UNIVERSIDADE FEDERAL DE OURO PRETO - UFOP

INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS

DEPARTAMENTO DE COMPUTAÇÃO

MARCELO TRAJANO ALVES JÚNIOR

Orientador: Prof. Dr. Guilherme Tavares de Assis

DESENVOLVIMENTO DE UM COLETOR TEMÁTICO DE PÁGINAS DA WEB BASEADO EM GÊNERO E CONTEÚDO

Ouro Preto, MG

2021

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

MARCELO TRAJANO ALVES JÚNIOR

DESENVOLVIMENTO DE UM 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 - UFOP 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
2021

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

FOLHA DE APROVAÇÃO

Marcelo Trajano Alves Júnior

Desenvolvimento de um 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 18 de Agosto de 2021.

Membros da banca

Guilherme Tavares de Assis (Orientador) - Doutor - Universidade Federal de Ouro Preto
Jadson Castro Gertrudes (Examinador) - Doutor - Universidade Federal de Ouro Preto
Andrea Gomes Campos Bianchi (Examinadora) - Doutora - 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 18/08/2021.

Documento assinado eletronicamente por **Guilherme Tavares de Assis**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 18/08/2021, às 12:39, 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 **0206594** e o código CRC **DD8A3F6A**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho aos meus pais que sempre acreditaram em mim e não mediram esforços para investir nos meus estudos.

Agradecimentos

Agradeço a minha mãe, paraibana, forte, braba, mas que sempre me deu muito carinho e força para que eu buscasse meus sonhos.

Agradeço ao meu pai, um grande exemplo, por todo os conselhos e acolhimento em todos esses anos da minha vida.

Agradeço ao meu orientador, Guilherme Tavares de Assis, por toda a paciência, dedicação e conselhos dados por inúmeras horas dedicadas para a conclusão deste trabalho.

Agradeço a minha namorada Terciana e família, por todo o apoio dado em todos esses anos de curso e, também, por todo o suporte dado durante o período de pandemia, me acolheram como parte de sua família e me trouxeram muito conforto em todos os momentos.

Por fim, agradeço a todos os meus amigos que fiz nessa minha passagem por Ouro Preto, em especial aos remanescentes do 17.2, sem vocês a caminhada teria sido muito mais difícil.

Tenho esta vida, que usarei para crescer.
Quem eu era antes, já não me consigo lembrar.
(KRAKAUER, 2012)

Resumo

Coletores temáticos são utilizados com um propósito de coletar páginas na *Web* que satisfaçam alguma propriedade particular e que sejam relevantes a um tópico de interesse específico, sendo importantes para uma grande variedade de aplicações. Para situações particulares, foi proposta e desenvolvida uma abordagem para coleta temática onde o tópico de interesse pode ser expresso por termos que descrevem o gênero e o conteúdo das páginas da *Web* desejadas. Visando aperfeiçoar a eficiência e a eficácia de tal abordagem original para coleta temática baseada em gênero e conteúdo, foram propostas, desenvolvidas e validadas as seguintes melhorias: uma nova política de localização de páginas relevantes baseada em *Link Context*, uma estratégia para a determinação semiautomática de páginas-semente, uma estratégia para a definição automática de limites de similaridade e uma estratégia de aperfeiçoamento automático dos conjuntos originais de termos de gênero e conteúdo. Nesse contexto, este trabalho propõe desenvolver uma primeira versão completa e funcional de um coletor temático, denominado Yucca, seguindo a abordagem original para coleta temática baseada em gênero e conteúdo e integrando as melhorias mencionadas, para que possa ser utilizada por distintos usuários de uma forma simples e robusta. Para validar o Yucca, experimentos foram realizados envolvendo a coleta de páginas da *Web* referentes a três tópicos de interesse distintos e atuais. De uma forma geral, o Yucca apresentou-se como um coletor temático eficaz, já que os níveis de precisão alcançados, pelos processos de coleta realizados, foram bem satisfatórios, chegando a ser superiores a 73% ao considerar 10 páginas retornadas como relevantes pelo coletor.

Palavras-chave: Processos de coleta temática. Coletor temático. Termos de gênero. Termos de conteúdo. Limite de similaridade. Expansão de termos.

Abstract

Focused crawlers are generally used to crawl pages that satisfy some particular property and that are relevant to a specific topic of interest and are important for a wide variety of applications. For particular situations, a focused crawling approach was proposed and developed where the topic of interest can be expressed by terms that describe the genre and content of the desired web pages. In order to improve the efficiency and effectiveness of such an original genre-aware approach to focused crawling, the following improvements have been proposed, developed and validated: relevant page location policy based on Link Context, semi-automatic seed page determination, automatic similarity threshold definition and automatic refinement of genre and content term sets. In this context, this work proposes to develop a complete and functional version of a crawler, called Yucca, following the original genre-aware approach to focused crawling and the improvements already developed and validated, so that it can be used by different users in a simple and robust way. To validate Yucca, experiments were performed involving the crawling of web pages referring to three distinct and current topics of interest. In general, Yucca presented itself as an effective focused crawler, since the levels of precision achieved by the crawling processes carried out were quite satisfactory, reaching more than 73% when considering 10 pages returned as relevant by the crawler.

Keywords: Focused crawling processes. Focused crawler. Genre terms. Content terms. Similarity threshold. Term expansion.

Lista de Ilustrações

Lista de Tabelas

Sumário

1 Introdução

Atualmente, existem mais de 1.83 bilhão de sites na internet e essa quantidade cresce a cada ano de forma exponencial, segundo AHLGREN (2021); com isto, torna-se necessária a criação de novas técnicas de Recuperação de Informação, no intuito de buscar acelerar e facilitar a coleta de páginas da *Web* e, consequentemente, a busca de informações desejadas por usuários. Para isto, como visto em BHATT; VYAS; PANDYA (2015), máquinas de busca constituem-se de ferramentas básicas para se buscar algo de interesse na internet a partir de repositórios que são gerados por meio de coletores *Web* (*Web Crawlers*) tradicionais: um coletor *Web* tradicional serve para coletar páginas da *Web* começando por páginas-semente e seguindo as ligações contidas nelas, visitando assim outras páginas até percorrer um número suficiente de páginas ou alcançar algum objetivo determinado.

Entretanto, segundo COSTA; ASSIS; SOUZA (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; DOM (1999) e JIANG et al. (2013), 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 JOHNSON; TSIOUTSIOULIKLIS; GILES (2003), PANT; SRINIVASAN (2006) e ALMPANIDIS; KOTROPOULOS; PITAS (2007), 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%.

1.1 Justificativa

Distintamente de estratégias de coleta temática existentes na literatura, vistas em KUMAR et al. (2018) e HOSSEINKHANI; TAHERDOOST; KEIKHAEE (2019), visando melhorar a eficiência e a eficácia de processos de coleta temática, foi proposta e desenvolvida uma abordagem em ASSIS et al. (2007, 2008, 2009), voltada para atender situações específicas. De uma forma geral, tal abordagem consiste em considerar as evidências de gênero e conteúdo presentes em

[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.

uma determinada página e estabelece um grau de similaridade entre tais evidências e o tópico específico de interesse. Por gênero, de acordo com ASSIS et al. (2009), entende-se o tipo, a categoria ou o estilo de texto de documentos específicos; por conteúdo, entende-se como o assunto ou tópico que se deseja coletar. Logo, tal trabalho desenvolvido 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.

Especificamente, tal abordagem original para coleta temática proposta é útil em situações onde um tópico de interesse possa ser expresso por meio de dois conjuntos distintos de termos: o primeiro descrevendo aspectos de gênero das páginas desejadas e o segundo referente ao assunto ou conteúdo descrito nessas páginas. Por meio de experimentos realizados, tal abordagem para coleta temática baseada em gênero e conteúdo, por ser mais específica, apresentou níveis de revocação e precisão entre 85% e 100%: níveis bem satisfatórios em relação a outras estratégias de coleta temática existentes na literatura.

Além disso, melhorias na abordagem original para coleta temática baseada em gênero e conteúdo foram propostas, desenvolvidas e validadas, no intuito de aperfeiçoar a eficiência e a eficácia da abordagem em questão. Tais melhorias envolvem (a) a utilização de características estruturais de links, presentes em uma página visitada pelo coletor em um processo de coleta temática, visto em MANGARAVITE; ASSIS; FERREIRA (2012), (b) a determinação semiautomática das páginas-semente a serem utilizadas em um processo de coleta temática, visto em MANGARAVITE; ASSIS; FERREIRA (2014), (c) a determinação automática do limite de similaridade a ser considerado em um processo de coleta temática, visto em SIQUEIRA et al. (2016), e (d) a expansão dos termos de gênero e conteúdo por meio de uma técnica automática de expansão de termos, visto em COSTA; ASSIS; SOUZA (2017).

Entretanto, não há um coletor, ou seja, uma ferramenta funcional propriamente dita, que realize processos de coleta temática seguindo a abordagem original e todas as melhorias já aplicadas na abordagem. Há uma versão *Web* inicial e não funcional de tal coletor, proposta por DINIZ; ASSIS (2018), que considera a abordagem original e algumas das melhorias citadas.

1.2 Objetivos Geral e Específicos

Este trabalho possui, como objetivo geral, o desenvolvimento e a validação da primeira versão completa e funcional de um coletor temático de páginas da *Web* baseado em gênero e conteúdo, denominado Yucca, para que possa ser utilizada por distintos usuários de uma forma simples e robusta. Para tanto, são consideradas a abordagem original para coleta temática baseada em gênero e conteúdo, vista em ASSIS et al. (2007, 2008, 2009), onde o tópico de interesse do usuário pode ser expresso por termos que descrevem o gênero e o conteúdo das páginas da *Web* desejadas, e as melhorias de eficiência e eficácia já aplicadas na abordagem e devidamente validadas, vistas em MANGARAVITE; ASSIS; FERREIRA (2012, 2014), SIQUEIRA et al. (2016)

e COSTA; ASSIS; SOUZA (2017).

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

- melhoria da eficiência em processos de coleta temática, ou seja, da localização mais rápida das páginas desejadas, por meio da utilização de características estruturais de *links* e da geração semi-automática das páginas-semente;
- melhoria da eficácia em processos de coleta temática, ou seja, da corretude das páginas coletadas como relevantes, por meio do aprimoramento da heurística relativa ao cálculo de similaridade, da geração automática de limites de similaridade e da expansão de termos;
- facilidade na intervenção do usuário, se for de interesse, quanto à definição de características de funcionalidade relativas ao processo de coleta que deseja realizar, por meio da geração de uma interface amigável para o Yucca;
- clareza no retorno das coleções geradas pelo Yucca, após processos de coleta temática realizados, por meio da criação de um componente para apresentação das páginas relevantes coletadas;
- certificação da qualidade de coleta temática, baseada em termos de gênero e conteúdo, a partir da análise de coleções geradas por meio de experimentos de validação do Yucca, considerando processos de coleta relativos a distintos e atuais tópicos de interesse.

