a partir de tópicos de interesse. Para este propósito, como visto em BHATT et al. (2015), máquinas de busca apresentam ferramentas básicas para se buscar algo de interesse na internet a partir de repositórios que são gerados por meio de coletores *Web* (*Web Crawler*) tradicionais, iniciando a partir das páginas-semente e avançando por páginas referenciadas nelas, com objetivo de alcançar o maior número possível de páginas.

Entretanto, segundo COSTA et al. (2017), máquinas de busca de propósito geral não resolvem bem o problema de localizar páginas da *Web* referentes a um tópico específico, já que as coleções de páginas geradas por elas são bem volumosas e, geralmente, as consultas dos usuários são curtas envolvendo pouca informação. Neste contexto, coletores temáticos (CHAKRABART et al. (1999); MANE et al. (2017)) servem para gerar coleções de páginas menores e restritas, já que apresentam o propósito maior de coletar páginas que sejam, da melhor forma possível, relevantes a um tópico ou interesse específico do usuário, a partir de uma especificação mais detalhada do que se deseja coletar.

Dessa maneira, visando a realização de processos eficazes e eficientes de coleta temática, ASSIS et al. (2009) propuseram e desenvolveram uma abordagem voltada para atender situações específicas. De uma forma geral, tal trabalho desenvolvido teve, como objetivo principal, estabelecer uma estrutura que permita a construção de coletores temáticos eficazes, eficiente e escaláveis, sem a necessidade de um treinamento a priori ou qualquer tipo de pré-processamento. Especificamente, a abordagem para a coleta temática proposta é útil em situações onde um tópico de interesse possa ser expresso por meio de dois conjuntos distintos de termos: o primeiro descrevendo aspectos de gênero (estilo) das páginas desejadas e o segundo referente ao assunto ou conteúdo (assunto) descrito nessas páginas.

Várias estratégias de coleta temática (PAVANI and SAJEEV (2017); KUMAR et al. (2018); HOSSEINKHANI et al. (2021)) utilizam classificadores de texto para determinar a relevância de uma página em relação a um tópico ou interesse específico do usuário, com um custo adicional para serem treinados; ademais, devido à generalidade das situações em que essas

estratégias são aplicadas, elas alcançam baixos níveis de revocação[1] e precisão[2], geralmente entre 40% e 70%.

1.1 Justificativa

Para que abordagens para coleta temática mais eficazes, como a baseada em gênero e conteúdo ASSIS et al.(2009; 2008; 2007), possam ser aplicadas, é necessário, geralmente, especificar termos (dados de entrada) que sejam significativos ao tópico de interesse: a má especificação de termos, necessários para a realização de um processo de coleta, pode levar a uma eficácia insatisfatória desse processo. Particularmente, na experimentação feita para validar a abordagem proposta para coleta temática baseada em gênero e conteúdo ASSIS et al.(2009; 2008; 2007), especialistas definiram os termos de gênero e conteúdo utilizados nos processos de coleta. Foi observada, na especificação dos conjuntos iniciais de termos para todos os tópicos de interesse, uma certa dificuldade dos especialistas em especificá-los com segurança e agilidade. Nesse caso, a geração semiautomática dos conjuntos iniciais de termos pode resultar em uma definição mais precisa e segura dos mesmos, resultando em um desempenho aprimorado dos processos de coleta em termos de eficácia. Isso significa que mais páginas relevantes ao tópico de interesse desejado podem ser localizadas e determinadas corretamente.

Ademais, outra importante justificativa para a realização desse trabalho é que existem poucas implementações de coletores temáticos, disponíveis na Web, que sejam bem quistos e utilizados. Muitos coletores temáticos foram produzidos com propósitos específicos ou foram descontinuados por falta de uso por usuários. Dessa forma, gerando uma versão mais eficaz de um coletor, que segue a abordagem para coleta temática baseada em gênero e conteúdo, e que trata a especificação semiautomática dos conjuntos iniciais de termos necessários para a realização de processos de coleta, o mesmo pode ter uma grande visibilidade acadêmica e comercial.

1.2 Objetivos Geral e Específicos

Este trabalho possui, como objetivo geral, propor, desenvolver e validar uma estratégia eficaz e robusta para a geração de conjuntos de termos iniciais que serão utilizados na execução de processos de coleta temática de páginas da *Web*. Para tanto, foi considerada a abordagem para coleta temática fundamentada em gênero e conteúdo ASSIS et al.(2009; 2008; 2007), onde o tópico de interesse do usuário pode ser descrito por meio de termos que descrevem o gênero e o conteúdo das páginas da *Web* desejadas.

De um modo geral, os objetivos específicos, a serem alcançados neste trabalho, são:

[1] De acordo com BROWNLEE (2020), revocação é uma métrica que consiste na fração de instâncias classificadas como corretas considerando o total de instâncias positivas que poderiam ser geradas.

[2] De acordo com BROWNLEE (2020), precisão é uma métrica que consiste na fração de instâncias classificadas como corretas considerando o total das instâncias classificadas como positivas.

- melhoria da eficácia em processos de coleta temática através da geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo relevantes, por meio de entrada de documentos relacionado ao tema de interesse;
- geração de uma nova versão do coletor existente, que segue a abordagem para coleta temática baseada em gênero e conteúdo, envolvendo a estratégia proposta e validada.

1.3 Método do Trabalho

Visando o alcance do objetivo geral deste trabalho, foi definida e desenvolvida uma arquitetura que descreve o funcionamento de uma estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo a serem utilizados utilizados em processo de coleta temática baseada em gênero e conteúdo.

Baseada em tal arquitetura, a estratégia foi desenvolvida e validada, considerando distintos e atuais tópicos de interesse e a medição da eficácia da estratégia, quanto à qualidade dos termos gerados pela mesma.

1.4 Organização do Trabalho

Os próximos capítulos deste trabalho constituem-se da seguinte forma: no Capítulo 2, é realizada a revisão bibliográfica que consiste no referencial teórico, para dar base ao desenvolvimento do trabalho, e em trabalhos relacionados, que apresentam soluções já desenvolvidas para o problema abordado neste trabalho; no Capítulo 3, é apresentada a metodologia deste trabalho, envolvendo a estratégia proposta para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo; no Capítulo 4, os experimentos práticos realizados são apresentados e os resultados obtidos são analisados; por fim, no Capítulo 5, são descritas conclusões deste trabalho e as perspectivas de trabalho futuro.

2 Revisão de Literatura

Este capítulo apresenta a revisão de literatura feita para a realização deste trabalho. Para tanto, encontra-se organizado da seguinte maneira: a Seção 2.1 aborda a fundamentação teórica necessária ao desenvolvimento deste trabalho e a Seção 2.2 apresenta os trabalhos diretamente relacionados.

2.1 Fundamentação Teórica

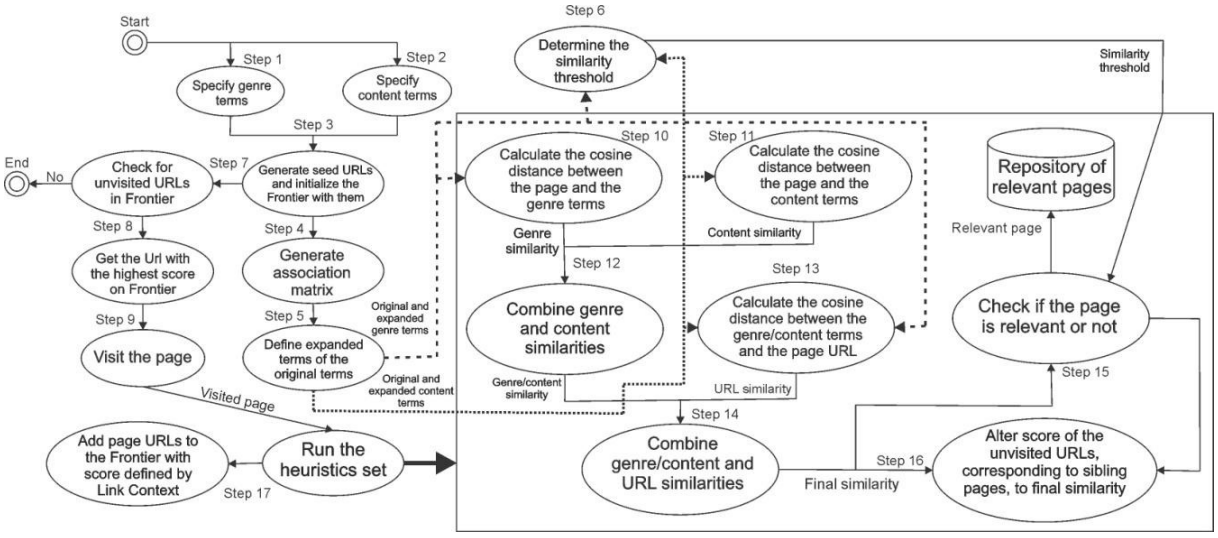
Nesta seção, é apresentado o suporte teórico necessário para o entendimento e o desenvolvimento deste trabalho. A Subseção 2.1.1 refere-se ao coletor temático Yucca, base para a estratégia da geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo proposta deste trabalho. A Subseção 2.1.2 refere-se ao algoritmo YAKE utilizado na estratégia proposta para a geração inicial dos possíveis termos desejados. A Subseção **??** descreve a técnica de *Lemmatização*. E por fim, a Subseção2.3 descreve a métrica utilizada nos experimentos realizados.

2.1.1 Coletor Temático Yucca

A abordagem original para coleta temática de páginas Web baseada em gênero, proposta em ASSIS et al.; JÚNIOR et al.(2009; 2008; 2007; 2021), estabelece um arcabouço que permite a construção de coletores temáticos eficazes, eficientes e escaláveis, que levam em consideração o gênero e o conteúdo das páginas desejadas. Mais especificamente, essa abordagem foi projetada para situações em que um tópico de interesse pode ser descrito por dois conjuntos distintos de termos: o primeiro conjunto expressa o gênero das páginas desejadas e o segundo descreve o conteúdo dessas páginas. Por gênero, de acordo com ASSIS et al., entende-se o tipo, a categoria ou o estilo de texto de documentos específicos; por conteúdo, entende-se como o assunto ou tópico que se deseja coletar.

A partir da abordagem original para coleta temática baseada em gênero e suas melhorias (MANGARAVITE et al.(2012; 2014); SIQUEIRA et al. (2016); COSTA et al. (2017); DINIZ and ASSIS (2018)), foi proposto e desenvolvido por JÚNIOR et al. (2021) um coletor funcional e completo, denominado Yucca, para coleta temática baseada em gênero e conteúdo. A Figura 2.1 apresenta a arquitetura de funcionamento do Yucca.

Figura 2.1 – Arquitetura de funcionamento do Yucca.



