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

THIAGO URZEDO DA SILVA PAULINO
Orientador: Prof. Dr. Guilherme Tavares de Assis

YUCCA: COLETOR TEMÁTICO DE PÁGINAS DA WEB BASEADO EM GÊNERO E CONTEÚDO

Ouro Preto, MG
2024

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

THIAGO URZEDO DA SILVA PAULINO

YUCCA: COLETOR TEMÁTICO DE PÁGINAS DA WEB BASEADO 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: Prof. Dr. Guilherme Tavares de Assis

Ouro Preto, MG
2024

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

FOLHA DE APROVAÇÃO

Thiago Urzedo da Silva Paulino

Yucca: Coletor temático de páginas da Web baseado 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 11 de Outubro de 2024.

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 11/10/2024.

Documento assinado eletronicamente por **Guilherme Tavares de Assis**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 11/10/2024, às 15:13, 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 **0789776** e o código CRC **F5F25A27**.

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 acreditou em mim e me deu apoiou em todas as fases da minha vida.

Agradecimentos

Agradeço a minha mãe, meu exemplo de vida, que mesmo com todas as dificuldades que passamos, sempre esteve ao meu lado me dando apoio para que eu continuasse meus estudos e buscasse meus sonhos.

Agradeço ao meu irmão, que sempre me apoiou e me ajudou bastante durante todos esses anos da minha vida.

Agradeço também ao restante da minha família, que desde a minha infância, sempre me ajudaram. Sem vocês, nada disso seria possível.

Agradeço ao meu orientador Guilherme, por toda a dedicação, paciência e conselhos para a conclusão deste trabalho.

Por fim, agradeço a todos meus amigos que me acompanharam durante toda essa longa jornada pela UFOP, todos foram essenciais para que eu chegasse onde estou hoje.

Resumo

Os coletores temáticos têm o propósito principal de facilitar a descoberta, organização e acesso a conteúdos relevantes na internet, agrupando informações relacionadas a um tema específico em um único lugar. Eles são utilizados de diversas maneiras, adaptando-se às necessidades e aos interesses individuais dos usuários. Nesse contexto, foi proposta em ASSIS *et al.* (2007), ASSIS *et al.* (2008) and ASSIS *et al.* (2009) uma abordagem de coleta temática em que o tópico de interesse do usuário pode ser expresso por meio de termos de gênero e conteúdo das páginas *Web* desejadas. Tal abordagem possibilita a construção de coletores temáticos eficazes, eficientes e escaláveis. Visando aperfeiçoar a eficiência e a eficácia de tal abordagem, foram propostas as seguintes melhorias: uma nova política de localização de páginas relevantes baseada em *Link Context*, proposto em MANGARAVITE, ASSIS e FERREIRA (2012); uma estratégia para a determinação semiautomática de páginas-semente, proposto em MANGARAVITE *et al.* (2014); uma estratégia para a definição automática de limites de similaridade, proposto em SIQUEIRA *et al.* (2016); uma estratégia de aperfeiçoamento automático dos conjuntos de termos de gênero e conteúdo, proposto em COSTA (2017); e uma estratégia para a geração semiautomática dos termos iniciais de gênero e conteúdo necessários para a realização de um processo de coleta, proposta em SILVA (2023). Dessa forma, este trabalho propõe o desenvolvimento e a validação de uma versão completa e funcional de um coletor temático, denominado Yucca, seguindo a abordagem original de coleta temática e suas melhorias citadas, além de solucionar necessidades e problemas ocorridos em implementações anteriores relativas ao Yucca, por meio de uma re-implementação do coletor original e suas melhorias, utilizando tecnologias atuais. Por meio dos processos de coleta realizados experimentalmente, considerando distintos tópicos de interesse, o Yucca apresentou-se como um coletor temático eficaz, já que os níveis de precisão alcançados, foram bem satisfatórios, chegando a ser superiores a 78% ao considerar 60 páginas retornadas como relevantes pelo coletor.

Palavras-chave: Coleta temática de páginas da web. Coletor temático. Termos de gênero. Termos de conteúdo.

Abstract

Foccused Crawlers have the primary purpose of facilitating the discovery, organization, and access to relevant content on the internet by grouping information related to a specific theme in one place. They are used in various ways, adapting to the individual needs and interests of users. In this context, an approach to thematic collection was proposed in ASSIS *et al.* (2007), ASSIS *et al.* (2008) and ASSIS *et al.* (2009) where the user's topic of interest can be expressed through genre and content terms of the desired web pages. Such an approach enables the construction of effective, efficient, and scalable thematic collectors. In order to enhance the efficiency and effectiveness of this approach, the following improvements have been proposed: a new policy for locating relevant pages based on Link Context, proposed in MANGARAVITE, ASSIS e FERREIRA (2012); a strategy for the semi-automatic determination of seed pages, proposed in MANGARAVITE *et al.* (2014); a strategy for automatically defining similarity limits, proposed in SIQUEIRA *et al.* (2016); a strategy for the automatic refinement of genre and content term sets, proposed in COSTA (2017); and a strategy for the semi-automatic generation of initial genre and content terms necessary for the collection process, proposed in SILVA (2023). Thus, this work proposes the development and validation of a complete and functional version of a focused crawler, called Yucca, following the original approach of focused collection and its cited improvements, in addition to addressing needs and issues encountered in previous implementations related to Yucca, through a re-implementation of the original crawler and its improvements, using current technologies. Through the collection processes conducted experimentally, considering different topics of interest, Yucca proved to be an effective focused crawler, as the precision levels achieved were quite satisfactory, reaching over 78% when considering 60 pages returned as relevant by the crawler.

Keywords: Focused Web Crawler. Genre terms. Content terms.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

CSV	*Comma-Separated Values*
DECOM	Departamento de Computação
HTML	*HyperText Markup Language*
PDF	*Portable Document Format*
PLN	Processamento de Linguagem Natural
TXT	Arquivo de texto
UFOP	Universidade Federal de Ouro Preto
URL	*Uniform Resource Locator*
XLSL	*Excel Spreadsheet (XML-based)*
XML	*eXtensible Markup Language*

Sumário

1 Introdução

Segundo AHLGREN (2023), atualmente existem mais de 1.97 bilhões de sites e cerca de 5.6 bilhões de usuários da internet e estes números crescem exponencialmente com o passar dos anos. Ademais, em 2023, somente na ferramenta de pesquisa do *Google*, há uma média de 8.5 bilhões de consultas diárias, oriundas de todas as partes do mundo. Tendo em vista estes fatores, as técnicas de recuperação de informação na internet tornam-se cada vez mais necessárias a fim de facilitar e refinar a busca e a coleta de dados.

Entretanto, segundo COSTA (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, como visto em CHAKRABARTI, BERG e DOM (1999) e JIANG, YU e LIN (2012), 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.

Várias estratégias de coleta temática, vistas em PANT e SRINIVASAN (2005), HOSSEINKHANI, TAHERDOOST e KEIKHAEE (2021) e SHRIVASTAVA, PATERIYA e KAUSHIK (2023), 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 de precisão[1] e revocação[2], geralmente entre 40% e 70%.

Diferente das abordagens que utilizam um classificador no processo de coleta e também de diversas outras, como FARAG, LEE e FOX (2018) e LEE *et al.* (2020) que consideram um conjunto único de termos de entrada, ASSIS *et al.* (2009) propôs uma abordagem diferente, onde conjuntos distintos de termos de gênero e de conteúdo são considerados para se realizar um processo eficaz e efetivo de coleta temática, não sendo necessário o uso de um classificador para que se obtenha um nível aceitável de precisão. Entende-se, como termo de gênero, o tipo, a categoria ou o estilo do texto e, como termo de conteúdo, o assunto ou o tópico das páginas a serem coletadas. A abordagem proposta teve, como objetivo principal, estabelecer um arcabouço que permita a construção de coletores temáticos eficazes, eficientes e escaláveis, sem a necessidade de um treinamento a priori ou qualquer tipo de pré-processamento.

A abordagem original já apresentou resultados expressivos, obtendo níveis de revocação

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

[2] 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.

e precisão entre 85% a 100% nos experimentos realizados. Diversas melhorias foram propostas, desenvolvidas e acrescentadas na abordagem posteriormente, como: a utilização de características estruturais de links, presentes em uma página visitada pelo coletor em um processo de coleta temática, visto em MANGARAVITE, ASSIS e FERREIRA (2012); a determinação semiautomática das páginas-semente a serem utilizadas em um processo de coleta temática, visto em MANGARAVITE *et al.* (2014); a determinação automática do limite de similaridade a ser considerado em um processo de coleta temática, visto em SIQUEIRA *et al.* (2016); a expansão dos termos de gênero e conteúdo por meio de uma técnica automática de expansão de termos, visto em COSTA (2017); e a geração semi-automática dos conjuntos iniciais de termos de gênero e conteúdo, visto em SILVA (2023). Todas essas melhorias aumentaram a eficácia e/ou a eficiência da abordagem original.

1.1 Justificativa

A partir da abordagem original e de suas melhorias posteriores, em DINIZ (2018), é proposta a primeira versão de um coletor temático baseado em gênero e conteúdo, denominado Yucca. Esta primeira versão foi utilizada apenas para facilitar os experimentos envolvendo as integrações das melhorias à abordagem original, não sendo, de fato, uma versão final e funcional do coletor. A fim de tornar o Yucca uma aplicação funcional, JUNIOR (2021) fez a sua primeira implementação de fato, tomando como base o trabalho de DINIZ (2018), porém ainda não devidamente completa e funcional. Posteriormente, SILVA (2023) acrescentou ao trabalho de JUNIOR (2021) a geração semi-automática dos termos de gênero e conteúdo e também melhorias na interface inicialmente proposta.

Apesar da existência das implementações citadas acima, nenhuma delas previu o uso de vários usuários simultâneos, visto que a entrada e saída dos dados eram gravados em um arquivo de texto, o que pode causar um problema de concorrência e/ou exibir dados incorretos ao usuário. Outro problema apresentado é a apresentação das coleções de páginas ao final do processo de coleta para o usuário, que não foi devidamente implementado, exibindo somente uma lista simples das páginas encontradas no processo de coleta. Além disso, a interface construída não é muito intuitiva, além de exigir configurações que talvez não sejam interessantes para um usuário leigo da aplicação. Este trabalho procura sanar estes problemas, por meio da reimplementação do coletor original e de todas as suas melhorias, utilizando tecnologias atuais, que tenha o mesmo ou um melhor desempenho quando comparado às implementações anteriores; ademais, este trabalho envolve a implementação de uma nova interface para o Yucca, que seja simples e de fácil usabilidade.

1.2 Objetivos Geral e Específicos

Este trabalho possui, como objetivo geral, o desenvolvimento e validação de uma versão completa e funcional do Yucca, um coletor temático de páginas da *Web* baseado em gênero e conteúdo, para que possa ser utilizado por múltiplos usuários de uma forma simples e robusta. Para tanto, foram consideradas a abordagem original para coleta temática baseada em gênero e conteúdo (ASSIS *et al.* (2007), ASSIS *et al.* (2008), ASSIS *et al.* (2009)) e suas melhorias desenvolvidas e validadas (MANGARAVITE, ASSIS e FERREIRA (2012), MANGARAVITE *et al.* (2014), SIQUEIRA *et al.* (2016), COSTA (2017) e SILVA (2023)). Ademais, também foram consideradas as implementações anteriores relativas ao Yucca (JUNIOR (2021) e SILVA (2023)), buscando corrigir os seus problemas existentes e validar cada um de seus componentes.

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

- implementação do Yucca, utilizando tecnologias atuais, que possibilita uma maior eficiência em processos de coleta temática;
- implementação de uma nova interface para o Yucca, que possibilita uma melhor usabilidade por parte do usuário, permitindo que, com o mínimo de interações possíveis, ele consiga gerar coleções de páginas de acordo com seus tópicos de interesse;
- possibilidade da configuração de parâmetros do processo de coleta, visando que usuários mais experientes possam modificar o resultado final das coleções geradas, sem interferir no entendimento da aplicação por parte de usuários leigos;
- clareza no retorno das coleções geradas pelo Yucca, após o processo de coleta temática, por meio da implementação de um componente para apresentação e exportação das páginas relevantes coletadas;
- certificação da qualidade da coleta temática, baseada em gênero e conteúdo, a partir da análise de coleções geradas relativas a distintos e atuais tópicos de interesse.

1.3 Método do Trabalho

Visando alcançar o objetivo geral deste trabalho, foi proposta uma nova e mais completa arquitetura de funcionamento do Yucca juntamente com a implementação de uma nova versão completa e funcional.

No intuito de validar o Yucca quanto ao seu funcionamento, experimentos práticos, considerando todas as características da nova arquitetura de funcionamento proposta para o mesmo, foram realizados envolvendo a coleta de páginas da *web* referentes a distintos e atuais tópicos de interesse e a consequente medição da precisão a partir dos resultados obtidos.

1.4 Organização do Trabalho

Os demais capítulos deste trabalho estão organizados da seguinte forma: no Capítulo 2, é apresentada a revisão de literatura, onde são apresentados trabalhos que dão suporte a este. No Capítulo 3, é apresentado o desenvolvimento da nova versão do Yucca, envolvendo sua arquitetura de funcionamento, características e interface. No Capítulo 4, são apresentados os experimentos práticos realizados, juntamente com a análise dos resultados obtidos. E, finalmente no Capítulo 5, são apresentadas as considerações finais, que incluem as conclusões deste trabalho e também as perspectivas de trabalho futuro.