1.3 Método do Trabalho

Visando o alcance do objetivo geral deste trabalho, foi implementada uma versão completa e funcional do Yucca considerando uma nova arquitetura de funcionamento, baseada na proposta apresentada em DINIZ; ASSIS (2018).

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

O restante deste trabalho monográfico encontra-se organizado como se segue. No Capítulo 2, é apresentada a revisão de literatura, relacionada ao tópico proposto e necessária para o entendimento deste trabalho. No Capítulo 3, o coletor temático proposto neste trabalho, envolvendo sua arquitetura de funcionamento, características e interface, é descrito. No Capítulo 4, os experimentos práticos realizados são apresentados e os resultados obtidos são analisados. E por

fim, no Capítulo 5, são apresentadas as conclusões deste trabalho e as perspectivas de trabalho futuro.

2 Revisão de Literatura

Este capítulo apresenta a revisão de literatura feita para a realização deste trabalho. Encontra-se organizado da seguinte forma: a Seção 2.1 apresenta a fundamentação teórica utilizada para o bom desenvolvimento deste trabalho e a Seção 2.2 apresenta os trabalhos diretamente relacionados ao objetivo geral deste trabalho.

2.1 Fundamentação Teórica

Nesta seção, é apresentado o suporte teórico necessário para o entendimento e o desenvolvimento deste trabalho. A Subseção 2.1.1 apresenta modelos de coleta de páginas na *Web*. A Subseção 2.1.2 apresenta a abordagem original elaborada por ASSIS et al. (2009) para a realização de processos de coleta temática baseada em gênero e conteúdo. As Subseções 2.1.3 a 2.1.6 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. Por fim, a Subseção 2.1.7 apresenta a ideia de uma primeira versão, definida por DINIZ; ASSIS (2018), do coletor temático proposto neste trabalho.

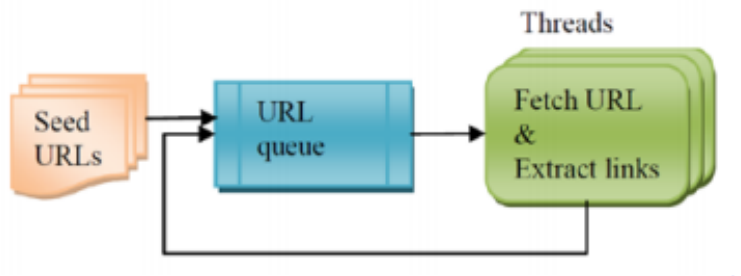
2.1.1 Modelos de Coleta de Páginas na *Web*

Segundo BHATT; VYAS; PANDYA (2015), máquinas de busca são ferramentas básicas, equipadas com vários algoritmos poderosos, para se consultar algo desejado na internet. Coletores tradicionais da *Web* constituem um dos principais componentes associados a máquinas de busca, visitando e baixando páginas para que possam ser armazenadas em um repositório e, assim, indexadas no intuito de que páginas desejadas por usuários, mediante consultas, possam ser mais facilmente localizadas.

Em termos gerais, um coletor tradicional possui, como objetivo principal, rastrear de forma rápida, eficaz e eficiente o maior número de páginas úteis, seguindo as estruturas dos *links* que as interconectam. Como ilustrado na Figura 2.1, o coletor inicia-se com um conjunto de páginas-semente e as URLs, que vão sendo encontradas, são inseridas em uma fila de URLs não visitadas pelo coletor. Em seguida, as URLs são retiradas da fila, na ordem em que se encontram, sendo buscadas as páginas na *Web* correspondentes às mesmas. Por fim, para uma determinada página visitada, os *links* são extraídos e inseridos na fila para que, em algum momento, sejam retirados pelo coletor.

Entretanto, com o aumento exponencial de páginas na *Web*, mais ênfase está sendo dada

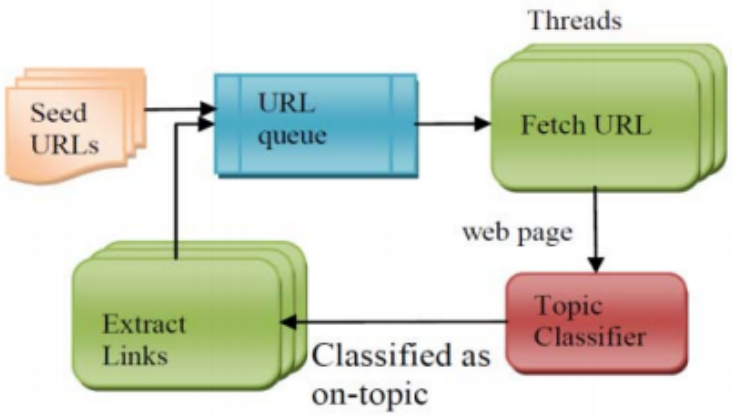
Figura 2.1 – Estrutura do coletor da *Web*



Fonte: BHATT; VYAS; PANDYA (2015)

a um modelo de coleta de páginas, denominado temático, que é mais robusto e que busca coletar páginas da *Web* que se relacionam a um tópico específico. Como pode ser visto na Figura 2.2, a sua estrutura diferencia-se de um modelo de coleta tradicional, visto na Figura 2.1, por conter um classificador de tópicos, onde se verifica se cada página visitada é relevante ou não a um tópico predefinido e desejado; uma vez sendo relevante, então o *link* é extraído e inserido na fila de URLs; caso contrário, a coleta não é processada nesta página.

Figura 2.2 – Estrutura de um coletor temático



Fonte: BHATT; VYAS; PANDYA (2015)

Ademais, GUPTA; ANAND (2015) reforçam que um processo de coleta temática é utilizado para encontrar páginas que satisfaçam alguma propriedade particular, relacionando-se a um tópico específico. Com este modelo, é possível buscar as páginas relevantes com um nível de eficácia alto e controlar, de forma mais eficiente, a extração dos *links*.

2.1.2 Abordagem Original para Coleta Temática Baseada em Gênero e Conteúdo

Como visto em LIMA (2018), a abordagem para coleta temática de páginas *Web* baseada em gênero e conteúdo, proposta em ASSIS et al. (2007, 2008, 2009), estabelece um arcabouço que permite a construção de coletores temáticos eficazes, eficientes e escaláveis, que levam em

consideração o gênero e o conteúdo das páginas desejadas. Mais especificamente, essa abordagem foi projetada para situações em que um tópico de interesse pode ser descrito por dois conjuntos distintos de termos: o primeiro conjunto expressa o gênero das páginas desejadas e o segundo descreve o assunto ou os aspectos de conteúdo dessas páginas.

Os coletores temáticos tradicionais guiados por classificadores geralmente analisam somente o conteúdo de uma página específica, diferentemente da abordagem proposta, não levando em consideração os dois tipos de informação. Portanto, páginas que fazem referência apenas aos termos de gênero ou apenas aos termos de conteúdo poderiam ser selecionadas por esses coletores, gerando erros de precisão; além disso, páginas relevantes, que especificam de maneira pobre o conteúdo de interesse da coleta, seriam classificadas como pertencentes à categoria das "não relevantes", consequentemente gerando erros de revocação.

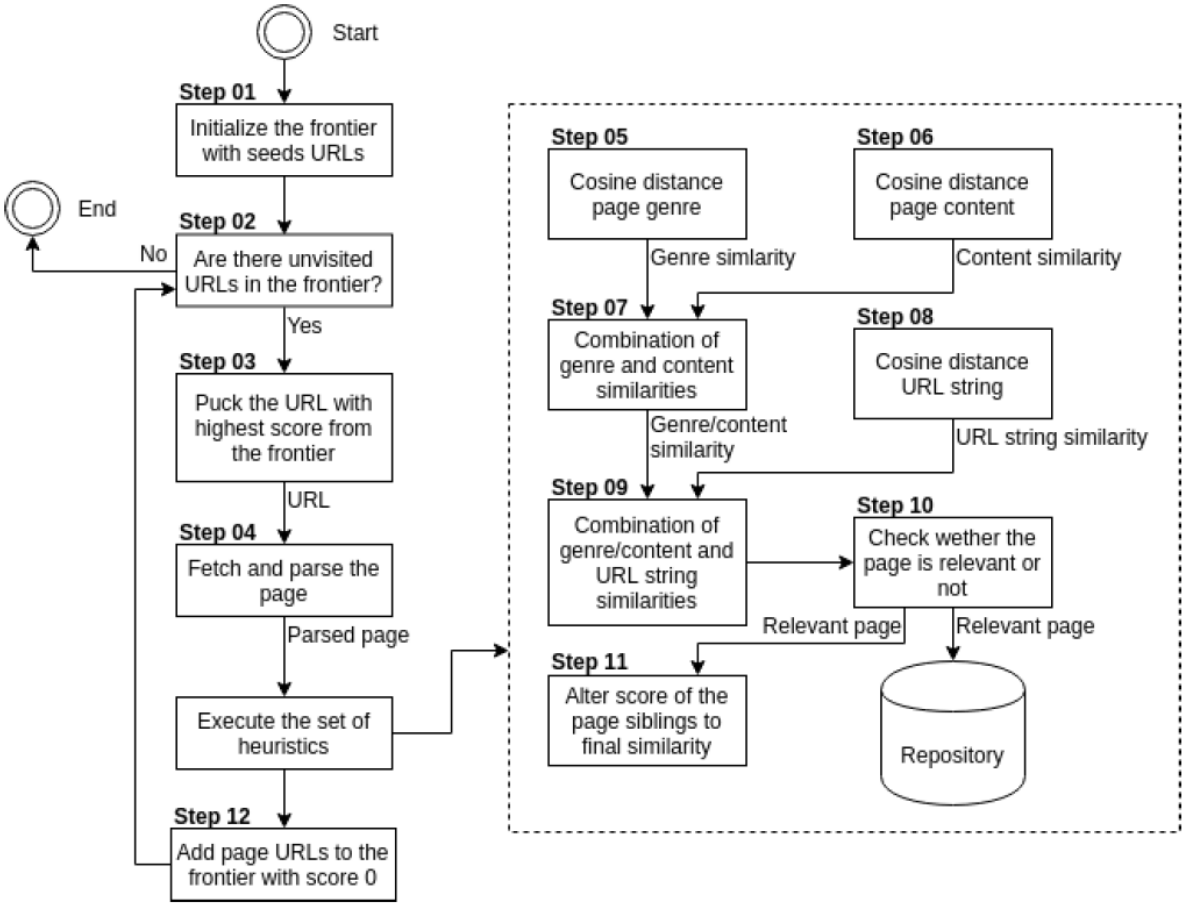
A Figura 2.3 sumariza os principais passos da abordagem proposta em ASSIS et al. (2009). Percebe-se que o coletor temático, construído de acordo com a abordagem baseada em gênero e conteúdo, analisa separadamente os termos referentes ao gênero e ao conteúdo do tópico de interesse (Passos 05 e 06); para tanto, os termos de gênero e conteúdo, assim como o limite de similaridade para se identificar as páginas da *Web* relevantes, correspondem a parâmetros de entrada para funcionamento da arquitetura.