Fonte: JÚNIOR et al. (2021).

De acordo com a Figura 2.1, visando um determinado processo de coleta a ser realizado para um tópico específico de interesse, inicialmente são especificados os termos de gênero (Passo 01) e de conteúdo (Passo 02), sendo estas tarefas do usuário.

Em seguida, no Passo 03, são geradas semiautomaticamente as páginas semente utilizando os termos de gênero e conteúdo especificados; tais páginas-semente inicializam a lista de URLs não visitadas, presentes no *Frontier*, com pontuação de prioridade de visita pelo Yucca igual a 1. Considerando as páginas-semente geradas, o Passo 04 gera uma matriz de associação para definição, em seguida (Passo 05), dos termos similares aos termos originais.

Dando sequência, no Passo 06, é especificado e determinado automaticamente o limite de similaridade utilizando os termos especificados pelo usuário nos Passos 01 e 02 e os termos similares no Passo 05. Dando início ao processo de coleta propriamente dito, enquanto há URLs não visitados no *Frontier* (Passo 7), aquela com maior pontuação de visita é desenfileirada do *Frontier* (Passo 08) e a página correspondente é visitada pelo Yucca (Passo 09); esta visita consiste em analisar a relevância da mesma, por meio de um conjunto de heurísticas de cálculo de similaridade, quanto ao tópico específico de interesse.

Assim, nos Passos 10 e 11, são calculadas as distâncias de cosseno entre a página visitada e os termos originais e similares de gênero e conteúdo, respectivamente, combinando e gerando, no Passo 12, a similaridade de gênero e conteúdo. Em seguida, no Passo 13, é calculada a distância de cosseno entre os termos originais e similares de gênero e conteúdo e a URL da página visitada, combinando-a, no Passo 14, com a similaridade calculada de gênero e conteúdo (Passo 12), gerando assim a similaridade final da página visitada em relação ao tópico específico de interesse.

Caso tal similaridade final seja superior ao limite de similaridade gerado automaticamente

(Passo 15), a página visitada é considerada relevante e, assim, é armazenada no repositório de páginas relevantes ao tópico específico de interesse; ademais (Passo 16), de acordo com a política de enfileiramento definida para o *Frontier*, altera-se a pontuação de visita das URLs ainda não visitadas, que sejam correspondentes às páginas irmãs da página visitada, para o valor da similaridade final calculada.

Por fim (Passo 17), não vinculado à execução de heurísticas para cálculo de similaridade, adiciona-se ao *Frontier* as URLs presentes na página visitada com pontuações de visita definidas pela similaridade entre os termos de gênero e conteúdo e os *link contexts* das URLs em questão.

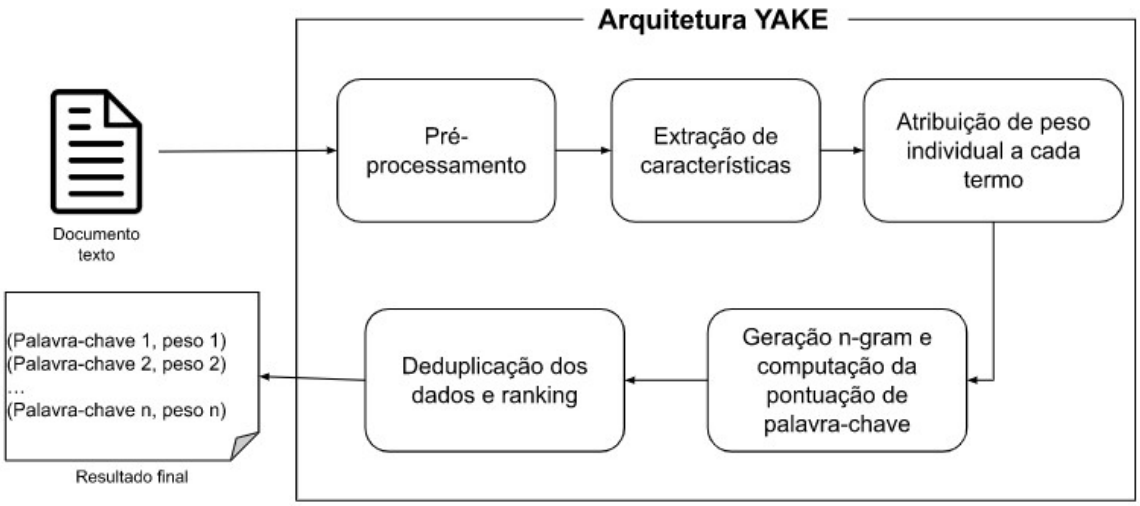
De uma forma geral, observa-se que, para a realização de processos de coleta temática seguindo a abordagem proposta, é preciso especificar muito bem conjuntos de termos de gênero e conteúdo que expressam o tópico de interesse desejado. Nas experimentações realizadas para validar a abordagem proposta, para cada tópico de interesse considerado, especialistas definiram os termos de gênero e conteúdo utilizados nos processos relativos de coleta. Dessa forma, a fim de melhorar o processo de entrada dos conjuntos de termos de gênero e conteúdo necessários para um processo eficaz de coleta relativo ao tema desejado, este trabalho propõe uma estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo.

2.1.2 Algoritmo YAKE

De acordo com CAMPOS et al. (2020), o YAKE é um algoritmo de extração de palavras-chave em documentos de forma automática; consiste em uma abordagem não supervisionada, fazendo uso de recursos textuais estatísticos extraídos de um único documento no intuito de selecionar as palavras-chave mais relevantes de seu texto. As tarefas de ler, extrair e sumarizar termos e palavras-chave de grandes coleções tornam-se desafiadoras e custosas à medida que a informação gerada dessas coleções cresce. A necessidade de automatizar tais tarefas requer o desenvolvimento de sistemas de extração de palavras-chave com a habilidade de identificá-las dentro do texto de forma automática.

De uma forma geral, considerando um documento texto fornecido, a arquitetura de funcionamento do YAKE, ilustrada na Figura 2.2, consiste em cinco etapas principais a serem detalhadas nas subseções seguintes: pré-processamento (Subseção 2.1.2.1); extração de características (Subseção 2.1.2.2); atribuição de peso individual a cada termo (Subseção 2.1.2.3); geração de *n-gram* e computação da pontuação de palavras-chave candidatas (Subseção 2.1.2.4); e deduplicação dos dados e *ranking* (Subseção 2.1.2.5).

Figura 2.2 – Arquitetura de funcionamento do YAKE.



Fonte: CAMPOS et al. (2020)

Neste trabalho, a fim de obter o conjunto inicial de palavras-chave e seus respectivos pesos, o algoritmo YAKE foi utilizado para estabelecer o conjunto inicial de palavras-chave de documentos de entrada, especificados por usuários, e extrair o conjunto inicial a fim de auxiliar nos passos posteriores referente à estratégia proposta (vide Capítulo 3).

2.1.2.1 Pré-processamento

A primeiro etapa consiste em dividir todo documento em frases. O algoritmo YAKE utiliza o segmentador de frases (*segtok*) baseado em regras, no qual os delimitadores serão as pontuações encontradas no texto (LEITNER (2015)). Após a divisão das frases, é aplicada a técnica de *tokenização* que divide as frases em termos individuais sempre que um espaço vazio ou um caractere especial delimitador for encontrado.

2.1.2.2 Extração de características

Na segunda etapa, é realizada uma análise estatística que foca especialmente na estrutura, frequência de termos e co-ocorrência. Segundo CAMPOS et al. (2020), a extração de características consiste em um conjunto de cinco recursos estatísticos do termo candidato à palavra-chave.

1. *Casing* (T_{Case}): localiza termos acrônimos ou termos em que a primeira letra encontra-se em maiúscula e que não são encontrados em início de uma frase.

2. *Term position* ($T_{Position}$): calcula a média das posições da palavra no documento. Desse modo, os termos que se encontram no início do documento tendem a ser mais relevantes do que termos que estão no final do texto.

3. *Term frequency normalization* (TF_{Norm}): indica a frequência dos termos presentes em um documento. Para evitar alta frequências de *stopwords*[1], são conjugados a frequência com a média de todas as frequências dos termos.

4. *Term relatedness to context* (T_{Rel}): quantifica o significado de um termo em relação ao seu contexto. Há a tendência de termos que se encontram à esquerda e à direita do termo candidato serem mais irrelevantes se aparecem constantemente próximos a palavras diferentes, obtendo uma pontuação maior, como o caso das *stopwords*.

5. *Term different sentence* ($T_{Sentence}$): mede a frequência com que um termo aparece em frases diferentes. Para realizar o cálculo, considera o número de frases que contém o termo dividido pelo número total de frases que contém o documento.

2.1.2.3 Cálculo da pontuação de termos

Uma vez que os recursos da etapa anterior são processados, para cada termo presente no documento, é possível gerar a pontuação ou relevância dele no documento a partir da seguinte equação:

$$S(t) = \frac{T_{Rel} \cdot T_{Position}}{T_{Case} + \frac{TF_{Norm}}{T_{Rel}} + \frac{TF_{Sentence}}{T_{Rel}}} \tag{2.1}$$

De acordo com a equação 2.1, quanto menor o resultado, mais chances do termo ser uma palavra-chave; ou seja, uma pontuação menor indica uma alta relevância. Para que isso ocorra, os recursos TF_{Norm} e $T_{Sentence}$ são divididos pelo T_{Rel} para que os termos com maior frequência tenham um peso menor.

2.1.2.4 Geração n-gram e computação da pontuação da palavra-chave

Para estruturar o conjunto final do algoritmo YAKE, deve-se considerar que palavras-chave podem ter mais de uma palavra. Sendo assim, para determinar a pontuação de cada palavra-chave, considera-se uma janela deslizante de tamanho máximo 3 para gerar uma sequência contígua de palavras. Para evitar a formação de sequências não relacionadas em *n-gram*[2], as palavras-chave não podem começar ou terminar com *stopwords*. O peso final da palavra-chave é determinada pela equação:

[1] Stopwords são palavras que podem ser consideradas irrelevantes para um conjunto de palavras. Exemplos: “a”, “e”, “ os”, para, etc.

[2] n-gram trata-se a uma sequência n de palavras. Exemplo: “Buscando” tem 1-gram; na sequência “Buscando termos” temos 2-gram.

$$S(kw) = \frac{\prod_{k \in kw} S(t)}{KF(km) \cdot (1 + \sum_{k \in kw} S(t))} \tag{2.2}$$

Onde *km* representa palavras-chave que contém *n-gram* e *S(kw)* é sua pontuação final. O numerador refere-se ao produto dos pesos de cada palavra presente e o denominador considera a soma dos pesos individuais e as pondera com a frequência da palavra-chave (*KF(kw)*). Uma vez obtidos os pesos, o YAKE irá produzir uma lista de palavras-chave formada até *3-gram*, de modo que quando menor o peso mais relevante será a palavra-chave.