2 Revisão de Literatura

Este capítulo contém a revisão de literatura realizada para a confecção deste trabalho, sendo organizado da seguinte maneira: Seção 2.1, onde é apresentada a fundamentação teórica utilizada como base para o desenvolvimento deste trabalho, e Seção 2.2, onde são apresentados trabalhos que possuem temas relacionados a este.

2.1 Fundamentação Teórica

Nesta seção, são apresentados trabalhos já desenvolvidos que dão suporte a este trabalho. Na Subseção 2.1.1, é apresentada a concepção original da estratégia de coleta temática baseada em gênero e conteúdo, elaborada por ASSIS *et al.* (2009), e são descritas algumas melhorias em seu funcionamento, realizadas em trabalhos posteriores. Já na Subseção 2.1.2, são apresentadas as implementações existentes e relativas ao coletor temático Yucca, mostrando funcionamento, resultados obtidos e dificuldades enfrentadas.

2.1.1 Abordagem original de coleta temática baseada em gênero e conteúdo

De acordo com ASSIS *et al.* (2007), ASSIS *et al.* (2008), ASSIS *et al.* (2009), coletores temáticos pertencem a uma importante classe de programas, que tem, como principal objetivo, navegar pelas páginas da internet que são relevantes a um tópico específico ou de interesse do usuário. Diferente dos coletores tradicionais que são utilizados por mecanismos de busca, os coletores temáticos geram uma coleção mais enxuta de páginas, reduzindo o consumo de recursos e favorecendo a escalabilidade do processo de coleta, visto que não precisam cobrir toda a internet.

O maior desafio de identificar páginas relevantes na internet, geralmente está relacionado com a escolha de heurísticas apropriadas que dão o direcionamento adequado ao processo de coleta, envolvendo determinar o quão relevante é uma certa página ao tópico de interesse. A maioria das estratégias de buscas atuais levam em consideração somente o conteúdo das páginas e dependem de classificadores de texto para determinar o seu grau de relevância, o que gera um número acentuado de resultados falso positivos, além de possuir o custo adicional de treinar o classificador.

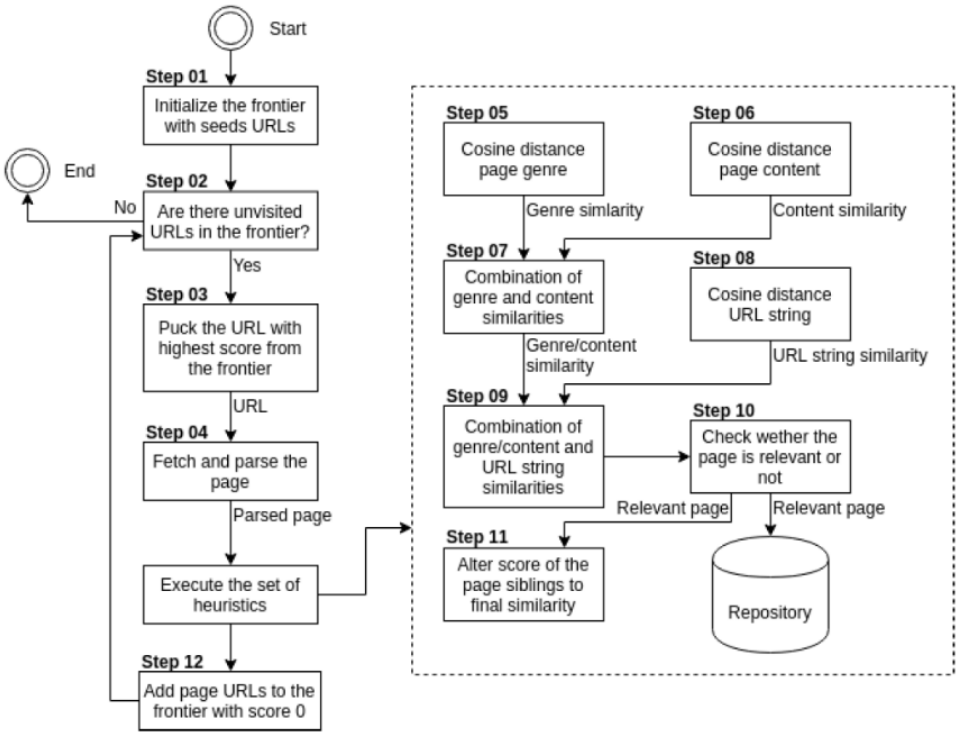
Com base nisso, foi proposto um coletor temático, com o objetivo de ser eficaz, eficiente e escalável, que leva em consideração o gênero (tipo de texto) e o conteúdo (assunto) das páginas, para determinar a relevância de suas informações de acordo com o interesse do usuário. A Figura 2.1 descreve os principais passos do coletor proposto em ASSIS *et al.* (2009) que, de forma geral, consiste em:

- Passo 01: inicializa-se uma fila de de URLs não visitadas, denominada de *Frontier*, com as URLs das páginas-semente, que são as páginas encontradas a partir dos termos de gênero e conteúdo definidos pelo usuário, com prioridade igual a 1 (um). Esta pontuação indica a prioridade de visita das páginas;
- Passo 02: é feita uma verificação se ainda existem URLs não visitadas no *Frontier*. Caso não existam, o processo é encerrado;
- Passo 03: remove do *Frontier* a URL com maior pontuação;
- Passo 04: é feita uma busca e analise do conteúdo da página removida no passo anterior;
- Passo 05: calcula a distância de cosseno[1] entre os termos de gênero e o resultado obtido no passo 04;
- Passo 06: calcula a distância de cosseno entre os termos de conteúdo e o resultado obtido no passo 04;
- Passo 07: Combinar as similaridades de gênero e conteúdo, obtidas nos passos 05 e 06;
- Passo 08: calcula a distância de cosseno entre os termos de gênero e conteúdo e a URL que foi removida no passo 03;
- Passo 09: combina as similaridades de gênero e conteúdo, obtida no passo 07, com a similaridade da URL, obtida no passo 08;
- Passo 10: a partir das similaridades obtidas nos passos anteriores, é feita uma verificação que identifica se a página visitada é realmente relevante aos tópicos desejados, por meio de um limite de similaridade pré-estabelecido. Caso a página seja relevante, ela é armazenada no repositório de páginas relevantes;
- Passo 11: caso a página no passo anterior for considerada relevante, as prioridades de visita das páginas irmãs, que são as URLs contidas no *Frontier*, são alteradas para a similaridade obtida no passo anterior;
- Passo 12: adiciona ao *Frontier* as URLs contidas na página visitada no passo 04, com a prioridade igual a 0 (zero). Em seguida, retorna ao passo 02.

Nota-se que o coletor construído de acordo com a abordagem de gênero, segue os mesmos passos básicos de um coletor tradicional guiado por um classificador. No entanto, de acordo com ASSIS *et al.* (2009), pelo fato de explorar tanto as informações de gênero e conteúdo como

[1] Em ASSIS *et al.* (2009), distância de cosseno é uma função derivada do modelo vetorial clássico. Segundo SOUZA e DORNELES (2017), o modelo vetorial representa documentos como vetores de termos, podendo-se classificar os documentos por meio de atribuições de pesos para o índice de termos nas consultas. Deste modo, é possível gerar um *ranking* de similaridade dos documentos.

Figura 2.1 – Arquitetura de funcionamento do coletor baseado em gênero e conteúdo



Fonte: ASSIS *et al.* (2009).

também possuir uma política dinâmica de fila, melhorou consideravelmente os resultados obtidos, atingindo níveis F1[2] superiores a 88%, para todos os tópicos de interesse considerados.

As seguintes Subseções 2.1.1.1 a 2.1.1.4 descrevem as melhorias propostas e validadas na abordagem original quanto, respectivamente, ao uso de Link Context em processos de coleta temática, à geração semiautomática das páginas-semente, à determinação automática de limites de similaridade a serem usados em processos de coleta temática, e à expansão automática dos conjuntos de termos de gênero e conteúdo fornecidos pelos usuários.

2.1.1.1 Uso de *Link Context*

Segundo ASSIS *et al.* (2009), o nível de eficiência de um coletor temático, dado um processo de coleta, está relacionado à proporção de páginas relevantes coletadas na internet em relação ao número de páginas visitadas pelo coletor. Além disto, TAYLAN *et al.* (2011) completam que é necessário que um processo de coleta seja realizado de forma rápida e isto depende da quantidade de URLs relevantes que são inseridas no *Frontier*.

Desta forma, visando a melhoria de eficiência da abordagem de coleta temática baseada em gênero e conteúdo, apresentada na Figura 2.1, sem perda na estabilidade e na eficácia da mesma, foi proposta em MANGARAVITE, ASSIS e FERREIRA (2012) a utilização de *Link Context*, mais precisamente texto de âncora, título do link e URL, para melhorar o processo de

[2] De acordo com CHRISTEN, HAND e KIRIELLE (2023), F1 (também conhecido como *F-measure* ou *F-score*) é uma métrica de acurácia de um teste, que consiste na média harmônica entre precisão e revocação.

determinação das pontuações de prioridade de visita, determinantes da ordenação das URLs ainda não visitadas que se encontram no *Frontier* do coletor. De uma forma geral, para computar tais pontuações, foi utilizada a distância dos cossenos entre os termos de gênero e conteúdo, parâmetros de entrada da abordagem proposta em ASSIS *et al.* (2009), e os textos gerados pela utilização do *Link Context*.

A aplicação de tal técnica resultou na melhoria da política de visita do coletor, gerando um aumento de até 100% da eficiência na abordagem original baseada em gênero e conteúdo.

2.1.1.2 Geração Semiautomática de Páginas-semente

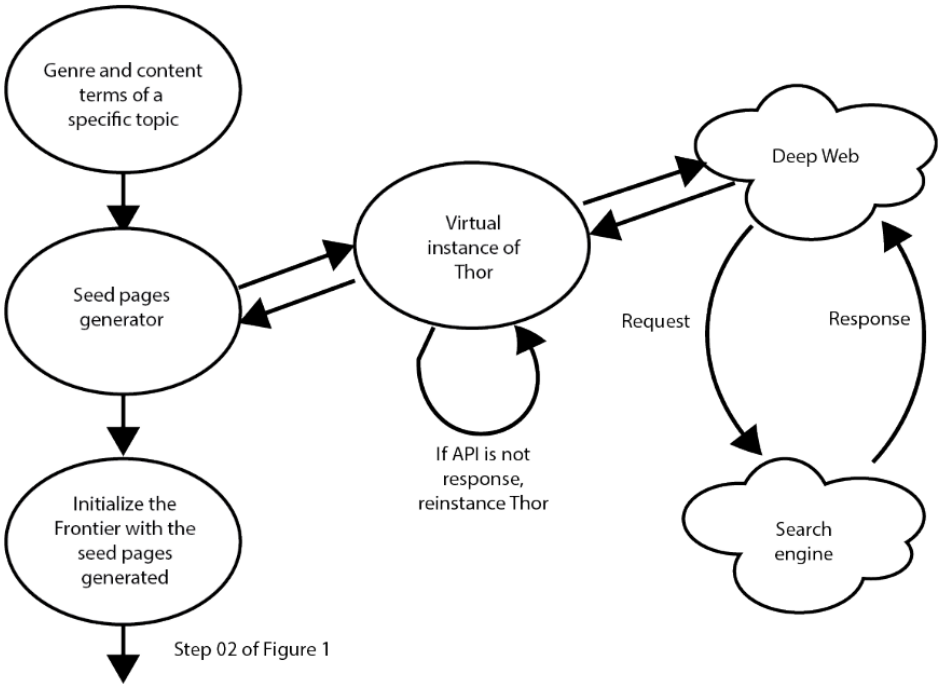
Com o intuito de também melhorar a eficiência da abordagem de coleta temática baseada em gênero e conteúdo, MANGARAVITE *et al.* (2014) propuseram uma estratégia para geração semiautomática de páginas-semente, relativas a um determinado tópico de interesse, de forma que as páginas relevantes ao tópico desejado sejam mais rapidamente localizadas pelo coletor. A arquitetura de funcionamento de tal estratégia pode ser observada na Figura 2.2, sendo que esta diz respeito apenas ao passo 01 do processo descrito pela Figura 2.1, alterando tal passo.

De acordo com a Figura 2.2, para se gerar semi automaticamente páginas-semente relativas a um determinado tópico de interesse, inicialmente, os termos de gênero e de conteúdo, especificados para tal tópico, são utilizados para se confeccionar uma consulta que é encaminhada a uma máquina de busca, mais especificamente, o Google. Para a confecção de tal consulta, foram propostas as seguintes heurísticas:

- *unionOR:* heurística que utiliza todos os termos de gênero e de conteúdo em uma única consulta, adicionando o conectivo lógico *OR*;
- *unionFirstOR e unionFirst:* heurísticas que utilizam somente o primeiro termo de gênero e de conteúdo em uma única consulta, adicionando ou não o conectivo lógico *OR*, respectivamente;
- *intersection e intersectionFirst:* heurísticas que realizam uma interseção entre todos ou apenas os primeiros termos de gênero e de conteúdo, respectivamente;
- *justContent e justContentOR:* heurísticas que utilizam apenas os termos de conteúdo em uma única consulta, adicionando ou não o conectivo lógico *OR*, respectivamente;
- *justGenre e justGenreOR:* heurísticas que utilizam apenas os termos de gênero em uma única consulta, adicionando ou não o conectivo lógico *OR*, respectivamente.