De uma forma geral, a arquitetura de funcionamento, ilustrada na Figura 2.3, consiste nos seguintes passos:

- Passo 01: inicializar a fila de URLs não visitadas, denominada *Frontier*, com as URLs das páginas-semente: todas com pontuação de prioridade de visita pelo coletor igual a 1 (um);
- Passo 02: verificar se existem URLs no *Frontier* que ainda não foram visitadas pelo coletor; caso não existam, o processo de coleta é encerrado;
- Passo 03: por meio da política de enfileiramento dinâmico, selecionar no *Frontier* a URL com maior pontuação;
- Passo 04: buscar e analisar a página referente à URL selecionada no Passo 03;
- Passo 05: calcular a similaridade, por meio da distância de cosseno (modelo vetorial[1]), entre os termos de gênero do tópico desejado e a página buscada no Passo 04;
- Passo 06: calcular a similaridade, por meio da distância de cosseno, entre os termos de conteúdo do tópico desejado e a página buscada no Passo 04;
- Passo 07: combinar os resultados das similaridades de gênero e conteúdo obtidos nos Passos 05 e 06;

[1] De acordo com CARDOSO (2004), o modelo vetorial consiste em um modelo básico da área de Recuperação de Informação, que representa documentos e consultas como vetores de termos, sendo capaz de determinar a similaridade entre tais vetores e, assim, gerar um *ranking* de documentos mais similares a determinadas consultas.

Figura 2.3 – Arquitetura de funcionamento da abordagem para coleta temática baseada em gênero e conteúdo



Fonte: ASSIS et al. (2009)

- Passo 08: calcular a similaridade, por meio da distância de cosseno, entre os termos de gênero e conteúdo do tópico desejado e a URL da página buscada no Passo 04;
- Passo 09: combinar a similaridade de gênero/conteúdo, obtida no Passo 07, com a similaridade da URL obtida no Passo 08;
- Passo 10: verificar se a página visitada em questão é relevante ao tópico desejado, ou seja, se a similaridade final entre tal página (obtido no Passo 09) e o tópico desejado é superior ao limite de similaridade pré-estabelecido; caso seja relevante, a página é armazenada em um repositório de páginas relevantes ao tópico desejado;
- Passo 11: se a página visitada em questão for relevante, alterar a pontuação de prioridade de visita às páginas irmãs, que correspondem a URLs não visitadas no *Frontier*, para a similaridade final obtida no Passo 09;
- Passo 12: adicionar as URLs da página visitada pelo coletor no *Frontier* com pontuação de prioridade de visita igual a 0 (zero); em seguida, retornar ao Passo 02.

De acordo com ASSIS et al. (2009), os experimentos realizados demonstraram que coletores temáticos, construídos de acordo com a abordagem baseada em gênero e conteúdo descrita na Figura 2.3, atingiram níveis de F1[2] superiores a 88% para todos os tópicos de interesse considerados.

2.1.3 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 *Web* 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*.

Dessa forma, visando a melhoria de eficiência da abordagem de coleta temática baseada em gênero e conteúdo, apresentada na Figura 2.3, sem perda na escalabilidade e na eficácia da mesma, foi proposta em MANGARAVITE; ASSIS; FERREIRA (2012) a utilização do *Link Context*, mais precisamente texto de âncora, título do *link* e URL, para melhorar o processo de 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 de cossenos entre os termos de gênero e de 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.

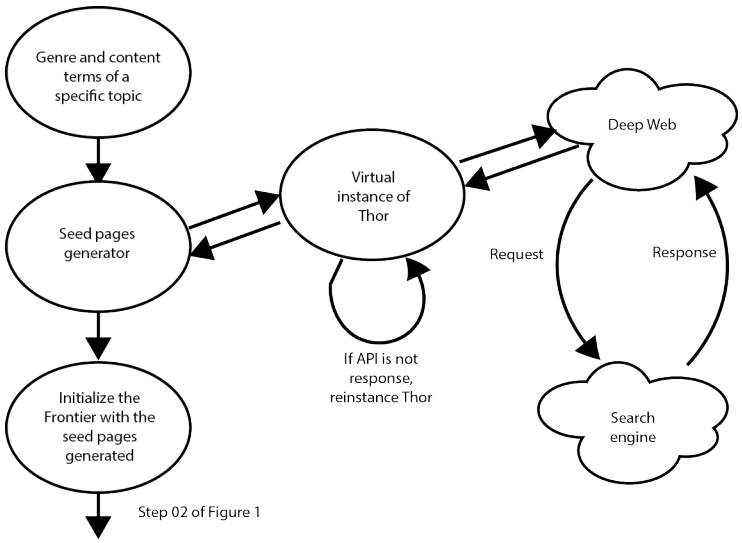
2.1.4 Geração Semiautomática de Páginas-semente

Com o intuito de também melhorar a eficiência da abordagem de coleta temática baseada em gênero e conteúdo, MANGARAVITE; ASSIS; FERREIRA (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 2.4, sendo que esta diz respeito apenas ao passo 01 do processo descrito pela Figura 2.3, alterando tal passo.

De acordo com a Figura 2.4, para se gerar semiautomaticamente 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:

[2] De acordo com BROWNLEE (2020), F1 é uma métrica que consiste na média harmônica entre a precisão e a revocação.

Figura 2.4 – Arquitetura de funcionamento da geração de páginas-semente



Fonte: , MANGARAVITE; ASSIS; FERREIRA (2014)

- *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%.

Uma vez definidas as páginas-semente para um processo de coleta temática, de acordo com a Figura 2.4, 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 2.3 (passo 02 em diante).

2.1.5 Determinação Automática de Limites de Similaridades

A abordagem para coleta temática de páginas *Web* proposta por ASSIS et al. (2007, 2008, 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.

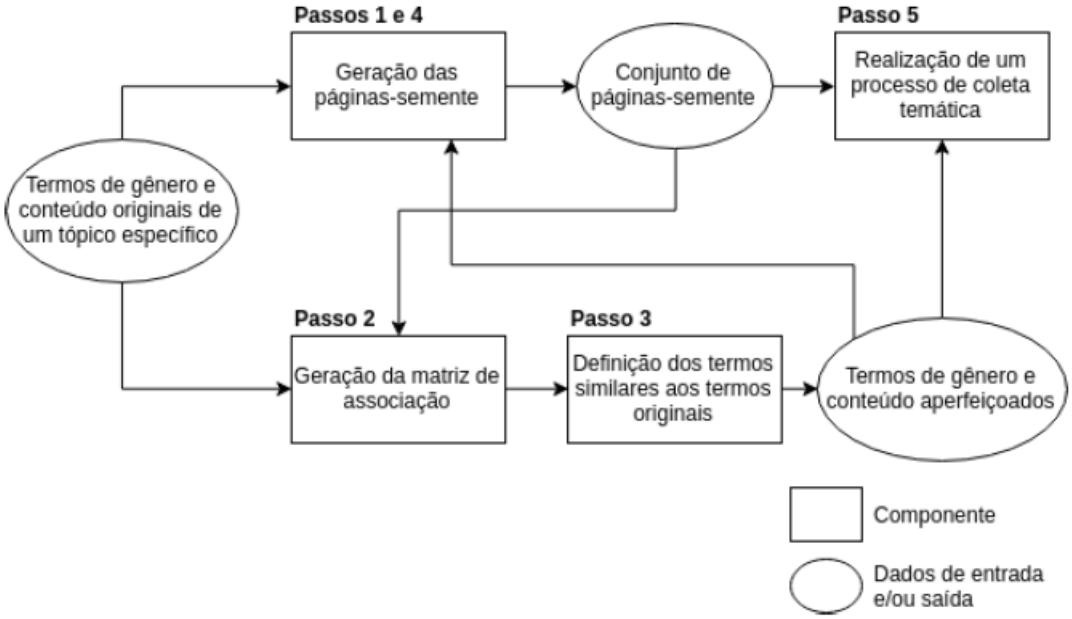
Para cada estratégia desenvolvida, foram realizados processos de coleta envolvendo três tópicos de interesse distintos. Por meio dos resultados obtidos, observou-se que os processos de coleta, relativos à estratégia baseada no método de agrupamento *K-Means*, foram os que apresentaram melhores eficácias, 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.

2.1.6 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 2.3, foram propostas por COSTA; ASSIS; SOUZA (2017) duas estratégias para expansão de tais conjuntos (vide Figuras 2.5 e 2.6).

Conforme mostrado pela arquitetura da Figura 2.5, a primeira estratégia visa expandir os

Figura 2.5 – Arquitetura de funcionamento da estratégia baseada em matriz de associação



Fonte: COSTA; ASSIS; SOUZA (2017)

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 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.

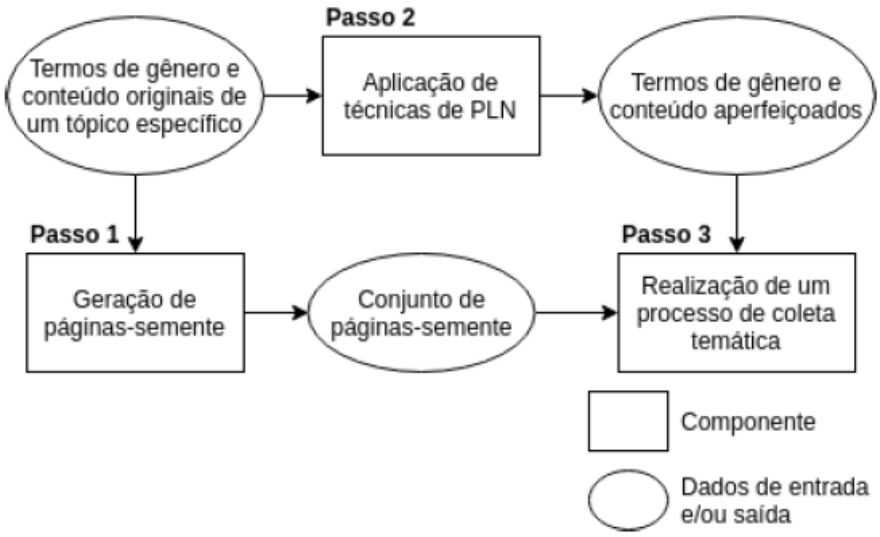
A segunda estratégia, apresentada na Figura 2.6, visa expandir os conjuntos de termos de gênero e conteúdo por meio da aplicação de técnicas de PLN. De modo geral, no passo 1, a partir dos conjuntos originais de termos de gênero e conteúdo, é gerado automaticamente o conjunto de páginas-semente necessário 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

[3] Uma matriz de associação, de acordo com CHARTREE; CANKAYA; PHITHAKKITNUKOON (2013), é uma estrutura de dados comumente utilizada para mensurar a relação existente entre os termos dos documentos de uma coleção.

termos, utilizando a métrica MenorDistância [4], foi a que apresentou melhores resultados quando comparada às demais estratégias propostas por COSTA; ASSIS; SOUZA (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.