2.1.2.5 Deduplicação dos dados e ranking

No último passo, dada a lista de palavras-chave encontrada na etapa anterior, será realizada a comparação entre os elementos da lista para que não haja redundância ou semelhanças entre si. Para que isso ocorra, uma medida de similaridade de distância é utilizada de modo que, quando são encontradas palavras-chave semelhantes, aquela de pontuação mais baixa é mantida e as restantes são removidas.

2.1.3 Lemmatização

Lematização é uma técnica utilizada no processo de encontrar a forma normalizada de uma palavra, sendo uma etapa de pré-processamento extremamente importante para aplicações de mineração de dados, recuperação de informação e processamento de linguagem natural. É também utilizada no fornecimento de uma forma eficiente para a geração de palavras-chave genéricas para motores de busca (Plisson et al., 2004).

A lematização é semelhante ao *stemming*, porém, não requer a produção do *stem* de uma palavra, requer a substituição do sufixo da palavra, tipicamente, pelo sufixo de outra palavra a fim de gerar a forma normalizada da palavra. Esse processo utiliza análise morfológica e de vocabulário na tentativa de remover finais de palavras flexionados, retornando-as à sua forma de dicionário. Além disso, a lematização também auxilia no processo de casar sinônimos pelo uso de tesauros, de modo que, por exemplo, pesquisas com palavras como “carro” também resultem em “automóvel”, entre outros sinônimos (Balakrishnan and Lloyd-Yemoh, 2014).

O emprego de ambas técnicas, *stemming* e lematização, presta um importante papel para aumentar a relevância e *recall* de um sistema de recuperação de informação. Quando usadas, essas técnicas reduzem o número de índices utilizados pelo sistema, uma vez que o sistema utilizará um único índice para apresentar um número de palavras semelhantes, que apresentam a mesma raiz ou *stem* (Balakrishnan and Lloyd-Yemoh, 2014).

2.1.4 Precisão ponderada

Para avaliação dos experimentos realizados quanto à eficacia da estratégia definida neste trabalho, levando em conta a posição de cada termo dentro de um determinado conjunto, a métrica mais apropriada ao escopo deste trabalho é a precisão ponderada (*Pp*). De acordo com Santos (2015) e considerando o contexto deste trabalho, a precisão ponderada é uma métrica que considera a ordem em que os resultados relevantes aparecem; desta forma, considerando um conjunto de termos, sendo de gênero ou de conteúdo, tal métrica é definida por:

$$Pp = \frac{\sum_{i=1}^{N} Ri * Pi}{\sum_{i=1}^{N} Pi} \tag{2.3}$$

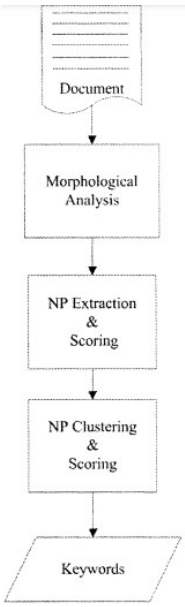
onde:

- *N* representa o total dos termos retornados, pela estratégia proposta, no *ranking* final;
- R*i* representa a i-ésima posição de um vetor R de pesos binários; possui o valor 0 caso o resultado na posição i do *ranking* gerado seja irrelevante e o valor 1 caso contrário;
- P*i* representa a i-ésima posição de um vetor P de pesos N -ários, associadas à posição i em que um determinado resultado se encontra no ranking final; possui valor igual a N $+1-i$. Dessa forma, quanto menor a posição no ranking, ou seja, o quanto antes o resultado aparecer na lista final de resultados, maior será o valor atribuído ao seu peso; por exemplo, caso sejam retornados 20 resultados, o resultado que ocupa a primeira posição terá peso de valor igual a 20, o segundo 19, o terceiro 18, e assim por diante.

2.2 Trabalhos Relacionados

BRACEWELL et al. (2005) desenvolveram uma abordagem para extração de palavras-chave em documento único e multilíngue com propósito de ser aplicável a vários idiomas e extrair palavras-chave identificadas em um único documento, não dependendo de métodos de aprendizagem de máquina. O único pré-requisito é que o idioma tenha um analisador morfológico e regras para encontrar frases nominais simples. Como pode ser visto na Figura 2.3, o algoritmo é dividido em três módulos, sendo eles: análise morfológica, extração e pontuação de frases nominais (NP) e agrupamento e pontuação de tais frases. Desta forma, uma vez fornecendo um documento, a abordagem gera conjunto de palavras-chave eficazes que, identificam exclusivamente um documento e não depende de métodos de aprendizado de máquina.

Figura 2.3 – Visão geral do algoritmo proposto.

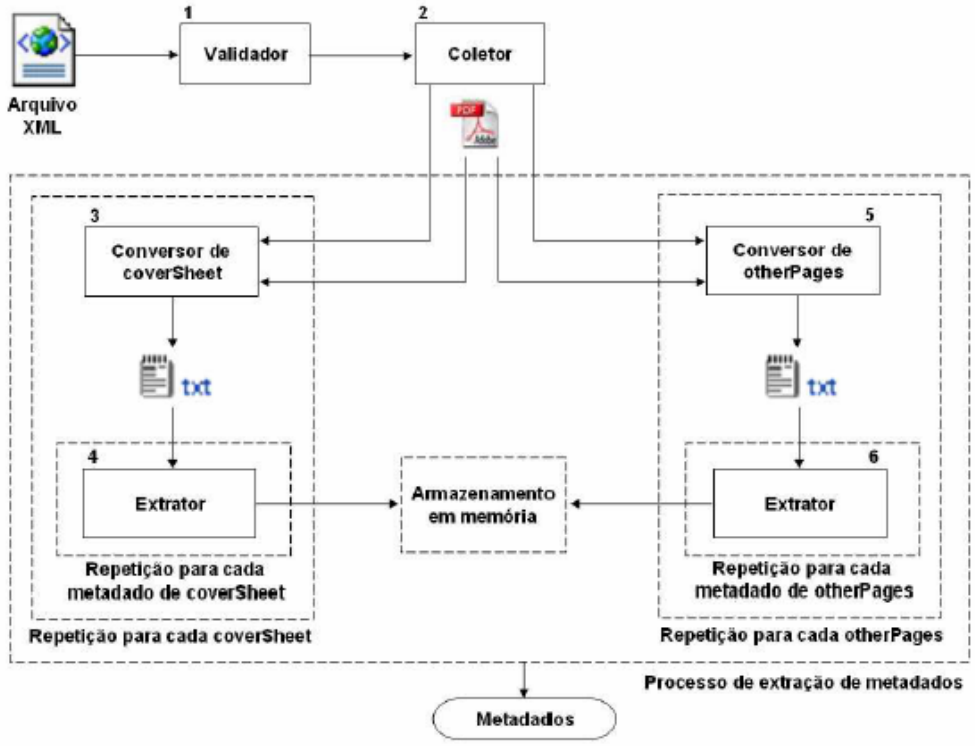


Fonte: BRACEWELL et al. (2005)

Nos experimentos realizados por BRACEWELL et al. (2005), considerando em documentos nos idiomas inglês e japonês, as palavras-chave extraídas pelo algoritmo proposto estavam no documento mais relevante 93,68% das vezes para o inglês e 96,62% das vezes para o japonês. Esses resultados demonstram que o algoritmo proposto no trabalho superam o algoritmo utilizado como referência. Entretanto, de acordo com a abordagem, não é possível utilizar mais do que um documento de entrada. Ademais, o conjunto final de palavras-chave apresentou em pequena quantidade palavras-chave; geralmente, em 9, 5 ou 4 palavras-chave.

MANICA et al. (2008) propôs um processo de extração de metadados a partir de documentos (EMP). O processo utiliza um documento XML como um *template*, onde são especificados campos de metadados que devem ser extraídos dos documentos, sendo o único processo em que necessita da intervenção do usuário. A Figura 2.4 apresenta a arquitetura proposta.

Figura 2.4 – Processo de extração dos metadados.



Fonte: MANICA et al. (2008)

De acordo com a Figura 2.4, primeiro passo é ler e validar o arquivo XML a fim de verificar a existência de elementos que, para serem usados corretamente, necessitam de outros. Logo após, o validador envia o documento XML ao coletor de informações, onde serão preparados os campos de metadados encontrados em uma única página e os encontrados em um intervalo de páginas que serviram de entrada para o EMP que, por sua vez, extrai os metadados descritos nos elementos *coverSheet* e *otherPages*. Para cada elemento especificados no arquivo XML, serão realizados os seguintes experimentos: transformação da página referente ao elemento *coverSheet* em um arquivo texto e conversão do intervalo de páginas referente ao elemento *otherPages* em um arquivo de texto. Em seguida, o conteúdo extraído é relacionado ao metadado correspondente a fim de montar o conjunto final que conterá todos os metadados extraídos. Para validar o EMP, foram submetidos 13 trabalhos de conclusão de curso, no formado *PDF* e um arquivo XML. Foram obtidos resultados satisfatórios em 10 dos trabalhos submetidos.

Este trabalho assemelha-se dos demais apresentados devido ao fato de que, uma vez fornecidos documentos de entrada, os trabalhos em questão extraem palavras-chave significativas com propósitos distintos. Já a diferença entre tais trabalhos está justamente nos propósitos, que são BRACEWELL et al. (2005) e MANICA et al. (2008) enquanto que, neste trabalho, a ideia é extrair palavras-chave de documentos que representem o gênero e conteúdo; no caso, tais palavras-chave são usadas para a realização de processos de coleta temática seguindo o funcionamento do Yucca, no intuito de que possam ser eficazes.