De acordo com os experimentos realizados, a melhor heurística para geração semiautomática de páginas-semente foi a *unionFirst*, que resultou em uma melhoria de eficiência na abordagem de coleta temática, proposta em ASSIS *et al.* (2009), de até 53%.

Figura 2.2 – Arquitetura de funcionamento da geração de páginas-semente



Fonte: MANGARAVITE *et al.* (2014).

Uma vez definidas as páginas-semente para um processo de coleta temática, de acordo com a Figura 2.2, as mesmas são utilizadas para inicializar o Frontier de URLs não visitadas pelo coletor. A partir daí, o processo de coleta segue o fluxo normal apresentado na Figura 2.1 (passo 02 em diante).

2.1.1.3 Determinação Automática de Limites de Similaridades

A abordagem para coleta temática de páginas Web proposta por ASSIS *et al.* (2007), ASSIS *et al.* (2008), ASSIS *et al.* (2009) utiliza a distância de cossenos para determinar a similaridade entre uma página da *Web* e os conjuntos de termos de gênero e conteúdo que representam as páginas de interesse. A medida de similaridade é utilizada para verificar se a página em questão é relevante ao tópico desejado; essa verificação ocorre por meio da comparação entre a medida de similaridade obtida e um limite de similaridade pré-estabelecido, intuitiva ou empiricamente, por um especialista. Nesse contexto, no trabalho desenvolvido por SIQUEIRA *et al.* (2016), foram desenvolvidas três estratégias para determinação automática do limite de similaridade utilizado em processos de coleta temática de páginas da *Web*.

A primeira estratégia definida busca determinar o limite de similaridade, para um tópico de interesse específico, por meio da média aritmética ou ponderada das similaridades entre as páginas-sementes e os termos de gênero e conteúdo. A segunda estratégia visa determinar o limite de similaridade mediante a aplicação de métodos de agrupamento sobre os valores de similaridade das páginas-semente; para tanto, foram considerados dois métodos de agrupamento clássicos: *K-Means* (método de particionamento) e *BIRCH* (método hierárquico). Por fim, a terceira estratégia

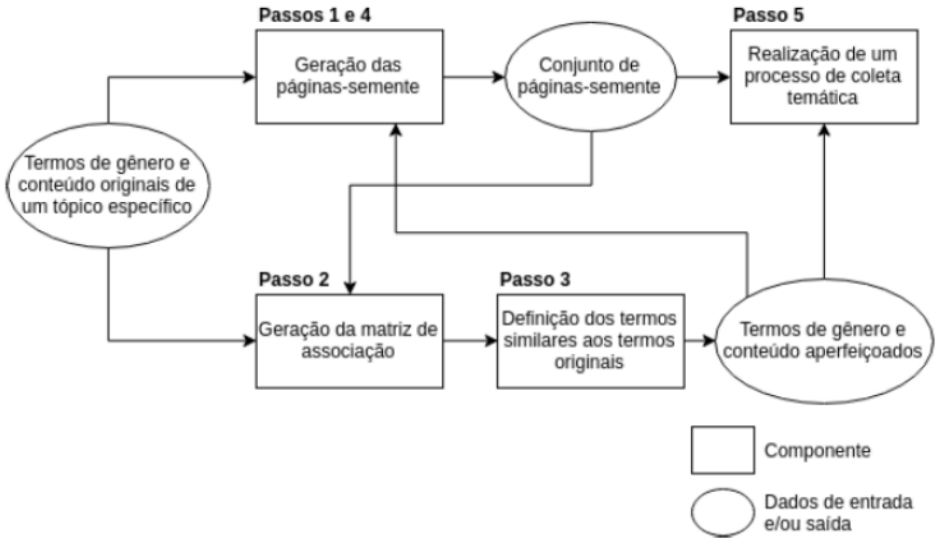
objetiva a determinação do valor do limite de similaridade por meio da maximização da métrica coeficiente de silhueta em *clusters* formados por páginas relevantes e não relevantes ao tópico em questão, de acordo com as similaridades das páginas-semente.

Para cada estratégia desenvolvida, foram realizados processos de coleta envolvendo três tópicos de interesse distintos. Por meio dos resultados obtidos, observou-se que os processos de coleta, relativos à estratégia baseada no método de agrupamento *K-Means*, foram os que apresentaram melhores eficácias, chegando a alcançar níveis de F1 bem próximos (diferença de apenas 5,4%) daqueles obtidos quando os limites de similaridade foram definidos por especialistas dos tópicos de interesse considerados.

2.1.1.4 Expansão Automática dos Conjuntos de Termos de Gênero e Conteúdo

Como visto em LIMA (2018), para que um processo de coleta temática baseada em gênero e conteúdo ocorra, conforme já mencionado, é necessário especificar o tópico de interesse desejado por meio da definição de conjuntos de termos de gênero e conteúdo que o representem. A eficácia de um processo de coleta está diretamente relacionada à qualidade dos conjuntos definidos de termos. Sendo assim, visando melhorar os conjuntos de termos fornecidos, como dados de entrada, para a abordagem original para coleta temática baseada em gênero e conteúdo apresentada na Figura 2.1, foram propostas por COSTA (2017) duas estratégias para expansão de tais conjuntos (vide Figuras 2.3 e 2.4).

Figura 2.3 – Arquitetura de funcionamento da estratégia baseada em matriz de associação



Fonte: COSTA (2017).

Conforme mostrado pela arquitetura da Figura 2.3, a primeira estratégia visa expandir os conjuntos de termos de gênero e conteúdo por meio da aplicação de uma técnica de expansão de consulta automática baseada no uso de matriz de associação[3]. Inicialmente, no passo 1, a

[3] Uma matriz de associação, de acordo com CHARTREE, CANKAYA e PHITHAKKITNUKOON (2013), é uma

partir dos conjuntos originais de termos de gênero e conteúdo, são geradas automaticamente páginas-semente que servem, no passo 2, para estabelecer a matriz de associação de termos. A partir da matriz estabelecida, o passo 3 define termos similares aos termos dos conjuntos originais, formando os conjuntos expandidos de termos de gênero e conteúdo. Esses conjuntos expandidos juntamente com as páginas-semente, obtidas por meio do passo 4 utilizando os próprios conjuntos expandidos, são utilizados como dados de entrada para a realização do processo de coleta desejado, conforme apresentado no passo 5.

A segunda estratégia, apresentada na Figura 2.4, visa expandir os conjuntos de termos de gênero e conteúdo por meio da aplicação de técnicas de Processamento de Linguagem Natural (PLN). De modo geral, no passo 1, a partir dos conjuntos originais de termos de gênero e conteúdo, é gerado automaticamente o conjunto de páginas-semente necessário para se iniciar um processo de coleta temática. No passo 2, são aplicadas técnicas de PLN (remoção de *stopwords* e técnica de *stemming*) sobre os conjuntos originais de termos para se obter os conjuntos expandidos de termos. Com os resultados dos passos 1 e 2, é ativado o processo de coleta desejado, conforme apresentado no passo 3.

Por meio da análise dos resultados dos experimentos descritos, como mencionado em LIMA (2018), foi possível perceber que a estratégia baseada em matriz de associação de termos, utilizando a métrica MenorDistância[4], foi a que apresentou melhores resultados quando comparada às demais estratégias propostas por COSTA (2017). Contudo, apesar de tal estratégia ter se sobressaído nos experimentos realizados, a melhoria apresentada por ela não foi tão satisfatória, uma vez que o melhor resultado, considerando a coleta de páginas relativas a um determinado tópico específico, promoveu um aumento na métrica F1 de apenas 6,29% ao se comparar com o valor de F1 obtido pelo processo de coleta, relativo ao mesmo tópico específico, cujos termos de gênero e conteúdo não foram expandidos.

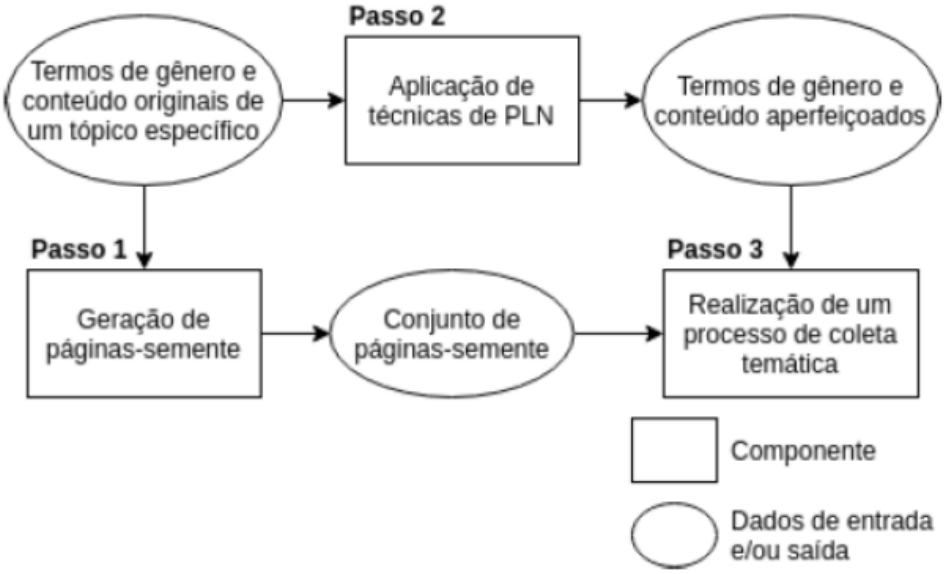
2.1.2 Implementações relativas ao Yucca

DINIZ (2018) propôs a primeira versão do coletor temático baseado em gênero e conteúdo, denominado Yucca, seguindo o que foi proposto em ASSIS *et al.* (2007), ASSIS *et al.* (2008), ASSIS *et al.* (2009) (vide Figura 2.1) e as integrações das melhorias propostas por MANGARAVITE, ASSIS e FERREIRA (2012), MANGARAVITE *et al.* (2014), SIQUEIRA *et al.* (2016) e COSTA (2017) (vide Subseções 2.1.1.1, 2.1.1.2 e 2.1.1.3, respectivamente). Porém, esta primeira versão foi desenvolvida apenas para facilitar a realização de experimentos envolvendo as integrações das melhorias citadas e não existindo, de fato, uma versão final e funcional do coletor propriamente dita em que um usuário possa utilizá-la.

estrutura de dados comumente utilizada para mensurar a relação existente entre os termos dos documentos de uma coleção.

[4] De acordo com LIMA (2018), a MenorDistância consiste em calcular a similaridade s_{ij}, entre dois termos t_i e t_j, pela soma normalizada das menores distâncias entre tais termos, considerando todas as páginas p que possuem estes termos.

Figura 2.4 – Arquitetura de funcionamento da estratégia baseada em PLN



Fonte: COSTA (2017).

De uma forma geral, nos experimentos realizados por DINIZ (2018), considerando processos de coleta relativos a dois tópicos de interesse, os resultados foram satisfatórios, destacando-se o *K-means* como heurística principal para a determinação automática de limite de similaridade, como já observado por SIQUEIRA *et al.* (2016); no caso, apresentou ganhos de até 13,21% na precisão ponderada[5] em relação às demais heurísticas. Ademais, quanto à determinação semiautomática de páginas-semente, não foi possível destacar a melhor forma de determiná-las, uma vez que os resultados obtidos nos processos de coleta realizados para os dois tópicos de interesse, embora satisfatórios, foram divergentes.

Apesar das integrações das melhorias citadas, nesta etapa não foi construída uma versão funcional e completa do coletor. Além disto, a estratégia de expansão automática de termos de gênero e conteúdo não foi integrada e um componente para apresentação das coleções geradas pelo Yucca não foi desenvolvido.

As seguintes Subseções 2.1.2.1 e 2.1.2.2 descrevem as implementações existentes relativas ao Yucca, sendo elas, respectivamente, a primeira versão funcional e não completa do Yucca e uma segunda versão envolvendo a melhoria de geração semiautomática dos conjuntos de gênero e conteúdo.

2.1.2.1 Primeira versão funcional do Yucca

Tomando como base o trabalho de ASSIS *et al.* (2007), ASSIS *et al.* (2008), ASSIS *et al.* (2009), as melhorias subsequentes descritas nas Subseções 2.1.1.1, 2.1.1.2, 2.1.1.3 e 2.1.1.4, e a proposta da versão inicial do coletor, descrita por DINIZ (2018), JUNIOR (2021) propôs uma

[5] De acordo com DINIZ (2018), precisão ponderada consiste na precisão das páginas coletadas considerando-se a ordem de relevância das mesmas em relação ao tópico desejado.