Figura 2.6 – Arquitetura de funcionamento da estratégia baseada em PLN



Fonte: COSTA; ASSIS; SOUZA (2017)

2.1.7 Versão Inicial do Coletor Temático Baseado em Gênero e Conteúdo

DINIZ; ASSIS (2018) propuseram a primeira versão do coletor temático baseado em gênero e conteúdo, seguindo o que foi proposto em ASSIS et al. (2007, 2008, 2009) (vide Subseção 2.1.1) e as integrações das melhorias propostas por MANGARAVITE; ASSIS; FERREIRA (2012, 2014), SIQUEIRA et al. (2016) e COSTA; ASSIS; SOUZA (2017) (vide Subseções 2.1.3, 2.1.4 e 2.1.5, 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 propriamente dita em que um usuário possa utilizá-la.

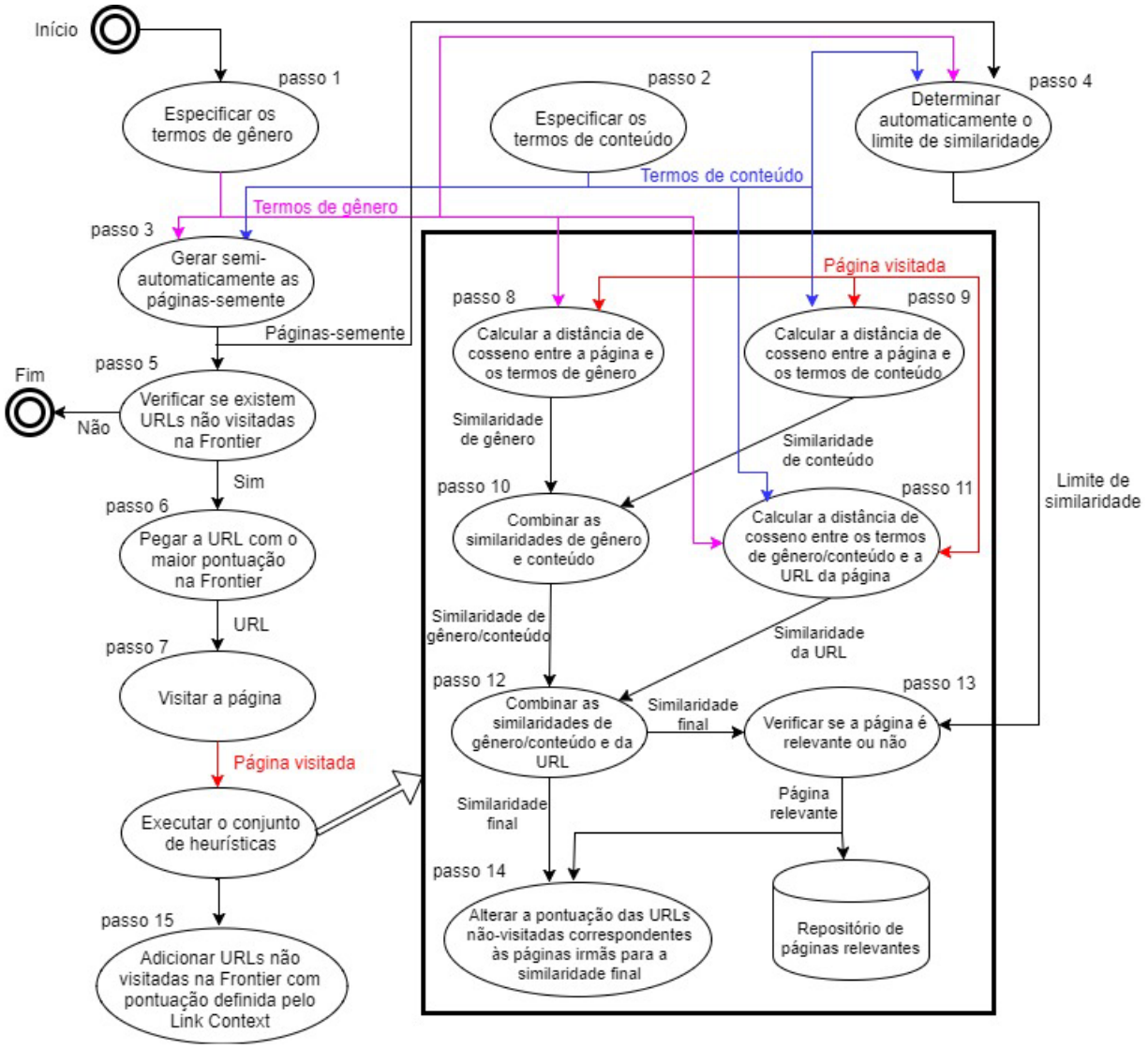
Para integrar as melhorias com a abordagem original (vide Figura 2.3) para coleta temática baseada em gênero e conteúdo, DINIZ; ASSIS (2018) propuseram uma nova arquitetura (vide Figura 2.7), seguindo o proposto em ASSIS et al. (2007, 2008, 2009) (vide Subseção 2.1.1) e acrescentando-se as seguintes melhorias: o uso do Link Context (vide Subseção 2.1.3), a geração

[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.

semiautomática de páginas-semente (vide Subseção 2.1.4) e a determinação automática de limites de similaridade (vide Subseção 2.1.5).

De uma forma geral, considerando um processo de coleta temática relativo a um determinado tópico de interesse, a arquitetura de funcionamento do coletor, ilustrada na Figura 2.7, consiste nos seguintes passos:

Figura 2.7 – Arquitetura de funcionamento do coletor



Fonte: DINIZ; ASSIS (2018)

- 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: inicializar a fila de URLs não visitadas, denominada *Frontier*, com as URLs das páginas-semente geradas semiautomaticamente: todas com pontuação de prioridade de visita pelo coletor igual a 1 (um);

- Passo 04: determinar automaticamente o limite de similaridade, utilizando os termos de gênero e conteúdo especificados pelo usuário nos Passos 01 e 02, respectivamente, e as páginassemente geradas no Passo 03;
- Passo 05: verificar se existem URLs na *Frontier* que ainda não foram visitadas pelo coletor; caso não existam, o processo de coleta é encerrado;
- Passo 06: por meio da política de enfileiramento dinâmico, selecionar na *Frontier* a URL com maior pontuação relativa à prioridade de visita;
- Passo 07: visitar a página referente à URL selecionada no Passo 06, no intuito de analisar a relevância da mesma, quanto ao tópico de interesse desejado, nos Passos 08, 09 e 11;
- Passo 08: calcular a similaridade, por meio da distância de cosseno, entre os termos de gênero especificados no Passo 01 e a página visitada no Passo 07;
- Passo 09: calcular a similaridade, por meio da distância de cosseno, entre os termos de conteúdo especificados no Passo 02 e a página visitada no Passo 07;
- Passo 10: combinar as similaridades de gênero e conteúdo obtidas nos Passos 08 e 09, por meio de uma média ponderada, gerando a similaridade de gênero/conteúdo;
- Passo 11: calcular a similaridade, por meio da distância de cosseno, entre os termos de gênero e conteúdo especificados nos Passos 01 e 02, respectivamente, e a URL da página visitada no Passo 07;
- Passo 12: combinar a similaridade de gênero/conteúdo, obtida no Passo 10, com a similaridade da URL obtida no Passo 11 por meio de uma média ponderada, gerando a similaridade final da página visitada no Passo 07;
- Passo 13: 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 12, é superior ao limite de similaridade determinado automaticamente no Passo 04; caso seja superior, a página visitada é considerada relevante e, assim, é armazenada no repositório de páginas relevantes ao tópico de interesse desejado;
- Passo 14: uma vez a página em questão sendo considerada relevante no Passo 13, alterar a pontuação de prioridade de visita das páginas irmãs da página visitada, que correspondem a URLs ainda não visitadas na *Frontier*, para a similaridade final obtida no Passo 12;
- Passo 15: 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 05.

Desta forma, em relação à abordagem original (vide Figura 2.3), foram acrescentados os Passos 01, 02 e 04 e modificados os Passos 01 (agora Passo 03) e 12 (agora Passo 15) para tratar as melhorias citadas.

De forma geral, nos experimentos realizados por DINIZ; ASSIS (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, não existe 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.

2.2 Trabalhos Relacionados

Buscando obter mais páginas relevantes em processos de coleta temática, LI; XING; ZHANG (2011) propuseram uma abordagem que utiliza um método de predição compreensiva, em que um algoritmo de partição de páginas particiona uma determinada página em blocos menores utilizando a URL, o texto âncora e o seu conteúdo para eliminar a interferência de tópicos distintos do interesse do usuário. Cada bloco é avaliado por um classificador já treinado, capaz de identificar a similaridade do bloco aos termos de conteúdo, analisando, assim, se a página é relevante ao tópico de interesse. Em experimentos realizados nessa abordagem, comparando com outra abordagem chamada *rule-based crawler* e utilizando as mesmas métricas e classificadores, percebeu-se que o coletor proposto obteve resultados superiores, devido à capacidade de eliminar interferências dos outros tópicos em seu algoritmo.