3 Estratégia Proposta

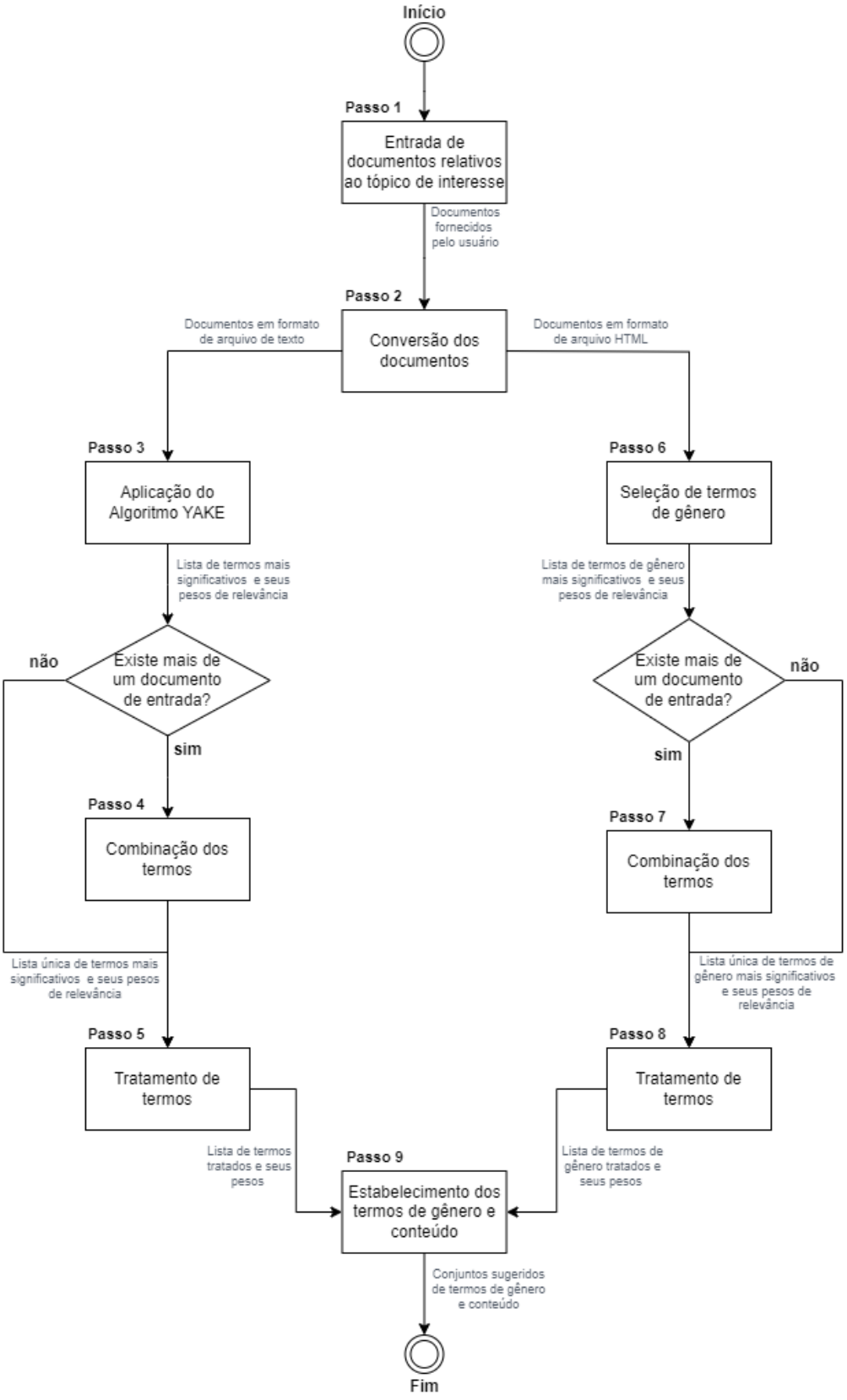
Como já mencionado, este trabalho possui como objetivo geral, a definição de uma estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo, utilizados em processos de coleta temática de páginas da *Web* que seguem a abordagem para coleta temática do Yucca (vide Subseção 2.1.1).

Desta forma, este capítulo apresenta a proposta desta estratégia, estando delineado da seguinte forma: a Seção 3.1 descreve a arquitetura da estratégia proposta para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo, e a Seção 3.2 apresenta a interface do Yucca adequado à estratégia proposta, envolvendo a parametrização necessária para executá-lo.

3.1 Arquitetura de Funcionamento

A arquitetura de funcionamento da estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo encontra-se na Figura 3.1.

Figura 3.1 – Arquitetura de funcionamento da estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo.



Fonte: Elaborada pela autora.

De acordo com a Figura 3.1, a estratégia proposta segue os seguintes passos:

- **Passo 01 - Entrada de documentos**: consiste em fornecer o(s) documento(s) relativo(s) ao tópico de interesse, sendo possível adicionar até quatro documentos (tarefa do usuário);
- **Passo 02 - Conversão dos documentos**: consiste em, a partir do(s) documento(s) fornecidos pelo Passo 01, converter o(s) documento(s) de entrada tanto para arquivo de texto, que serão encaminhados para o Passo 03, como também para arquivo(s) HTML, que serão encaminhados para o Passo 06;
- **Passo 03 - Aplicação do Algoritmo YAKE**: consiste em, a partir dos documentos convertidos em arquivos de texto no Passo 02, gerar uma lista de termos mais significativos e seus pesos de relevância por documento. Por meio da aplicação do algoritmo YAKE (Subseção 2.1.2), determina-se a primeira lista de possíveis termos relevantes que será encaminhada para o Passo 04. Quando há apenas um documento fornecido pelo usuário no Passo 01, a lista de termos gerada é então encaminhada diretamente para o Passo 05;
- **Passo 04 - Combinação dos pesos**: consiste em, quando houver mais de um documento de entrada, combinar as listas de termos mais significativos e seus pesos de relevância retornadas do Passo 03. Ao final deste passo é gerada uma lista única, de modo que, os termos em comum entre tais listas são unificados e um novo peso associado é gerado de acordo com cada ocorrência deste termo nas listas recebidas do passo anterior;
- **Passo 05 - Tratamento de termos**: consiste em tratar os termos por meio de duas etapas, sendo elas: aplicação da técnica de *lemmatização* e tratamento de frases. Este tratamento é aplicado a lista resultante do Passo 04 (ou do Passo 03, caso haja apenas um documento fornecido pelo usuário no Passo 01), e tem como resultado uma lista única de termos tratados, com seus respectivos pesos, que é encaminhada para o último passo (Passo 09);
- **Passo 06 - Seleção de termos de gênero**: consiste em, a partir dos documentos convertidos em arquivos de texto no Passo 03, consiste em selecionar os termos de gênero a partir da semântica das *tags* HTML[1] e estabelecer seus pesos de relevância por meio da taxa de frequência dos mesmos nos documentos considerados. Tais termos de gênero e seus pesos de relevância serão encaminhados para o Passo 07; quando há apenas um documento fornecido pelo usuário no Passo 01, os termos de gênero gerados e seus pesos de relevância são encaminhados diretamente para o Passo 08;
- **Passo 07 - Combinação dos pesos**: similarmente ao Passo 04, consiste em, quando houver mais de um documento de entrada, combinar as listas de termos mais significativos e seus pesos de relevância retornadas do Passo 03. Ao final deste passo é gerada uma lista

[1] O HTML semântico é uma técnica de desenvolvimento de páginas web que se baseia na utilização de *tags* HTML para descrever o significado do conteúdo presente na página.

única, de modo que, os termos em comum entre tais listas são unificados e um novo peso associado é gerado de acordo com cada ocorrência deste termo nas listas recebidas do passo anterior;

- **Passo 08 - Tratamento de termos**: consiste em tratar os termos por meio de duas etapas, sendo elas: aplicação da técnica de *lemmatização* e tratamento de frases; aplicada à lista resultante do Passo 07 (ou do Passo 06, caso haja apenas um documento fornecido pelo usuário no Passo 01), encaminhando uma lista única de termos tratados com seus respectivos pesos para o último passo (Passo 09);
- **Passo 09 - Estabelecimento dos termos de gênero e conteúdo**: consiste em, a partir da lista de termos com seus pesos de relevância dos Passos 05 e do Passo 07, compará-las no intuito de gerar os conjuntos finais de termos de gênero e conteúdo a serem sugeridos ao usuário para a realização de um processo de coleta pelo Yucca.

Particularmente, em relação aos Passos 04 e 07, são identificados os termos, em conjunto com seus respectivos pesos, que se repetem em mais de uma lista de termos; sendo que cada lista de termos diz respeito a um documento em específico. Para evitar repetições de termos na lista única a ser gerada, foi definida uma lista S_i de pesos para cada termo *i* repetido nas listas de entrada e três métodos para calcular o valor final do peso de tal termo *i*. O primeiro método compara os pesos S_i de um termo repetido *i* e retorna o menor peso como sendo o valor final do peso do termo *i*. O segundo calcula a média dos pesos S_i, como sendo o peso final do termo *i*. E o terceiro método estabelece o valor final do peso do termo *i* como sendo a média dos pesos S_i (método anterior), dividido pelo total de pesos contidos em S_i.

Além disso, nos Passos 05 e 08, é feito o tratamento de termos em duas etapas. A primeira etapa consiste na aplicação da técnica de *lemmatização* para o tratamento de radical dos termos. Nesse processo, os termos originais são representados por seus respectivos *lemmas* e comparados entre si. Isso possibilita a identificação de termos com o mesmo radical, sendo possível comparar os pesos e manter apenas aquele termo que possuir o menor peso e removendo os demais. Já na segunda etapa, o tratamento de frases tem como objetivo de eliminar repetições de termos em frases ou termos únicos presentes em uma lista. Para isso, o tratamento de frases utiliza um procedimento iterativo. Primeiramente, cada termo é comparado com o restante da lista e verifica-se se ele é único ou não. Caso seja único, verifica-se também se esse termo aparece em alguma frase pertencente à lista. Se for o caso, esta frase é removida da lista, reduzindo a duplicidade de informações. Por outro lado, quando um termo analisado é uma frase, sua relevância é comparada com a de um termo comum já presente na lista. Se a frase tiver um peso menor que esse termo comum, ela é mantida na lista, e os termos únicos em comum são removidos, a fim de evitar redundâncias.

Por último, no Passo 06, a ideia é tentar selecionar apenas possíveis termos de gênero com base na semântica das *tags* HTML dos documentos repassados. Para ser considerado um termo

de gênero, são levadas em consideração *tags* relacionadas à título e/ou ao palavras em negrito, indicando que a informação contida é provavelmente um termo de gênero. Em seguida, após listas de possíveis termos de gênero serem combinadas no Passo 07 e tratadas no Passo 08, qualquer termo de tal lista final, encontrado também na lista retornada pelo Passo 05, é considerado pela estratégia proposta um termo de gênero; por outro lado, todos os termos presentes na lista final de termos retornada pelo Passo 5 e que não se encontram na lista retornada pelo Passo 8 são considerados termos de conteúdo. Dessa forma, será retornado dois conjuntos de termos separados por gênero e conteúdo para que o usuário seja capaz de utilizar da maneira que preferir.