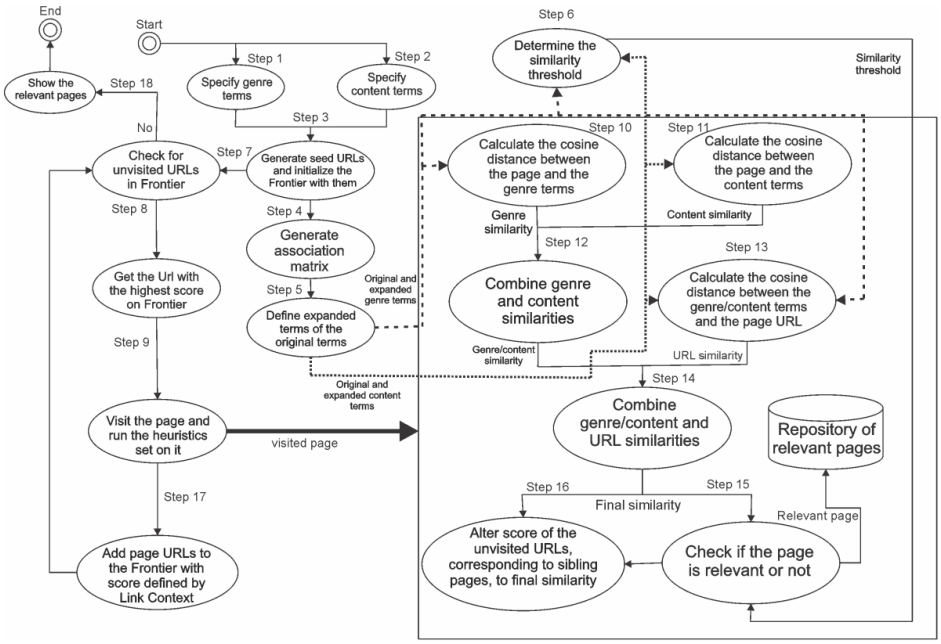
primeira versão funcional do Yucca, onde foi desenvolvida uma interface *Web* em que o usuário pudesse interagir.

A partir da abordagem descrita em DINIZ (2018), JUNIOR (2021) acrescentou ao Yucca a melhoria da expansão automática dos conjuntos de termos de gênero e conteúdo, descrita em LIMA (2018), e o componente relativo à apresentação das coleções geradas pelo Yucca. De uma forma geral, considerando um processo de coleta temática relativo a um determinado tópico de interesse, a arquitetura de funcionamento do Yucca tornou-se a ilustrada na Figura 2.5, consistindo nos seguintes passos:

- Passo 01: especificar os termos de gênero relativos ao tópico de interesse (tarefa do usuário);
- Passo 02: especificar os termos de conteúdo relativos ao tópico de interesse (tarefa do usuário);
- Passo 03: gerar semiautomaticamente as páginas-semente utilizando os termos de gênero e conteúdo especificados e inicializar o *Frontier* de URLs não visitadas pelo coletor com pontuação de prioridade de visita pelo coletor igual a 1 (um);
- Passo 04: gerar a matriz de associação de termos utilizando as páginas-semente geradas no Passo 03;
- Passo 05: gerar os termos expandidos dos termos originais de gênero e conteúdo, formando os conjuntos expandidos de termos;
- Passo 06: definir automaticamente o limite de similaridade, utilizando os termos de gênero e conteúdo especificados pelo usuário nos Passos 01 e 02, os conjuntos de termos expandidos no Passo 05 e as páginas-semente geradas no Passo 03;
- Passo 07: verificar se existem URLs na *Frontier* que não foram visitadas pelo coletor; caso não existam, o processo de coleta é encerrado;
- Passo 08: por meio da política de enfileiramento dinâmico, selecionar na *Frontier* a URL com maior pontuação relativa à prioridade de visita;
- Passo 09: Visitar a página referente à URL selecionada no passo 08, executando o conjunto proposto de heurísticas, no intuito de analisar a relevância da mesma, quanto ao tópico de interesse desejado, nos Passos 10, 11 e 13;
- Passo 10: calcular a similaridade da distância de cosseno entre a página visitada no Passo 09 e os termos de gênero especificados no Passo 01 e os termos de gênero expandidos no Passo 05;
- Passo 11: calcular a similaridade da distância de cosseno entre a página visitada no Passo 09 e os termos de conteúdo especificados no Passo 02 e os termos de conteúdo expandidos no Passo 05;

- Passo 12: combinar as similaridades de gênero e conteúdo obtidas nos Passos 10 e 11, por meio de média ponderada, gerando a similaridade de gênero/conteúdo;
- Passo 13: calcular a similaridade da distância de cosseno entre a URL da página visitada no Passo 09 e os termos de gênero e conteúdo especificados nos Passos 01 e 02 e os expandidos no Passo 05;
- Passo 14: combinar, por meio de média ponderada, a similaridade da URL obtida no Passo 13 com a similaridade de gênero/conteúdo obtida no Passo 12, gerando a similaridade final da página visitada no Passo 09;
- Passo 15: verificar se a página visitada em questão é relevante ao tópico de interesse desejado, ou seja, se a similaridade final entre tal página e tal tópico, obtida no Passo 14, é superior ao limite de similaridade determinado automaticamente no Passo 06; caso seja superior, a página visitada é considerada relevante e, assim, é armazenada no repositório de páginas relevantes ao tópico de interesse desejado;
- Passo 16: caso a página em questão seja considerada relevante no Passo 15, alterar a pontuação de prioridade de visita das páginas irmãs da página visitada, que correspondem a URLs não visitadas na *Frontier*, para a similaridade final obtida no Passo 14;
- Passo 17: adicionar as URLs da página visitada em questão na Frontier com pontuação de prioridade de visita definida pelo *Link Context*; em seguida, retornar ao Passo 07.
- Passo 18: apresentar ao usuário a coleção de páginas relevantes gerada no repositório.

Figura 2.5 – Arquitetura de funcionamento do Yucca



Fonte: JUNIOR (2021).

Dessa forma, em relação à arquitetura da abordagem original, descrita na Figura 2.1, foram acrescentados os Passos 01, 02, 04, 05 e 06 e modificados os Passos 01 (agora Passo 03) e 12 (agora Passo 17). Com isto, os Passos 06, 10, 11 e 13 utilizam agora tanto os termos especificados pelo usuário quanto os termos expandidos pela estratégia de expansão de termos por meio da expansão automática dos conjuntos de termos de gênero e conteúdo (vide Subseção 2.1.1.4).

Quanto ao funcionamento prático e geral do Yucca, foi também adicionada a possibilidade do usuário alterar características de alguns passos da arquitetura proposta, a saber: utilizar ou não a estratégia de expansão de termos (Passos 04 e 05); alterar a heurística a ser utilizada na determinação automática de limite de similaridade (Passo 06); alterar os pesos *default* associados aos termos de gênero, termos de conteúdo, URL da página visitada e combinação gênero/conteúdo para cálculo das médias ponderadas nos Passos 12 e 14; definir o número máximo de páginas-semente a serem utilizadas em um processo de coleta; e definir o número máximo de páginas a serem visitadas pelo coletor ao invés de um processo de coleta finalizar apenas quando não houver mais URLs não visitadas no Frontier (Passo 7).

Nos experimentos realizados por JUNIOR (2021), foram utilizados 3 (três) tópicos de interesse, com uma variação entre os pesos dos termos de gênero e conteúdo para o cálculo de limite de similaridade. Foi observado que, dependendo dos pesos dos termos de gênero e de conteúdo, os níveis de precisão[6] podem ser distintos, embora, independente de tais pesos, os níveis de precisão foram acima de 73% para até 10 (dez) páginas retornadas como relevantes pelo Yucca nos três tópicos. Além disso, ao considerar possíveis coleções de 60 (sessenta) páginas geradas pelo Yucca, os níveis de precisão foram superiores a 55%.

Apesar da versão proposta por JUNIOR (2021) ser funcional, ela não prevê a utilização de multi-usuários na aplicação, visto que todos as entradas feitas pelo usuário na interface da aplicação são armazenadas em arquivos de texto, para somente assim serem consumidas pelo coletor, que também salva seus dados de saída em arquivo de texto, que posteriormente é consumido pela interface e apresentados para o usuário. Isso gera problemas de concorrência na aplicação, tendo em vista que é esperado de uma aplicação *Web* a utilização por vários usuários simultaneamente. Ademais, as melhorias relativas aos Passos 03 a 06, apesar de terem sido integradas ao Yucca, não foram devidamente validadas nesta primeira versão funcional do mesmo. Outro ponto a ser mencionado é o fato do componente de apresentação de resultados (Passo 18) não ter sido devidamente implementado, apresentando somente uma lista simples das páginas obtidas no processo de coleta.

[6] JUNIOR (2021) considera como precisão, uma métrica que consiste na fração de páginas realmente relevantes ao tópico de interesse que foram retornadas pelo coletor, em relação a todas as páginas retornadas pelo mesmo.

2.1.2.2 Geração semiautomática dos conjuntos de gênero e conteúdo

Afim de reduzir o trabalho do usuário ao utilizar o Yucca, SILVA (2023) propôs uma estratégia, como uma nova melhoria no funcionamento do coletor, para gerar semiautomaticamente os termos de gênero e conteúdo que são utilizados como entrada para o coletor. Adaptando a aplicação apresentada por JUNIOR (2021), SILVA (2023) desenvolveu uma nova interface para aplicação, com a adição do módulo de geração semiautomática dos termos de gênero e conteúdo.

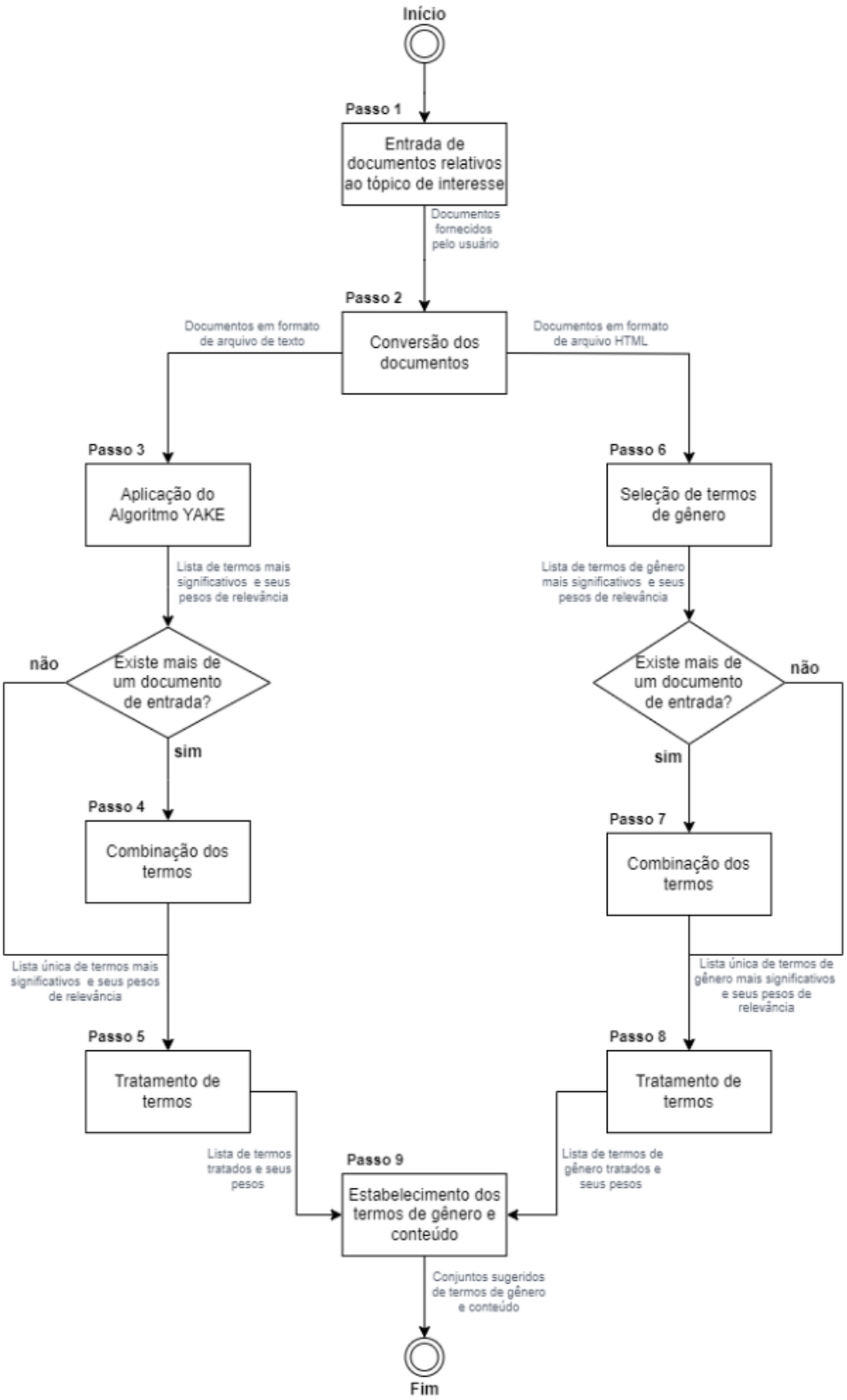
Deste modo, na etapa inicial da aplicação, onde era necessário o usuário submeter manualmente os conjuntos termos de gênero e conteúdo, foi desenvolvido um módulo onde é possível com que este procedimento seja feito de forma semiautomática, em que o usuário necessita apenas fornecer ao coletor URLs ou documentos em formato pdf que sejam relevantes ao tópico desejado. A partir destes documentos, é gerado um conjunto de termos de gênero e conteúdo, no qual o usuário deve selecionar os que lhe convêm ou adicionar outros manualmente, caso deseje.

A arquitetura de funcionamento da estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo encontra-se na Figura 2.6 e pode ser descrita nos seguintes passos:

- Passo 01: consiste em fornecer o(s) documento(s) relativo(s) ao tópico de interesse, sendo possível adicionar até quatro documentos (tarefa do usuário);
- Passo 02: 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: 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[7], 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: 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;

[7] De acordo com CAMPOS *et al.* (2020), o YAKE é um algoritmo de extração de palavras-chave em documentos de forma automática.

Figura 2.6 – Arquitetura de funcionamento da estratégia para a geração semiautomática dos conjuntos iniciais de termos de gênero e conteúdo.