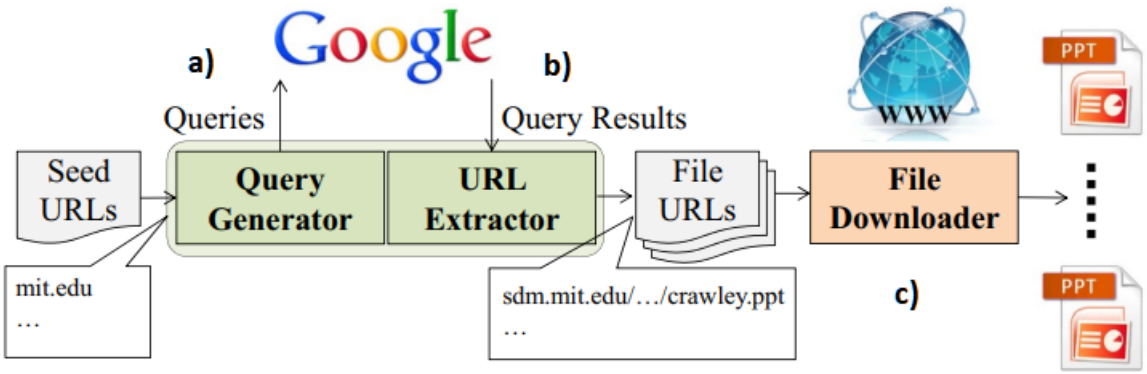
CHEN et al. (2012) utilizaram um algoritmo de reconhecimento baseado em análise de *links* para obter as página mais relevantes ao tópico de interesse desejado. Este algoritmo segue duas premissas: (a) se uma página A possui um *link* para uma página B, então a página B é uma recomendação da página A; e (b) se existem *links* que conectam as páginas A e B, então ambas páginas podem pertencer a um tópico em comum. Baseado nisto, CHEN et al. (2012) deduziram mais duas premissas, a saber: (a) se as páginas A e B apontam para as mesmas páginas, então estas duas páginas são consideradas relevantes, ou seja, quanto mais *links* duas páginas combinarem, maior será o grau de relevância entre elas; e (b) se uma página possui muitos *links*

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

apontando para o mesmo tópico, isso significa que esta página tem muita chance de ser relevante ao tópico também. Dessa forma, para se considerar que uma determinada página A visitada pelo coletor temático proposto é relevante ao tópico especificado, é necessário que a razão entre o número de *links* que estão em tal página A e o número de *links* que levam à mesma seja superior a um limite predefinido. Considerando tópicos médicos em seus experimentos, esta abordagem obteve um nível de precisão superior a 93% e de revocação superior a 83%, considerando limites de similaridade iguais a 0.5, 0.6, 0.7, 0.8 e 0.9.

Em LEE et al. (2019), um coletor temático, denominado *SlideCrawler*, baseado no formato do arquivo de mídia e no gênero[6], foi proposto objetivando coletar o máximo possível de arquivos de slides com conteúdo acadêmico, por meio do Google como ferramenta de coleta para gerenciar as consultas e realizar os *downloads* desejados. Como pode ser visto na Figura 2.8, este coletor possui: (a) um gerador de consultas predefinido anteriormente para especificar o formato do slide desejado (*filetype*: *ppt* ou *filetype*: *pptx*) e a universidade que se deseja consultar (ex: *site:mit.edu*); (b) um extrator de URLs que é responsável por extrair as URLs dos slides e remover possíveis duplicatas; e (c) um gerenciador de downloads que baixa os arquivos apontados pelo extrator. Nos experimentos realizados, o *SlideCrawler* foi capaz de baixar mais de 850 mil arquivos de slides acadêmicos com conteúdo diversos. Comparando-se com outro coletor chamado *Apache Nutch* [7], o *SlideCrawler* foi capaz de coletar 3.7 vezes mais arquivos de slides.

Figura 2.8 – Procedimento de coleta do SlideCrawler



Fonte: LEE et al. (2019)

Relacionando os trabalhos citados, foi produzido um comparativo ilustrado na Tabela 2.1, avaliando-se os pontos em comum em cada trabalho, de acordo com o que foi apresentado.

De acordo com a Tabela 2.1, observa-se, de uma forma geral, que as características funcionais "análise de *links*" e "uso de informações estruturais de páginas" foram abordados por quase todos os trabalhos relacionados. A ideia de utilizar estas duas características funcionais dá-se pela necessidade de melhorar a eficiência na abordagem de coleta, sem perder a escalabilidade e eficácia da mesma. Já a "utilização dos termos de gênero" é realizada apenas pelo coletor proposto

[6] Gênero, definido por LEE et al. (2019), refere-se apenas a textos acadêmicos.
[7] Segundo LEE et al. (2019), *Apache Nutch* é um *software* de coleta da *Web* de código aberto.

Tabela 2.1 – Comparativo de Trabalhos Relacionados

	Autores			
Características Funcionais	LI; XING; ZHANG	CHEN et al.	LEE et al.	Coletor Proposto
Análise de Links	X	X	X	X
Determinação semiautomática de páginas-semente				X
Determinação automática de limite de similaridade				X
Uso de termos de conteúdo	X	X		X
Uso de termos de gênero			X	X
Expansão de termos de gênero e conteúdo				**X**
Uso de informações estruturais de páginas	X		X	X

Fonte: Elaborado pelo autor.

e por LEE et al. (2019), sendo este gerado de forma limitada, apenas para textos acadêmicos, diferentemente da proposta deste trabalho, o qual se estende ao tipo, a categoria ou o estilo de texto de documentos específicos.

Os trabalhos citados não utilizaram determinação automática de páginas-semente, determinação semiautomática de limite de similaridade e expansão de termos de gênero e conteúdo, tornando-se limitados em processos de coleta. Desta forma, neste trabalho, diferentemente dos demais, todas essas características funcionais estão sendo utilizadas para realizar coletas mais relevantes e com menos ruídos, tornando-as mais eficientes e eficazes quando se deseja encontrar páginas relevantes a um determinado tópico especificado pelo usuário.

3 Proposta e Desenvolvimento do Yucca

Como já mencionado (vide Seção 1.2), este trabalho de monografia possui, como objetivo geral, o desenvolvimento e a validação da primeira versão completa e funcional do Yucca, coletor temático de páginas da *Web* baseado em gênero e conteúdo. Para tal, considera-se, como base, a abordagem original descrita na Figura 2.3, a arquitetura descrita na Figura 2.7 e as melhorias apresentadas nas Subseções 2.1.3, 2.1.4, 2.1.5 e 2.1.6.

Desta forma, este capítulo apresenta uma proposta de versão final e funcional do Yucca, estando delineado da seguinte forma: a Seção 3.1 descreve uma nova arquitetura de funcionamento do Yucca para a realização de processos de coleta temática baseada em gênero e conteúdo, e a Seção 3.2 apresenta a interface do Yucca, envolvendo a parametrização necessária para a executá-lo.

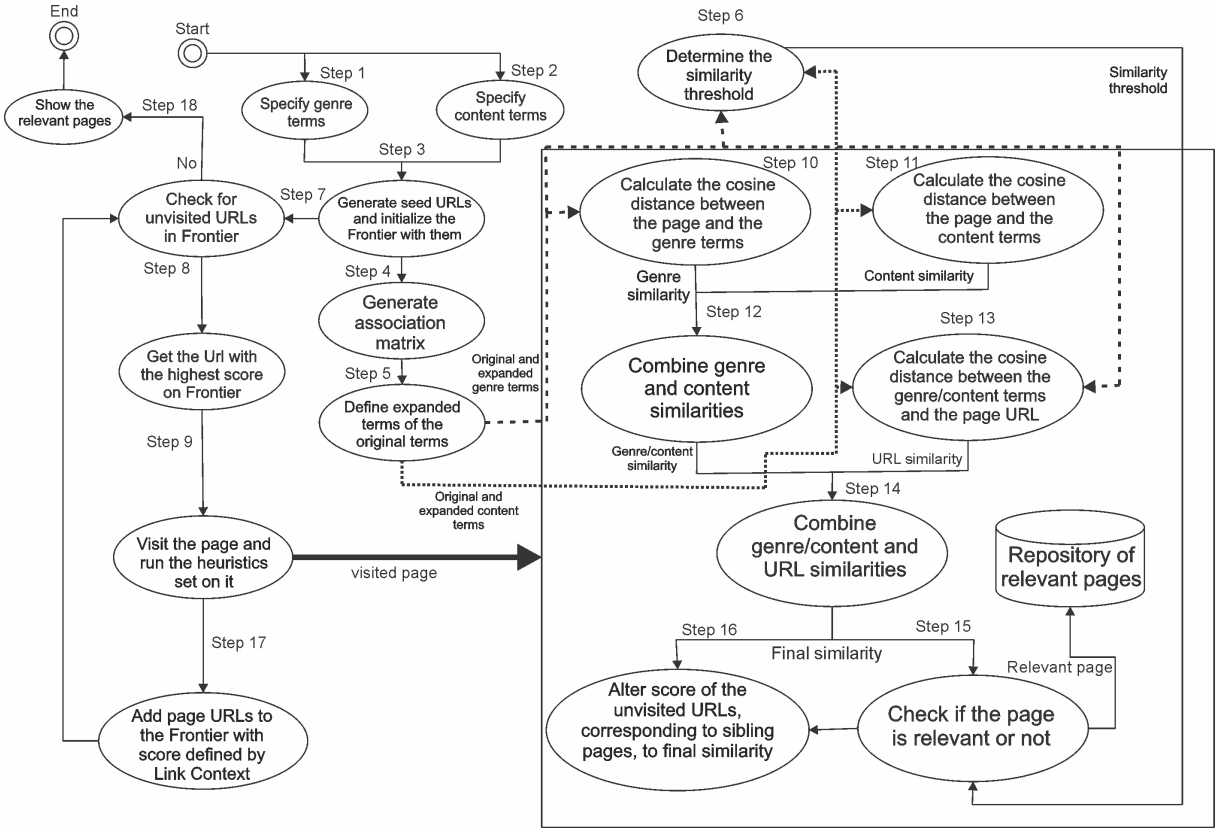
3.1 Arquitetura e Funcionamento do Yucca

A arquitetura de funcionamento do Yucca é apresentada na Figura 3.1. Em relação à arquitetura da abordagem descrita na Figura 2.7, acrescentou-se ao Yucca a melhoria da expansão automática dos conjuntos de termos de gênero e conteúdo, descrita na Subseção 2.1.6, 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, ilustrada na Figura 3.1, consiste 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;

Figura 3.1 – Arquitetura de funcionamento do Yucca



Fonte: Elaborado pelo autor

- 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.

Dessa forma, em relação à arquitetura descrita na Figura 2.7, foram acrescentados os Passos 04 e 05 após a geração semiautomática das páginas-semente do Passo 03. 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 2.1.6). Ademais, também foi adicionado o Passo 18 que corresponde a um componente para apresentação ao usuário da coleção de páginas relevantes gerada por meio de um processo de coleta realizado pelo Yucca.