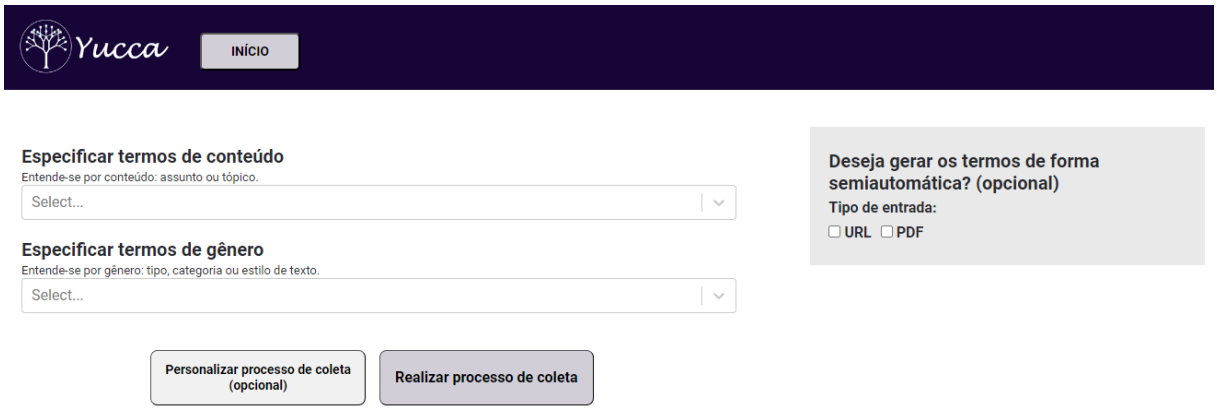
3.2 Nova interface e Parametrização do Yucca

Nesta seção, é apresentada a interface com a estratégia de geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo integrado ao Yucca, envolvendo a parametrização necessária para executá-lo.

A tela inicial do Yucca, como ilustrada na Figura 3.2, possui campos destinados ao preenchimento manual das especificações dos termos de gênero e conteúdo, ou a opção de preenchimento semiautomático, fazendo uso da estratégia sugerida neste trabalho. Além dos campos destinados à entrada dos termos de gênero e conteúdo, a tela inicial possui os seguintes botões: (a) "Personalizar processo de coleta"(opcional), permitindo a modificação das configurações padrão para a execução de um processo de coleta, caso seja do interesse do usuário; (b) "Realizar processo de coleta", destinado a realizar o processo completo de coleta conforme especificado pelo usuário; (c) "INÍCIO", um botão para ser utilizado em qualquer fluxo, permitindo retornar à tela inicial.

Para integrar a estratégia de geração semiautomática dos conjuntos iniciais dos termos de gênero e conteúdo ao Yucca, foi desenvolvido um componente que possibilita interações por parte do usuário. Como pode ser observado na Figura 3.2, o usuário deve inicializar selecionando o tipo de documento de entrada que pretende fornecer, sendo estes em formato *PDF* ou *URL*. Após especificar o formato dos documentos de entrada, o componente exibe as informações correspondentes ao tipo de entrada selecionado.

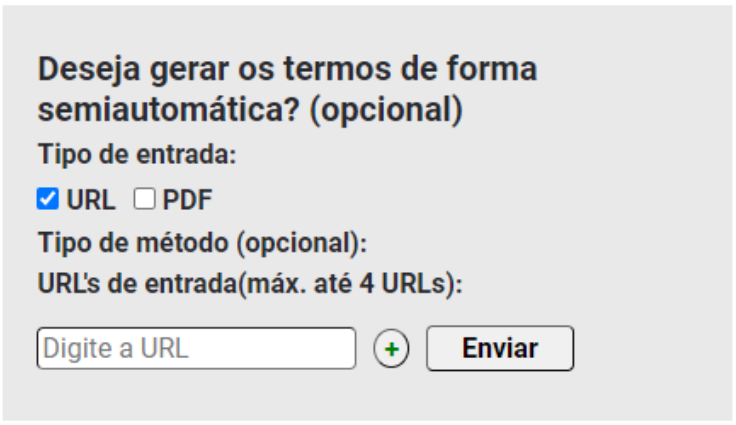
Figura 3.2 – Tela inicial do Yucca



Fonte: Elaborada pela autora.

A Figura 3.3 apresenta as informações relacionadas ao tipo de entrada *URL*. Para que o usuário informe as *URL*'s de entrada, é exibido um campo destinado à inserção da *URL*, acompanhado de um botão "+"que permite adicionar outras *URL*'s, sendo possível inserir até quatro delas por processo. Após preencher os campos com as *URL*'s desejadas, o usuário pode selecionar o botão "Enviar"para dar início ao processo de geração dos conjuntos iniciais dos termos de gênero e conteúdo.

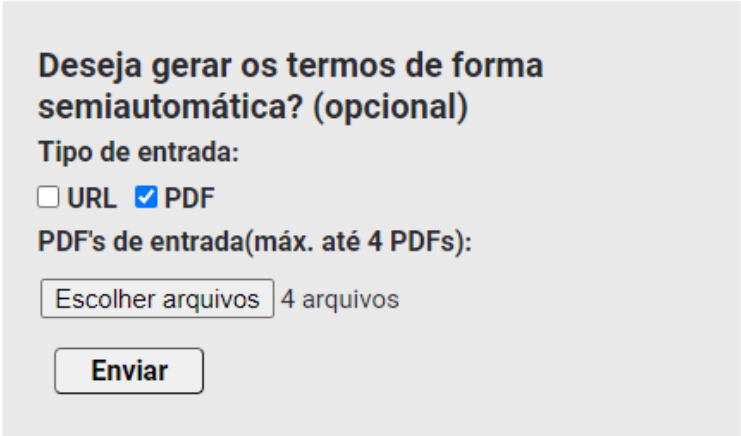
Figura 3.3 – Componente referente ao tipo de entrada em formato *URL*s



Fonte: Elaborada pela autora.

A Figura 3.4 apresenta as informações relacionadas ao tipo de entrada *PDF*, oferecendo ao usuário a possibilidade de selecionar até quatro documentos pertinentes ao seu tópico de interesse. Após inserir os documentos desejados, o usuário pode acionar o botão "Enviar", de maneira semelhante ao tipo de entrada *URL*, para iniciar o processo de criação dos conjuntos iniciais de termos de gênero e conteúdo.

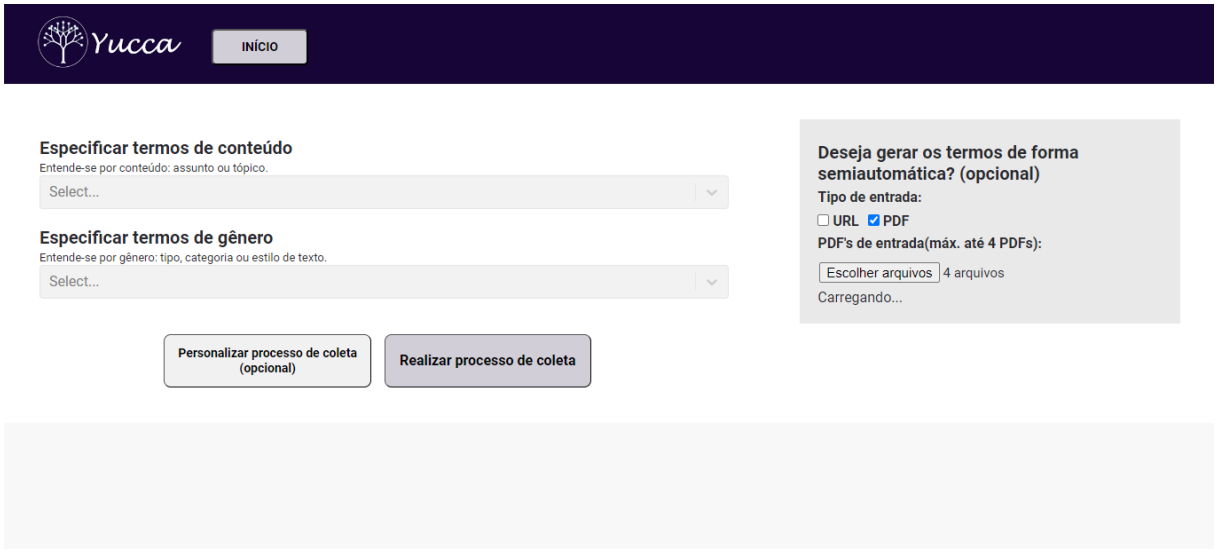
Figura 3.4 – Componente referente ao tipo de entrada em formato *PDF*s



Fonte: Elaborada pela autora.

Uma vez fornecidos os documentos de entrada referente ao tópico de interesse, o processo de gerar os conjuntos iniciais de termos de gênero e conteúdo é inicializado. Para comunicar ao usuário que o processo está em andamento, como ilustrado na Figura 3.5, os campos de especificações dos termos de gênero e conteúdo são desativados e o botão "Enviar"é ocultado, enquanto uma mensagem "Carregando"é exibida na tela inicial.

Figura 3.5 – Tela de carregamento da estratégia de geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo

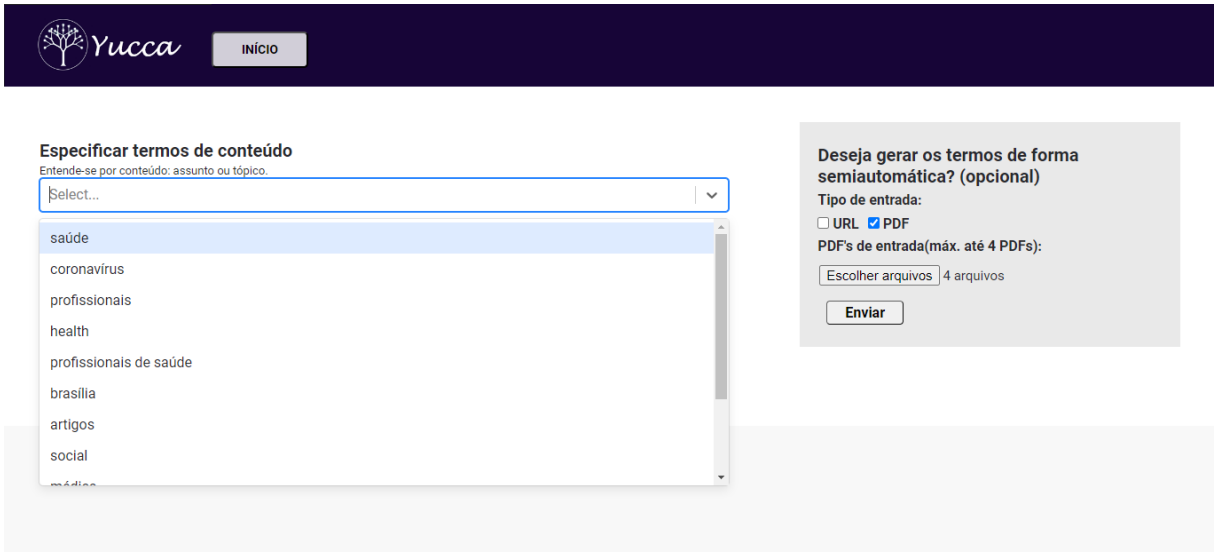


Fonte: Elaborada pela autora.