Fonte: SILVA (2023).

- Passo 05: tratar os termos por meio de duas etapas, sendo elas: aplicação da técnica de *lemmatização*[8] 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: a partir dos documentos convertidos em arquivos de texto no Passo 03, selecionar os termos de gênero a partir da semântica das *tags* HTML 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: similarmente ao Passo 04, 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 08: 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: 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 10: 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.

Com o objetivo de avaliar a eficácia da estratégia para geração dos conjuntos iniciais de termos de gênero e conteúdo, SILVA (2023) realizou experimentos considerando 3 (três) assuntos distintos, sendo eles: planos de ensino de Banco de Dados, artigos relacionados ao COVID 19 e

[8] De acordo com PLISSON, LAVRAC e MLADENIC (2004), lemmatização 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.

receitas de bolo de cenoura. Em todos os cenários, a estratégia demonstrou níveis de precisão[9] significativos para os termos gerados a partir de documentos relevantes fornecidos como dados de entrada, sendo iguais ou superiores a 75% nos testes conduzidos.

Contudo, a aplicação proposta por SILVA (2023) sofre do mesmo problema da aplicação de JUNIOR (2021), já mencionado na Seção 2.1.2.1, onde a mesma não prevê o uso de múltiplos usuários, visto que pode haver problemas de concorrência na leitura e escrita dos dados.

2.2 Trabalhos Relacionados

Em LEE *et al.* (2020), é proposto um coletor temático chamado de *SlideCrawler*, que tem, como objetivo, coletar arquivos de apresentações de slides de instituições acadêmicas, utilizando o mecanismo de busca do Google para realizar consultas na *Web* e realizar o download dos arquivos desejados. O processo de coleta é dividido em três etapas. Na primeira etapa, é utilizado um gerador de consultas, definido anteriormente, que irá gerar uma consulta no Google, especificando a instituição de ensino desejada e os formatos de arquivos desejados. No segundo passo, é feita a extração das URLs do resultado gerado no passo anterior, afim de selecionar apenas aquelas, que direcionam para o download de um arquivo no formato de apresentação de slides, e também de remover possíveis duplicatas. Já na terceira etapa, é feito o download dos arquivos obtidos no passo anterior. Nos experimentos realizados, o *SlideCrawler* mostrou-se bastante eficiente, conseguindo fazer o download de mais de 800 mil arquivos de apresentação de slides de mais de 500 universidades distintas. Quando comparado com outro coletor já conhecido na literatura, o *Apache Nutch*, o *SlideCrawler* foi capaz de coletar aproximadamente 3.7 vezes mais arquivos de apresentações de slides.

Após analisar algumas das técnicas atuais utilizadas para a construção de coletores temáticos, GUO (2021) propôs uma implementação simples, utilizando algumas dessas técnicas, onde é possível coletar várias páginas de um mesmo site, afim de alimentar um baco de informações. O processo de coleta tem início com o fornecimento de URLs semente para o coletor, que checa se na raiz do site há um arquivo de configuração chamado *Robots.txt*. Caso esse arquivo esteja presente, o escopo do processo de coleta é determinado pelo conteúdo do arquivo; caso contrário, o coletor somente tenta acessar a URL fornecida. Se a URL for acessível e não houver nenhum erro ao recuperar seu conteúdo, as URLs irmãs contidas na página são recuperadas e armazenadas numa fila para poderem ser acessadas futuramente; caso contrário, a URL é armazenada numa lista de URLs que apresentaram algum tipo de erro, para que não seja necessário acessá-la novamente, caso a mesma encontre-se em alguma outra página. Este tipo de implementação, além de ser bem simples, não requer um *hardware* muito avançado para sua execução, como demonstrado nos testes, onde um grande número de páginas foi coletado em um período razoável de tempo.

[9] SILVA (2023) considera como precisão, uma métrica chamada de precisão ponderada, que dá a média ponderada de precisão com pesos iguais à ordem em que termos foram encontrados.

Visando contornar problemas de conexão e de velocidade no processo de coleta de páginas, KUMAR e AGGARWAL (2023) propuseram uma abordagem onde as páginas, classificadas como relevantes no processo de coleta, são armazenadas em uma base de dados e atualizadas de acordo com um escalonador, fazendo com que as buscas futuras do mesmo tópico sejam retornadas a partir da base de dados, dispensando um novo processo de coleta na *Web*. No processo de coleta de novos tópicos, para cada página encontrada, é calculado uma pontuação, baseando-se na quantidade de vezes que a palavra de entrada é encontrada dentro da página. As páginas então são enviadas a um classificador previamente treinado, que identifica se a página é ou não relevante. Após isso, as páginas relevantes são armazenadas na base de dados, onde serão atualizadas de acordo com a frequência de atualização de cada site. Segundo os testes realizados, o coletor proposto mostrou-se 275% mais eficiente em termos de velocidade, se comparado a outros modelos clássicos de coletores.

Na Tabela 2.1, são apresentadas as funcionalidades propostas pelos trabalhos citados acima, comparando-as com as do Yucca. É possível observar que nenhum deles faz uso de ambos termos de gênero e conteúdo, utilizando apenas um dos dois. Também é visto que os trabalhos não possuem métodos para a expansão automática dos termos fornecidos como entrada e para a geração semiautomática dos mesmos. O trabalho que mais se aproxima do Yucca é o proposto em KUMAR e AGGARWAL (2023), que também apresenta uma técnica para determinar o limite de similaridade, gera de forma semiautomática as páginas-semente e faz a análise dos *links* encontrados; por outro lado, distancia em outros pontos como, por exemplo, no uso de um classificador para análise das páginas coletadas. Além disso, todos os trabalhos fazem o *download* das páginas e/ou arquivos coletados, ao contrário do Yucca, que somente gera uma coleção de *links* ao final de um processo de coleta.

Tabela 2.1 – Comparativo de funcionalidades

	Autores			
Funcionalidades	LEE *et al.* (2020)	GUO (2021)	KUMAR e AGGARWAL (2023)	Yucca
Análise de links	x	x	x	x
Determinação automática de limite de similaridade			x	x
Download de páginas e/ou arquivos coletados	x	x	x	
Expansão automática de termos de gênero e conteúdo				x
Geração semiautomática de páginas semente	x		x	x
Geração semiautomática de termos de gênero e conteúdo				x
Uso de classificador			x	
Uso de termos de conteúdo		x	x	x
Uso de termos de gênero	x			x

Fonte: Elaborado pelo autor.

Como visto em todos os trabalhos analisados, fazer a análise dos *links* é importante para que se tenha uma melhor eficácia no processo de coleta; além disso, pôde-se perceber nos

trabalhos LEE *et al.* (2020), KUMAR e AGGARWAL (2023) e também na Seção 2.1.1.2 que a geração semiautomática de páginas-semente também gera um ganho de eficiência no processo de coleta. Outras funcionalidades que também podem trazer ganhos significativos de eficiência para um processo de coleta são a determinação automática de limite de similaridade e a expansão de termos de entrada, como visto nas Seções 2.1.1.3 e 2.1.1.4 respectivamente. Além disso, o fato de se utilizar ambos termos de gênero e conteúdo faz com o Yucca tenha ótimos resultados, atingindo níveis de F1 superiores a 88%, como pode ser visto desde a concepção original, descrita na Seção 2.1.1. Também é possível observar que quando aplicada técnicas de geração semiautomática destes termos, apesar de não interferir nos resultados obtidos pelo coletor, há uma melhora na experiência de uso do coletor pelo usuário, como pode ser visto na Seção 2.1.2.2. Logo, devido a todas essas funcionalidades presentes no Yucca, observa-se que o mesmo pode proporcionar a geração de coleções de páginas relevantes a um determinado tópico de interesse desejado, de forma eficaz e eficiente. No entanto falta aprimorar a experiência do usuário quanto a utilização do Yucca, como visto nas Seções 2.1.2.1 e 2.1.2.2, possibilitando a utilização do mesmo por multi-usuários simultaneamente e uma interface simples e amigável, afim de não gerar erros de exibição de dados e nenhum tipo de dúvidas quanto à sua utilização.

3 Desenvolvimento

Como mencionado na Seção 1.2, este trabalho de monografia possui, como objetivo geral, o desenvolvimento e a validação de uma versão completa e funcional do Yucca, coletor temático de páginas *Web* baseado em gênero e conteúdo, buscando sanar os problemas apresentados nas implementações anteriores (vide Seção 1.1) e adicionando novas funcionalidades, como por exemplo, a possibilidade do usuário especificar manualmente páginas semente e o limite de similaridade que serão utilizados no processo de coleta e também, um componente para apresentação de resultados, que permite que o usuário consiga visualizar com clareza e exportar a coleção gerada ao final do processo de coleta. Para tal, considera-se, como base, a abordagem descrita na Figura 2.5, que envolve a própria abordagem original de coleta e as melhorias apresentadas nas Subseções 2.1.1.1, 2.1.1.2, 2.1.1.3 e 2.1.1.4.

Desta forma, este capítulo apresenta uma proposta para uma versão final e funcional do Yucca, estando organizado da seguinte maneira: a Seção 3.1 apresenta as tecnologias e recursos que foram utilizados para a implementação do Yucca, a Seção 3.2 descreve uma nova arquitetura de funcionamento do Yucca e a Seção 3.3 apresenta a nova interface do Yucca, adequada à nova arquitetura proposta e envolvendo a parametrização necessária para executá-lo.

3.1 Tecnologias e recursos utilizados

Para garantir um bom nível de performance e facilitar manutenções futuras, foram escolhidas tecnologias atuais e amplamente utilizadas para o desenvolvimento do Yucca. No *frontend*, optou-se pelo ReactJS Facebook (2013), uma biblioteca renomada pela sua eficiência e flexibilidade na construção de interfaces dinâmicas. Já no *backend*, o *microframework* Flask Ronacher (2010) foi selecionado por sua simplicidade e robustez na implementação de APIs e processamento de dados. Nas Subseções 3.1.1 e 3.1.2, essas tecnologias são detalhadas em maior profundidade e na Subseção 3.1.3, é explicado como é feita a integração entre as duas.

3.1.1 ReactJS

O ReactJS, desenvolvido e mantido pelo Facebook desde 2013, é uma biblioteca JavaScript amplamente utilizada para a construção de interfaces de usuário interativas e eficientes. Sua principal característica é a utilização de um conceito chamado "componentes", onde a interface é fragmentada em partes reutilizáveis, tornando o código mais organizado e fácil de manter. Além disso, o React utiliza o *Virtual DOM*, que otimiza a atualização da interface de acordo com as mudanças no estado da aplicação, resultando em uma experiência de usuário mais fluida e performática.

A tecnologia permitiu a construção de uma interface dinâmica, onde o usuário pode interagir com os diferentes componentes do Yucca, que serão apresentados na Seção 3.3. A capacidade do React de lidar com grandes volumes de dados e atualizações rápidas foi essencial para garantir uma navegação eficiente e responsiva, mesmo com um número elevado de páginas sendo coletadas.

3.1.2 Flask

O Flask é um *microframework* escrito em Python e lançado em 2010. O Flask é conhecido por sua simplicidade e flexibilidade, permitindo que desenvolvedores criem aplicações web robustas com uma estrutura mínima. Embora simples, o Flask oferece uma série de funcionalidades essenciais para o desenvolvimento de *APIs RESTful*, como o roteamento de URLs, suporte a *templates* e integração com bancos de dados.

No *backend* do Yucca, o Flask foi utilizado para implementar a lógica de coleta das páginas e processamento dos dados. Ao receber uma requisição do *frontend* (via React), o Flask executa as etapas do processo de coleta, que serão explicados na Seção 3.2.

3.1.3 Integração entre as Tecnologias

A comunicação entre o *frontend* (ReactJS) e o *backend* (Flask) foi feita via requisições HTTP utilizando o protocolo *REST*. O React envia as URLs a serem processadas através de requisições *POST* para o servidor Flask, que por sua vez realiza a coleta dos dados e retorna o resultado em formato JSON. Esta arquitetura cliente-servidor permitiu uma clara separação de responsabilidades, tornando o sistema mais modular e permitindo futuras expansões, como a adição de novas melhorias e funcionalidades.