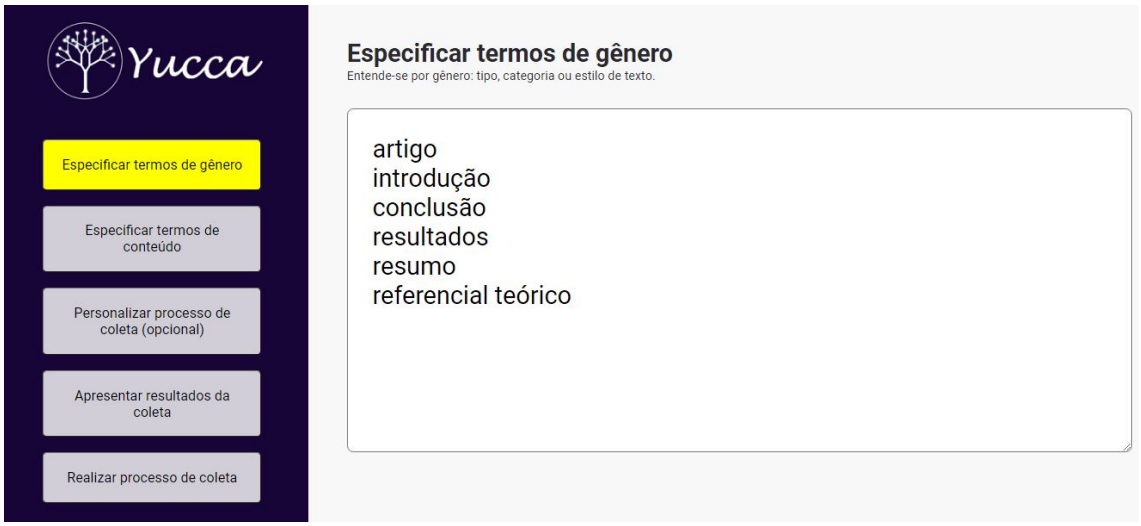
Além disto, na utilização do Yucca, existe a possibilidade do usuário alterar características de alguns passos da arquitetura proposta, a saber: utilizar ou não a estratégia de expansão de termos (Passos 04 e 05); alterar a heurística a ser utilizada na determinação automática de limite de similaridade (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).

3.2 Interface e Parametrização do Yucca

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

A tela inicial do Yucca, apresentada na Figura 3.2, apresenta as seguintes opções de menu principal: (a) "Especificar termos de gênero", voltada para o fornecimento dos termos de gênero relativos ao tópico de interesse desejado; (b) "Especificar termos de conteúdo", voltada para o fornecimento dos termos de conteúdo relativos ao tópico de interesse desejado; (c) "Personalizar processo de coleta", voltada para a alteração das configurações *default* para a execução de um processo de coleta, se for de interesse do usuário; (d) "Apresentar resultados da coleta", voltada para a apresentação da coleção gerada; e (e) "Realizar processo de coleta", voltado para realizar todo o processo de coleta com o que foi especificado pelo usuário. No caso, as telas das Figuras 3.2 e 3.3 ilustram, respectivamente, as opções "Especificar termos de gênero" e "Especificar termos de conteúdo". Como exemplo, estão especificados termos de gênero, relacionados a termos associados à localização de artigo, e de conteúdo, relacionados aos sintomas causados pela COVID-19, associados ao tópico de interesse "artigos relacionados a sintomas causados pela COVID-19".

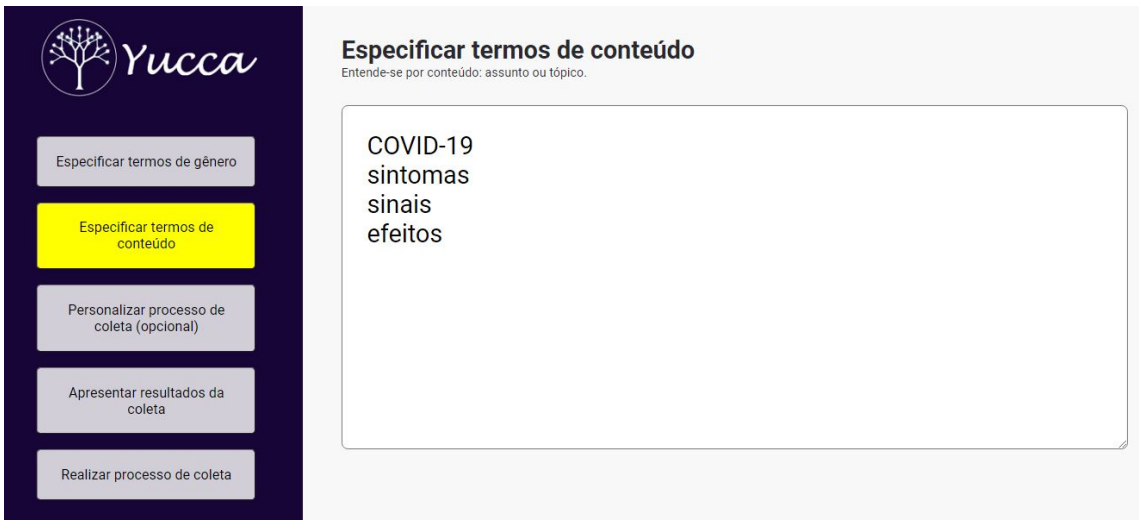
Figura 3.2 – Tela para especificação de termos de gênero



Fonte: Elaborado pelo autor

Uma vez fornecidos os termos de gênero e conteúdo desejados por meio das duas primeiras opções do menu principal do Yucca, o usuário pode personalizar ou não o processo de coleta por meio da opção "Personalizar processo de coleta" do menu principal. Uma vez selecionando tal opção, a tela da Figura 3.4 é apresentada, sendo utilizada para o fornecimento das configurações do processo de coleta a ser realizado. Esta figura apresenta 8 parâmetros utilizados pelo Yucca, inicializados com valores *default*, que o usuário pode especificar; são eles: 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 combinação gênero/conteúdo, o peso da

Figura 3.3 – Tela para especificação de termos de conteúdo



Fonte: Elaborado pelo autor

URL, a utilização ou não da expansão de termos e a heurística a ser utilizada para a determinação automática do limite de similaridade. É importante reforçar que a integração da melhoria relativa à expansão automática dos conjuntos de termos de gênero e conteúdo (vide Subseção 2.1.6) à abordagem, correspondente aos Passos 04 e 05 da Figura 3.1, está relacionada ao botão "Expansão de Termos", o qual o usuário poderá deixá-lo ativado ou não, caso deseje ou não que seja realizado a expansão dos termos de gênero e conteúdo no processo de coleta em questão.

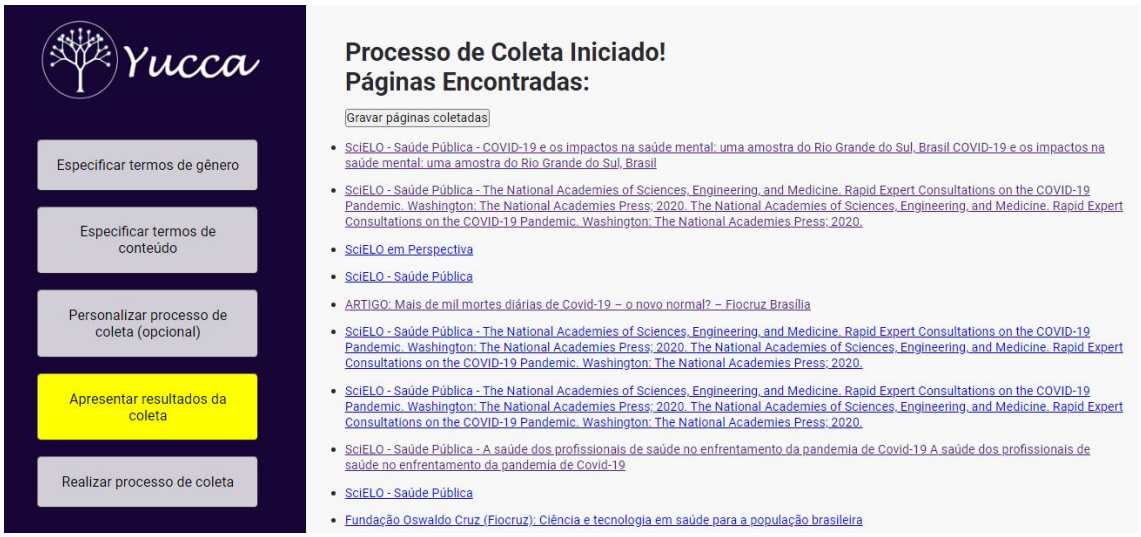
Figura 3.4 – Tela de configurações gerais do Yucca

Fonte: Elaborado pelo autor

Por fim, para que o processo desejado de coleta seja realizado, deve ser escolhida a opção

"Realizar processo de coleta" do menu principal do Yucca. Enquanto é realizado o processo de coleta, as páginas coletadas como relevantes pelo Yucca são retornadas ao usuário por meio da opção "Apresentar resultados da coleta", como pode ser visto na Figura 3.5. Nesta opção existe ainda a opção de gravar as páginas que foram coletadas por meio do botão "Gravar páginas coletadas", sendo gerado um arquivo JSON contendo a url, o título e a similaridade dos dados coletados.

Figura 3.5 – Resultados obtidos com a coleta do Yucca



Fonte: Elaborado pelo autor

4 Experimentação Prática

Neste capítulo, são apresentados e analisados os experimentos de validação da primeira versão funcional do Yucca, seguindo a arquitetura proposta na Figura 3.1. A Seção 4.1 descreve a métrica utilizada para avaliar a eficácia do Yucca. A Seção 4.2 descreve os experimentos realizados. Por fim, a Seção 4.3 apresenta e avalia os resultados obtidos por meio dos experimentos realizados.

4.1 Métrica de Avaliação

Para avaliação dos experimentos realizados, foi utilizada a métrica de precisão. De acordo com BROWNLEE (2020) e considerando o contexto deste trabalho, precisão é uma métrica que consiste na fração de páginas realmente relevantes ao tópico desejado, que foram retornadas pelo coletor temático, em relação a todas as páginas retornadas pelo mesmo. Ou seja:

$$precisao = \frac{TP}{TP + FP}$$

onde:

- TP é o número de páginas da *Web* visitadas pelo coletor, que são realmente relevantes ao tópico de interesse da coleta, e que foram classificadas como relevantes pelo coletor;
- FP é o número de páginas da *Web* visitadas pelo coletor, que não são relevantes ao tópico de interesse da coleta, e que foram erroneamente classificadas como relevantes pelo coletor.

4.2 Descrição dos Experimentos

Para que fosse possível avaliar a primeira versão funcional do Yucca, foram realizados processos de coleta considerando 3 tópicos atuais e distintos, a saber:

- artigos relacionados a sintomas causados pela COVID-19;
- artigos relacionados a racismo estrutural;
- artigos relacionados a aquecimento global.

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); nota-se que os termos de gênero são os mesmos para os 3 tópicos já que o gênero é o mesmo: artigos.

- Conjuntos de páginas-semente, para cada tópico especificado, obtidos por meio da aplicação da melhor heurística proposta por MANGARAVITE; ASSIS; FERREIRA (2014) (vide Subseção 2.1.4);
- Número máximo de páginas visitadas: 3000;
- 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.

Tabela 4.1 – Conjunto de termos que definem o tópico "artigos relacionados a sintomas causados pela COVID-19".