Por fim, os campos de especificações dos termos de gênero e conteúdo são preenchidos com os conjuntos de termos obtidos por meio da estratégia. No campo "Especificar termos de

gênero", são preenchidos os termos de gênero mais significativos retornados. Esse processo também é replicado para o campo "Especificar termos de conteúdo", onde são inseridos os termos de conteúdo mais relevantes identificados. O usuário tem a opção de selecionar um ou mais termos retornados, além de poder adicionar novos termos de acordo com sua preferência. Na Figura 3.6 é possível observar os termos de conteúdo referente ao tópico de interesse "artigos relacionados ao COVID-19"retornados pela estratégia de geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo.

Figura 3.6 – Tela com os termos gerados a partir da estratégia de geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo



Fonte: Elaborada pela autora.

4 Resultados

Neste capítulo, são apresentados e analisados experimentos de validação da estratégia proposta e apresentada no Capítulo 3, voltada à estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo a serem usados em processos de coleta temática baseada em gênero e conteúdo ASSIS et al.(2009; 2008; 2007). A Subseção 2.1.4 descreve as métricas utilizadas para validar a eficácia da estratégia proposta. A Subseção 4.1 descreve os conjuntos de teste utilizados e os experimentos realizados. Por fim, a Subseção 4.2 apresenta e avalia os resultados obtidos por meio dos experimentos realizados.

4.1 Descrição dos Experimentos

Para realizar os experimentos da estratégia para a geração de conjuntos de termos iniciais de gênero e conteúdo, considerando duas formas de entrada de dados (*PDF* ou *URL*), foram utilizadas documentos de distintos tópicos de interesse, a saber:

- planos de ensino relacionados a Banco de Dados;
- artigos relacionados ao COVID-19;
- receitas de bolo de cenoura.

As Tabelas 4.1 e 4.2 descrevem as fontes dos documentos utilizadas para conduzir os experimentos relacionados aos tópicos planos de ensino de Banco de Dados e artigos relacionados ao COVID-19. Neste caso, os documentos, utilizados como dados de entrada para a estratégia proposta neste trabalho, encontram-se no formato PDF e estão disponibilizados na *web*.

Tabela 4.1 – Coleção de planos de ensino relacionados a Banco de Dados

Fonte
http://www3.decom.ufop.br/decom/disciplina_ementa/bcc321/
https://planos.inf.ufsc.br/modulos/planos/pdf.php?id=86
https://engcomp.dainf.ct.utfpr.edu.br/downloads/IF63C%20-%20Estruturas%20de%20Dados%201%20-%20web.pdf
http://www.ifba.edu.br/professores/pablovf/repositorio/planodeensino-INF007-novo.pdf

Tabela 4.2 – Coleção de artigos relacionados ao COVID-19

Fonte
https://acervomais.com.br/index.php/saude/article/view/4128/2188
https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/12947/10878
https://www.scielo.br/j/csc/a/bbcZzgN6Sns8mNPjKfFYRhb/?format=pdf&lang=pt
https://apsemrevista.org/aps/article/view/69/49

A Tabela 4.3 descrevem as fontes dos documentos utilizadas para conduzir o experimento relacionados ao tópico de receitas de bolo de cenoura. Neste caso, os documentos, utilizados como dados de entrada para a estratégia proposta neste trabalho, encontram-se no formato *URL* e estão disponibilizados na *web*.

Tabela 4.3 – Coleção de receitas de bolo de cenoura

Fonte
https://www.receiteria.com.br/receita/bolo-de-cenoura-com-cacau/
https://www.receitasnestle.com.br/receitas/bolo-de-cenoura-com-cobertura-de-brigadeiro
https://comidinhasdochef.com/bolo-de-cenoura-fofinho-de-liquidificador
https://www.panelinha.com.br/receita/Bolo-de-cenoura-com-cobertura-de-chocolate

Com o propósito de realizar uma análise comparativa dos resultados relativos aos processos de geração semiautomática dos termos de gênero e conteúdo, foram utilizadas as métricas descritas na Seção 2.1.4. Os valores de tais métricas foram calculados a partir dos resultados obtidos por meio da execução da arquitetura proposta (vide Figura 3.1), variando os métodos de combinação de pesos.

Os resultados experimentais retornados pela estratégia proposta foram armazenados em *logs* contendo, para cada execução da estratégia, os conjuntos de termos de gênero e conteúdo com seus respectivos pesos.

Para a realização dos experimentos, foi utilizado um notebook com as especificações a seguir: sistema operacional Windows 10, processador Intel(R) Core(TM) i5-6200U, frequência de 2.40 GHz e 12GB de RAM, obtendo um tempo médio de execução de aproximadamente 1 minuto para cada experimento.

4.2 Análise dos Resultados Obtidos

Nesta seção, são apresentados e analisados os resultados obtidos por meio da experimentação prática realizada, envolvendo a descrição experimental apresentada na Seção 4.1. A Tabela 4.4 exibe os resultados dos casos de testes realizados.

Tabela 4.4 – Resultados dos casos de testes realizados

Tópico de interesse	**Método**	**Quantidade de termos de conteúdo retornados**	**Quantidade de termos de conteúdo relevantes**	**Quantidade de termos de gênero retornados**	**Quantidade de termos de gênero relevantes**	**Precisão ponderada (termos de gênero)**	**Precisão ponderada (termos de conteúdo)**
Planos de ensino de Banco de Dados	Menor peso	12	10	12	12	1.0	0.74445
	Média dos pesos	12	10	12	12	1.0	0.9
	Média modificada dos pesos	12	11	12	12	1.0	0.86667
Artigos relacionados ao COVID-19	Menor peso	12	8	12	12	1.0	0.74445
	Média dos pesos	12	6	12	12	1.0	0.55556
	Média modificada dos pesos	12	8	12	12	1.0	0.733333
Receitas de bolo de cenoura	Menor peso	12	10	11	7	0.80	0.87778
	Média dos pesos	12	9	11	7	0.80	0.822222
	Média modificada dos pesos	12	9	11	7	0.80	0.788889

Para um melhor entendimento da Tabela 4.4, é composta pelas seguintes informações associadas a um caso de teste:

- **Método:** indica o método relativo ao critério de pontuação dos termos de gênero e conteúdo visto nos Passos 04 e 07 da arquitetura proposta (vide Figura 3.1), podendo ser método do menor peso, método da média dos pesos ou método da média modificada dos pesos;
- **Quantidade de termos retornados:** indica a quantidade de termos relevantes ou não que retornados pela estratégia proposta para cada tópico/método, podendo ser no máximo 12 termos;
- **Quantidade de termos relevantes:** indica a quantidade total de termos relevantes dentre os termos retornados pela estratégia para cada tópico/método, descartando os termos que não são considerados relevantes;
- **Precisão ponderada:** indica o cálculo da métrica de precisão ponderada para validar a relevância dos termos alcançados em cada tópico/método, considerando a ordem dos termos relevantes dentro do conjunto de termos. Se os termos relevantes ocuparem posições superiores no conjunto, o resultado tende a ser mais elevado, variando numa escala de 0 a 1.0.

Ao analisar os resultados apresentados na Tabela 4.4, fica evidente que os três métodos avaliados apresentaram resultados bastante similares e significativos. Conforme evidenciado na Tabela 4.4, foram considerados 12 termos de conteúdo retornados para todos os tópicos de interesse pela estratégia de geração semiautomática dos conjuntos de termos de gênero e conteúdo. Em relação ao tópico de interesse do planos de ensino de Banco de dados, tanto os métodos de menor quanto de média dos pesos proporcionaram 10 termos de conteúdo relevantes. Por outro lado, através do método da média modificada dos pesos, foram identificados 11 termos de conteúdo relevantes. No que diz respeito aos artigos relacionados ao COVID-19, tanto o método menor peso quanto a média modificada dos pesos foram considerados 8 termos de conteúdo relevantes. Em contraposição, através do método da média dos pesos, somente 6 termos de conteúdo foram considerados significativos. E no último tópico de interesse, relacionados a receitas de bolo de cenoura, constatou-se que o método de menor peso retornou 10 termos de conteúdo considerados relevantes. Em contraste, os métodos da média e da média modificada dos pesos apresentaram 9 termos de conteúdo relevantes.

Em relação aos conjuntos de termos de gênero, foram considerados 12 termos de gêneros retornados, tendo alcançado o número máximo de termos de gênero relevante, pelos três métodos proposto considerados, relevantes no contexto a artigos relacionados ao COVID-19 e aos planos de ensino de Banco de Dados. Em contrapartida, em relação ao tópico de interesse relacionado a receita de bolo de cenoura, apenas 11 termos de gênero foram retornados, dos quais 7 foram

considerados significativos através dos métodos do menor peso, média e média modificada dos pesos.

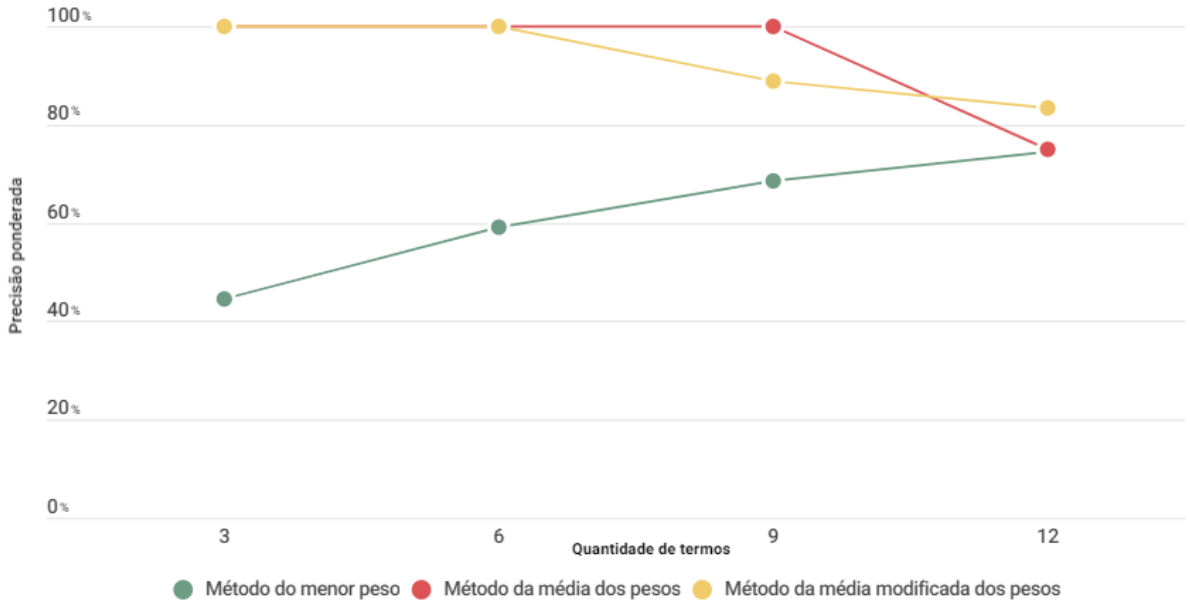
Tabela 4.5 – Resultados da média da precisão ponderada considerando cada tópico de interesse