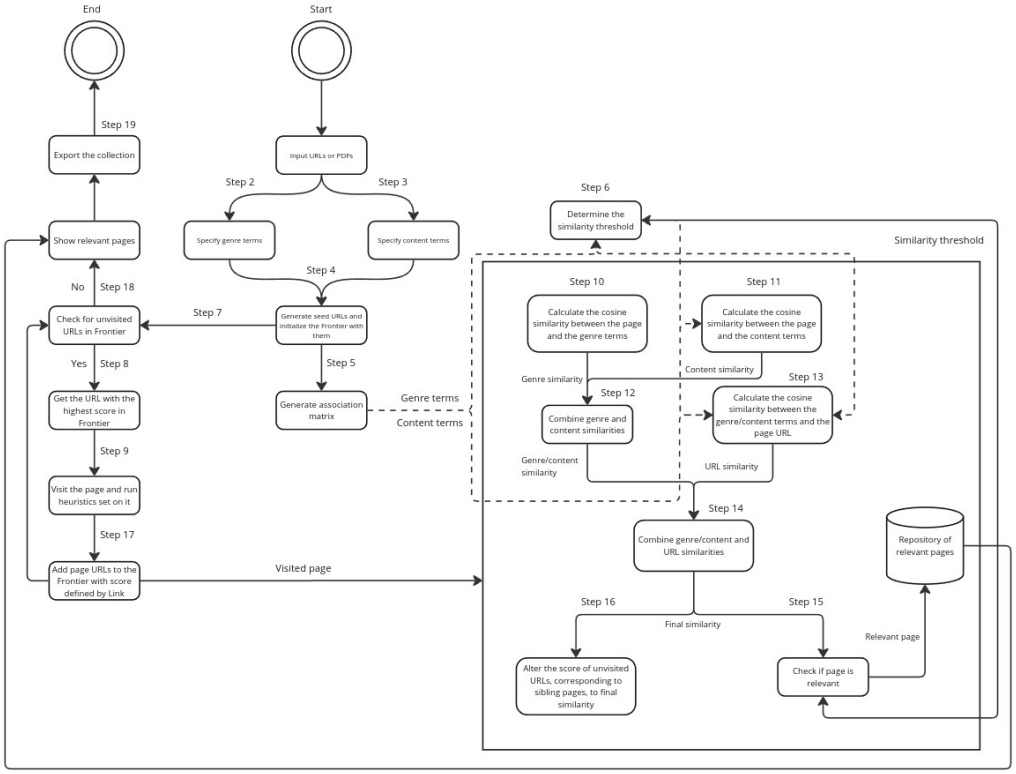
3.2 Arquitetura de Funcionamento do Yucca

A arquitetura de funcionamento do Yucca, é apresentada na Figura 3.1. Em relação à arquitetura da abordagem descrita na Figura 2.5, acrescentou-se ao Yucca a melhoria de geração semiautomática de termos de gênero e conteúdo, apresentada na Subseção 2.1.2.2, e a apresentação das coleções geradas pelo mesmo.

De uma forma geral, considerando o processo de coleta temática relativo a um determinado tópico de interesse, a arquitetura de funcionamento do Yucca, ilustrada na Figura 3.1, consiste nos seguintes passos:

- Passo 01: inserir URLs ou PDFs relativos ao tópico de interesse, afim de se gerar, semi-automaticamente, uma lista inicial de termos de gênero e conteúdo;

Figura 3.1 – Arquitetura de funcionamento do Yucca.



Fonte: Elaborado pelo autor.

- Passo 02: especificar os termos de gênero relativos ao tópico de interesse (tarefa do usuário), considerando ou não a lista inicial de termos de gênero gerada semi-automaticamente (Passo 01);

- Passo 03: especificar os termos de conteúdo relativos ao tópico de interesse (tarefa do usuário), considerando ou não a lista inicial de termos de conteúdo gerada semi-automaticamente (Passo 01);

- Passo 04: gerar semiautomaticamente as páginas-semente utilizando os termos de gênero e conteúdo especificados (Passos 02 e 03) e inicializar o *Frontier* de URLs não visitadas pelo coletor com pontuação de prioridade de visita pelo coletor igual a 1 (um);

- Passo 05: gerar a matriz de associação de termos utilizando as páginas-semente geradas no Passo 04;

- Passo 06: definir automaticamente o limite de similaridade, utilizando os termos de gênero e conteúdo especificados pelo usuário nos Passos 02 e 03 e as páginas-semente geradas no Passo 04;

- Passo 07: verificar se existem URLs na *Frontier* que não foram visitadas pelo coletor; caso não existam, o processo de coleta é encerrado;

- Passo 08: por meio da política de enfileiramento dinâmico, selecionar na *Frontier* a URL com maior pontuação relativa à prioridade de visita;
- Passo 09: Visitar a página referente à URL selecionada no Passo 08, executando o conjunto proposto de heurísticas, no intuito de analisar a relevância da mesma, quanto ao tópico de interesse desejado, nos Passos 10, 11 e 13;
- Passo 10: calcular a similaridade da distância de cosseno entre a página visitada no Passo 09 e os termos de gênero especificados no Passo 02;
- Passo 11: calcular a similaridade da distância de cosseno entre a página visitada no Passo 09 e os termos de conteúdo especificados no Passo 03;
- Passo 12: combinar as similaridades de gênero e conteúdo obtidas nos Passos 10 e 11, por meio de média ponderada, gerando a similaridade de gênero/conteúdo;
- Passo 13: calcular a similaridade da distância de cosseno entre a URL da página visitada no Passo 09 e os termos de gênero e conteúdo especificados nos Passos 02 e 03;
- Passo 14: combinar, por meio de média ponderada, a similaridade da URL obtida no Passo 13 com a similaridade de gênero/conteúdo obtida no Passo 12, gerando a similaridade final da página visitada no Passo 09;
- Passo 15: verificar se a página visitada em questão é relevante ao tópico de interesse desejado, ou seja, se a similaridade final entre tal página e tal tópico, obtida no Passo 14, é superior ao limite de similaridade determinado automaticamente no Passo 06; caso seja superior, a página visitada é considerada relevante e, assim, é armazenada no repositório de páginas relevantes ao tópico de interesse desejado;
- Passo 16: caso a página em questão seja considerada relevante no Passo 15, alterar a pontuação de prioridade de visita das páginas irmãs da página visitada, que correspondem a URLs não visitadas na *Frontier*, para a similaridade final obtida no Passo 14;
- Passo 17: adicionar as URLs da página visitada em questão na Frontier com pontuação de prioridade de visita definida pelo *Link Context*; em seguida, retornar ao Passo 07;
- Passo 18: uma vez finalizado um processo de coleta, seja pela ausência de URLs não visitadas ou por um critério previamente estabelecido de parada, apresentar ao usuário a coleção de páginas relevantes gerada no repositório;
- Passo 19: caso seja do interesse do usuário, exportar a coleção gerada em diferentes formatos de arquivo.

Desta forma, em relação à arquitetura descrita na Figura 2.5, foi acrescentado o Passo 01, que corresponde a geração semiautomática de termos de gênero e conteúdo, apresentada na

Subseção 2.1.2.2, antes da especificação dos termos de gênero e conteúdo pelo usuário (Passos 02 e 03). O antigo passo 05, que se refere à estratégia de expansão de termos, apresentada na Subseção 2.1.1.4, foi removido, tendo em vista a inclusão do Passo 01, que o torna redundante devido às similaridades nas estratégias de geração de novos termos. No Passo 18, houve também uma modificação: foi acrescentada a possibilidade do usuário configurar um critério de parada para o coletor como, por exemplo, o número de páginas visitadas em um processo de coleta. Ademais, foi acrescentado o Passo 19, que possibilita a exportação da coleção obtida em um arquivo compactado, contendo distintos formatos de arquivo (XLSX, CSV e TXT) com a coleção gerada, por meio de parâmetros utilizados para a sua geração, para que seja mais fácil o armazenamento e tratamento das páginas em questão por parte do usuário.

3.3 Nova interface do Yucca

Nessa seção, é apresentada a nova interface do Yucca, envolvendo a parametrização necessária para executá-lo.

Na Figura 3.2, é apresentada a tela inicial do Yucca, que tem como objetivo ser bem simples, porém preservando a robustez do sistema, para que não cause nenhum tipo de dúvida aos usuários durante sua utilização. Para ajudar nessa função, foram adicionados alguns botões informativos, que mostram uma breve explicação do componente ao se passar com o *mouse* ou clicar sobre os mesmos. Nessa tela, inicialmente, o usuário pode inserir quantos termos de gênero e conteúdo quiser, por meio das caixas de texto especificadas.

Figura 3.2 – Tela inicial do Yucca.

Fonte: Elaborado pelo autor.

Também é possível observar na Figura 3.2 que existem botões do tipo *toggle*, que habilitam ou desabilitam alguns componentes onde o usuário pode interagir com o processo de coleta.

O primeiro deles refere-se ao componente de geração semiautomática de termos de gênero e conteúdo. Quando habilitado, como apresentado na Figura 3.3, o usuário tem a possibilidade de inserir URLs ou PDFs, para que possam ser usados na geração de novos termos de gênero e conteúdo. Há uma sugestão de inserir no máximo dez URLs ou PDFs, devido ao tempo necessário para a geração dos termos, mas nada impede que o usuário ultrapasse esse valor. Esse componente foi totalmente reestilizado, quando comparado a versão proposta em SILVA (2023), contendo também novas funcionalidades. Foi implementado uma função que remove as URLs ou PDFs que foram adicionados na lista, ao clicar no botão com o ícone de um "X"vermelho, ao lado direito de um dos elementos adicionados; ademais, caso o usuário clique no botão "Limpar", todos os elementos adicionados na lista de URLs e PDFs serão removidos. Por fim, ao clicar no botão "Enviar", caso a lista contenha pelo menos um elemento, termos de gênero e conteúdo serão semiautomaticamente gerados, como explicado na Seção 2.1.2.2, e estes termos serão adicionados à lista das caixas de textos especificadas, como pode ser observado na Figura 3.4

Figura 3.3 – Componente de geração semiautomática de termos de gênero e conteúdo.

Fonte: Elaborado pelo autor.

A partir desta lista de termos de gênero e conteúdo, o usuário pode selecionar quantos quiser para servir de entrada para o processo de coleta, como pode ser visto na Figura 3.4, onde são apresentados uma lista de termos de conteúdo obtidos a partir das URLs fornecidas com o tópico de interesse "Receita de bolo". Caso haja necessidade, o usuário também tem a opção de inserir novos termos manualmente, não ficando restrito somente aos termos semiautomaticamente gerados.

O segundo botão do tipo *toggle*, apresentado na Figura 3.2, habilita um componente que possibilita o usuário inserir páginas semente manualmente, a serem utilizadas juntamente com as páginas semente obtidas no Passo 04 da Seção 3.2. Como pode ser visto na Figura 3.5, o usuário pode inserir quantas URLs quiser, tendo a possibilidade de excluir uma a uma, caso clique no botão com o ícone de um "X"vermelho, ou excluir todas, caso clique no botão "Limpar".

Figura 3.4 – Geração de termos de gênero e conteúdo.

Fonte: Elaborado pelo autor.

Figura 3.5 – Configuração das páginas-semente de um processo de coleta.

Fonte: Elaborado pelo autor.

Já o terceiro botão, habilita um componente que possibilita o usuário inserir manualmente o limite de similaridade do processo de coleta. Quando habilitado, o Passo 06 da Seção 3.2, que faz o cálculo do limite de similaridade automaticamente, será desconsiderado e o valor inserido pelo usuário será utilizado para o mesmo propósito. Como pode ser visto na Figura 3.6, existe apenas uma caixa de texto simples, onde o usuário deve especificar um valor compreendido entre 0 e 1.

Figura 3.6 – Configuração do limite de similaridade de um processo de coleta.

Fonte: Elaborado pelo autor.

Caso seja necessário, o usuário também pode customizar o processo de coleta, alterando os parâmetros necessários utilizados no mesmo. A partir da tela inicial (Figura 3.2), ao clicar no botão "Configurar coleta", um *Modal* é aberto, como pode ser visto na figura 3.7, onde o usuário pode fazer a customização dos parâmetros ao seu gosto. Para salvar as alterações, basta clicar no botão "Salvar"; caso o usuário deseje voltar aos valores padrão da coleta, basta clicar no botão "Restaurar valores padrão". Como pode ser visto na Figura 3.7, os parâmetros configuráveis são: o número máximo de páginas-semente a serem coletadas, o número máximo de páginas visitadas, o peso dos termos de gênero, o peso dos termos de conteúdo, o peso da URL, o peso da combinação gênero/conteúdo e a heurística a ser utilizada para a determinação automática do limite de similaridade.

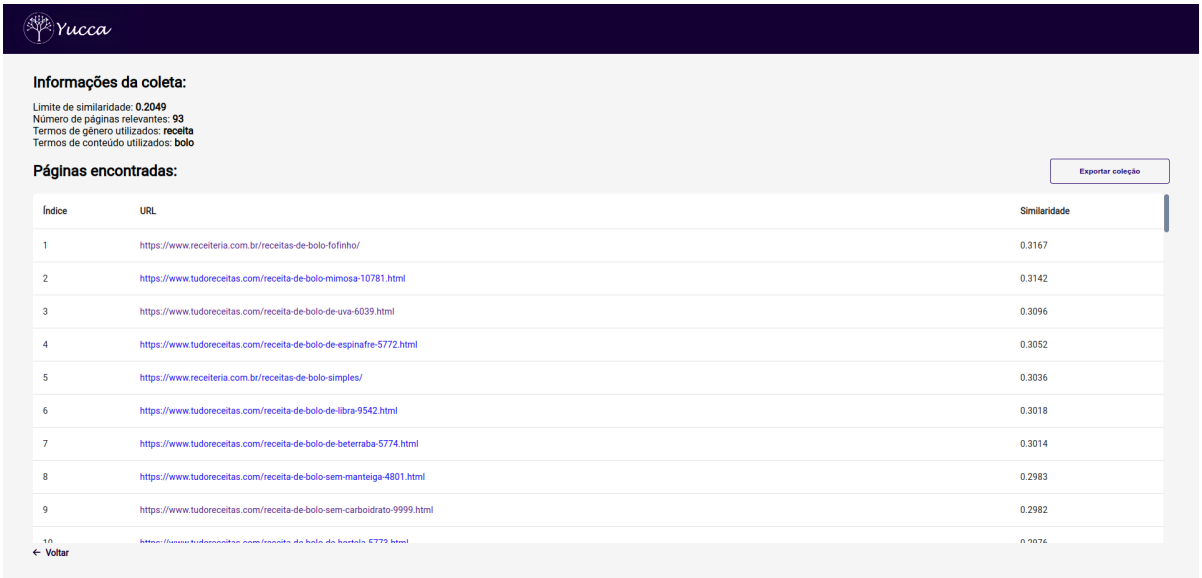
Após especificados os termos de gênero e conteúdo desejados, a partir da tela inicial (Figura 3.2), ao clicar no botão "Iniciar coleta", o processo de coleta temática terá o seu início. Quando o processo for finalizado, o usuário será redirecionado à tela de apresentação de resultados, que pode ser vista na Figura 3.8, onde foi realizado um processo de coleta com o tópico de interesse "Receita de bolo"e com limite de cem páginas visitadas. Nesta tela, são apresentadas algumas informações da coleta, como o limite de similaridade calculado (ou o limite inserido pelo usuário), o número de páginas consideradas relevantes pelo Yucca e também os termos de gênero

Figura 3.7 – Configuração dos parâmetros de um processo de coleta.