Termos de Gênero	**Termos de Conteúdo**
artigo	COVID-19
introdução	sintomas
conclusão	sinais
referencial teórico	efeitos
resumo	-
resultado	-

Fonte: Elaborado pelo autor

Tabela 4.2 – Conjunto de termos que definem o tópico "artigos relacionados a racismo estrutural".

Termos de Gênero	**Termos de Conteúdo**
artigo	racismo estrutural
introdução	preconceito
conclusão	discriminação racial
referencial teórico	-
resumo	-
resultado	-

Fonte: Elaborado pelo autor

Para cada tópico particularmente, foi realizado um estudo da importância dos termos de conteúdo e gênero, por meio dos seguintes pesos: gênero 0.3 e conteúdo 0.7; gênero 0.4 e conteúdo 0.6; gênero 0.5 e conteúdo 0.5; gênero 0.6 e conteúdo 0.4; e gênero 0.7 e conteúdo 0.3.

Ademais, visando análise das páginas retornadas como relevantes pelo Yucca, ao longo da execução de cada processo de coleta, foi armazenado um *log* contendo as seguintes informações sobre cada página da *Web* visitada:

Tabela 4.3 – Conjunto de termos que definem o tópico "artigos relacionados a aquecimento global".

Termos de Gênero	**Termos de Conteúdo**
artigo	aquecimento global
introdução	clima
conclusão	mudanças climáticas
referencial teórico	camada de ozônio
resumo	efeito estufa
resultado	temperatura
-	meio ambiente

Fonte: Elaborado pelo autor

- identificador da página visitada, atribuído automaticamente pelo coletor;
- título da página visitada;
- URL da página visitada;
- valor de similaridade calculado entre a página visitada e os termos originais e expandidos de gênero e conteúdo para o tópico de interesse.

Por fim, para a realização dos experimentos, foi utilizado um laptop com as seguintes características: sistema operacional Windows 10, processador Intel(R) Core(TM) i5-6200U, frequência de 2.40 GHz e RAM de 16GB; com um tempo médio de 2 horas para cada experimento.

4.3 Análise dos Resultados Obtidos

Nesta seção, são apresentados e analisados os resultados obtidos por meio da experimentação prática realizada, envolvendo a descrição experimental apresentada na Seção 4.2.

Considerando todos os processos de coleta realizados, a Tabela 4.4 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. Observe que tais valores são apresentados para cada caso de teste realizado para um mesmo tópico, variando o peso dos termos de gênero e conteúdo.

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 do tipo artigo e relacionada ao tópico especificado, ela foi considerada relevante. A Tabela 4.5 apresenta as URLs de algumas páginas visitadas pelo Yucca para o tópico "artigos relacionados a sintomas causados pela COVID-19", correspondendo a artigos considerados relevantes e não relevantes.

Baseando-se nos dados da Tabela 4.4 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

Tabela 4.4 – Tabela com os resultados dos casos de testes realizados

Tópico	Caso de Teste	Limite de similaridade	Quantidade de páginas visitadas	Quantidade de páginas relevantes
Artigos relacionados a sintomas causados pela COVID-19	Gênero: 0.3 Conteúdo: 0.7	0.2468	1603	373
	Gênero: 0.4 Conteúdo: 0.6	0.1664	1733	488
	Gênero: 0.5 Conteúdo: 0.5	0.1381	1432	605
	Gênero: 0.6 Conteúdo: 0.4	0.2468	1625	482
	Gênero: 0.7 Conteúdo: 0.3	0.2663	1524	343
Artigos relacionados a racismo estrutural	Gênero: 0.3 Conteúdo: 0.7	0.1960	1543	757
	Gênero: 0.4 Conteúdo: 0.6	0.1381	1328	496
	Gênero: 0.5 Conteúdo: 0.5	0.1610	1423	498
	Gênero: 0.6 Conteúdo: 0.4	0.1381	1338	481
	Gênero: 0.7 Conteúdo: 0.3	0.1587	1645	746
Artigos relacionados a aquecimento global	Gênero: 0.3 Conteúdo: 0.7	0.2168	2113	1044
	Gênero: 0.4 Conteúdo: 0.6	0.1781	2365	1270
	Gênero: 0.5 Conteúdo: 0.5	0.1664	2017	994
	Gênero: 0.6 Conteúdo: 0.4	0.1472	1930	1074
	Gênero: 0.7 Conteúdo: 0.3	0.1067	1982	1190

Fonte: Elaborado pelo autor

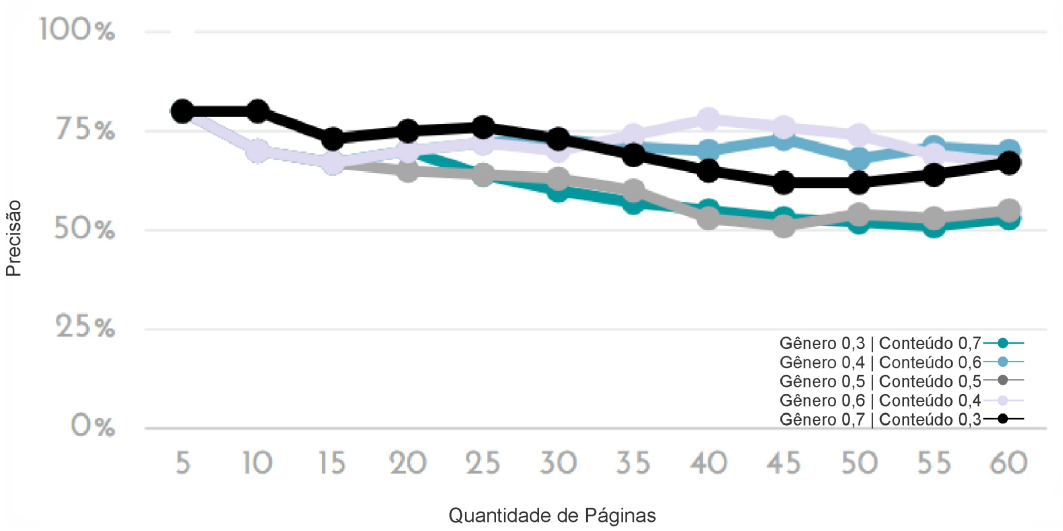
Tabela 4.5 – Exemplos de URLs visitadas pelo Yucca.

Páginas relevantes	**Páginas não relevantes**
https://scielosp.org/article/ csc/2020.v25n9/3517-3554/pt/	https://www.nucleodoconhecimento .com.br/saude/hemorragicas-ginecologicas
https://scielosp.org/article/ csc/2020.v25n9/3401-3411/	https://scielosp.org/article/ csc/2020.v25n9/3677-3684/pt/
https://www.nucleodoconhecimento .com.br/saude/risco-preexistentes	https://www.nucleodoconhecimento .com.br/saude/papiloma-escamoso

Fonte: Elaborado pelo autor

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 COVID-19

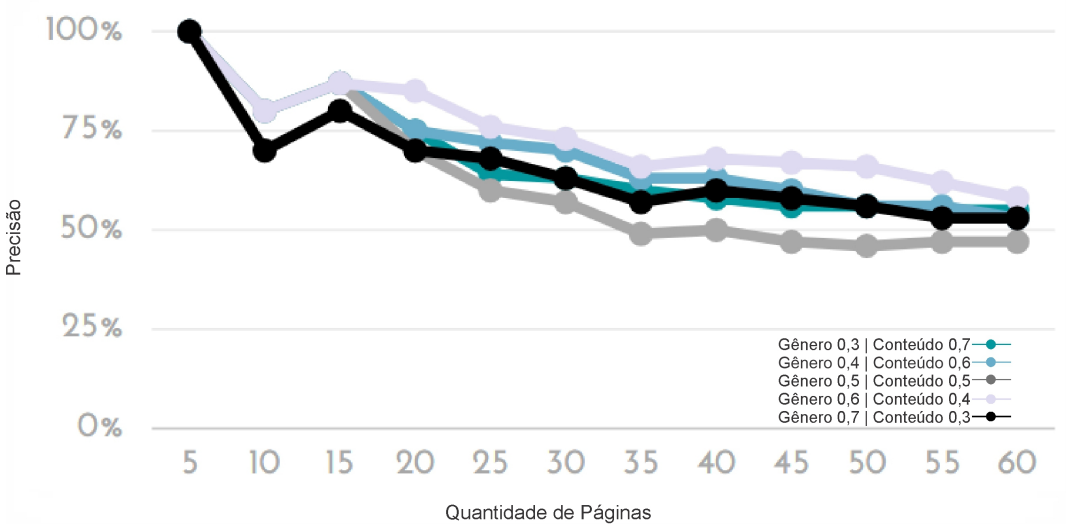


Fonte: Elaborado pelo autor

Como pode ser visto na Figura 4.1, relacionada aos processos de coleta relativos ao tópico "artigos relacionados a sintomas causados pela COVID-19", o caso de teste 4, associado aos pesos 0.6 para gênero e 0.4 para conteúdo, obteve uma precisão melhor do que os demais casos de testes, mantendo-se um nível médio de precisão de 72% 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, o teste 5, associado aos pesos 0.7 para gênero e 0.3 para conteúdo, apresenta-se com uma precisão média superior aos demais.

Já quanto ao tópico "artigos relacionados a racismo estrutural", como pode ser visto na Figura 4.2, o caso de teste 4, associado aos pesos 0.6 para gênero e 04 para conteúdo, mostrou-se superior aos demais testes, mantendo-se um nível médio de precisão de 74% quando se considera as 60 páginas retornadas e possuindo um nível médio de precisão acima de 80% quando se considera as 10 primeiras páginas retornadas com maior similaridade pelo Yucca.