Tópico de Interesse	Média da precisão ponderada (termos de gênero)	Média da precisão ponderada (termos de conteúdo)
Planos de ensino de Banco de dados	1.0	0.84
Artigos relacionados ao COVID-19	1.0	0.68
Receitas de bolo de cenoura	0.80	0.83

Com base nos dados apresentados na Tabela 4.4 e na análise realizada em relação a cada tópico de interesse, foi conduzida uma avaliação da média da precisão ponderada referente aos métodos de menor peso, média e média modificada dos pesos, como demonstrado na Tabela 4.5. Para os tópicos relacionados a planos de ensino de Banco de Dados e artigos sobre COVID-19, foi alcançado nível médio da precisão ponderada de 100%, enquanto o tópico de receitas de bolo de cenoura obteve 83%. É válido destacar que esses valores podem variar conforme o formato dos documentos de entrada, sendo que os planos de ensino de Banco de Dados e os artigos relacionados ao COVID-19 estão no formato *PDF*, enquanto as receitas de bolo de cenoura estão no formato de *URL*. No que se refere às médias da precisão ponderada relacionadas aos termos de conteúdo, verificou-se um nível médio da precisão ponderada de 84% para o tópico de planos de ensino de Banco de Dados, 68% para os artigos sobre COVID-19 e 83% para as receitas de bolo de cenoura.

Com base nas informações da Tabela 4.4 e na análise realizada em relação a relevância dos termos presentes no conjunto de termos de gênero e conteúdo, as Figuras 4.1, 4.2, 4.3 e 4.4 demonstram, para cada assunto, os níveis de precisão ponderada obtidos, levando em conta diferentes quantidades de termos relevantes retornados, com incrementos de 3 em 3.

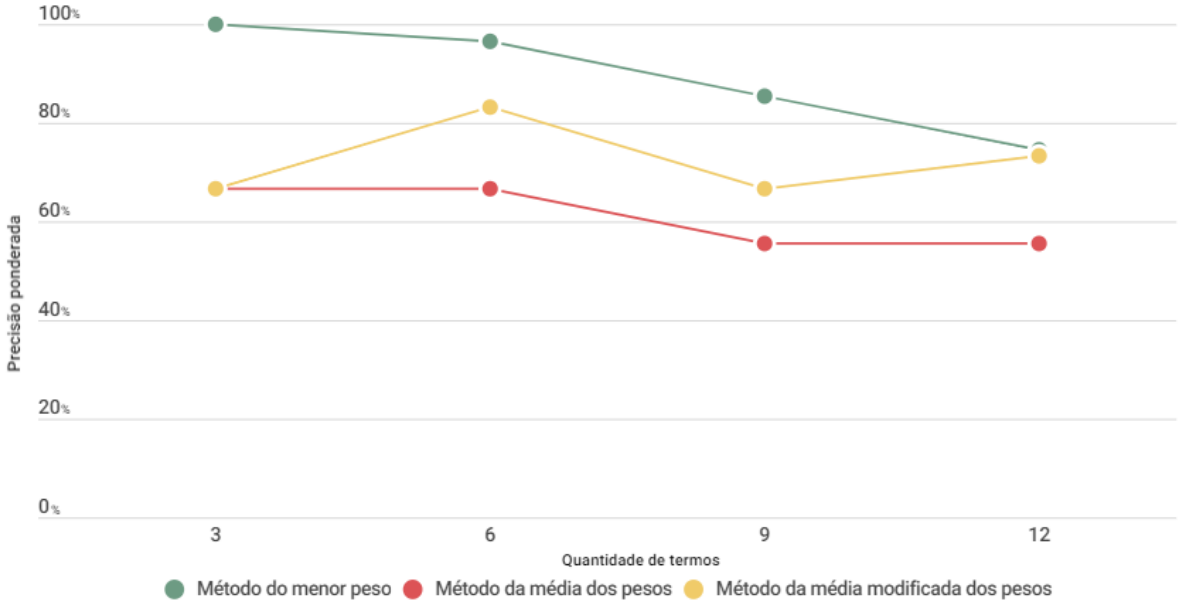
Figura 4.1 – Níveis de precisão ponderada relacionados ao tópico planos de ensino de Banco de dados



Fonte: Elaborada pela autora.

Conforme ilustrado na Figura 4.1, que diz respeito aos conjuntos iniciais de termos de conteúdo, em relação ao tópico planos de ensino de Banco de dados, o método da média dos pesos obteve uma precisão ponderada melhor do que os demais métodos, mantendo-se um nível médio de precisão de 84% quando se leva em conta os 12 termos retornados. Por outro lado, o método que teve o resultado menos satisfatório foi o método do menor peso, com uma média de precisão de 70% ao se levar em consideração os mesmos termos de conteúdo retornados. No entanto, ao restringir a análise aos três primeiros termos de conteúdo retornados, os métodos de média e média modificada dos pesos atingiram uma precisão ponderada de 100% em relação aos termos de conteúdo.

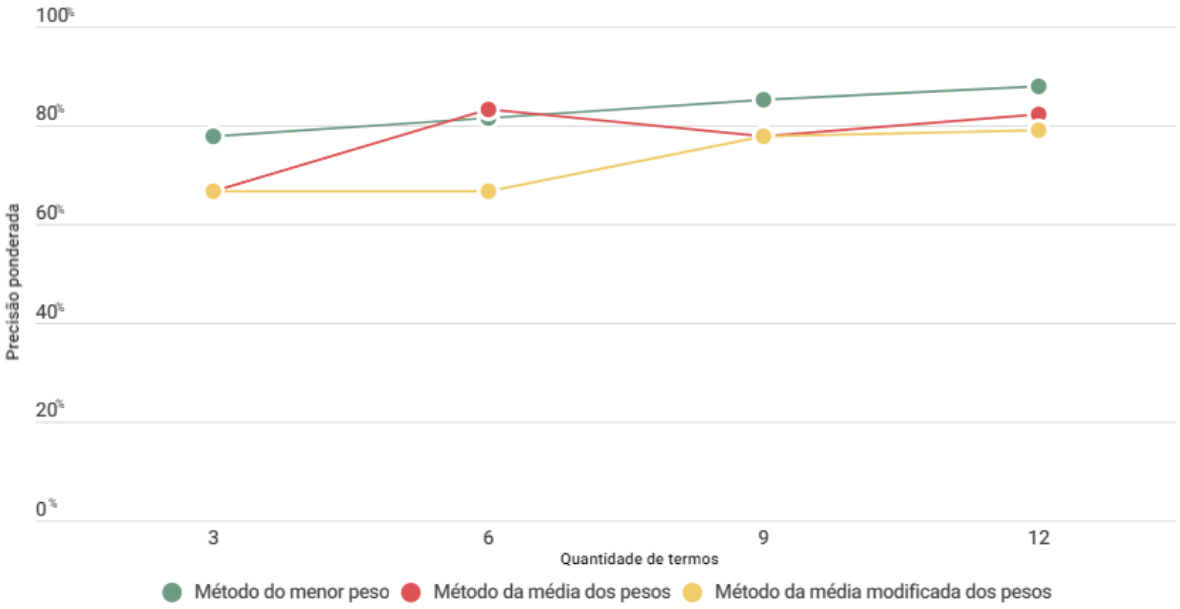
Figura 4.2 – Níveis de precisão ponderada ao tópico de artigos relacionados ao COVID-1



Fonte: Elaborada pela autora.

Em relação ao tópico artigos relacionados ao COVID-19, como visto na Figura 4.2, destaca-se a curva de precisão ponderada associada ao método do menor peso, atingindo uma precisão ponderada de 79% ao considerar 12 termos de conteúdo retornados. Por outro lado, o método que sobressaiu com o menor nível de precisão ponderada relacionados aos 12 termos de conteúdo retornados foi o método da média dos pesos. Já ao considerar os três primeiros termos retornados, o método do menor peso, destaca-se com uma precisão ponderada de 100%.

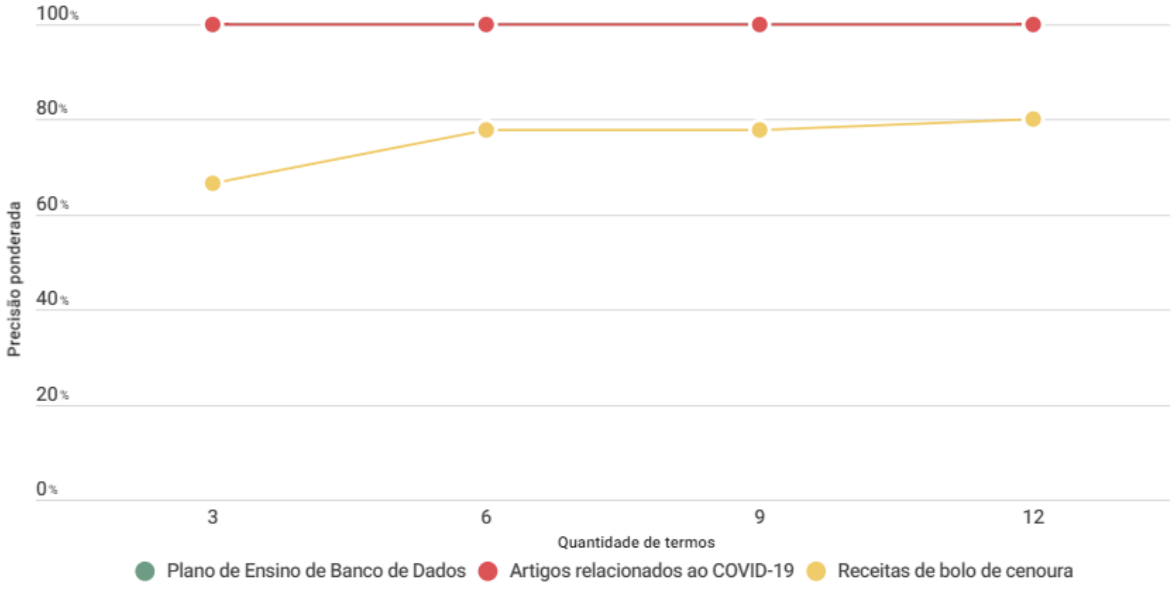
Figura 4.3 – Níveis de precisão ponderada ao tópico de receitas de bolo de cenoura



Fonte: Elaborada pela autora.