Fonte: Elaborado pelo autor.

e conteúdo utilizados. Logo abaixo, é apresentada uma tabela, contendo a coleção de páginas relevantes encontradas pelo Yucca, ordenadas por sua similaridade. O usuário tem a opção de exportar a coleção obtida, clicando no botão "Exportar coleção", onde é feito o *download* de um arquivo compactado, contendo arquivos nos formatos XLSX, CSV e TXT, com a coleção obtida. Caso o usuário clique no botão escrito "Voltar", que se encontra logo abaixo da tabela, ou no logotipo do Yucca, que se encontra na barra superior da tela, ele será redirecionado para a tela de início (Figura 3.2).

Figura 3.8 – Apresentação dos resultados



Fonte: Elaborado pelo autor.

4 Experimentação Prática

Neste capítulo, são apresentados e analisados os experimentos de validação de uma nova versão funcional do Yucca, seguindo a arquitetura proposta na Figura 3.1. A Seção 4.1 descreve os experimentos realizados e a Seção 4.2 apresenta e avalia os resultados obtidos por meio dos experimentos realizados.

4.1 Descrição dos Experimentos

Para avaliar a nova versão funcional do Yucca, foram realizados processos de coleta considerando 3 tópicos distintos, sendo eles:

- receitas de bolo de chocolate;
- artigos relacionados ao aquecimento global;
- notícias relacionadas aos conflitos em Israel.

Os processos de coleta realizados possuíram as seguintes características:

- Conjuntos de termos de gênero e conteúdo que definem os tópicos especificados (vide Tabelas 4.1, 4.2 e 4.3); para os dois primeiros tópicos, foi utilizado o componente de geração semiautomática de geração de termos (vide Subseção 2.1.2.2) para criar os termos automaticamente, tomando com base as fontes descritas nas Tabelas 4.4 e 4.5 que, de forma manual, foram escolhidos os mais adequados para o processo de coleta; já para o terceiro tópico, os termos foram escolhidos manualmente sem a utilização do componente de geração semiautomática de geração de termos;
- Conjuntos de páginas-semente, para cada tópico especificado, obtidos por meio da melhor heurítica proposta por MANGARAVITE *et al.* (2014) (vide Subseção 2.1.1.2);
- Número máximo de páginas visitadas: 1000;
- Número máximo de páginas-semente: 50;
- Peso da URL: 0.5;
- Peso da combinação gênero/conteúdo: 0.5;
- Dentre as páginas retornadas pelo Yucca, ordenadas em ordem decrescente de acordo com suas similaridades, quantidade máxima de páginas avaliadas para se calcular a precisão: 60;

- Para cada tópico, foi realizado um estudo da importância de conteúdo e gênero, por meio dos seguintes pesos: gênero 0.7 e conteúdo 0.3; gênero 0.5 e conteúdo 0.5; e gênero 0.3 e conteúdo 0.7.

Ao final de cada processo de coleta, a coleção gerada foi exportada visando a análise das páginas retornadas pelo Yucca.

Tabela 4.1 – Conjunto de termos que definem o tópico "receitas de bolo de chocolate".

Termos de Gênero	**Termos de Conteúdo**
modo preparo	bolo
ingredientes	chocolate

Tabela 4.2 – Conjunto de termos que definem o tópico "artigos relacionados ao aquecimento global".

Termos de Gênero	**Termos de Conteúdo**
consequências	aquecimento global
causas	efeito estufa
-	mudanças climáticas
-	dióxido de carbono

Tabela 4.3 – Conjunto de termos que definem o tópico "notícias relacionadas aos conflitos em Israel".

Termos de Gênero	**Termos de Conteúdo**
notícia	conflitos
reportagem	Israel
matéria	Gaza
-	Palestina
-	Irã

Tabela 4.4 – PDFs de receitas de bolo de chocolate

Fonte
https://www.armovos.com/receitas/Bolo%20de%20Chocolate.pdf
https://www.ribeiraopreto.sp.gov.br/portal/pdf/educacao1089202210.pdf
https://www.acucarcaravelas.com.br/assets/pdf/receitas/bolo-simples-de-chocolate-.pdf

Tabela 4.5 – URLs de artigos relacionados ao aquecimento global

Fonte
https://www.wwf.org.br/natureza_brasileira/reducao_de_impactos2/clima/mudancas_climaticas2/
https://fia.com.br/blog/aquecimento-global/
https://www.todamateria.com.br/aquecimento-global/

Para a realização dos experimentos, foi utilizado um *desktop* com as seguintes especificações: sistema operacional Ubuntu 24.04, processador AMD(R) Ryzen(TM) 5 5600 com

frequência de 3.5 GHz e RAM de 32GB com frequência de 3200MHz. Cada experimento obteve um tempo médio de aproximadamente 30 minutos de execução.

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 da Seção 4.1.

Considerando todos os processos de coleta realizados, a Tabela 4.6 apresenta, para cada tópico, o limite de similaridade atingido, a quantidade de páginas visitadas e a quantidade de páginas retornadas e, consequentemente, consideradas relevantes pelo Yucca. Tais valores são apresentados para cada caso de teste realizado para um mesmo tópico, variando o peso dos termos de gênero e conteúdo.

Tabela 4.6 – Resultado dos experimentos realizados

Tópico	**Caso de teste**	**Limite de similaridade**	**Quantidade de páginas visitadas**	**Quantidade de páginas relevantes**
Receitas de bolo de chocolate	Genero: 0.7 Conteúdo: 0.3	0.1219	1000	936
	Genero: 0.5 Conteúdo: 0.5	0.1407	1000	848
	Genero: 0.3 Conteúdo: 0.7	0.1513	1000	739
Artigos relacionados a aquecimento global	Genero: 0.7 Conteúdo: 0.3	0.1364	1000	553
	Genero: 0.5 Conteúdo: 0.5	0.1409	1000	608
	Genero: 0.3 Conteúdo: 0.7	0.1361	1000	633
Notícias relacionadas aos conflitos em Israel	Genero: 0.7 Conteúdo: 0.3	0.1270	1000	810
	Genero: 0.5 Conteúdo: 0.5	0.1285	1000	723
	Genero: 0.3 Conteúdo: 0.7	0.1365	1000	970

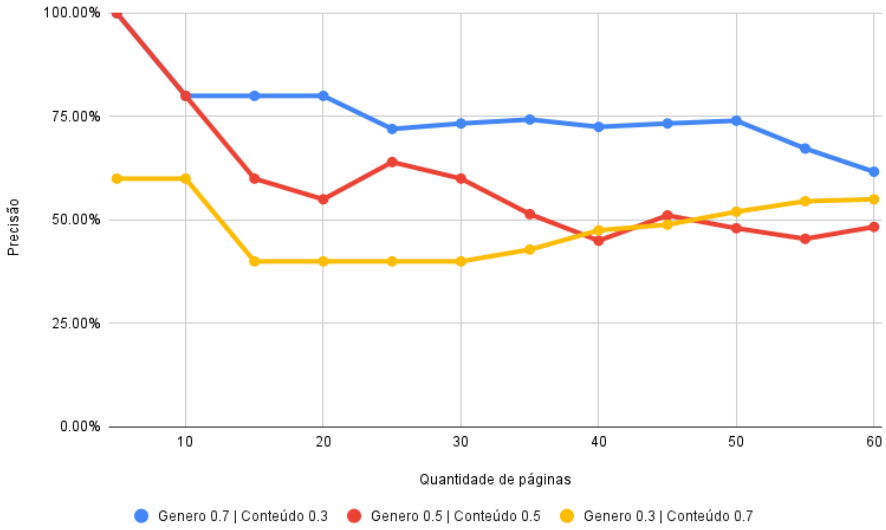
Para verificar se uma determinada página retornada pelo Yucca é realmente relevante para um determinado tópico de interesse, foi realizada uma análise da própria página associada à URL retornada pelo coletor. Caso a página seja relacionada ao tópico especificado e tenha o mesmo gênero ao qual foi pesquisado, ela foi considerada relevante. A Tabela 4.7 apresenta as URLs de algumas páginas consideradas relevantes pelo Yucca para o tópico "artigos relacionados ao aquecimento global", por exemplo, correspondendo a artigos considerados relevantes e não relevantes na análise das páginas.

Tabela 4.7 – Exemplos de URLs visitadas pelo Yucca

Páginas relevantes	**Páginas não relevantes**
https://solucoes.edp.com.br/blog/aquecimento-global-o-que-e-causas-e-consequencias-edp/	https://t.me/share/url?url=https://exame.com/esg/aquecimento-global-o-que-e-causas-e-consequencias/&text=Aquecimento Global: o que é, causas, efeitos, consequências
https://exame.com/esg/aquecimento-global-o-que-e-causas-e-consequencias/	https://www.ecodebate.com.br/tag/aquecimento-global/page/2/
https://www.ecycle.com.br/aquecimento-global/	https://www.todamateria.com.br/exercicios-sobre-efeito-estufa/

Baseando-se nos dados da Tabela 4.6 e na análise feita quanto à real relevância das páginas retornadas, as Figuras 4.1, 4.2 e 4.3 apresentam, para cada tópico, os níveis de precisão obtidos considerando distintas quantidades de páginas relevantes retornadas pelo Yucca, em ordem decrescente de similaridade ao tópico desejado: 5 a 60 páginas relevantes retornadas, de 5 em 5.

Figura 4.1 – Níveis de precisão relacionados ao tópico receitas de bolo

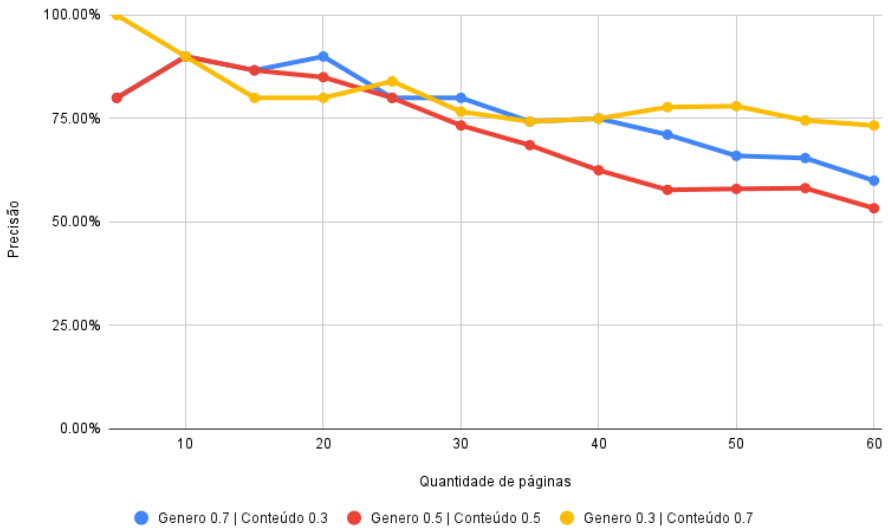


Fonte: Elaborado pelo autor.

Como pode ser visto na Figura 4.1, relacionada ao tópico "receitas de bolo de chocolate", o primeiro caso de teste, associado aos pesos 0.7 para gênero e 0.3 para conteúdo, obteve uma precisão melhor do que os demais casos de teste, mantendo-se um nível médio de precisão de 61% quando se considera as 60 páginas retornadas. Entretanto, quando se considera apenas as 10 primeiras páginas retornadas com maior similaridade pelo Yucca, caso comum em uma máquina de busca, os dois primeiros casos, associado aos pesos 0.7 para gênero e 0.3 para o primeiro e 0.5 para gênero e 0.5 para o segundo, tem um empate, com uma precisão média de 80%.

Para o tópico "artigos relacionados ao aquecimento global", como pode ser visto na Figura 4.2, o terceiro caso de teste, associado aos pesos 0.3 para gênero e 0.7 para conteúdo, mostrou-se superior aos demais testes, com um nível de precisão médio de 73% quando se considera as 60 páginas retornadas; para as 10 primeiras páginas, todos tem o mesmo nível de precisão, sendo este 90%.