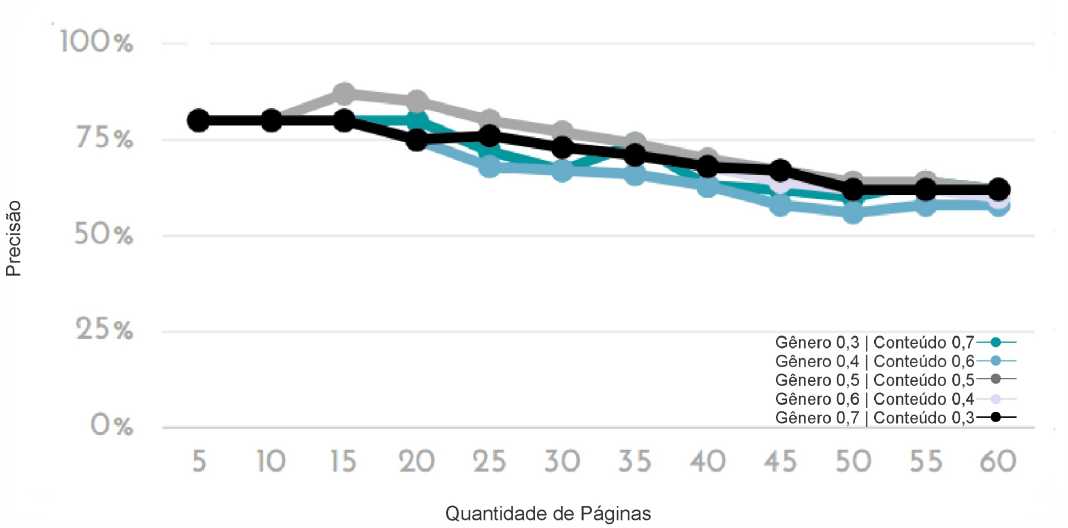
Figura 4.2 – Níveis de precisão relacionados ao tópico racismo estrutural



Fonte: Elaborado pelo autor

Em relação ao tópico "artigos relacionados a aquecimento global", como visto na Figura 4.3, as curvas de precisão do gráfico mostraram-se muito similares; entretando, o teste 3, associado aos pesos 0.5 para gênero e 0.5 para conteúdo, mostrou-se um pouco superior aos demais, mantendo-se um nível médio de precisão de 74% quando se considera as 60 páginas retornadas. Já quando se considera as 10 primeiras páginas retornadas com maior similaridade pelo Yucca, todos os testes possuem semelhantes médias de precisão.

Figura 4.3 – Níveis de precisão relacionados ao tópico aquecimento global

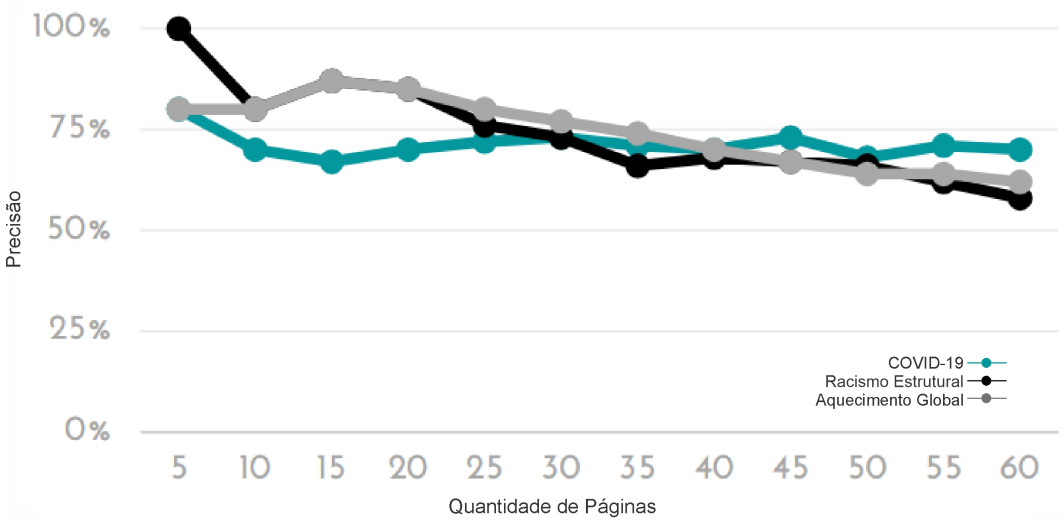


Fonte: Elaborado pelo autor

De uma forma geral, observando-se as Figuras 4.1, 4.2 e 4.3 é possível observar que, independente dos pesos de termos de conteúdo e de gênero, as 10 páginas mais relevantes coletadas pelo Yucca obtiveram um nível médio de precisão superior a 73% para todos os casos de testes associados aos 3 tópicos.

A Figura 4.4 apresenta a melhor curva de precisão obtida pelos três tópicos. Comparativamente, observa-se que essas curvas mantiveram-se muito próximas e com bons níveis de precisão, demonstrando, dessa forma, resultados precisos e satisfatórios para os tópicos considerados. Particularmente, ao se considerar as primeiras 30 páginas retornadas como relevantes pelo 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 sobre o trabalho desenvolvido (vide Seção 5.1) e as perspectivas de trabalho futuro (vide Seção 5.2).

5.1 Conclusão

Como apresentado, este trabalho propôs e desenvolveu uma versão completa e funcional de um coletor temático baseado em gênero e conteúdo, denominado Yucca, considerando a abordagem original proposta em ASSIS et al. (2007, 2008, 2009) e as melhorias feitas por MANGARAVITE; ASSIS; FERREIRA (2012, 2014), SIQUEIRA et al. (2016) e COSTA; ASSIS; SOUZA (2017).

Buscando avaliar uma primeira versão do Yucca, como visto, foram realizados experimentos considerando 3 tópicos distintos, e, em todos os tópicos, os resultados de eficácia mostraram-se bem satisfatórios, com níveis de precisão semelhantes. Particularmente, foi possível observar que, dependendo dos pesos dos termos de gênero e de conteúdo, os níveis de precisão podem ser distintos, embora, independente de tais pesos, os níveis de precisão foram acima de 73% para até 10 páginas retornadas como relevantes pelo Yucca nos três tópicos. Além disso, ao considerar possíveis coleções de 60 páginas geradas pelo Yucca, os níveis de precisão foram superiores a 55%.

Ademais, embora os resultados demonstrados aqui aparentam ser inferiores à abordagem original (vide Subseção 2.1.2), não há mais a necessidade do usuário definir limite de similaridade e páginas-semente, que influenciam diretamente a eficácia. Como também, não há mais a necessidade do usuário definir muitos termos de gênero e conteúdo, já que eles podem ser expandidos automaticamente, facilitando a coleta.

5.2 Trabalhos Futuros

Nesta seção, são apresentadas algumas perspectivas de trabalho futuro. Desta forma, pretende-se: (1) realizar novos experimentos de validação do Yucca utilizando, inclusive, outras métricas de validação como revocação e F1; (2) realizar estudos sobre a experiência do usuário quanto ao uso do Yucca, a fim de analisar a sua usabilidade; (3) realizar experimentos comparando com estudos de outros autores, utilizando as mesmas métricas; (4) realizar experimentos para calcular a média de precisão utilizando os mesmos pesos, a fim de analisar a constância dos resultados (5) propor, desenvolver e validar um componente de geração semiautomática de termos de gênero e conteúdo.

Referências

AHLGREN, M. *100 + estatísticas e fatos da internet para 2021*. 2021. Disponível em: <https://www.websitehostingrating.com/pt/internet-statistics-facts/>.

ALMPANIDIS, G.; KOTROPOULOS, C.; PITAS, I. Combining text and link analysis for focused crawling—an application for vertical search engines. *Information Systems*, Elsevier, v. 32, n. 6, p. 886–908, 2007.

ASSIS, G. T. D.; LAENDER, A. H.; GONÇALVES, M. A.; SILVA, A. S. D. Exploiting genre in focused crawling. In: SPRINGER. *Proceedings of the 14th Symposium on String Processing and Information Retrieval*. Santiago, Chile, 2007. p. 62–73.

ASSIS, G. T. D.; LAENDER, A. H.; GONÇALVES, M. A.; SILVA, A. S. D. A genre-aware approach to focused crawling. *World Wide Web*, Springer, v. 12, n. 3, p. 285–319, 2009.

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

BHATT, D.; VYAS, D.; PANDYA, S. Focused web crawler. In: *Advances in Computer Science and Information Technology (ACSIT)*. [S.l.: s.n.], 2015. v. 2, n. 11, p. 1–6.

BROWNLEE, J. How to calculate precision, recall, and f-measure for imbalanced classification. 01 2020. Disponível em: <https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification>.

CARDOSO, O. N. P. Recuperação de informação. *INFOCOMP Journal of Computer Science*, v. 2, n. 1, p. 33–38, 2004.

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

CHARTREE, J.; CANKAYA, E. C.; PHITHAKKITNUKOON, S. Query expansion using association matrix for improved information retrieval performance. *In Proceedings of the International Conference on Information and Knowledge Engineering*, WorldComp, 2013.

CHEN, Z.; LIU, J.; ZHAI, H.; JIANG, L.; CAO, B. Web page recognition algorithm based on link analysis in theme search engine. In: *2012 Second International Conference on Cloud and Green Computing*. IEEE, 2012. Disponível em: <https://doi.org/10.1109%2Fcgc.2012.42>.

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

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

GUPTA, A.; ANAND, P. Focused web crawlers and its approaches. *International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, p. 619–622, 2015.

HOSSEINKHANI, J.; TAHERDOOST, H.; KEIKHAEE, S. Anton framework based on semantic focused crawler to support web crime mining using svm. *Annals of Data Science*, 2019. ISSN 2198-5804,2198-5812. Disponível em: <http://doi.org/10.1007/s40745-019-00208-5>.

JIANG, J.; SONG, X.; YU, N.; LIN, C.-Y. Focus: learning to crawl web forums. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 25, n. 3, p. 1293–1306, 2013.

JOHNSON, J.; TSIOUTSIOULIKLIS, K.; GILES, C. L. Evolving strategies for focused web crawling. In: *ICML*. [S.l.: s.n.], 2003. p. 298–305.

KRAKAUER, J. *Na Natureza Selvagem*. São Paulo: Companhia das Letras, 2012.

KUMAR, M.; BINDAL, A.; GAUTAM, R.; BHATIA, R. Keyword query based focused web crawler. *Procedia Computer Science*, v. 125, p. 584–590, 2018. ISSN 1877-0509. Disponível em: <http://doi.org/10.1016/j.procs.2017.12.075>.

LEE, J.-G.; BAE, D.; KIM, S.; KIM, J.; YI, M. Y. An effective approach to enhancing a focused crawler using google. *The Journal of Supercomputing*, Springer US, 2019. ISSN 0920-8542,1573-0484. Disponível em: <http://doi.org/10.1007/s11227-019-02787-9>.

LI, X.; XING, M.; ZHANG, J. . A comprehensive prediction method of visit priority for focused crawler. *2nd International Symposiumon In Intelligence Information Processing and Trusted Computing (IPTC)*, p. 27–30, 2011.

LIMA, C. O. A. Aperfeiçoamento automático dos conjuntos de termos utilizados pelo yucca: Um coletor temático baseado em gênero. *UFOP*, v. 1, p. 56, 2018.

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

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

PANT, G.; SRINIVASAN, P. Link contexts in classifier-guided topical crawlers. *Knowledge and Data Engineering, IEEE Transactions on*, IEEE, v. 18, n. 1, p. 107–122, 2006.

SIQUEIRA, G. O. d.; ASSIS, G. T. d.; FERREIRA, A. A.; SILVA, A. S. N. e.; MANGARAVITE, V.; PADUA, F. L. C. 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, 2016.

TAYLAN, D.; POYRAZ, M.; AKYOKUS, S.; GANIZ, M. C. Intelligent focused crawler: Learning which links to crawl. p. 504–508, 6 2011.