Por fim, em relação ao tópico receitas de bolo de cenoura, conforme observado na Figura 4.3, os métodos da média e média modificada dos pesos destacaram-se, atingindo um nível de precisão de 80% ao levar em conta 12 termos de conteúdo. Ao analisar somente os 3 primeiros termos de conteúdo, os três métodos exibiram níveis de precisão bastante similares entre si. De modo geral, fica evidente que o método que obteve os resultados mais notáveis em relação aos termos de conteúdo foi o método da média modificada dos pesos para os casos de testes em questão.

Figura 4.4 – Níveis de precisão ponderada relacionados aos termos de gênero



Fonte: Elaborada pela autora.

Em relação ao conjuntos de termos de gênero, a Figura 4.4, exibe as curvas que representam os níveis de precisão ponderada alcançados nos casos de teste. Ao comparar essas curvas, é notável que elas se mantiveram bastante próximas ou até mesmo coincidentes, mantendo padrões elevados de precisão. Isso evidencia resultados consistentes e satisfatórios, com uma média de precisão ponderada de 100% ao considerar os 12 termos de gênero obtidos. Contudo, apenas no tópico relacionado a "receitas de bolo de cenoura", os três métodos alcançaram um nível de precisão ponderada de 66% ao considerar os três primeiros termos de gênero retornados.

Tabela 4.6 – Resultados da média da precisão ponderada considerando cada método

Método	Média da precisão ponderada (termos de gênero)	Média da precisão ponderada (termos de conteúdo)
Menor peso	0.78	0.93
Média dos pesos	0.75	0.93
Média modificada dos pesos	0.80	0.93

De uma forma geral, observando-se a Tabela 4.6 e para os exemplos considerados, a média da precisão ponderada para os termos de gênero foi de 93% para todos os métodos. E em relação aos termos de conteúdo, a média do nível de precisão ponderada do método de menor peso foi de 78%, enquanto o método da média dos pesos obteve 75%, e o método da média modificada alcançou 80%. Portanto, é evidente que os métodos apresentaram médias de precisão ponderada bastante próximas. No entanto, destaca-se que o método da média modificada obteve

o resultado mais favorável para a obtenção vinculada aos pesos, mas devido à proximidade dos resultados obtidos, independentemente do tema, essa circunstância pode ser alterada.

5 Considerações Finais

Neste capítulo, são apresentadas as conclusões sobre o trabalho desenvolvido (vide Seção 5.1) e as perspectivas de trabalho futuro (vide Seção 5.2).

5.1 Conclusão

Como já apresentado, este trabalho propôs desenvolver e validar uma estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo a serem utilizados em processos de coleta temática de páginas da *Web* propostos em ASSIS et al.(2009; 2008; 2007).

Com o objetivo de avaliar a eficácia da estratégia para geração dos conjuntos iniciais de termos de gênero e conteúdo, como visto, foram realizados experimentos considerando 3 assuntos distintos sendo eles: planos de ensino de Banco de Dados, artigos relacionados ao COVID-19 e receitas de bolo de cenoura. Em todos os cenários, a estratégia demonstrou níveis de precisão significativos para os termos gerados a partir de documentos relevantes fornecidos como dados de entrada. Além disso, quando o usuário submete múltiplos documentos, foram desenvolvidos métodos também eficazes para combinar os pesos dos termos comuns entre esses documentos. Esse processo resulta em uma lista única de termos relevantes, cujos pesos são ajustados para as etapas posteriores, conforme detalhado na Seção 3.1.

Em relação aos conjuntos de termos de conteúdo gerados a partir dos documentos de entrada, dos 12 termos resultantes de cada caso de teste, uma média de 9 termos relevantes foi identificada. Nos testes conduzidos utilizando os métodos de menor peso, média dos pesos e média modificada dos pesos, foram obtidos os seguintes níveis médios de precisão ponderada para cada método: 78% para o método de menor peso, 75% para a média dos pesos e 80% (melhor valor) para o método de média modificada dos pesos. Ao contrário dos conjuntos de termos de conteúdo, os conjuntos de termos de gênero demonstraram resultados uniformes em relação aos três métodos propostos, independentemente dos tópicos considerados. Por fim, o método da média modificada destaca-se com resultado mais favorável entre os métodos.

De maneira geral, este trabalho desenvolveu e validou uma estratégia para geração dos conjuntos iniciais de termos de gênero e conteúdo para o coletor temático Yucca (vide Subseção 2.1.1). Dessa forma, usuários possuem agora duas alternativas para iniciar um processo de coleta com Yucca: manualmente, ao fornecer termos de gênero e conteúdo relacionados ao tópico de interesse desejado, ou, de maneira semiautomática, ao indicar documentos relevantes ao tópico de interesse para gerar conjuntos iniciais de termos de gênero e conteúdo, os conjuntos iniciais de termos de gênero e conteúdo são sugestões e podem ser adaptados pelo usuário antes de se

iniciar um processo de coleta. No que diz respeito às restrições dessa abordagem, é evidente que a precisão ponderada diminui quando os documentos são fornecidos em formato de *URL*. Isso ocorre devido ao fato de que cada *URL* possui uma configuração específica em relação às tags *HTML* utilizadas em sua construção.

5.2 Trabalhos Futuros

Nesta seção, são apresentadas algumas perspectivas de trabalho futuro, a saber: (1) realizar testes da estratégia juntamente com o Yucca; (2) possibilitar outros formatos de documentos, como dados de entrada, para a estratégia proposta de geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo; (3) realizar novos experimentos com outros tópicos de interesse no intuito mais parâmetros de precisão ponderada; (4) realizar estudos sobre a experiência do usuário quanto ao uso completo do Yucca na realização de processos de coleta, a fim de analisar a sua usabilidade.

Referências

M. AHLGREN. Mais de 100 estatÍsticas, fatos e tendÊncias da internet para 2023. Disponível em: https://www.websiterating.com/pt/research/internet-statistics-facts/. Acesso em: 24 janeiro 2023, 2023.

G. T. ASSIS, A. HF LAENDER, M. A. Gonçalves, and A. S. da Silva. Exploiting genre in focused crawling. In *String Processing and Information Retrieval*, pages 62–73. Springer, 2007.

G. T. de ASSIS, A. HF LAENDER, A. S. da SILVA, and M. A. GONÇALVES. The impact of term selection in genre-aware focused crawling. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1158–1163. ACM, 2008.

G. T. de ASSIS, A. HF LAENDER, M. A. GONÇALVES, and A. S. da. SILVA. A genre-aware approach to focused crawling. *World Wide Web*, 12(3):285–319, 2009.

Vimala Balakrishnan and Ethel Lloyd-Yemoh. Stemming and lemmatization: A comparison of retrieval performances. 2014.

D. BHATT, D. A. VYAS, and S. PANDYA. Focused web crawler. *algorithms*, 5:18, 2015.

D. B. BRACEWELL, F. REN, and S. KURIOWA. Multilingual single document keyword extraction for information retrieval. In *2005 international conference on natural language processing and knowledge engineering*, pages 517–522. IEEE, 2005.

J. BROWNLEE. How to calculate precision, recall, and f-measure for imbalanced classification. *Machine Learning Mastery*, 2020.

R. CAMPOS, V. MANGARAVITE, A PASQUALI, A JORGE, C. NUNES, and A. JATOWT. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2019.09.013. URL https://www.sciencedirect.com/science/article/pii/S0020025519308588.

S. CHAKRABART, M. V. D. BERG, and B. DOM. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11):1623–1640, 1999.

G. G. COSTA, G. T. ASSIS, and M. V. O. SOUZA. Automatic improvement of terms used in focused crawling processes on web page. In *Proceedings of the 16th International Conference WWW/Internet*. WWW, 2017.

D. DINIZ and G. T. ASSIS. Yucca: Um coletor temático de páginas da web baseado em gênero. UFOP, 2018.

J. HOSSEINKHANI, H. TAHERDOOST, and S. KEIKHAEE. Anton framework based on semantic focused crawler to support web crime mining using svm. *Annals of Data Science*, 8 (2):227–240, 2021.

M. T. A. JÚNIOR, M. F. P. REZENDE, and G. T. de ASSIS. Development of a focused web page crawler based on genre and content. *Proceedings of the International Conferences on WWW/Internet 2021 and Applied Computing 2021*, 2021.

M. KUMAR, A. BINDAL, R. GAUTAM, and R. BHATIA. Keyword query based focused web crawler. *Procedia Computer Science*, 125:584–590, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2017.12.075. URL https://www.sciencedirect.com/science/article/pii/S1877050917328399. The 6th International Conference on Smart Computing and Communications.

F. LEITNER. segtok - a segmentation and tokenization library. Disponível em: https://fnl.es/segtok-a-segmentation-and-tokenization-library.html. Acesso em: 24 agosto 2022, 2015.

R. S. MANE, G. B. BAGGA, D. U. BHUTE, A. D. NIKAM, and S. GAIKWAD. Focus: Learning to crawl web forums. 2017.

V. MANGARAVITE, G. T. de ASSIS, and A. A. FERREIRA. Improving the efficiency of a genre-aware approach to focused crawling based on link context. In *Web Congress (LA-WEB), 2012 Eighth Latin American*, pages 17–23. IEEE, 2012.

V. MANGARAVITE, G. T. de ASSIS, and A. A. FERREIRA. Semi-automatic generation of seed pages in genre-aware focused crawling. In *Proceedings of the 13th International Conference WWW/Internet (ICWI)*, pages 51–58. WWW, 2014.

E. MANICA, C. R. CERVI, and R. de M. GALANTE. Um processo automático para extraçãoo de metadados de documentos pdf usando um template xml. *Escola Regional de Banco de Dados (ERBD 2008)*, 4, 2008.

K. PAVANI and GP. SAJEEV. A novel web crawling method for vertical search engines. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1488–1493. IEEE, 2017.

Joël Plisson, Nada Lavrac, Dunja Mladenic, et al. A rule based approach to word lemmatization. In *Proceedings of IS*, volume 3, pages 83–86, 2004.

Filipe Eduardo Mata Dos Santos. Proposta e desenvolvimento de um serviÇo de busca na web por documentos similares aos trabalhos catalogados na biblioteca digital do curso de ciÊncia da computaÇÃo da universidade federal de ouro preto. 2015.

G. O. de SIQUEIRA, G. T. de ASSIS, A. A. FERREIRA, A. S. N. e SILVA, V.r MANGARAVITE, and F. L. C. PÁDUA. Automatic determination of similarity threshold for focused crawling processes on web pages. In *Proceedings of the 15th International Conference WWW/Internet (ICWI)*. WWW, 2016.