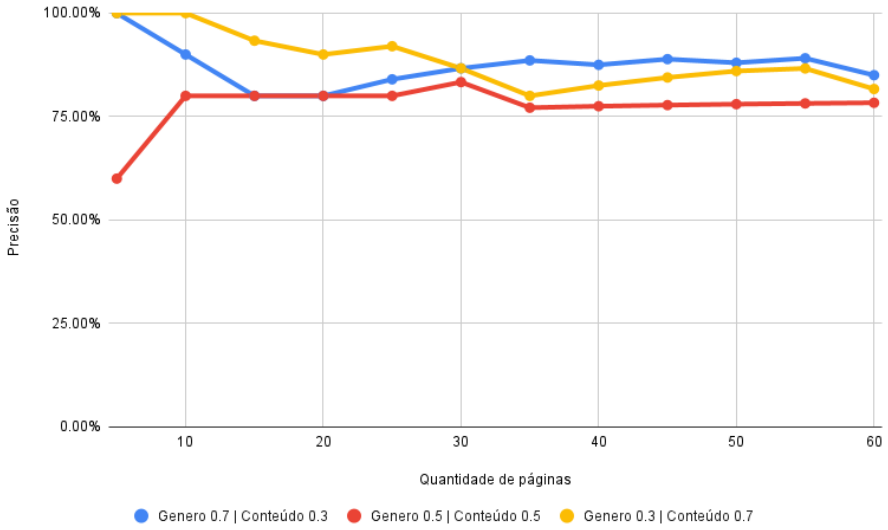
Figura 4.2 – Níveis de precisão relacionados ao tópico aquecimento global



Fonte: Elaborado pelo autor.

Já para o tópico "notícias relacionadas aos conflitos em Israel", como pode ser visto na Figura 4.3, o primeiro caso de teste, associado aos pesos 0.7 para gênero e 0.3 para conteúdo, novamente tem uma precisão melhor quando comparado aos demais, mantendo-se um nível médio de precisão de 85% quando se considera as 60 páginas; para as 10 primeiras páginas, o terceiro caso de teste tem um nível de precisão maior, mantendo-se em 100% de precisão.

Figura 4.3 – Níveis de precisão relacionados ao tópico conflitos em Israel

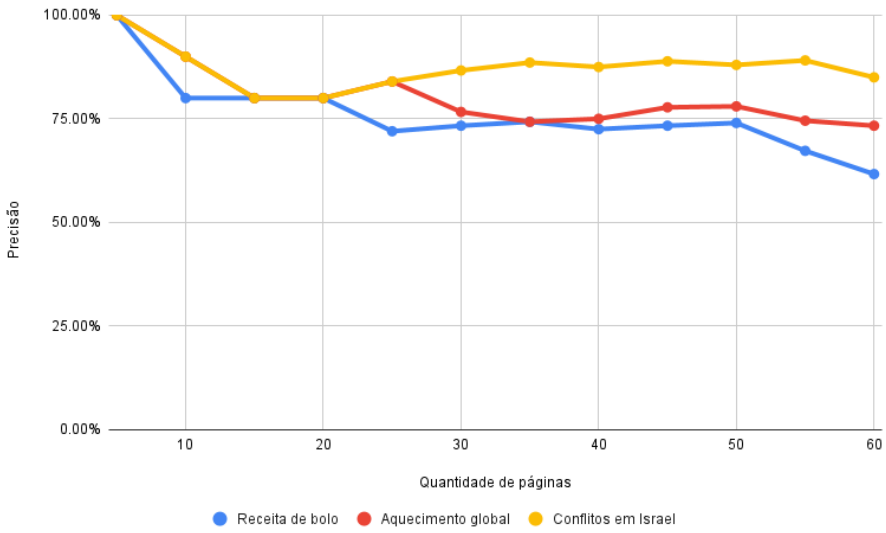


Fonte: Elaborado pelo autor.

De uma forma geral, observando-se as Figuras 4.1, 4.2 e 4.3, é possível observar que os níveis de precisão, quando comparados entre tópicos distintos, podem variar bastante, mas quando comparados entre si, a variação dos pesos de gênero e conteúdo produzem níveis de precisão bem parecidos, quando são comparados ao se considerar as 60 primeiras páginas. Observando a Figura

4.4, onde são apresentadas as melhores curvas de precisão obtidas pelos três tópicos, observa-se que todas mantiveram-se em níveis relativamente próximos e com bons níveis de precisão, demonstrando, dessa forma, resultados precisos e satisfatórios para os tópicos considerados. Particularmente, ao se considerar as 50 primeiras página retornadas como relevantes para o Yucca, ou seja, uma quantidade significativa para uma coleção de páginas relevantes, não houve nível de precisão abaixo de 70%.

Figura 4.4 – Melhores níveis de precisão de cada tópico



Fonte: Elaborado pelo autor.

5 Considerações Finais

Neste capítulo, são apresentadas as conclusões finais sobre o trabalho desenvolvido. A Seção 5.1 apresenta as conclusões obtidas com o desenvolvimento do trabalho e a Seção 5.2 discorre sobre perspectivas de trabalho futuro.

5.1 Conclusão

Como apresentado, este trabalho propôs desenvolver uma versão completa e funcional do Yucca, um coletor temático baseado em gênero e conteúdo, considerando a abordagem inicial, proposta em ASSIS *et al.* (2007), ASSIS *et al.* (2008), ASSIS *et al.* (2009), e suas melhorias propostas em MANGARAVITE, ASSIS e FERREIRA (2012), MANGARAVITE *et al.* (2014), SIQUEIRA *et al.* (2016), COSTA (2017) e SILVA (2023), além de sanar os problemas de implementação mencionados em DINIZ (2018), JUNIOR (2021) e SILVA (2023).

Buscando avaliar uma versão completa e funcional do Yucca, como visto no Capítulo 4, foram realizados experimentos considerando 3 tópicos distintos. Em todos eles, os resultados de eficácia mostraram-se bem satisfatórios, com os melhores resultados ficando entre 78% e 85% de precisão. Particularmente, foi possível observar que tópicos distintos podem gerar níveis de precisão distintos e que a variação dos pesos de gênero e conteúdo produzem níveis de precisão parecidos, quando são comparados ao se considerar o nível de precisão das 60 primeiras páginas coletadas pela nova versão do Yucca.

Embora os resultados obtidos apresentem ser inferiores à abordagem original (vide Subseção 2.1.1), houve a implementação de várias melhorias no processo de coleta, apresentadas no Capítulo 3, facilitando a usabilidade do Yucca por parte do usuário. Além disso, foi possível observar que as páginas *web* atuais são bem mais complexas e não tão bem estruturadas quando comparadas as existentes na época do desenvolvimento da abordagem original, dificultando o processo de coleta das mesmas.

5.2 Perspectivas de Trabalho Futuro

Nesta seção, são apresentadas algumas perspectivas de trabalhos futuros. Desta forma, pretende-se: (1) hospedar o Yucca em um servidor *web*, para que possa ser utilizado pelo público geral; (2) desenvolver e validar a estratégia de expansão automática de termos (vide Subseção 2.1.1.4), afim de se verificar ou não a obtenção de níveis melhores de precisão no processo de coleta; (3) implementar melhorias na interface do Yucca, para melhorar usabilidade do usuário; (4) validar o funcionamento e usabilidade do Yucca com diversos usuários, afim de se corrigir

eventuais problemas encontrados; (5) realizar experimentos comparativos com estudos de outros autores, utilizando distintas métricas.

Referências

AHLGREN, M. **100+ Internet Statistics, Facts And Trends For 2023**. 2023. Disponível em: <https://www.websiterating.com/research/internet-statistics-facts/>. Acesso em: 26 de dezembro 2023.

ASSIS, G. *et al.* The impact of term selection in genre-aware focused crawling. In: . [S.l.: s.n.], 2008. p. 1158–1163.

ASSIS, G. T. D. *et al.* Exploiting genre in focused crawling. In: ZIVIANI, N.; BAEZA-YATES, R. (Ed.). **String Processing and Information Retrieval**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 62–73. ISBN 978-3-540-75530-2.

ASSIS, G. T. D. *et al.* A genre-aware approach to focused crawling. **World Wide Web**, v. 12, n. 3, p. 285–319, Sep 2009. ISSN 1573-1413.

BROWNLEE, J. How to calculate precision, recall, and f-measure for imbalanced classification. **Machine Learning Mastery**, v. 1, 2020.

CAMPOS, R. *et al.* Yake! keyword extraction from single documents using multiple local features. **Information Sciences**, v. 509, p. 257–289, 2020. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025519308588>.

CHAKRABARTI, S.; BERG, M. V. D.; DOM, B. Focused crawling: a new approach to topic-specific web resource discovery. **Computer networks**, Elsevier, v. 31, n. 11-16, p. 1623–1640, 1999.

CHARTREE, J.; CANKAYA, E. C.; PHITHAKKITNUKOON, S. Query expansion using association matrix for improved information retrieval performance. In: . [s.n.], 2013. Disponível em: <https://api.semanticscholar.org/CorpusID:14328156>.

CHRISTEN, P.; HAND, D. J.; KIRIELLE, N. A review of the f-measure: Its history, properties, criticism, and alternatives. 2023.

COSTA, G. F. **Aperfeiçoamento automático dos conjuntos de termos utilizados em processos de coleta temática de páginas da Web baseada em gênero**. 73 p. Monografia — Universidade Federal de Ouro Preto, Ouro Preto, MG, Brasil, 2017.

DINIZ, D. N. **Um coletor temático de páginas da Web baseado em gênero**. 39 p. Monografia — Universidade Federal de Ouro Preto, Ouro Preto, MG, Brasil, 2018.

Facebook. **React: A JavaScript library for building user interfaces**. 2013. [Acesso em: 26 de dezembro de 2023]. Disponível em: <https://react.dev/>.

FARAG, M. M. G.; LEE, S.; FOX, E. A. Focused crawler for events. **International Journal on Digital Libraries**, v. 19, n. 1, p. 3–19, Mar 2018. ISSN 1432-1300. Disponível em: <https://doi.org/10.1007/s00799-016-0207-1>.

GUO, H. Research on web data mining based on topic crawler. **Journal of Web Engineering**, v. 20, n. 4, p. 1193–1206, 2021.

HOSSEINKHANI, J.; TAHERDOOST, H.; KEIKHAEE, S. Anton framework based on semantic focused crawler to support web crime mining using svm. **Annals of Data Science**, v. 8, n. 2, p. 227–240, Jun 2021. Disponível em: <https://doi.org/10.1007/s40745-019-00208-5>.

JIANG, J.; YU, N.; LIN, C.-Y. Focus: learning to crawl web forums. In: **Proceedings of the 21st International Conference on World Wide Web**. [S.l.: s.n.], 2012. p. 33–42.

JUNIOR, M. T. A. **Desenvolvimento de um coletor temático de páginas da Web baseado em gênero e conteúdo**. 33 p. Monografia — Universidade Federal de Ouro Preto, Ouro Preto, MG, Brasil, 2021.

KUMAR, N.; AGGARWAL, D. Learning-based focused web crawler. **IETE Journal of Research**, Taylor & Francis, v. 69, n. 4, p. 2037–2045, 2023. Disponível em: <https://doi.org/10.1080/03772063.2021.1885312>.

LEE, J.-G. *et al.* An effective approach to enhancing a focused crawler using google. **The Journal of Supercomputing**, v. 76, 10 2020.

LIMA, C. O. A. **Aperfeiçoamento automático dos conjuntos de termos utilizados pelo Yucca: Um coletor temático caseado em gênero**. 56 p. Monografia — Universidade Federal de Ouro Preto, Ouro Preto, MG, Brasil, 2018.

MANGARAVITE, V. *et al.* Semi-automatic generation of seed pages in genre-aware focused crawling. In: . [S.l.: s.n.], 2014.

MANGARAVITE, V.; ASSIS, G. T.; FERREIRA, A. A. Improving the efficiency of a genre-aware approach to focused crawling based on link context. In: **2012 Eighth Latin American Web Congress**. [S.l.: s.n.], 2012. p. 17–23.

PANT, G.; SRINIVASAN, P. Link contexts in classifier-guided topical crawlers. **IEEE Transactions on knowledge and data engineering**, IEEE, v. 18, n. 1, p. 107–122, 2005.

PLISSON, J.; LAVRAC, N.; MLADENIC, D. A rule based approach to word lemmatization. In: . [s.n.], 2004. Disponível em: <https://api.semanticscholar.org/CorpusID:15628229>.

RONACHER, A. **Flask Documentation**. 2010. [Acesso em: 26 de dezembro de 2023]. Disponível em: <https://flask.palletsprojects.com/>.

SHRIVASTAVA, G. K.; PATERIYA, R. K.; KAUSHIK, P. An efficient focused crawler using lstm-cnn based deep learning. **International Journal of System Assurance Engineering and Management**, v. 14, n. 1, p. 391–407, Feb 2023. ISSN 0976-4348. Disponível em: <https://doi.org/10.1007/s13198-022-01808-w>.

SILVA, M. F. D. **Geração semiautomática dos conjuntos iniciais de termos utilizados em processos de coleta temática de páginas da Web baseada em gênero e conteúdo**. 33 p. Monografia — Universidade Federal de Ouro Preto, Ouro Preto, MG, Brasil, 2023.

SIQUEIRA, G. O. D. *et al.* Automatic determination of similarity threshold for focused crawling processes on web pages. In: **WWW. Proceedings of the 15th International Conference WWW/Internet (ICWI)**. Mannheim, Germany: [s.n.], 2016. p. 95–102.

SOUZA, R.; DORNELES, C. Analisando a eficácia do modelo vetorial de busca na ordenação de questionários. In: SBC. **Anais do XIII Simpósio Brasileiro de Sistemas de Informação**. [S.l.], 2017. p. 563–570.

TAYLAN, D. *et al.* Intelligent focused crawler: Learning which links to crawl. In: **2011 International Symposium on Innovations in Intelligent Systems and Applications**. [S.l.: s.n.], 2011. p. 504–508.